HØGSKOLEN I OSLO
OG AKERSHUS

Elisa Storchi

Søk etter relevante deler i tekster med blandet innhold:

gjenfinningsforsøk med metadata fra NRKs radioprogram "Ukeslutt"

Masteroppgave 2014
Master i bibliotek- og informasjonsvitenskap
Høgskolen i Oslo og Akershus, Institutt for arkiv- bibliotek- og informasjonsfag

Sammendrag

I denne oppgaven har jeg evaluert gjenfinningen av relevante deler i tekster med blandet innhold i ett gjenfinningssystem. Jeg har utført et gjenfinningsforsøk med metadata fra NRKs radioprogram "Ukeslutt". Programmene er enheten for registreringen i den manuelle indekseringen av metadata i NRKs database. Programmene inneholder flere nyhetsinnslag som er de semantiske enhetene i programmene.

Formålet er å evaluere gjenfinningseffektivitet når det søkes i innslag delt hver for seg i forhold til søk i hele programmer, og se om inndelingen etter semantiske enheter forbedrer søkeresultatene.

I evalueringen brukes standardmål for gjenfinningseffektivitet, presisjon og fullstendighet, og resultatene for gjenfinningen analyseres kvalitativt.

Resultatene viser at det er mulig å finne fram til de enkelte innslagene eller temaene også der søkeenheten er programmene. Forskjellene mellom gjenfinningseffektivitet for søk i innslag og søk i programmer er marginale. Analysen viser også at indekseringspraksis og flere valg knyttet til gjenfinningssystemet påvirker gjenfinningseffektivitet. Det er en kombinasjon av flere faktorer som påvirker søkeresultatene.

Abstract

In this thesis I evaluate the retrieval of relevant parts of texts with mixed content in a retrieval system. I have performed a retrieval experiment with metadata from the NRK radio programme "Ukeslutt". The programmes are units in the manual indexing of metadata made by NRK. They contain several news items that are semantic units in the programmes.

The purpose is to evaluate the retrieval effectiveness when searching for news items indexed as units compared with search in complete programmes, and see if the division by semantic units improves search results.

The evaluation uses standard measures of retrieval effectiveness, precision and recall, and the results of the retrieval was analyzed qualitatively.

The results show that it is possible to find the news items and the single topics also when programmes are units for search. The difference between retrieval effectiveness for searching in news and search in programmes are marginal. The analysis also shows that indexing practices and several choices related to the retrieval system influences the retrieval effectiveness. A combination of different factors affects the search results.

Høgskolen i Oslo og Akershus, Institutt for arkiv- bibliotek- og informasjonsfag
Oslo 2014

Forord

Å skrive masteroppgave har vært utfordrende, men givende. Jeg har fått muligheten til å utdype det som er essensen i bibliotekfaget for meg. Gjenfinningen i metadata og i fulltekst som mange bibliotekarer jobber med i hverdagen, har vært fokus for meg lenge. Teoriene for hvordan gjenfinningen kan forbedres og utvikles videre er mange, men disse jobber alle mot et felles mål: å gjøre dataene mest mulige tilgjengelige for brukerne.

Takk til David Massey som har gitt meg tilgang til datamaterialet fra NRKs SIFT-basen, og som har kommet med nyttige innspill for beskrivelsen av metadataene.

En stor takk til min veilder Ragnar Nordlie for faglige og hyggelige samtaler, og spesielt for de mange konstruktive tilbakemeldingene som hjalp meg videre i arbeidet med oppgaven.

Takk til gode kollegaer ved Arbeiderbevegelsens arkiv og bibliotek som har holdt ut med mine opp- og nedturer i hele studieperioden. Spesielt takk til Jorunn og Kjersti.

Og til slutt takk til familien. Takk til Tore som både har veiledet meg gjennom java-programmeringen og lest korrektur. Takk til Andreas og Sofia for forståelse og tålmodighet.

Innholdsfortegnelse

Figurer og tabeller

1 Innledning

I dag er grensene mellom metadata, innholdsbeskrivende fulltekst og digitalisert fulltekst visket bort. Dette gjelder både ved søking på nettet og i flere gjenfinningssystemer. Eksempler er Nasjonalbibliotekets Bokhylla, Digitalarkivet fra Arkivverket, Google scholar og fulltekst tidsskriftdatabaser. Når man har tilgang til fullt digitaliserte tekster på nettet hvor innholdet kan bestå av ulike temaer blandet i samme tekst, blir det sentralt at gjenfinningssystemene klarer å finne fram til de spesifikke emnene.

Allerede i 1931 skrev Ranganhatan at boklig materiale som inneholder heterogene tekster, er en hindring for brukerne til å finne den informasjonen de trenger. De må bruke mye tid til å bla seg gjennom store mengde tekst. I sin forklaring til den fjerde biblioteksloven, "Save the time of the reader", skriver han at løsningen er å bruke kvalifisert bibliotekpersonale til å indeksere denne typen dokumenter. (Ranganathan, 2006, s. 337, 351-355)

Nå finnes dokumentene i fulltekst på nettet, men det er fortsatt slik at gjenfinningssystemene bør få fram akkurat den informasjonen brukerne trenger for at de skal slippe å lete gjennom en stor mengde tekst der det muligens bare er en liten del som er relevant. Min oppgave har denne problematikken som utgangspunkt.

Passage retrieval er en teknikk for å finne en løsning på dette. Ved *passage retrieval* deles tekstene i mindre enheter kalt for *passages*, og man evaluerer gjenfinningen for de delene av den større fullteksten som er relevante i forhold til søkespørsmålet. Resultatene er ikke så entydige, men flere studier viser noe forbedring i gjenfinningseffektivitet ved bruk av *passage retrieval* fremfor fulltekstsøk i hele dokumenter.

Jeg har brukt metadata for NRKs program "Ukeslutt", der de forskjellige nyhetsinnslagene sendt i samme program, er beskrevet samlet. Helt ulike temaer fra innslagene er beskrevet i samme innholdsmetadata med både en innholdsbeskrivende fulltekst og et kontrollert vokabular. Temaet for masteroppgaven er å se på hva slags konsekvenser denne typen indekseringen har for gjenfinningen, og om *passage retrieval* kan være til hjelp for å finne de enkelte innslagene.

Metadataene til "Ukeslutt" er hentet fra NRKs SIFT-base der flere tv- og radioprogrammer har vært indeksert siden andre halvdel av 80-tallet. Gjenfinningen av nyhetsinnslagene, av de omtalte og medvirkende personene i programmene, og av tekniske metadata knyttet til programmene er viktige både for videre medieproduksjon, men også for almenheten som skal finne fram i NRKs historiske arkiv og i de nyere programmene som er digitalt tilgjengelige på NRKs nettsider.

1.1 Problemstilling

Jeg ønsker å se på hvordan gjenfinningen i en datasamling påvirkes av at ulike emner er indeksert samlet i de innholdsbeskrivende metadataene.
Metadataene er hentet fra NRKs SIFT-base. Fra basen er det valgt innholdsbeskrivende metadata til radioprogrammet "Ukeslutt". Det kontrollerte vokabularet og den innholdsbeskrivende fullteksten er indeksert i samme felt for innhold. Det enkelte radioprogram er enheten for søking i basen. Dette innebærer at metadataene om ulike nyhetsinnslag, og dermed ulike temaer, er samlet og ikke kan søkes adskilt.

Den overordnede problemstillingen er:

Hvordan påvirkes gjenfinningen ved søk i metadata hvor ulike temaer er indeksert samlet sammenliknet med et søk hvor de er semantisk inndelt?

Noe mer konkret skal følgende forskningsspørsmål besvares:

1. Hvordan påvirkes gjenfinningen i innholdsmetadata til NRKs radioprogram "Ukeslutt" med programmene som søkeenhet i forhold til innslagene?

2. Hvilken betydning har det kontrollerte vokabularet i forhold til den innholdsbeskrivende fullteksten for gjenfinningen av innslagene og programmene?

1.2 Terminologi

Her følger definisjon av enkelte relevante termer og begreper, og hvordan de er brukt i denne oppgaven. Enkelte av de valgte termene er brukt i ulike betydninger innenfor ulike fagområder eller har vært gjenstand for en del diskusjon avhengig av forskjellige teoretiske utgangspunkter. Termene blir her definert for å gi en avklaring, og de vil ikke bli diskutert i dybden.

Emneord og tagger
Diskusjonen om bruk av emneord eller tagger er omfattende i forskningslitteraturen, og ikke sentral for denne oppgaven. Jeg har brukt disse termene slik som NRK har brukt dem for å definerer de to

indekseringsmetodene og omleggingen av deres indekseringspraksis fra 2012. Taggene blir valgt fra en ordliste som nylig er oppdatert og som er delvis bygget på de gamle emneordene. (Søbak, 2013, s. 35-36) I min oppgave er derfor emneord og tagger betraktet som to kontrollerte vokabularer som kan brukes i en sammenligning med den innholdsbeskrivende fullteksten.

Indeksering

Jeg bruker termen indeksering i to forskjellige, men relaterte betydninger. Der jeg skriver om indekseringen ved NRKs SIFT-base, referer termen til den manuelle registreringen av metadata, enten disse er i fulltekst eller med et kontrollert vokabular. Denne indekseringen har skapt en beskrivelse av NRKs programmer som blir brukt som grunnlagsmateriale for min oppgave. Når termen indeksering er brukt i forbindelse med gjenfinningssystemet Lucene, menes med indeksering den automatiske formaliseringen gjort av systemet for å lage en dokumentrepresentasjon av datamaterialet. (Lancaster, 2003, s. 6-7, 282-283)

Innholdsbeskrivende fulltekst

Jeg referer til innholdsbeskrivende fulltekst som en sammenhengende tekst med hele setninger som beskriver hva som ble sagt og skjedde i NRKs programmer. Den innholdsbeskrivende fullteksten er metadata, og ikke en ordrett og fullstendig gjengivelse av dialoger og andre ord sagt i løpet av programmene. Derfor er dette omtalt som innholdsbeskrivende fulltekst og ikke bare med termen fulltekst. Samtidig skiller denne beskrivelsen seg fra et kontrollert vokabular som beskriver dokumentene ved hjelp av termer eller tall som er laget på forhånd, er en del av en struktur og har regler for hvordan termene skal brukes i beskrivelsen. (Lancaster, 2003, s. 19) Eksempler på det kontrollerte vokabularet i programmene er bruken av emneord, tagger og klassifikasjonsnumre.

Gjenfinningseffektivitet

I evalueringen av gjenfinningssystemer skilles det mellom systemets *efficiency* og *effectiveness*. På norsk blir begge oversatt med termen effektivitet eller gjenfinningseffektivitet.
Systemets *efficiency* er systemets evne til å løse indekserings- og søkeoppgavene med minst mulig bruk av datamaskinens ressurser. *Efficiency* måles i bruk av lagringsplass, hastighet ved indeksering og søk. Systemets *efficiency* er ikke relevant for denne oppgaven.
Effectiveness ser i stedet på hvordan et system kan hjelpe brukeren til å finne fram til relevante metadata med minst mulig jobb. *Effectiveness* er en evaluering av blant annet de valgte indekseringsmetodene, av gjenfinningsmodellene og grensesnittet som resultatene blir presentert

med. (Baeza-Yates & Ribeiro-Neto, 2011, s. 337)
Når jeg bruker den norske termen gjenfinningseffektivitet eller effektivitetsmål, er det i samme betydning som den engelske termen *effektiveness*.

Søkespørsmål, spørsmål og søk.

Termen søkespørsmål blir brukt i oppgaven som en oversettelse av den engelske termen *query*. Søkespørsmål er en spesifikk forespørsel som sendes til en database eller et gjenfinningssystem. Det kan bestå av nøkkelord eller uttrykkes i naturlig språk (*Natural language query*). Søkespørsmålet er en første formalisering av brukerens behov og brukerens spørsmål (*Question*). Søkespørsmålet blir formalisert ytterligere av gjenfiningssystemet til en sekvens av indekstermer, og blir til en spørsmålsrepresentasjon. Spørsmålsrepresentasjonen brukes for å gjennomføre sammenligningen med den indekserte datasamlingen.
Med termen søk (*search*) menes i stedet den mer omfattende prosessen som starter fra et informasjonsbehov som uttrykkes gjennom et søkespørsmål, og som fører til en treffliste av dokumenter fra datasamlingen.

2 Bakgrunn og litteratur

I denne oppgaven vil jeg undersøke hvordan gjenfinningen i innholdsbeskrivende metadata blir når ulike temaer er indeksert samlet i forhold til når de er indeksert hver for seg. Det teoretiske rammeverket for oppgaven er knyttet til faget Gjenfinningsteorier (*Information Retrieval*), og mer spesifikt Gjenfinningsevaluering (*Retrieval evaluation*). Relevant tidligere forskning for oppgaven er knyttet til gjenfinningsforsøk med *passage retrieval,* og til sammenligninger av søk i kontrollerte vokabularer og søk i fulltekst.

2.1 Gjenfinningsteorier

Den engelske termen *Information retrieval* (IR), på norsk Informasjonsgjenfinning, ble brukt for første gang av Calvin N. Mooers i 1950. Den ble tatt i vanlig bruk på slutten av 50- og begynnelsen av 60-tallet da man begynte med de første databaserte gjenfinningseksperimentene i dokumentsamlinger og de første studiene om informasjonsgjenfinning ble publisert. (Baeza-Yates & Ribeiro-Neto, 2011, s. 1-2; Kelly & Ruthven, 2011, s. 1-2)
Informasjonsgjenfinning "embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, tecnique, or machines that are employed to carry out the operation" ifølge Mooers. (Mooers, 1951, s. 25) Tefko Saracevic peker på viktigheten av Mooers definisjon fordi for første gang blir søkeprosessen eksplisitt nevnt som et mål for forskningen, noe som ikke var tilfellet i tidligere studier der indekseringsmetodene stod sentralt. (Saracevic, 2007, s. 1917)
Siden starten av forskningen innenfor informasjonsgjenfinning har det vært flere forsøk på å lage teoretiske modeller som skal bidra til å forklare og evaluere de forskjellige aspektene ved fagområdet. G.G. Chowdhury definerer to hovedgrupper blant de teoretiske modellene: brukersentrerte/kognitive gjenfinningsmodeller og systemsentrerte gjenfinningsmodeller. (Chowdhury, 2010, s. 204-205) Baeza-Yates og Ribeiro-Neto nevner på samme måte to retninger innenfor IR-forskningen, det menneskesentrerte og datamaskinsentrerte synspunktet. (Baeza-Yates & Ribeiro-Neto, 2011, s. 1) Systemsentrerte, eller datamaskinsentrerte gjenfinningsmodeller fokuserer på systemenes gjenfinningsalgoritmer, indekseringsmetoder, grensesnitt og rangeringsalgoritmer, dvs. det som er direkte knyttet til systemene. De brukersentrerte modellene var i starten for det meste opptatte av å sette brukeren i sentrum, og legge systemene til side. Dette

var en reaksjon på at i de første studiene var brukerperspektivet for det meste oversett. Etter hvert fikk brukersentrerte modeller et mer holistisk synspunkt, og tok seg både av gjenfinningsmekanismene og de menneskelige aspektene ved søkingen og gjenfinningen, de kontekstuelle aspektene rundt gjenfinningen og interaksjonene mellom menneskene og datamaskinene. (Chowdhury, 2010, s. 204-205, 251; Saracevic, 2007, s. 1925-1926)
Systemsentrerte og brukersentrerte gjenfinningsmodeller er viktige for min oppgave fordi de vektlegger forskjellige aspekter av gjenfinningen og gjenfinningsevalueringen. Valg av gjenfinningsmodell er derfor knyttet til valg av evalueringsmetode noe jeg skal diskutere nærmere i metodekapittel.

2.1.1 Systemsentrerte gjenfinningsmodeller

Systemsentrerte gjenfinningsmodeller konsentrerer seg om to hovedoppgaver ifølge Baeza-Yates og Ribeiro-Neto. Den ene oppgaven er å definere et logisk rammeverk som gjør det mulig å formalisere både informasjonen i dokumentene og i spørsmålene som blir stilt til systemene ved en søkeprosess. Gjennom formaliseringen oppnår man en dokument- og en spørsmålsrepresentasjon. Den andre oppgaven til gjenfinningsmodellene er å definere en rangeringsfunksjon som skal sammenlikne dokument- og spørsmålsrepresentasjon. Rangeringsfunksjonen brukes til å lage en rangert liste av dokumentrepresentasjonene som skal være systemets svar på søkene representert ved spørsmålsrepresentasjonen. (Baeza-Yates & Ribeiro-Neto, 2011, s. 57-59)
De klassiske gjenfinningsmodellene for ikke-strukturerte tekster er den boolske modellen, vektormodellen og den probabilistiske modellen. (Baeza-Yates & Ribeiro-Neto, 2011, s. 59-60) Modellene oppsummerer ulike løsninger for formaliseringen av dokumentene og spørsmålene, og for sammenligningen av dokument- og spørsmålsrepresentasjonene.
Den boolske modellen er den eldste og enkleste. Den er en *exact match* modell der alle kravene fra spørsmålsrepresentasjonen skal tilfredsstilles i dokumentrepresentasjonen, for at dokumentet skal bli med i trefflisten. Det er derfor ikke mulig å rangere dokumentene etter relevans siden alle dokumentene har samme karakteristikker i forhold til spørsmålsrepresentasjonen. (Nicholas J. Belkin & Croft, 1987, s. 113-114)
I den boolske modellen kombineres søkeordene med de logiske operatorene AND, NOT og OR. Kombinasjonen av termene med operatoren AND gir en treffliste der alle termene skal være med for at dokumentet skal bli vurdert som relevant. Operatoren NOT gir en treffliste der termer ikke skal være i dokumentrepresentasjonen for at dokumentet vurderes som relevant. Med OR-

operatoren kan enten den ene eller den andre termen være med i dokumentrepresentasjonen for at dokumentrpresentasjonen vurderes som relevant. Operatorene kan kombineres og gir ulike muligheter, men det er likevel flere svakheter med modellen. Chowdury nevner blant annet at modellen krever for mye forhåndskunnskap fra brukeren. Ikke alle brukerne er i stand til å bruke operatorene korrekt, og de kan heller ikke vite på forhånd hvor lang trefflisten et gitt spørsmål kan gi. Er spørsmålet for vidt, kan det bli flere sider med treff som ikke er rangert etter relevanskriterier. Mangelen på rangeringsmuligheter gjør at dokumenter som behandler temaer i ulik grad, blir sidestilt. (Chowdhury, 2010, s. 205-207; Salton & McGill, 1983, s. 24-28)

Vektor- og sannsynlighetsmodellene er egenskapbaserte *partial match* modeller. De er kalt for egenskapsbaserte fordi de baserer seg på egenskapene til dokumentene, og ikke på strukturen til dokumentene. De er *partial match* modeller fordi de lager en rangering av dokumentene ved å gradere resultatene. Noen dokumentrepresentasjoner blir vurdert med en større likhetsgrad til spørsmålsrepresentasjonen enn andre, i motsetning til *exact match* modellene. (Nicholas J. Belkin & Croft, 1987, s. 112-117)

I vektormodellene blir dokument- og spørsmålsrepresentasjonene gjort om til vektorer. Vektorene til de to representasjonene består av vekter tilegnet hver indeksterm. Indekstermene er sett som gjensidige uavhengige av hverandre og de vektes ut fra TF-IDF vekter. TF-IDF vekter tar hensyn til termfrekvensen i dokumentet (TF) og til invers dokumentfrekvensen i samlingen (IDF). Termfrekvensen referer til antall forekomster av en term i ett dokument. En term som forekommer flere ganger antas å være viktigere for å representere dokumentet enn en term som har lav forekomst. Invers dokumentfrekvensen i en samling viser til termens evne til å skille mellom ulike dokumenter i samlingen. Hvis termen finnes i flere dokumenter i samlingen, har den lav diskrimineringsverdi fordi den ikke hjelper til å skille dokumentene fra hverandre. En term med lav diskrimineringsverdi skal vektes mindre enn en term som har høy verdi. (Salvesen, 1994, s. 30-31)

Vektormodellene beregner graden av likheten mellom vektorene til dokument- og spørsmålsrepresentasjonen ved å se på avstanden mellom de to representasjonene i en vektorrom. For å måle graden av likheten brukes cosinus-likheten, dvs. cosinus av vinkelen mellom de to vektorene.

Ifølge Baeza-Yates og Ribeiro-Neto er vektormodellen en enkel og rask modell, samtidig som bruken av termvekting forbedrer gjenfinningskvaliteten. Modellen gjør det mulig å sortere dokumentene etter likhetsgraden med spørsmålsrepresentasjonen, og likhetsformelen inneholder normaliseringsmuligheter av dokumentenes lengde. (Baeza-Yates & Ribeiro-Neto, 2011, s. 77-79)

Den siste klassiske gjenfinningsmodellen er den probabilistiske eller sannsynlighetsmodellen.

Det finnes flere probabilistiske metoder. *Probability ranking principle* lager et estimat for hvor sannsynlig det er at dokumenter i en samling er relevante i forhold til et gitt spørsmål, og rangerer dokumentene i trefflisten etter estimatet for relevansen. Estimatet er avhengig av egenskapene ved dokument- og spørsmålsrepresentasjoner, og det tas ikke hensyn til andre faktorer utenfor selve samlingen og spørsmålet. I tillegg tar modellen for gitt at det finnes en viss mengde dokumenter som er relevante for brukeren i samlingen. (Baeza-Yates & Ribeiro-Neto, 2011, s. 79-82; Büttcher, Clarke & Cormack, 2010, s. 258-261) Resultater fra forskning knyttet til modellen, viser ifølge Chowdury at denne modellen ikke gir bedre resultater enn vektormodellen. (Chowdhury, 2010, s. 207-208)

2.1.2 Brukersentrerte gjenfinningsmodeller

Det finnes flere brukersentrerte gjenfinningsmodeller. Blant de mest kjente er Belkins, Saracevics og Ingwersens modeller.

Belkins modell ASK (*Anomalous state of knowledge*) er et eksempel på en brukersentrert modell som er knyttet til det kognitive synspunktet. Belkin undestreker viktigheten av brukerne. Det er menneskene som engasjerer seg og starter en søkeprosess når de erkjenner mangelen på kunnskap som de trenger for å oppnå sine mål. Ifølge Belkin forutsetter en del gjenfinningssystemer at brukeren kan uttrykke det han mangler kunnskap om for at systemet skal finne fram til informasjonen. Brukerens mangel på kunnskap gjør at han må gå gjennom en interaktiv prosess for å finne ut hvordan han skal søke på noe han ikke kjenner i utgangspunktet. Belkin og hans medarbeidere foreslo derfor en gjenfinningsmodell som tar hensyn til prosessen der brukerens behov for informasjon og kunnskap forandrer seg gjennom søkeprosessen. (N.J. Belkin, Oddy & Brooks, 1982, s. 62-70; Kelly & Ruthven, 2011, s. 27)

I Saracevics modell er brukeren og datamaskinen sentrale aspekter som interagerer med hverandre. Ulike nivåer eller strata karakteriserer brukeren og datamaskinen. Brukeren har et kognitivt nivå, et affektivt nivå og et situasjonsnivå. Datamaskinen har et *engineering* nivå, et prosesseringsnivå og et innholdsnivå. Interaksjonen foregår som en kontinuerlig prosess på ulike nivåer. Interaksjonen er gjort mulig av et grensesnitt mellom brukeren og datamaskinen. Saracevic bruker samme stratifiserte modell for å beskrive andre aspekter av gjenfinningen, som relevansbegrepet (Saracevic, 1997, s. 321-324) Relevansbegrepet er viktig også for systemsentrerte modeller der man er avhengig av å bygge en fasit av relevante dokumenter for å evaluere gjenfinningen med presisjon- og fullstendighetsmålene. Hvilke dokumenter er relevante i forhold til søkene i min

oppgave diskuteres i metodekapittelet.

Peter Ingwersen er kjent for sin holistiske modell hvor man tar hensyn til både systemet og brukerne. Modellen ble presentert først i 1992 i boken "Information Retrieval Interaction", og videre utarbeidet sammen med Kalervo Järvelin i boken "The Turn". (Chowdhury, 2010, s. 256-257; Ingwersen & Järvelin, 2005, s. VII) Saracevic ser på Ingwersens modell som en løsning på konflikten som hadde foregått i flere år mellom de som prioriterte å forske på gjenfinningssystemene og utelukket brukerne, og de som i stedet plasserte brukerne i sentrum uten å ta hensyn til systemene. (Saracevic, 2007, s. 1925-1926)

Ingwersen og Järvelin velger et kognitiv og holistisk synspunkt. De ser på informasjongjenfinningen som en kognitiv prosess der interaksjon og kommunikasjon ikke er avgrenset til brukersentrerte tilnærminger til informasjonen. Menneskene er ikke de eneste aktørene i informasjonsprosessene. Informasjonsprosessene kan bli mediert på forskjellige nivåer av datamaskiner eller andre involverte aktører. Enhver aktør i en informasjonsprosess tilhører et sosiologisk, organisatorisk og kulturelt miljø som utgjør konteksten for selve prosessen. Informasjonen er kontekst og situasjonsavhengig. (Chowdhury, 2010, s. 256-257; Ingwersen & Järvelin, 2005, s. 25-27) Den kognitive og holistiske tilnærmingen ser på informasjonsgjenfinningen fra et interaktiv perspektiv, der man har kontinuerlige kognitive prosesser som befinner seg i en kontekst. Den kontinuerlige interaktiviteten mellom aktørene tilsier at det ikke er mulig å finne frem til en endelig harmoni mellom de ulike strukturene. Informasjonen er ikke lenger noe objektivt som kan gis til brukeren i en endelig versjon. Systemene har som funksjon å hjelpe brukerne på vei i en prosess. (Ingwersen & Järvelin, 2005, s. 28-30; Kelly & Ruthven, 2011, s. 10-11)

2.2 Gjenfinningsevaluering

Gjenfinningsevaluering er en systematisk vurdering av kvaliteten til gjenfinningssystemet og av de valgte gjenfinningsmetodene. Evalueringen innebærer at man velger noen kriterier som skal måles mot en standard som er definert på forhånd. Hvilken standard og hvordan man måler kvaliteten har vært et sentralt spørsmål i forskningen om gjefinningsevaluering siden det startet på 50-tallet. Her presenteres noen av de mest kjente evalueringsmål for systemsentrerte gjenfinningsmodeller og noe av forskningen som bidro til å etablere disse målene.

2.2.1 Cranfield-eksperimentene og presisjon- og fullstendighetsmålene

De første eksperimentene innenfor gjenfinningsevaluering er Cranfield-eksperimentene gjennomført fra 50-tallet på College of Aeronautics i Cranfield i Storbritannia, og ledet av Cyril Cleverdon. Disse eksperimentene bidro til å definere de mest kjente evalueringsmålene for gjenfinningssystemene, presisjon (*Precision*) og fullstendighet (*Recall*).
Målene regnes ut fra en rangert treffliste av dokumenter, der dokumentene som er vurdert som mer relevante av systemet blir rangert øverst. Presisjonen er antall relevante dokumenter funnet i forhold til det totale antall dokumenter funnet. (Fig. 1) Presisjonen sier noe om hvor god gjenfinningssystemet er til å unngå irrelevante dokumenter. Fullstendighet er antall relevant dokumenter funnet i forhold til det totale antall relevante dokumenter. (Fig. 1) Fullstendigheten sier noe om hvor god gjenfinningssystemet er til å finne alle relevante dokumenter. (Baeza-Yates & Ribeiro-Neto, 2011, s. 134-139; Salton & McGill, 1983, s. 162)

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Fig.1. Formel for beregning av presisjon og fullstendighet (*Information retrieval*)

Cranfield-eksperimentene beviste også inversforholdet mellom de to målene. I tillegg prøvde man å finne ut av effekten av søk i en samling indeksert med forskjellige indekseringemetoder. Disse ga uventede resultater, der søk i ikke-kontrollerte og ikke-prekoordinerte indekser ga de beste resultatene.
Studiene ble i ettertid kritisert av forskjellige grunner. De første studiene om presisjon og fullstendighet var basert på søkespørsmål konstruert i laboratoriet uten å ta hensyn til faktiske brukerbehov. (Chowdhury, 2010, s. 296-298) Relevansbegrepet som ble brukt for å vurdere om dokumentene svarte til søkespørsmålene var altfor enkel. Brukerbehovet er vurdert som statisk og dokumentenes relevans uavhenging av andre dokumenters relevans. Til tross for dette, ga disse studiene måleenheter som har gjort det mulig å sammenlikne og vurdere forskjellige gjenfinningssystemer og metoder senere. I tillegg åpnet Cranfield-eksperimentene for automatisk

indeksering av samlingene takket være resultatene fra sammenligningene av forskjellige indekseringsmetoder. (Baeza-Yates & Ribeiro-Neto, 2011, s. 132-134)

Presisjon- og fullstendighetsmål i ulike varianter er fortsatt mye brukt i forskningen for å evaluere gjenfinningseffektivitet. Målene gir en verdi for hvert treff i en treffliste. Noen ganger er det mer hensiktsmessig å ha en samlet verdi for alle søkene som gjør det mulig å sammenlikne treffene fra ulike gjenfinningsalgoritmer. Varianter av presisjon og fullstendighet som gir unike verdier er: *Average Precision at n* (P@n), *Mean Average Precision* (MAP) og *R-Precision*.

P@n er gjennomsnittlig presisjonsverdi ved det *n*'te dokumentet for alle søkespørsmålene stilt til et system. *N*-verdien defineres på forhånd, og den er avhengig av hvor langt i en trefflisten man tror brukerne er villige til å lete for å få svar på deres spørsmål. Hvis man velger et lav tall for *n*, viser P@n om systemet er i stand til å rangere de relevante dokumentene øverst. Hvis man velger et høyt tall for *n*, viser P@n om systemet klarer å finne fram til mange av de relevante dokumentene.

MAP er gjennomsnittverdi for presisjon. Man summerer først alle presisjonsverdiene for de relevante dokumentene, deler dem på antall relevante dokumenter for hvert søk, og man regner gjennomsnittet for alle søk. Problemet med evalueringsmålet er at den tar hensyn til presisjonsverdier for dokumenter som ikke befinner seg blant de første treffene i trefflisten, og som brukeren sannsynligvis ikke kommer til å se på.

R-precision er verdi for presisjon på R-posisjon i trefflisten, hvor R er det totale antallet relevante dokumenter. Målet sier noe om hvor mye gjenfinningssystemet avviker fra det perfekte systemet som ville ha rangert alle de relevante dokumentene først. Ulempen er at når det totale antall relevante dokumenter er høy, sier målet veldig lite om hvor god systemet er til plassere de relevante dokumentene øverst på trefflisten.

2.2.2 TREC-eksperimentene

Etter Cranfield-eksperimentene fulgte flere andre eksperimenter slik som Medlars-, SMART- og STAIRS-eksperimentene, alle med utgangspunkt i systemsentrerte gjenfinningsmodeller. Disse eksperimentene bidro til forskningen, men lite til den praktiske uviklingen av gjenfinningssystemene. Ifølge Chowdury var dette en konsekvens av at man brukte for små datasamlinger og at eksperimentene var altfor isolerte fra hvordan en søkeprosess foregår i virkeligheten. Som en konsekvens ble det startet *The Text Retrieval Conference* (TREC) i USA på 1990-tallet. TREC-eksperimentene var basert på større datasamlinger, og de tok seg av flere aspekter ved gjenfinningsevalueringen. (Chowdhury, 2010, s. 302-312) Fra de første "ad hoc tasks"

hvor man sendte en del søk mot en fast database med dokumenter, ble det etter hvert laget gjenfinningsforsøk i forskjellige retninger. De introduserte blant annet gjenfinningseksperimenter av webbdata og av videomateriale, av datasamlinger i ulike språk, av veldig store datasamlinger og studier innenfor datalingvistikk. (Baeza-Yates & Ribeiro-Neto, 2011, s. 161-165; Sparck Jones, 2000, s. 44)

"Ad hoc tasks" til TREC-eksperimentene vurderer resultatene av søkene med ulike presisjon- og fullstendighetsmål ved hjelp av en evauleringspakke kalt "trec_eval". For å kunne regne ut presisjon- og fullstendighetsmålene trenger man en fasit med relevante dokumenter som man kan sammenlikne systemenes søkeresultat med. En slik fasit var ikke mulig å lage manuelt siden samlingene er veldig store. Derfor ble fasiten laget ved å plukke ut de toppscorende treffene etter å ha søkt i alle samlinger i de ulike systemene. Det er disse treffene som blir vurdert manuelt og ikke hele datasamlingene. Metoden brukt er kalt for *pooling method*. Metoden antar at søketreffene som er funnet av ulike systemer og samlinger inneholder en stor del av de relevante treffene, og at de som ikke er øverst i trefflistene ikke er relevante. Metoden er mulig fordi eksperimentene blir gjort i stor skala, både når det gjelder mengde metadata, antall involverte institusjoner og gjenfinningssystemer. (Baeza-Yates & Ribeiro-Neto, 2011, s. 163; Chowdhury, 2010, s. 315-316)

TREC-eksperimentene nærmet seg også etter hvert brukersentrerte gjenfinningsmodeller ved å lage interaktive søk der brukerne fikk en mer sentral plass. "Interactive tracks" konsentrerte seg om forsøk der mennesker kunne interagere med gjenfinningssystemet for å vurdere relevansen av dataene. Fra 2001 begynte TREC å foreta også noen studier med observasjoner, der forskerne brukte offentlige tilgjengelige data på nettet. (Ingwersen & Järvelin, 2005, s. 177-178)

2.3 Tidligere forskning

Flere studier understreker hvordan gjenfinningen i fulltekstsøk er basert på en sammenligning av en dokumentrepresentasjon for et helt dokument med en spørsmålsrepresentasjon, selv om det ofte bare er en liten del av dokumentrepresentasjonen som er relevant for søkespørsmålet. (Bendersky & Kurland, 2010, s. 158; Callan, 1994, s. 302; Kaszkiel & Zobel, 2001, s. 344; Lin, 2009, s. [2]) At dokumentrepresentasjonen har dokumentet som enhet for søket, har konsekvenser for rangeringen av treffene når bare en del av dokumentrepresentasjonen er relevant. Denne problematikken gjelder flere typer fulltekstdokumenter, uavhengige om det er e-bøker, tidsskriftartikler eller andre typer dokumenter. (Craig & Miles, 2013, s. 8)

Passage retrieval er en teknikk brukt for å forbedre gjenfinningen i disse tilfellene. Wade og Allan definerer *passage retrieval* slik: "the task of retrieving only the portions of a document that are relevant to a particular information need" (Wade & Allan, 2005, s. [1]) *Passage*s kan defineres som en kontinuerlig sekvens av tekst som er en del av et større fulltekstdokument. (Kaszkiel & Zobel, 2001, s. [1]) *Passage* har ikke noe tilsvarende term på norsk, og mye av den publiserte forskningslitteraturen er på engelsk. Termen kan heller ikke oversettes med ordet "avsnitt" som tilsvarer *paragraph* på engelsk. *Paragraphs* er knyttet til dokumentenes struktur og er en type *passage.* Jeg kommer derfor til å bruke det engelske uttrykket for *passage* i oppgaven.
Min oppgave skal sammenligne gjenfinningen i programmer med gjenfinningen i innslag. Programmene er dokumenter av blandet innhold hvor innslagene er de semantiske enhetene. Innslagene er en del av et større dokument og er beskrevet med en fulltekst og et kontrollert vokabular. De er *passages* i mitt gjenfinningsforsøk.
Studier av p*assage retrieval* har funnet sted siden 70-tallet*,* men interessen økte spesielt på 2000-tallet når det ble vanlig med søk i fulltekstdatabaser i steden for søk i sammendrag av dokumentene (Wade & Allan, 2005, s. [1]) *Passage retrieval* kan ses fra ulike perspektiver og det er flere aspekter av gjenfinningen som belyses av forskningen. Jeg har valgt å presentere noen studier i *passage retrieval* som fokuserer på karakteristikkene av de ulike typer *passages*, studier som ser på rangeringen av *passages* i trefflistene, de som undersøker ulike gjenfinningsmodeller for *passage retrieval og* en studie som bruker *passage retrieval* for å forbedre rangeringen av dokumentene i trefflistene.
Metadatasamlingen for programmet "Ukeslutt" består både av en innholdsbeskrivende fulltekst og et kontrollert vokabular. Jeg har derfor også ett avsnitt som belyser noen av aspektene ved søk i fulltekst og i kontrollert vokabular, og som er av interesse for gjenfinningsevalueringen.

2.3.1 Bruk av ulike typer *passages*

Kaszkiel og Zobel (2001) ønsker å evaluere gjenfinningen av ulike typer *passages* og mener at en sammenligning av disse ikke er mulig ved å bruke tidligere gjennomførte studier fordi dataene er fra helt forskjellige datasamlinger. De lager derfor et gjenfinningseksperiment der de deler samme datasamling i forskjellige typer *passages*.
De bruker J.P. Callans (1994, s. 302) inndeling av *passages* i tre hovedgrupper for deres datasamling:

- *Discourse passages.* Hvert dokument blir delt etter strukturelle og logiske komponenter slik som setningene, avsnitt, seksjoner eller strukturen i filene. Fordelen med denne typen inndelingen er at den er intuitiv ifølge Kaszkiel og Zobel. Setninger, avsnitt og seksjoner pleier å gjenspeile semantiske enheter slik som ideer, emner, temaer. Problemet med inndelingen er at den forutsetter en høy grad av konsistens mellom forfatterne av tekstene for at *passages* skal kunne representere samme type enhet. (Kaszkiel & Zobel, 2001, s. 348)

- *Semantic passages.* Hvert dokument blir delt etter tematiske enheter. Dette er vanskelig å oppnå ved en automatisk inndeling. Ifølge Kaszkiel og Zobel er det i 2001 ennå ikke oppnådd så gode resultater for en automatisk inndeling av teksten i forhold til en manuell inndeling. (Kaszkiel & Zobel, 2001, s. 348-349)

- *Window-based passages*. Hvert dokument blir delt etter en fast forhåndsbestemt lengde uavhengig av tekstens struktur. Metoden kan brukes der strukturen i teksten ikke er så lett å kjenne igjen, eller der den semantiske inndelingen skiller seg for mye fra strukturen i teksten. Fordelen med metoden er at den er lett å lage, men den kan bli vanskelig å forstå hvis den blir presentert som resultat for søk fordi den er en statisk inndeling. (Kaszkiel & Zobel, 2001, s. 349-350)

Datasamlingen de bruker består av 5 ulike samlinger fra TREC-eksperimentene. De bruker to typer søkespørsmål mot samlingene, korte og lange, for å se på eventuelle forskjeller i rangeringene av treffene. Hele dokumenter blir funnet på basis av relevansen til deres *passages*, og der det er flere relevante fra samme dokument blir det på basis av *passage* som har høyest relevans. De velger vektormodellen som gjenfinningsmodell. Gjenfinningseffektiviteten blir målt med presisjon og fullstendighet, og ved hjelp av statistikk for å se om forskjellene er statistisk signifikante. For å få til en samlet verdi av presisjonen i de rangerte trefflistene bruker de P@5, P@10, P@20, P@30, P@100. (Kaszkiel & Zobel, 2001, s. 346, 350) Til tross for en del variasjoner i de ulike testene, er deres konklusjon at *passage retrieval* er mer effektiv enn søk med fulltekstdokumenter. Ingen av de ulike *passages* skiller seg spesielt ut for å være bedre enn de andre i de forskjellige forsøkene. De innfører derfor en ny type inndeling kalt for *Variable-length arbitrary passage,* hvor størrelsen og inndelingen av *passages* gjøres etter at en evaluering av søket er gjort. Denne nye inndelingen ga litt bedre resultater enn de andre typer *passages*. (Kaszkiel & Zobel, 2001, s. 359-362)

Fernández, Azzopardi og Losada (2011; 2012) konsentrerer seg om s*entence retrieval*, der *passages* tilsvarer setninger i dokumentene. I studiet fra 2012, mener Fernández og Losada at mange av gjenfinningseksperimentene som bruker *sentence retrieval* egentlig er tilpasninger av de generelle gjenfinningsmetodene for søk i dokumentene. Disse beregner likhet mellom søkespørsmålene og *passages* ut fra deres felles termer. Dette ser de som problematisk for søk i så korte tekster som setninger som inneholder langt færre termer enn lengre dokumenter. I noen studier er det brukt spørsmålsutvidelse som en løsning for dette, noe de er kritiske til. De utvikler derfor en modell som bruker informasjon som er spørsmålsuavhengig (*Query-independent information*) for å forbedre gjenfinningen. De bruker *opinion-based features*, *named entities* og *length based features* som uavhengige egenskaper for vekting. Datasamlingen er hentet fra TREC-eksperimentene fra 2002-2004, og de bruker en vektormodell for gjenfinningen. (Fernández & Losada, 2012, s. 1203-1205) Resultatene viser at av de tre valgte utvidelsene for vekting, er det bare *opinion-based features* og *length based features* som forbedrer gjenfinningen i *sentence retrieval*. (Fernández & Losada, 2012, s. 1226-1227)

Initiative for the Evaluation of XML Retrieval (INEX) ble startet i 2002 for å bidra med bygging av samlinger, infrastruktur og nye evalueringsmetoder innenfor gjenfinningsevaluering av systemer som søker i XML-innhold. INEX har hatt fokuset på gjenfinningen i deler av dokumentene hvor xml-strukturen i dokumentene er enhetene for søket. (Fuhr, 2005, s. V; Ruthven, 2008, s. 66-67) Før 2007 var det bare gjenfinningen i XML-elementene, som var fokuset for eksperimentene, men fra 2007 ble andre former for *passage retrieval* introdusert sammen med søk i boksider. Jenkinson og Trotman (Fuhr, 2008, s. 426-439) utfører en evaluering av *passage retrieval* som en del av "Ad hoc track" til INEX 2007 med datasamlingne hentet fra Wikipedia. I eksperimentet definerer de *passages* som deler av dokumentene med en fast lengde på ca. 300 termer og som er uavhengige av dokumentenes XML-struktur. Disse tilsvarer *window-based passages* i J.P. Callans definisjon av *passages*. Andre INEX-eksperimenter bruker XML-elementer som *passages*, dvs *discourse passages* i Callans definisjoner. (Callan, 1994, s. 302) Forskerne velger *window-based passages* for at det skal bli mulig å gjøre en sammenligning med *passages* basert på XML-strukturen. Ved en sammenligning kan man se om XML-strukturen i dokumentene faktisk har verdi for å finne det som er relevant i et dokument.
Hvordan de definerer starten og slutten av *passages* siden denne ikke er definert av dokumentenes struktur blir viktig for gjenfinningen. De velger *passages* ut fra det området som har høyest densitet av relevans i en tekst. Det relevante området og teksten før og etter den, er den relevante *passage.*

(Fuhr, 2008, s. 428-431) Eksperimentet sammenlikner ikke direkte ulike typer *passages*, men ser blant annet om rotlemmatisering av termene har innflytelse for søk i *passages* og for å definere området (*Window*) for *passages*. Hypotesen deres er at rotlemmatisering ikke har noe særlig innflytelse i større tekster der de ulike språklige variantene av en term har mer sannsynlighet for å bli dekket i teksten. Den bør i stedet spille en rolle for mindre tekster som *passages*, hvor sannsynligheten for at flere varianter av samme term finnes, er mindre. Eksperimentet viser i stedet at rotlemmatisering ikke har noe særlig funksjon for valg av området for *passages*. De ser et behov for mer testing innefor temaet. (Fuhr, 2008, s. 434)

2.3.2 Rangeringen av *passages* i trefflistene

Rangeringen av *passages* i trefflistene kan bli problematisk når rangeringen skal sammenliknes med søk i de samme dokumentene behandlet som en helhet. *Passages* må i disse tilfellene representeres med sine dokumenter for at de to trefflistene skal være sammenlignbare og for å regne presisjon og fullstendighet mot en felles fasit. Så lenge bare en *passage* er funnet, kan denne representeres med sitt dokument, men når to eller flere *passage*s fra samme dokument er i samme treffliste, må dokumentet få en plass i trefflisten som tar hensyn til posisjonen for de ulike *passages* den skal representere. I mitt gjenfinningsforsøk skal også innslagene representeres med sine programmene i trefflistene for å kunne sammenligne treff i innslagene med treff i programmene. I tillegg kan flere innslag fra samme program bli funnet i ett søk.
Et gjenfinningseksperiment som belyser problematikken ble utført i 2009 av Jimmy Lin. (Lin, 2009) Studiet skulle først og fremst sammenlikne søk i fullteksten og i sammendragene av tidsskriftartikler, men han inkluderte også *passage retrieval*. Målet var å se om å søk i fullteksten er mer effektivt enn søk i sammendraget til artiklene. Han brukte to forskjellige gjenfinningsmodeller, Okapi *bm25* og Lucenes egen vektorbaserte modell, og søkte i tre indekserte datasamlinger. Metadataene er hentet fra *Genomics Track Evaluation* (TREC 2007). De samme metadataene er delt i tre samlinger. Den ene består av titlene og sammendragene til tidsskriftartikler, den andre av fullteksten til artiklene, og den tredje av fullteksten til artiklene med avsnittene (*paragraphs*) som enhet for indekseringen. Han bruker avsnitt som er en type *discourse passage*, for gjenfinningseksperimentet. Trefflisten for datasamlingen indeksert etter avsnittene, måtte modifiseres slik at de valgte effektivitetsmålene kunne regnes mot en felles fasit som var basert på id-numrene til hele dokumenter. Lin valgte å bruke to metoder som først ble foreslått av Hearst og Plaunt (1993, s. 64):

- *Maximum of supporting spans*: med denne metoden gir man dokumentet score for *passage* som er rangert først i trefflisten. Alle *passages* fra samme dokument samles derfor på posisjonen til den mest relevante *passage* i den rangerte listen. Metoden favoriserer de artiklene som bare har én *passage* som er relevant, og som er rangert høyt oppe i trefflisten. (Lin, 2009, s. [5])

- *Sum of supporting spans*: med denne metoden summeres score for alle *passages* funnet i trefflisten, og dokumentet som representerer dem blir rangert etter den nye score. Denne metoden favoriserer artikler som har flere *passages* i trefflisten. (Lin, 2009, s. [5])

Gjenfinningseffektiviten i eksperimentet ble målt med MAP, P@20 og IP@R50 (Interpolert presisjon med en fullstendighet på 50%). Søket i sammendragene ga ikke bedre resultater enn fulltekstsøket, og eksperimentet bekreftet at fulltekstsøk er like effektivt. *Passage retrieval* ga også bedre resultater enn søk i sammendrag. Det var spesielt den første metoden *Maximum of supporting spans* som ga bedre resultater. En mulig forklaring for dette kan finnes i lengdenormaliseringen. (Lin, 2009, s. [8-10])
Ifølge Lin har *passage retrieval* lite effekt på presisjonen og mer effekt på fullstendigheten. At resultatene for fullstendigheten er bedre ses gjennom MAP-verdien. Han forklarer dette med at sentrale temaer for en tidsskriftartikkel vanligvis blir repetert flere ganger gjennom selve artikkelen. Ved deling i *passages* får man flere tekster om de temaene som er repetert flere ganger, og disse har en større sjanse for å bli funnet igjen ved søk enn når man har hele fullteksten. (Lin, 2009, s. [8-10])

Bendersky og Kurland (2010) bruker en probabilistisk modell for rangeringen av de enkelte *passages* i stedet for de tradisjonelle og mest brukte rangeringsmetodene. De mener at denne probabilistiske modellen forener flere av de tidligere brukte modellene hvor man valgte den ene eller den andre rangeringsmetode. De slår sammen opplysninger hentet direkte fra dokumentet som helhet med opplysninger fra *passages* til dokumentet, og bruker disse i sin probabilistiske beregning av rangeringen. For at dette skal fungere, lager modellen et estimat av dokumentets homogenitet. Tanken er at når et dokument er homogent, kan flere termer fra dokumentet som helhet brukes for å representere det. Det er i stedet det motsatte når et dokument inneholder flere ulike emner. Styrken i modellen ifølge forskerne, er at de ikke tar for gitt at alle dokumentene i en samling er like

homogene, noe som har vært problematisk i tidligere studier. De mener at deres modell gir bedre effektivitet. (Bendersky & Kurland, 2010, s. 158-161, 165)

2.3.3 Bruk av forskjellige gjenfinningsmodeller

Flere av studiene i *passage retrieval* bruker ulike gjenfinningsmodeller i samme eksperiment for å se eventuelle variasjoner for gjenfinningen.

J. Lin (2009) velger en probabilistisk modell (Okapi bm25*)* og Lucenes gjenfinningsmodell som er vektorbasert, for sin sammenligning av søk i fulltekstartikler, i sammendragene og søk i artiklene delt i *passages*. I analysen sammenlikner han hovedsakelig søkene i de forskjellige metadataene, men han viser også til noen forskjeller i effektivitetsmålene ved de to gjenfinningsmodellene. Søket i sammendrag får for eksempel bedre verdier av MAP med *bm25* enn søket i de samme dokumentene med Lucenes gjenfinningsmodell. (Lin, 2009, s. [5-6])

Wade og Allan (2005) utførte et gjenfinningseksperiment der de sammenlikner syv forskjellige gjenfinningsmodeller. Disse er ulike varianter av vektor- og språkmodeller. (Wade & Allan, 2005, s. [2-5]) Dataene er hentet fra *TREC High Accuracy Retrieval from Documents-track* (HARD-track) for 2003 og 2004. Eksperimenter innenfor HARD-track ble gjennomført mellom 2003 og 2005, og tok i bruk forskjellige gjenfinningsteknikker som blant annet *passage retrieval* for å oppnå høy nøyaktighet i gjenfinningen. (*TREC tracks*, 2002)
Wade og Allan bruker samme inndeling i *passages* som er brukt i TREC HARD-track, der *passages* er definert som en kontinuerlig og ikke-tom tekststreng som tilhører ett dokument. *Passages* kan være av svært ulik lengde. Relevansvurderingene av treffene ble gjort ved at en gruppe personer gikk manuelt gjennom trefflistene, og pekte på hvilke dokumenter og hvilke områder i disse som var relevante i forhold til søkespørsmålene. Resultatene av søketreffene viser til hele dokumenter, og det blir brukt samme metode som Jimmy Lin kaller for *Maximum of supporting spans* for rangering av *passages*. (Wade & Allan, 2005, s. [3])
Resultatene viser at de gjenfinningsmodellene som gir best resultater, har tendenser til å gjøre det med alle de valgte målene. Samtidig er ikke resultatene veldig gode for noen av modellene eller målene. Ifølge forskerne viser resultatene at det fortsatt gjenstår en del arbeid med *passage retrieval*. De foreslår spesielt at videre eksperimenter bør gjøres med spesielle oppgaver som mål,

og at det bør brukes datasamlinger med lengre dokumenter som består av forskjellige emner. (Wade & Allan, 2005, s. [7])

2.3.4 *Passage retrieval* for rerangering av treffene

Passage retrieval er også brukt for å forbedre rangeringen av søketreffene. Eyal Krikon og Oren Kurland bruker i 2011 både søk i hele dokumenter, clustermetode og *passage retrieval* for å forbedre rangeringen av søketreff. Datasamlingen og søkespørsmålene til studiet er hentet fra TREC-eksperimentene. Fokuset er på verdiene av P@5 og P@10 siden de er opptatt av toppen av trefflistene. (Krikon & Kurland, 2011, s. 601-602) De bruker P@n verdier ved søk i hele dokumenter, søk med clustermetode og søk med *passage retrieval*. Disse blir sammenliknet med P@n verdiene der treffene er rangert på nytt med ulike kombinasjoner av metodene. De finner ut at enhver ulik kombinasjon av to av metodene er bedre en rangeringen ved bruk av bare en av metodene. Å bruke alle de tre metodene samtidig gir bedre resultater enn både å bruke én eller to av metodene av gangen. Dette viser at det er en potensiell forbedring av gjenfinningen ved å integrere de ulike metodene, selv om det fortsatt er et forbedringspotensiale. (Krikon & Kurland, 2011, s. 604-605) Resultatene støtter også deres antakelse at *cluster* og *passage retrieval* er komplementære metoder. Clustermetoden har fokuset på å utvide dokumentrepresentasjonen ved å se på dokumentene i sin kontekst. *Passage retrieval* fokuserer på at et lengre dokument består av ulike deler der bare en del av hele teksten kan være relevant i forhold til et søk. Der den ene metoden utvider, snevrer den andre inn dokumentrepresentasjonen, og bruk av begge metodene i gjenfinningen forbedrer søkeresultatene. (Krikon & Kurland, 2011, s. 594-595, 612)

2.3.5 Søk i kontrollerte vokabularer og i fulltekst

Interessen for sammenligningen av søk i kontrollerte vokabularer og i fulltekst begynte allerede i 1950-årene når de første søkene for fulltekst ble mulige. Flere studier har forsøkt å sammenligne gjenfinningen basert på fulltekst med gjenfinningen basert på det kontrollerte vokabularet. Dette har ifølge Lancaster ofte vært mislykket. En del av forsøkene peker på at fullstendigheten er bedre der teksten er lenger, mens verdiene på presisjon blir dårligere. Hovedgrunnen er at en lenger tekst har større sjanser for å inneholde flere termer fra søkespørsmålene, enn en kortere tekst. Ifølge Lancaster sier ikke dette noe om forskjellene mellom søk i fulltekst og kontrollert vokabular, men noe om uttømmenheten i indekseringen. (Lancaster, 2003, s. 252-254)

Det samme påpekes også av Hemminger et al. i deres gjenfinningsforsøk hvor søk i fulltekst og søk i metadataene sammenlignes. Deres eksperiment skal derfor ikke bare se om søk i fulltekst gir flere resultater i en *exatct match* gjennfinningsmodell, men også om disse er bedre enn søk i metadata (Hemminger, Saelim, Sullivan & Vision, 2007, s. 2347) Tittel og sammendrag for tidsskriftartiklene er metadataene i gjenfinningsforsøket, mens fullteksten er selve teksten i artiklene. Vurderingene av relevans for forsøket er basert på de samme metodene som er brukt i *TREC Genomics Track ad hoc retrieval* fra 2006. Søk i fullteksten får ikke overraskende bedre resultater for fullstendighet, og noe dårligere resultater for presisjon i forhold til søk i metadataene. (Hemminger et al., 2007, s. 2348) Samtidig får treffene for søk i metadata bedre resultater på relevans enn treffene fra søk i fulltekst. Det er ifølge forfatterne en svakhet med undersøkelsen at det brukes en *exact match* modell. Mer avanserte modeller kan gi bedre resultater for søk i fulltekst. I tillegg inneholder ikke metadataene noe kontrollert vokabular, men bare tittel og sammendrag. Disse er metadata i naturlig språk slik som fullteksten. (Hemminger et al., 2007, s. 2350-2351)

I en teoretisk artikkel peker Jeffrey Beall utelukkende på svakhetene ved fulltekstsøk i forhold til søk i kontrollert vokabular. Han nevner problemene med synonymer, homonymi, og med å identifisere det konseptuelle innholdet for en tekst som typiske svakheter for fulltekstsøk. (Beall, 2008) Lancaster nevner også noen av de samme ulempene for søk i fulltekst, men han balanserer bildet ved å beskrive både fordeler og ulemper med søk i fulltekst og i kontrollerte vokabularer. Fullteksten er som oftest lenger, og har derfor flere termer man kan få treff i. Fullteksten inneholder mer spesifikke termer, og termer som er moderne og i bruk da teksten ble skrevet. Dette kan gi større redundans i gjenfinningssammenheng. Det kontrollerte vokabularet kan i stedet bidra til bedre konsistens mellom forskjellige dokumenter. I tillegg bidrar det til å identifisere de overordnede konseptuelle termene som oppsummerer betydningen av en hel fulltekst. (Lancaster, 2003, s. 265-270) I mitt forsøk skal jeg se om det er mulig å finne ut hvilken funksjon det kontrollerte vokabularet har for gjenfinningen i forhold til fullteksten i metadataene for radioprogrammet "Ukeslutt". Da er det interessant å se om det kontrollerte vokabularet og fullteksten bidrar til gjenfinningen, slik som Lancaster hevder.

3 Metode

Masteroppgaven har som formål å evaluere gjenfinningseffektivitet ved søk i en datasamling med et gitt gjenfinningssystem. Målet er å se på eventuelle konsekvenser for gjenfinningen når datasamlingen blir indeksert med forskjellige enheter for søk. I den ene indeksen er metadataene om ulike nyhetsinnslag fra samme program registrert samlet, i den andre er nyhetsinnslagene indeksert hver for seg. Datasamlingen som er valgt som case, er en del av dataene i NRKs SIFT-base.
Jeg skal diskutere de metodologiske sidene og noen praktiske valg knyttet til metadatasamlingen og indekseringen av metadataene, søkespørsmålene, evalueringsmålene, fasit for søkene og gjenfinningssystemet.

3.1 Metodologisk og teoretisk tilnærming

Masteroppgaven er tenkt som et gjenfinningsforsøk der man stiller et visst antall søkespørsmål til to datasamlinger som inneholder de samme dataene indeksert med to forskjellige metoder. Søkeresultatene blir analysert ved hjelp av kjente effektivitetsmål, og sammenliknet. Fokuset i oppgaven er på hvordan de to representasjonsmodellene påvirker gjenfinningen og på evalueringen av systemets evne til å finne til emnespesifikke data. Effektivitetsmålene gir kvantitative data, men de analyseres kvalitativt ved en fortolkning av resultatene. Størrelsen på datasamlingen og utvalget av søkespørsmål er altfor begrenset for å kunne trekke statistiske og overførbare konklusjoner som er karakteristiske for kvantitative analyser. Resultatene kan derfor ikke generaliseres til å gjelde hele NRKs metadatasamling i SIFT-basen. Fokuset er på en kvalitativ dybdeanalyse av kvantitative data. Selv om det ikke er mulig å generalisere funnene ved en kvalitativt analyse, skal jeg se på eventuelle fellesaspekter og generelle tendenser. (Johannessen & Tufte, 2002, s. 77-81; Ringdal, 2007, s. 91-93)
Det blir ikke involvert brukere i søkeprosessen eller i evalueringen av søkeresultatene.
Med utgangspunkt i Ingwersen og Järvelins teorier om et rammeverk for interaktiv informasjonsgjenfinning, vil mitt gjenfinningsforsøk plassere seg i det de kaller for *test-collection approach* der man måler fullstendighet, presisjon, effektivitet og kvalitet på informasjonen og prosessen. (Ingwersen & Järvelin, 2005, s. 322) Faktorer som søkekonteksten, oppgavekonteksten og den sosioorganisatoriske og kulturelle konteksten blir ikke tatt hensyn til her. Masteroppgaven kan dermed plasseres teoretisk blant evalueringer av systemsentrerte gjenfinningsmodeller.

Jeg utelukker brukerperspektivet som Ingwersen og de andre teoretikerne av de brukersentrerte gjenfinningsmodellene påstår er sentrale faktorer i en søkeprosess. Ifølge Järvelin utelukker en del studier brukeren av økonomiske grunner og for å ha bedre kontroll på designet av undersøkelsene. (Kelly & Ruthven, 2011, s. 117-118) Også i mitt tilfelle er disse noen av grunnene for valget om å ikke involvere brukerne. Gjenfinningssystemet har ikke en søkeapplikasjon som kan brukes eksternt, og det er derfor ikke mulig for brukere å søke direkte i systemet. Tidsperspektivet og størrelsen på gjenfinningsforsøket er en annen avgjørende faktor for valget. En del av kriteriene for valg av spørsmålene kombinert med samlingsstørrelsen spiller også en rolle for valget av den metodologiske tilnærmingen. Datasamlingen er ikke så veldig stor i forhold til andre samlinger med fulltekstdata. Spørsmålene må velges ut fra om temaene er behandlet i samlingen. Dette fører til at man ikke kan ta utgangspunkt i reelle brukerbehov eller brukeroppgaver.
Det involveres ikke reelle brukere og det skapes derfor et fiktivt brukerbehov gjennom søkespørsmålene. Dette gjør at min oppgave skaper en virkelighet som er reell bare i eksperimentet. Dette er ifølge Wallace og Van Fleet (2012) både fordelen og ulempen med disse metodene. Det er en fordel fordi den skapte virkeligheten er mer kontrollerbar enn virkeligheten som analyseres ved bruk av andre forskningsmetoder. Ulempen er at denne virkeligheten er skapt som et bilde av den reelle verden, og den er avhengig av en representasjon som ikke nødvendigvis representerer verden slik som den er. (Wallace & Van Fleet, 2012, s. 235-236) Fidel definerer eksperimentene som studier som introduserer kunstige komponenter ved undersøkelsen. De holder enkelte elementer ved eksperimenter under kontroll for å sette fokus på de elementene som skal undersøkes. (Kelly & Ruthven, 2011, s. 63-64) Analysen skal ta hensyn til at eksperimentet er basert på en virkelighet som er skapt for formålet med studiet, og ikke nødvendigvis representerer den virkeligheten som brukerne befinner seg i når de søker i materialet fra NRKs databaser.

3.2 Datamateriale

Valget og forarbeidet med datamaterialet er knyttet til en del metodologiske, men også praktiske avgjørelser som måtte tas underveis i prosessen for å kunne utføre gjenfinningseksperimentet. Her følger en beskrivelse av metadataene og de metodologiske begrunnelser for valgene tatt i denne prosessen.

3.2.1 Valg av datasamlingen

Datasamlingen jeg bruker i gjenfinningseksperimentet er et utvalg av metadata registrert i NRKs SIFT-base. Instituttet for arkiv-, bibliotek- og informasjonsfag ved HiOA har fått tilgang til basen i forbindelse med prosjektet *Transforming the Organization and Retrieval of Cultural Heritage* (TORCH). Basen inneholder metadata om flere av NRKs radio- og tvprogrammer registrert tom 2013. Den består av over 750.000 registrerte poster.

Jeg har valgt å jobbe med metadataene fra radioprogrammet "Ukeslutt" som er registrert i SIFT-basen. Valget av programmet "Ukeslutt" er basert på to kriterier. Det ene er at jeg trengte datamateriale med tilstrekkelig innholdsbeskrivende metadata. De innholdsbeskrivende metadataene skulle være en kombinasjon av emneord og en lenger beskrivende tekst som kunne brukes som fulltekst. "Ukeslutt" har i gjennomsnitt 1141 tegn for hvert registrert program, og er et av programmene med flest antall tegn i gjennomsnitt. I tillegg har flere av postene emneord og tagger i innholdsbeskrivelsen i tillegg til en innholdsbeskrivende fulltekst.

Det andre kriteriet for valget av datasamlingen er knyttet til type innholdsmetadata som er brukt for å beskrive de forskjellige programmene.

Innholdsmetadata som beskriver bilde- og videomateriale skiller seg ut fra metadata som tradisjonelt blir brukt for skrevne dokumenter. Innholdsmetadata til skriftlig materiale representerer som oftest temaet som behandles i dokumentet. Innholdsmetadata til bilde- og videomateriale kan representere ulike sider ved materialet: de kan representere materialets hovedemne eller emner slik som er vanlig for skriftlig materiale, eller de kan representere hva materialet tilfeldigvis illustrerer, eller beskrive miljøet rundt det som skjer i materialet. (Lancaster, 2003, s. 200-202) I NRKs SIFT-base finnes det programmer som har en innholdsbeskrivelse som tilsvarer Lancasters beskrivelse av indekstermer for bilde- og videomateriale. Det mest tydelige eksempelet på dette er metadataene om "Radioteater" hvor innholdsbeskrivelse består av beskrivelser av scenesetting og skuespillernes bevegelser.

"Ukeslutt" er et av programmene hvor postene har innholdsmetadata som mest ligner på metadata til skriftlig materiale, og er derfor valgt for gjenfinningsforsøket. Dette gjør det mulig å søke i datamateriale på politiske begivenheter i Norge eller internasjonalt, på begivenheter fra kulturlivet, idrettsarrangementer eller andre temaer knyttet til samfunnet generelt.

Jeg har vurdert om jeg kunne tatt metadata fra flere nyhetsprogrammer og avgrenset det på en tidsperiode. Dette ville ha komplisert bearbeidelsen av dataene for søking, fordi man kan ha hatt ulike indekseringspraksis for de ulike programmene.

Datasamlingen er ikke så stor siden den består av metadata fra bare ett av NRKs programmer som er registrert i basen. Dette er et av aspektene som Cranfield-eksperimentene og en del av den senere forskningen, ble kritisert for. Derfor skal dataene analyseres kvalitativt, og overførbarhet av funnene til hele NRKs samling er noe begrenset. Fordelen med en slik samling er at jeg har hatt større kontroll over gjenfinningsforsøket i forhold til en samling som er et tilfeldig utvalg av data. (Ringdal, 2007, s. 110-112)

3.2.2 Beskrivelse og valg av metadataene

Datasamlingen består av 1666 poster, hvor hver post inneholder metadata for en radiosending av "Ukeslutt". Utvalget er resultatet av et søk på alle poster i SIFT-basen som har tekststrengen "Ukeslutt" registrert i tittelfeltet.
Det er 17 mulige felter med metadata i beskrivelsen av hvert program, men ikke alle er fylt ut i alle postene. Feltene inneholder ulike typer metadata. Disse kan beskrives ved hjelp av NISOs inndeling av metadata i tre kategorier: administrative, strukturelle og deskriptive metadata. Administrative metadata er knyttet til institusjonen og deres håndtering av metadatasamlingen, og kan ha opplysninger om opprettelsen og rettigheter for dokumentet. Strukturelle metadata har med organisering av dokumentene å gjøre. Deskriptive metadata er beskrivelser av dokumentene slik som titler, sammendrag eller annen fulltekst, emneord, opphavspersoner. (NISO, 2004, s. 1) I Fig. 2 er det en kort beskrivelse av feltene, og hvilken type metadata de representerer.

Feltnavn	Kort forklaring	Type metadata
KEYS	Nøkkel - unik identifikator for posten	Administrative metadata
Reg-dato	Datoen posten er registrert	Administrative metadata
Reg-nr	Registreringsnummer (Samme verdi som KEYS)	Administrative metadata
Signatur	Ukjent	
Avl-dato	Avleveringsdato	Administrative metadata
Korrektur	Ukjent	
Arkivnr	Arkivnummer	Administrative metadata
Tittel	Tittel på programmet (Både programnavn og tittel spesifikk for det enkelte programmet)	Deskriptive metadata
Opptaksdato	Dato for når programmet ble laget	Deskriptive metadata
Sendedato	Dato for når programmet ble sendt	Deskriptive metadata
Kuttnr	Ukjent	
Redaksjon	Forkortelser for redaksjonen	Deskriptive metadata
Tid	Lengde på programmet	Deskriptive metadata
Medium	Medium programmet er lagret i	Administrative metadata
Prog-leder	Programleder (Ett eller flere navn)	Deskriptive metadata
Innhold	Innholdsbeksrivende fulltekst, emneord eller tagger, og klassifikasjonsnummer	Deskriptive metadata
Sist-endret	Ukjent	

Fig.2. Metadatafeltene for radioprogrammet "Ukeslutt"

Dataene dekker sendingene fra 1977 til 2013. Det er ikke alltid samsvar mellom datoene registrert i feltet "Sendedato" og "Reg-dato", spesielt for de eldste programmene. Registreringen av enkelte av dataene i SIFT-basen har da enten foregått mye senere enn sendingene, eller dataene har blitt overført fra et annet system.

Jeg har valgt å bare bruke feltet "Innhold" i mitt gjenfinningsforsøk. I innholdsfeltet finnes de innholdsbeskrivende metadataene for de ulike nyhetsinnslagene som blir sendt i løpet av et program. Innholdsfeltet inneholder både en fulltekstbeskrivelse, emneord eller tagger, og klassifikasjonsnummer for de ulike innslagene. Nyhetsinnslagene kan ses som semantiske enheter som representerer en sak, men enheten for registreringen er programmene og ikke innslagene. Emneord eller tagger som tilhører forskjellige innslag er også samlet i dette feltet. Emneord og klassifikasjonsnummer er brukt i dataene registrert frem til 2011. Fra 2012 innførte NRK en ny indekseringspraksis. Den nye praksisen innebærer blant annet at emneord og klassifikasjonsnumrene blir erstattet med tagger. (Søbak, 2013, s. 30-36)

3.2.3 Forarbeidet med metadataene

Forarbeidet med metadataene har hatt som mål å lage to samlinger med de samme innholdsbeskrivende metadataene, men indeksert etter ulike enheter for søk. Den ene samlingen gjenspeiler indekseringen i SIFT-basen til NRK, der hver post har innholdsmetadata fra de ulike nyhetsinnslagene som tilhører samme radioprogram samlet. Den andre samlingen bruker nyhetsinnslagene som enheter og ulike innslag fra samme program blir delt i ulike felter. Innslagene kan betraktes som semantiske enheter i programmene siden hvert innslag handler om en sak eller et tema. Inndelingen etter innslag tilsvarer derfor inndelingen i *semantic passages* beskrevet av Callan og av Kaszkiel og Zobel. (Callan, 1994, s. 302; Kaszkiel & Zobel, 2001, s. 348-349)

Alle postene fra SIFT-basen ble levert samlet i én tekstfil. Dataene ble lest inn i en tabell i en Oracle database ved hjelp av et java-program. I tabellen utgjør de opprinnelige postene for hvert program en rad, og feltene fra postene ("Innhold", "Sendedato, "Keys", mm.) er kolonner.

I tillegg ble det laget en ny tabell hvor hvert program ble splittet opp i innslag. Fulltekst, emneord og tagger ble registrert i egne kolonner for å kunne analysere deres rolle i gjenfinningsforsøket.

Det var ikke noen tydelig struktur i feltet "Innhold" som kunne brukes for å automatisk dele hvert program i de respektive innslagene. Det var heller ikke noe gjenkjennelig struktur eller repetisjon av tegn som gjorde det mulig å plukke emneord og tagger automatisk fra fulltekstbeskrivelsen. Registreringen i SIFT-basen har foregått gjennom flere år, og antageligvis gjennomført av ulike personer og uten en streng struktur innenfor hvert felt.

Det finnes flere algoritmer som er utarbeidet for å automatisk dele *passages* etter semantiske enheter der det ikke er en struktur i metadataene som skiller mellom ulike emner. Kaszkiel og Zobel refererer til flere algoritmer som forsøker å dele tekster i semantiske enheter, men konkluderer at det ikke er bevis på hvor vellykket disse algoritmene er. (Kaszkiel & Zobel, 2001, s. 348-349)

Andre typer inndelinger er også vurdert, men strukturen og karakteristikkene ved metadataene egner seg ikke til disse inndelingene. *Discourse passages* baserer inndelingen på strukturelle og logiske komponenter slik som setninger, avsnitt og seksjoner. Disse antas å være tematiske enheter i en større fulltekst. (Kaszkiel & Zobel, 2001, s. 348) I metadataene fra "Ukeslutt" er det ikke konsekvent bruk av skilletegn for avsnitt eller setninger slik det er i fulltekstbaser med tidsskriftartikler.

I *Window-based passages* deles fullteksten i *passages* med en fast forhåndsbestemt lengde uavhengige av tekstens struktur. (Kaszkiel & Zobel, 2001, s. 349) Lengden på fulltekstbeskrivelsene for hvert innslag i NRKs database er veldig varierende, fra ett ord til flere setninger. Det er ikke

mulig å anta en fast lengde som skal gjelde for alle programmene.
På grunn av begrensningene i indekseringen av metadataene, og mangelen på en sikker automatisk metode for en semantisk inndeling, måtte delingen av programmene i innslag gjøres manuelt.
Jeg har brukt følgende inndelingskriterier for å få til en samling der innslagene er enhetene for søket:

- Et program med ulike innslag som utgjør ulike semantiske enheter er delt i ulike poster (rader i databasen). Emneord og tagger blir registrert i samme rad for innslaget de refererer til, men i egne kolonner.

- Et program med ulike innslag som handlet om samme tema og hvor alle innslag hadde felles emneord er ikke delt i flere rader. Spesielt de eldste programmene av "Ukeslutt" hadde et gjennomgående tema for hele programmet.

- Musikkinnslagene er ofte brukt som pause fra et nyhetsinnslag til et annet. Musikkinnslagene er ikke tatt ut i egne rader, men holdt sammen med andre innslag. Dette bør ikke påvirke resultatet av søkingen siden det er flest engelske sanger med engelske termer, og det er få ord knyttet til musikkinnslagene. Ikke noen emneord er knyttet til dem heller.

Jeg har forsøkt å bruke disse inndelingskriteriene mest mulig konsekvent gjennom samlingen, men skillelinjene mellom nyhetsinnslagene har ikke alltid vært like klare. Jeg har måttet ta noen skjønnsmessige vurderinger underveis. Enkelte nyhetsinnslag i samme program handlet om svært ulike sider av samme tema, og de har hatt forskjellige medvirkende personer som har bidratt til de ulike innslagene. Å vurdere når nyhetsinnslagene er samme semantiske enhet innebærer en vurdering som kunne ha blitt unngått med en tydelig struktur i indekseringen.
I noen av programmene er den innholdsbeskrivende fullteksten bygget tett på fremføringen av radiosendingen. Da begynner beskrivelsen med en innledende presentasjon av alle nyhetsinnslagene som følger ordlyden i programmet, og etter det en mer utfyllende beskrivelse av hvert av innslagene. Disse presentasjonene har vært vanskelig å dele.
En annen utfordring har vært å dele innslagene fra 2012 og koble dem til sine tagger. Taggene er samlet enten i starten eller på slutten av hele programmet, i stedet for å være i nærheten av de innslagene de beskriver. Koblingen mellom fullteksten og taggene måtte rekonstrueres. Det har ikke

alltid vært så åpenbart, spesielt der ikke alle aspekter ved en sak dekkes av fullteksten, men er dekket av taggene. Det var i enkelte tilfeller helt nødvendig å ha kjennskap til de enkelte hendelsene for å klare det.
Den manuelle inndelingen har også gjort det mulig å oppdage variasjoner i indekseringen i feltet for "Innhold". For enkelte programmer finnes det både en innholdsbeskrivende fulltekst og emneord, andre har en innholdsbeskrivende fulltekst, men uten emneord eller tagger, og en del har ikke noen av delene. De programmene som ikke har noen av metadataene, er ikke nødvendigvis tomme for tegn og ville ikke blitt oppdaget ved en automatisk inndeling. I feltet for "Innhold" blir det enkelte ganger bare referert til tittelen av programmet. Feltet er ikke tomt ut fra en automatisk analyse av feltene, men den har heller ikke noe innholdsbeskrivelse som hjelper gjenfinningen av innholdet. I hvilken grad de ulike tilfellene forekommer og hvilke konsekvenser dette har for gjenfinningen, er en del av analysen.
Resultatet av forarbeidet med metadataene er en tabell i databasen med 1666 rader, en for hvert "Ukeslutt"-program, og en tabell i databasen med 5883 rader, en for hvert innslag.

3.3 Søkespørsmålene

Sammenligning av de to samlingene med metadata baseres på et gjenfinningsforsøk. Hvem, hvordan og hvilke spørsmål man stiller til metadataene har en viktig rolle for gjenfinningsresultatet. (Salton & McGill, 1983, s. 173) Jeg skal se på de metodologiske aspektene knyttet til utformingen og utvelgelsen av spørsmålene, og hvordan dette er løst i praksis.

3.3.1 Utforming av søkespørsmålene

Spørsmålene er ikke hentet fra reelle brukerbehov, og er ikke stilt av brukere. Hovedfokus i oppgaven er en analyse av gjenfinningseffektivitet ved søk i to indekser der de samme metadataene har enten programmene eller innslagene som enheter for søket. Involvering av reelle brukere eller brukerbehov ville ha betydd flere variabler, og mindre kontroll på forsøket. Kontrollen over hvilke spørsmål som skulle stilles til systemet, er vurdert som viktigst, til tross for at dette svekker forsøkets nærhet til virkelige brukerbehov. (Ringdal, 2007, s. 110-111)
Realisme i søkespørsmålene er forsøkt ivaretatt ved å bruke beskrivelser av fakta og hendelser fra Wikipedia, i stedet for at jeg skulle formulere søkespørsmålene til systemet. Jeg hadde for mye kunnskap om hvilke emneord og tagger som er brukt, og om hvordan fullteksten var strukturert for

å formulere disse på egen hånd, etter å ha jobbet manuelt med dataene.
Wikipedia har lister over norske og internasjonale hendelser på norsk sortert etter år. (*Wikipedia : 20. århundre*) Listene fra Wikipedia er ikke like omfattende for hvert år. For de første årene er det langt færre hendelser beskrevet enn for de siste. Mengden øker betraktelig spesielt fra slutten av 2000-tallet. For alle årene henvises i tillegg til en del kategorier som også har kronologiske lister over hendelser. Jeg har holdt meg til den generelle listen over hendelsene for hvert år, og ikke gått videre til disse ytterligere kategoriene som det er lenker til. De generelle listene hadde allerede en viss bredde i innholdet for hva som blir beskrevet og mengden var overkommelig.
De generelle listene har også en inndeling i underkategorier. De mest brukte underkategoriene, med variasjoner for de ulike årene, er "Hendelser" (for de første årene kalt "Begivenheter"), "Idrett", "Personer" (herunder "Fødsler" og "Dødsfall"), "Nobelprisvinnere" og "Musikk". De beskrivelsene som passet best til mine metadata var fra kategoriene "Hendelser" og "Idrett".
Beskrivelsene fra kategoriene "Personer" og "Nobelprisvinnere" består bare av personnavn og en dato uten en beskrivelse av hendelsen som er underforstått i kategoriinndelingen. Da blir beskrivelsen av hendelsen altfor mangelfull for å kunne lage et søkespørsmål av den. Det er derfor ikke valgt noen fakta fra "Personer".
"Musikk"-kategorien er også utelatt. Musikk er for det meste brukt som pauseinnslag i "Ukeslutt" og som forklart i § 3.2.3 er den ikke prioritert i inndelingen av programmene etter innslagene. Da er det heller ikke hensiktsmessig å bruke eventuelle spørsmål om det.
Beskrivelsene i Wikipedia består for det meste av en eller to setninger. Siden beskrivelsene er relativt korte, har jeg valgt å bruke dem direkte som søkespørsmål, og ikke bare som kilde for innhold i spørsmålene. Setningene blir formalisert av gjenfinningssystemet på samme måte som metadataene i samlingen. Dette innebærer fjerning av stoppord og rotlemmatisering. Spørsmålene blir behandlet av systemet som en vektor av term-vekter slik som dokumentrepresentasjonen. Ifølge Baeza-Yates åpner vektormodellene for å sammenlikne lengre tekster når begge er representert som vektorer hvor den ene er brukt som søkespørsmål. I min oppgave blir setningene fra Wikipedia formalisert til spørsmålsrepresentasjoner og sammenliknet med innholdsmetadataene i samlingen. Bruken av slike spørsmål i naturlig språk kan betraktes som *natural language queries* selv om det ikke stilles direkte spørsmål til systemet. (Baeza-Yates & Ribeiro-Neto, 2011, s. 262)

3.3.2 Søkespørsmålenes innhold og utvelgelse av spørsmålene

Jeg har brukt to generelle kriterier knyttet til hendelsenes innhold for å plukke ut mulige fakta i

Wikipedia. Kriteriene har som mål å lage liste med søkespørsmål som dekker ulike aspekter ved datasamlingen.

- Hele tidsperioden som jeg har datamaterialet fra, dvs. 1977-2013, skulle være representert i spørsmålene slik at jeg kunne se på eventuelle forskjeller i indekseringen av datamateriale over tid.

- Flere typer hendelser og fakta skulle være representert. Temaer som er forsøkt dekket er norske eller internasjonale hendelser, krig og terrorhandlinger, naturkatastrofer eller klimaforhold, idrettsrelaterte hendelser, ulykker, politiske hendelser og andre samfunnsmessig relevante hendelser.

Gjennomgangen av Wikipedias sider med de nevnte kriteriene, resulterte i en første liste på 130 søkespørsmål. (Vedlegg 1) Hendelsene skal være behandlet i min datasamling, for at man skal kunne gjennomføre en sammenligning av gjenfinningseffektivitet med de to indeksene. Dette ble hovedkriteriet for å velge et mindre utvalg av søkespørsmål som kunne brukes i den kvalitative analysen av gjenfinningseffektivitet. Jeg kunne ikke vite på forhånd hvilke hendelser som fantes i metadataene, og har derfor søkt med alle de 130 spørsmålene mot de to datasamlingene. Søkingen gjennom programmet "MittSøk" (Vedlegg 3) går gjennom alle spørsmålene i begge indeksene samtidig. Alle resultatene blir lagret i en resultat-tabell i databasen. Fra resultatene kunne jeg se om hendelsene var omtalt i radioprogrammet.
Andre kriterier for det endelige utvalget har handlet om å finne frem til søkespørsmål som kunne vise ulike aspekter av samlingens metadata og konsekvenser for gjenfinningen. Søkespørsmålene skulle fortsatt dekke hendelser fra ulike år for å se på endringer i indekseringen av dataene over tid, og dekke temaer av ulik karakter. I tillegg brukte jeg to nye kriterier:

- Det skulle være en kombinasjon av spørsmålene med få relevante innslag i datamaterialet og spørsmål med mange relevante innslag. I og med at programmet sendes en gang i uken, blir en del fakta omtalt bare en gang i hele datasamlingen. Jeg har forsøkt å balansere det med enkelte mer generelle spørsmål som har flere relevante treff i basen fordi det er en viss repetisjon av hendelsene.

- Enkelte spørsmål skulle bestå av en lengre tekst og andre av en kort tekst, for å se på

eventuelle konsekvenser for gjenfinningen.

Resultatet er en liste på 16 søkespørsmål. Disse er uthevet i vedlegg 1 som inneholder listen for alle de 130 spørsmålene.

3.4 Valg av evalueringsmål

Hovedproblemstillingen er å se hvordan effektiviteten i et gitt gjenfinningssystem påvirkes i to ulike representasjonsmodeller.
Aktuelle evalueringsmål for gjenfinningseffektivitet er presisjon og fullstendighet. Disse er standardmål som er brukt i mange studier knyttet til gjenfinningsevaluering, fra Cranfield- til TREC-eksperimentene og i studier for evaluering av *passage retrieval*. Kvantitative evalueringer bruker spesielt MAP og P@n for å vurdere gjenfinningseffektiviteten. Disse målene oppsummerer resultatene for alle søkene i en samling og gjør det mulig å få en samlet forståelse av resultatene. Resultatene i kvantitative evalueringer blir supplert med statistisk beregning av signifikans. Det gjøres ikke noen analyse av verdiene for presisjon og fullstendighet for hvert av søkene på grunn av den store datamengden. (Bendersky & Kurland, 2010, s. 167; Kaszkiel & Zobel, 2001, s. 346-347; Lin, 2009, s. [5])
Min oppgave skal gi en kvalitativ analyse av resultatene. Verdiene av presisjon og fullstendighet i de enkelte søkene bidrar til en bedre dybdeforståelse av rangeringene av programmene for hvert søk. Det begrensede utvalget av søkespørsmål gjør det mulig å se etter karakteristikker og tendenser for alle de 16 søkeresultatene, og en sammenligning av resultatene for de to indeksene. Analysen av målene og trefflistene brukes til å vise eventuelle felles trekk ved treffene og til å forklare årsakene til resultatene.
Et java-program beregner presisjon og fullstendighet for de første 100 treffene for alle de 16 søkespørsmålene. Resultatene blir til en ny tabell i databasen. I analysen tas det ikke hensyn til flere enn de første 20 treffene i resultatene siden brukerne ikke pleier å bla gjennom lange trefflister og stopper med de første treffene. (Baeza-Yates & Ribeiro-Neto, 2011, s. 268) Unntaksvis har jeg sett på resultatene for de relevante programmene lenger ned i trefflistene for å finne forklaringer for resultatene.
Selv om det er trefflistene, med deres rangeringer og verdiene for presisjon og fullstendighet som blir sentrale for analysen, har jeg også beregnet noen av standardmålene for presisjon og fullstendighet som er brukt i kvantitative gjenfinningsforsøk. Disse har en sekundær rolle i mitt

forsøk, men kan bidra med å bekrefte eller avkrefte eventuelle felles tendenser i søkene. I tillegg bruker jeg flere av målene fordi de viser ulike egenskaper ved gjenfinningen. De valgte målene er *Average Precision at n* (P@n), *Mean Average Precision* (MAP), samt *R-Precision*.

P@n er gjennomsnittverdien av presisjonen ved det *n'*te dokumentet for alle spørsmålene. P@n skal vise hvor god systemet er til å plassere relevante dokumenter på toppen av trefflistene. P@n for de to ulike basene kan sammenliknes for å se hvilken indeks som oppnår best resultater. Vanlige verdier for *n* er 5, 10 og 20. (Baeza-Yates & Ribeiro-Neto, 2011, s. 140) Jeg skal beregne verdien av presisjon for $n = 3, 5, 10$ og 20, men med noen justeringer avhengig av antall relevante programmer som finnes for de enkelte søkespørsmålene. Flere av søkespørsmålene har veldig få relevante programmer. Disse blir ikke tatt med i beregningen av P@10 eller P@20. De ville gitt et feil bilde av søkeresultatene, og ikke klart å vise hva systemet klarer å rangere øverst i trefflisten. Av samme grunn har jeg brukt $n = 3$ i tillegg til de mer vanlige verdiene for *n*. Da får jeg et mål som er representativt for søk som har færre enn 5 mulige relevante programmer.

MAP beregner gjennomsnittverdi av presisjon for alle spørsmålene og tar derfor hensyn til flere av de relevante treffene, ikke bare de som er plassert øverst i trefflisten. Når relevante treff er utenfor de første 100 treffene i trefflisten, er presisjonsverdien satt til 0. Det antas at brukerne ikke kommer til å se på disse, og deres presisjon har derfor lite verdi. (Baeza-Yates & Ribeiro-Neto, 2011, s. 141) Ifølge Büttcher, Clarke og Cormack (2010, s. 71) har MAP vært et populært mål fordi det har blitt brukt mye i TREC-eksperimentene, men ulempene har sjelden blitt vurdert. MAP er en gjennomsnitt av alle gjennomsnittverdier for presisjon for alle søk. Konsekvensen er at den ikke bidrar så godt til en forståelse for hva som faktisk skjer i gjenfinningen, i motsetning til de andre målene. (Büttcher et al., 2010, s. 71) Gjenfinningseksperimenter innenfor *passage retrieval* som bruker TREC-data, har også brukt MAP som et av målene for evaluering av gjenfinningseffektivitet. (R. Fernández et al., 2011, s. 363-364; Krikon & Kurland, 2011, s. 601-602; Lin, 2009, s. [4-5]) MAP er brukt i min analyse, men slik som de andre målene som oppsummerer resultatene har den en begrenset rolle for analysen.

R-precision er verdi av presisjonen for hvert spørsmål ved R-funnede dokumenter, hvor R er det totale antallet relevante dokumenter for søket. *R-precision* regnes for hvert av spørsmålene i gjenfinningsforsøket slik at man kan se tendensene gjennom de ulike søkene. Dette anbefales også av Baeza-Yates og Ribeiro-Neto (2011, s. 141), som mener at en eventuell gjennomsnittsverdi for alle søkene gir et upresist tall for vurderingen. Målet er noe mer representativt for mitt gjenfinningsforsøk enn P@n og MAP på grunn av karakteristikkene ved fasiten. Svakheten ved målet er når det er et høyt tall av relevante dokumenter, dvs. høy R-verdi, sier målet lite om hva

som skjer i starten av trefflisten. Siden flere av mine søk har få relevante programmer, er målet representativt for resultatene også på toppen av trefflistene.

Det er en del problemer knyttet til presisjon og fullstendighet og deres bruk i gjenfinningsevalueringen, uavhengig om det blir brukt presisjon og fullstendighet for hvert av treffene eller mål som oppsummerer dem, dvs. MAP, P@n, *R-precision*. Presisjon og fullstendighet er sterkt knyttet til fasiten med relevante dokumenter. (Swanson, 1988, s. 557) Hvilke og hvor mange dokumenter i samlingen som blir vurdert som relevante, brukes i beregningen for fullstendighet og presisjon. (Fig. 1) De relevante dokumentene må oppdages i sin helhet for at det skal gis en troverdig verdi av fullstendighet. Det kan bli problematisk i store samlinger. I tillegg er byggingen av fasiten problematisk i forhold til hva som oppfattes som relevant av forskjellige brukere. (Chowdhury, 2010, s. 289) Formlene for presisjon og fullstendighet tar heller ikke hensyn til at enkelte dokumenter kan være mer relevante enn andre i forhold til de enkelte søkene. (Chowdhury, 2010, s. 289) For brukerne kan resultatene oppfattes annerledes avhengig av hvilke dokumenter fra fasiten som blir funnet, mens dokumentene er likestilt i evalueringen av gjenfinningseffektivitet.

Både Salton og McGill, og Lancaster peker på søkerne som en viktig faktor for resultatene av presisjon og fullstendighet i et gjenfinningsforsøk. (Lancaster, 2003, s. 145; Salton & McGill, 1983, s. 173)

Flere faktorer bidrar derfor til resultatene av evalueringen. Presisjon og fullstendighet kan ikke ses som uavhengige evalueringsmål. Valg av fasiten og søkespørsmålene, men også valg knyttet til gjenfinningsmodellen har mye å si for resultatene som ikke kan ses isolert fra de valgene som er gjort. En kvalitativ analyse gjør det mulig å ta hensyn til dette, og kan vise hvordan de metodologiske valgene er med på å forklare resultatene.

3.5 Fasit

3.5.1 Relevans og tolkning av spørsmålene for å bygge fasiten

Beregningene av presisjon- og fullstendighet forutsetter at det er en fasit med dokumenter fra samlingen som er vurdert som relevante i forhold til søkespørsmålene. Slik som jeg allerede har pekt på i forbindelse med diskusjonen om presisjon og fullstendighet, er fasiten et av de problematiske sidene ved denne typen gjenfinningsevaluering. Fasiten tar ikke hensyn til ulike oppfatninger av relevans, som kan variere mellom ulike brukere. I tillegg er alle dokumentene i

fasiten like relevante i beregning av presisjon og fullstendighet. Relevansbegrepet har vært et viktig tema i gjenfinningsammenheng, og mye diskutert i litteraturen. Cranfield-eksperimentene ble kritisert for hvordan dokumentene ble vurdert som relevante. (Lesk, 2005, s. 211) Saracevic undestreker viktigheten av begrepet innenfor informasjonsgjenfinning, og definerer den ved å beskrive ulike manifestasjoner av relevansen. (Saracevic, 2007) Relevansen er ikke et entydig begrep, og jeg hadde heller ikke brukere som kunne vurdere relevansen av søketreffene.
Mine spørsmål er beskrivelser av hendelser. Tanken bak fasiten er å finne mer informasjon i datasamlingen om fakta som beskrives i Wikipedia. Når hendelsene som er beskrevet i søkespørsmålene finnes i datamaterialet, bør dokumentene kunne vurderes som relevante. Utgangspunktet er at man ønsker seg mer om det samme tema, men det er flere tilfeller hvor jeg har måttet tolke beskrivelsene. Tolkningen har direkte konsekvenser for resultatene av evalueringen.
Et eksempel er fasit for søk med lange spørsmål. Lange spørsmål viser som oftest til ulike sider ved en sak. Spørsmål nr 22 ("Norge blir klare for VM i fotball 1994 etter å ha slått Polen 3-0 på bortebane. Jostein Flo, Jan Åge Fjørtoft og Ronny Johnsen skåret målene") og nr 120 ("Terrorangrepene i Norge 2011 i Regjeringskvartalet og på Utøya. 77 mennesker omkommer, mange blir skadet, og det ble store materielle ødeleggelser") illustrerer problemet. For spørsmål nr 22 handler vurderingen om alle innslag om Norges kvalifisering til fotball-VM og deltakelse, skulle være med, eller bare innslagene om den ene kvalifiseringskampen. For spørsmål nr 120, handlet vurderingene om hvorvidt flere av innslagene om rettsaken og konsekvensene for ofrene, skulle inkluderes i tillegg til innslagene som beskriver selve hendelse.
Et annet eksempel som viser til noen av problemene med tolkning av spørsmål og valg av fasit illustreres med spørsmål nr 27 "Norge stemmer nei til EU-medlemskap". Flere av hendelsene fra Wikipedia er beskrivelser av resultatene av en prosess som har foregått over tid. Programmene i NRK kan ha innslag både om resultatet og hva som har skjedd før og rundt hendelsen. Slik er det med EU-avstemningen. I datasamlingen var det mange innslag med debatter i perioden før folkeavstemningen som kunne være like relevante som akkurat innslaget som handler om resultatet. Spørsmålene kunne ikke tolkes for strengt. Konsekvensen ville vært at de aller fleste bare ville fått ett eller to innslag som fasit. Jeg har derfor valgt å utvide fasiten med programmer som beskriver ulike sider eller konsekvenser for saken beskrevet i søkespørsmålet. Der søkespørsmålet beskriver konklusjonen for en prosess er også programmer som beskriver debatter og prosessen tatt med i fasiten. For de eksemplene som er nevnt over, betyr dette at innslag som handler om Norges deltakelse i fotball VM er vurdert som relevante for spørsmål nr 22, at innslag om konsekvensene for terrorangrepene i 2011 er relevante for spørsmål nr 120, og programmer med EU debatter før

folkeavstemningen i 1994 er med i fasiten for spørsmål nr 27. Med en kvalitativ analyse kan det tas hensyn til tolkningene gjort i søkespørsmålene, og resultatene for gjenfinningseffektivitet kan analyseres i forhold til det.

3.5.2 Søk etter relevante innslag for fasiten

Det er en forutsetning for å bruke effektivitesmål at fasiten inneholder alle de relevante dokumentene som finnes i samlingen. Jeg har brukt ulike teknikker for å sikre fullstendighet i fasiten:

- Jeg har brukt dato for når de enkelte hendelsene har funnet sted. Hendelsene er som oftest omtalt i radioprogrammene før eller rett etter de har skjedd, noe avhengig av temaet. Da kunne jeg sjekke innslagene rundt datoen for hendelsen. Det var f eks ikke noe interessant å se på programmer med en sendedato som er før en terrorhandling har skjedd. Det var interessant å se på tidligere sendinger i forbindelse med politisk valg eller folkeavstemning.

- Jeg har laget noen veldig vide spørsmål som jeg har stilt direkte til tabellene i databasen med SQL-spørringer for å finne fram til flere relevante dokumenter. Enkelte hendelser blir tatt opp igjen etter en viss tid eller blir aktuelle igjen. Dette skulle også dekkes. Jeg har sett gjennom relativt store trefflister og plukket ut relevante dokumenter.

- For noen få søkespørsmål har jeg gått gjennom hele metadatasamlingen fordi det var hendelser som repeteres jevnlig. Et eksempel på dette er spørsmålene nr 63 og 87. (Vedlegg 4)

3.6 Valg av rangeringsmetode for innslagene

Fasiten er bygget ved å finne de relevante innslagene for de enkelte søkespørsmålene, men de representeres med nøkkelen for de programmene de er en del av. Dette brukes også av Kaszkiel og Zobel (2001) og Lin (2009) for å kunne regne gjenfinningseffektivitet mellom hele dokumenter og fasit.
I trefflisten fra søk i indeksen av innslag har hvert innslag egen nøkkel. Trefflisten for

nyhetsinnslagene må referere tilbake til nøklene for programmene for at det skal kunne sammenliknes med fasiten som består av nøklene for programmene. Dette er ikke problematisk når bare ett av innslagene fra ett program er i trefflisten, men når det er treff på flere innslag fra samme program må disse representeres samlet i trefflisten. Det finnes flere mulige rangeringsmetoder som er brukt i ulike studier om *passage retrieval* som illustrert i § 2.3.2. Jeg har valgt å bruke metoden kalt for *"Maximum of supporting spans"* av Jimmy Lin, hvor innslag fra samme program representeres med innslaget som er rangert øverst i trefflisten. (Lin, 2009, s. [5]) Programmer med bare ett relevant innslag favoriseres fremfor programmer med flere relevante innslag. Innslag som ikke har høyest score, men som likevel er relevante bidrar ikke til å få programmet høyere opp i trefflisten med denne metoden.
Begrunnelsen for valget er at inndelingen av innslagene skulle representere semantiske enheter. Når ulike innslag handler om samme tema og er indeksert med felles emneord, er de holdt sammen. Bare unntaksvis burde det være tilfeller hvor det er mange relevante innslag fra samme program som er delt. Den kvalitative analysen skal se på eventuelle konsekvenser for valget av rangeringsmetoden og om datamaterialet og inndelingen i innslag fungerer i dette tilfellet.

3.7 Innholdsbeskrivende fulltekst og kontrollert vokabular

I oppgaven skal jeg se på funksjonen av emneordene og taggene i forhold til den innholdsbeskrivende fullteksten i gjenfinningssammenheng. Trefflistene for de 16 søkene analyseres kvalitativt. Det er derfor mulig å se i metadataene for de enkelte innslagene og analysere hvordan enkelte av emneordene og taggene bidrar til resultatene for presisjon og fullstendighet.
For å få til et mer sammensatt bilde av de ulike metadataene i samlingen skal analysen også bestå av en beskrivelse av metadataene utenom de enkelte søkene. Denne beskrivelsen er ikke hovedmålet for oppgaven, men enkelte tall om mengde og typer metadata kan være med på å forklare resultatene for gjenfinningen.
I analysen skal jeg diskutere tall knyttet til mengden innholdsmetadata som faktisk finnes i samlingen og hvordan innholdsbeskrivende fulltekst og kontrollert vokabular er prioritert i forhold til hverandre. I tillegg skal jeg se på likhetsgraden mellom termene i den innholdsbeskrivende fullteksten og termene i det kontrollerte vokabularet. Dette kan være med på å definere de ulike metadataenes rolle for gjenfinningen. Sammenligningen av termene innebærer noen metodologiske valg som skal forklares her.

3.7.1 Sammenligning av termene fra det kontrollerte vokabularet og den innholdsbeskrivende fullteksten

Sammenligningen av emneord og tagger med den innoldsbeskrivende fullteksten skal ta hensyn til hvordan gjenfinningssystemet oppfatter at to termer er like, og ikke nødvendigvis hvordan man har tenkt likhet mellom termer ved den manuelle indekseringen. Det systemet oppfatter som likt har direkte konsekvenser for resultatet av gjenfinningen, og er derfor viktigere i denne sammenheng enn hva målet for den manuelle indekseringen har vært med termene.
Sammenligningen mellom termene gjøres ikke fra indeksene i Lucene, men ved å bruke termene i tabellene i databasen. Sammenligningen er gjort med en SQL-spørring. Jeg har likevel forsøkt å gjenskape noen av forandringene av termene som skjer når Lucene indekserer. Sammenligningen skal være mest mulig lik systemets oppfatning av likhet mellom termene. Jeg har derfor tatt følgende valg:

- Systemet fjerner stoppord ved indekseringen. Stoppordene er fjernet også for sammenligningen og sammensatte emneord og tagger som inneholder artikler, preposisjoner og konjunksjoner, mister disse.

- Ved indekseringen blir hvert ord indeksert uavhengig av andre ords posisjon i teksten. Søkespørsmålene har verken nærhetsoperatorer eller frasesøk. Termene blir behandlet uavhengig av hverandre både ved indekseringen og ved søk. Sammensatte emneord og tagger blir derfor splittet til frittstående termer.

Disse valgene kan eksemplifiseres med emneordene fra programmet med nøkkel "1990/01625". Programmet hadde følgende emneord og klassifikasjonsnumre etter hverandre:

> "K355 - Forsvar (eo) + K355.343 - Etterretningstjeneste : militærvesen (eo) + K331 -Arbeiderbevegelsen (eo) + K329.5 - Det norske Arbeiderparti (eo) + K327.12 - Overvåking (eo) + K335 - Kommunisme (eo) + K329.7 - Norges Kommunistiske Parti (eo)".

Klassifikasjonsnumrene, tegn som "+" og "(eo)", ":", og artikkelen "Det" ble fjernet. Termene ble splittet. Resultatet kan ses i Fig. 3. Det er disse termene som blir sammenliknet med den innholdsbeskrivende fullteksten.

"1990/01625"	Forsvar
"1990/01625"	Etterretningstjeneste
"1990/01625"	militærvesen
"1990/01625"	Arbeiderbevegelsen
"1990/01625"	norske
"1990/01625"	Arbeiderparti
"1990/01625"	Overvåking
"1990/01625"	Kommunisme
"1990/01625"	Norges
"1990/01625"	Kommunistiske
"1990/01625"	Parti

Fig.3. Eksempel på forandringer i det kontrollerte vokabularet

Tilpasningene fører til at enkelte emneord og tagger får en mye mer generell betydning, som f eks ved splittelsen av korporasjonsnavn. Graden av likhet mellom det kontrollerte vokabularet og den innholdsbeskrivende fullteksten øker. Sannsynligheten for at termen "Arbeiderparti" finnes i fullteksten er større enn for at hele korporasjonsnavnet "Det Norske arbeiderparti" finnes der. Det er likevel dette som gjenfinningssystemet oppfatter som likt, og som teller i rangeringene av trefflistene og ved beregning av gjenfinningseffektivitet.
Sammenligningen gjøres mot tabellen av innslagene, ikke av programmer. Det er i denne tabellen at emneordene og taggene er skilt fra den innholdsbeskrivende fullteksten, og man kan gjøre en sammenligning. I enkelte programmer var det kontrollerte vokabularet samlet for hele programmet i starten eller på slutten av innholdsbeskrivelsen. Hvis disse beskrev flere innslag som skulle deles, måtte jeg repetere dem for hvert av de innslagene de gjaldt for. Enkelte av tallene for det kontrollerte vokabularet kan derfor bli høyere enn om jeg hadde regnet antallet mot programmene. Dette må tas hensyn til spesielt med taggene, hvor det er generell praksis å ha taggene samlet ett sted. Dette gjelder i stedet veldig få av programmene med emneord. Emneordene er for det meste registrert i forbindelse med de innslagene de handler om, og ikke samlet for hvert program.

3.8 Implementering i Lucene og valg av gjenfinningsmodell

Indekseringen av dataene fra NRK har foregått ved hjelp av Lucene, et *open source information retrieval library*. Lucene er brukt for den automatiske indekseringen av dataene og for søking.

Lucene er ikke en fullt utviklet søkeapplikasjon, og søkingen har derfor foregått med java-programmer som kaller metoder i Lucene. (McCandless, Hatcher & Gospodnetić, 2010, s. 6-7)

3.8.1 Indeksering av metadataene i Lucene

Den automatiske indekseringen av metadataene innebærer en analyse av teksten i programmene hvor teksten deles i atomiske enheter kalt for *tokens*. *Tokens* blir til termer i indeksen. (McCandless et al., 2010, s. 110-116) Hvilken analyse av datamaterialet man velger påvirker gjenfinningen. Det finnes flere forhåndsdefinerte *analyzers* i Lucene. *StandardAnalyzer* kan brukes når dataene ikke har spesielle behov ved indekseringen. En *StopAnalyzer* fjerner stoppord og har noen basisfunksjoner som orddeling og omgjøring av store boktaver til små bokstaver. De generelle *analyzers* baserer seg på engelske termer, men det finnes også *analyzers* for andre språk. (*Class StopwordAnalyzerBase : org.apache.lucene.analysis..util*, cop. 2000-2014; McCandless et al., 2010, s. 127-128) Jeg har valgt å bruke *NorwegianAnalyzer* for min samling som for det meste inneholder norske termer, med unntak av musikkinnslagene. *NorwegianAnalyzer* fjerner norske stoppord, rotlemmatiserer termene og omgjør alle store bokstaver til små bokstaver. (*Class NorwegianAnalyzer : org.apache.lucene.analysis.no*, cop. 2000-2014)

Stoppordene, vanligvis artikler, preposisjoner og konjunksjoner, blir fjernet fra indeksene. Disse har liten diskrimineringsverdi. De bidrar ikke til å skille mellom dokumentene fordi de har en høy frekvens gjennom alle dokumentene i en samling. De gjør dokumentene likere. (Salton & McGill, 1983, s. 66-69) Det er mulig å utvide stoppordene til å gjelde også enkelte adverb, verb og adjektiver, men dette kan redusere fullstendigheten i gjenfinningen. (Baeza-Yates & Ribeiro-Neto, 2011, s. 63, 226) Jeg har ikke utvidet *NorwegianAnalyzer* for å fjerne adverb, verb og adjektiver. Det er ikke nødvendig å fjerne flere ordklasser i en liten samling. Innoldsbeskrivelsene i NRKs samling er ikke så lange, og disse ordklassene kan bidra i gjenfinningen siden de også er med i søkespørsmålene.

Fullstendigheten økes ved bruk av *analyzers* for rotlemmatisering. Lucene har denne muligheten med *PositionalPorterStopAnalyzer* som bruker Porter-algoritmen for stemming. Ifølge Baeza-Yates er Porter-algoritmen en av de mest brukte rotlemmatiseringsalgoritmene, med en enkel formel for fjerning av stemmer for engelsk språk og som gir gode resultater. (Baeza-Yates & Ribeiro-Neto, 2011, s. 226-227; McCandless et al., 2010, s. 138; Porter, 1979, s. 313-316) Rotlemmatiseringen for det norske språket er en del av *NorwegianAnalyzer*.

Indekseringen gjøres gjennom programmet "LagIndex". (Vedlegg 2) Resultatet for den automatiske indekseringen av de to basene inndelt etter programmer og etter innslag, er to indekser av dokumentrepresentasjoner. Termene i indeksene er like, men de er delt i forskjellige enheter. Den ene indeksen er en indeks av programmer og den andre en indeks av innslag.

3.8.2 Søkespørsmålene i Lucene

Søkespørsmålene blir analysert av Lucene ved hjelp av en *QueryParser* som gjør dem til spørsmålsrepresentasjoner. *QueryParser* lager indekstermer av søkespørsmålene etter den samme *NorwegianAnalyzer* som er brukt for metadataene i samlingen. (Vedlegg 3) Det er mulig å bruke en annen *analyzer* for spørsmålene enn for metadata i samlingen, men dette er anbefalt bare når man har utviklet spesielle typer *analyzers* for samlingen. (McCandless et al., 2010, s. 114-115) *QueryParser* bruker automatisk OR-forbindelse mellom termene i spørsmålene hvis ikke annet er spesifisert. (*Class QueryParser : org.apache.lucene.queryparser.classic*, cop. 2000-2014) Det er dette jeg bruker i gjenfinningsforsøket. OR-forbindelsen mellom spørsmålstermene gir et stort antall treff for søkene. Enkelte av spørsmålene er lange, og de ville ikke gitt noen treff i samlingen med frasesøk eller AND-forbindelse. Ulempen med valget av OR-forbindelsen kan være at enkelte av de korte søkespørsmålene med få termer får med veldig mye støy i trefflistene. Systemet rangerer programmene etter likhetsgraden med søkespørsmålene, og dette bør løse problemet med at OR-forbindelsen ikke stiller strenge krav og at programmer med lite til felles med søkene er tatt med i trefflistene.

Bruken av nærhetsoperatorer, som er en mykere variant av frasesøk, passet heller ikke for samlingen. Flere av programmene fra de siste to årene har tagger for flere av innslagene samlet og ikke i nærheten av innholdsbeskrivelsen av de innslagene de referer til. Et eksempel som illustrerer dette er programmet med nøkkel "2012/09043"

> " [...] Tags: norske muslimer + Syria + krigføring + raseriutbrudd + fredsforhandlinger + farc-geriljaen + Colombia + Hurdal + fredsmekler + ylvis-brødrene + videohyllest + big-bird + debatt + presidentkandidatene + one-liners + 22.juliforklaringer + politi + embetsverk + politikere + regjeringen + bursdag + fest + gaver Rubrikk: **20 norske muslimer kjemper i Syria. - Noen av dem som drar til Syria fortjener honnør, sier generalmajor Robert Mood. **Daværende PST-sjef Janne Kristiansen ble varslet om 22.juliterroren av en journalist. - blablabla, sier journalisten. **200.000 har sett musikkvideoen som hyller fredsmeklaren og supermannen Jan Egeland. stempel Dette er noe av det du får høre i Ukeslutt hvor vi også skal hjelpe sinnataggen Liv Signe Navarsete i gang med sinnemestring... DAB tekst: Sendt: 12:30:56 -12:31:02 Lengde: 00:06 Tittel: Norske muslimer i Syria Reporter: Ida Thune Øritsland Sendt: 12:33:04 -12:33:30 Lengde: 00:26 Tittel: Norske muslimer i Syria Reporter: Ida Thune Øritsland Medvirkende: Tags: DAB tekst: Sendt: 12:33:54 -12:37:45 Lengde: NaN:00 Tittel: Robert Mood Reporter: Tags: Medvirkende: DAB tekst: Sendt: 12:44:54

-12:47:27 Lengde: 02:54 Tittel: Love that girl Artist: Raphael Saadiq Sendt: 12:41:27 -12:42:22 Lengde: 00:55
[...]
Tittel: Homeless (Radio Edit) Artist: Maria Mena Sendt: 13:14:43 -13:20:54 Lengde: NaN:00 Tittel: Harald Stanghelle om 22.juliforklaringer Reporter: Tags: Medvirkende: DAB tekst: Sendt: 13:24:55 -13:25:21 Lengde: 00:26 Tittel: 22.juli forklaringene - telefon Reporter: Ida Thune Øritsland Medvirkende: Tags: DAB tekst: Sendt: 13:24:55 -13:27:28 Lengde: 03:39 Tittel: Lazin' in the Sunshine Artist: Jonathan Jeremiah Sendt: 13:22:39 -13:24:53 Lengde: 02:14 Tittel: Regjeringen har bursdag Reporter: Ida Thune Øritsland Medvirkende: Tags: DAB tekst: Sendt: 13:24:57 -13:27:12 Lengde: 02:15 Tittel: Gaver til Regjeringa Reporter: Ida Thune Øritsland [...]"

Et søk med nærhetsoperatorer ville fått problemer i dette programmet. Taggene for det siste innslaget om "gaver til Regjeringen" står helt øverst i feltet for innhold sammen med taggene for alle de andre innslagene. Taggene ville ikke blitt koblet til riktig innslag med nærhetsoperatorer, men i stedet til de første innslagene som handlet om Syria. En OR-forbindelse i søkespørsmålene løser dette fordi den ikke tar hensyn til termenes posisjon i teksten.

3.8.3 Gjenfinningsmodellen

Gjenfinningen foregår ved å sammenlikne indekstermene fra indeksen av programmer og fra indeksen av innslag med termene i spørsmålsrepresentasjonen. (McCandless et al., 2010, s. 110) Hvordan sammenligningen gjøres av systemet er avhengig av hvilken *Similarity*-klasse som er valgt for søket. *Similarity*-klassene tilsvarer forskjellige gjenfinningsmodeller. Alle er *partial match* modeller som lager en rangering av dokumentene etter deres likhet med søkespørsmålene. De mulige klassene i Lucene er *DefaultSimilarity*, *BM25Similarity*, *MultiSimilarity*, *PerFieldSimilarityWrapper*, *SimilarityBase*. (*Class Similarity : org.apache.lucene.search.similarities*, cop. 2000-2014)

DefaultSimilarity-klassen blir automatisk brukt av systemet når ikke annet er spesifisert. Jeg har valgt å bruke denne klassen i gjenfinningen. Den er en del av *TFIDFSimilarity*-klassen, og følger prinsippene for vektorbaserte gjenfinningsmodeller.

DefaultSimilarity lager rangering av dokumentene i trefflistene ved å beregne en score for hvert av dokumentene når de blir sammenliknet med søkespørsmålet. Formelen som beregner score har følgende faktorer: termfrekvensen (TF), invers dokumentfrekvens (IDF), lengdenormalisering (length Norm), spørsmålnormalisering (queryNorm), *boost* og *coord*. (*Class DefaultSimilarity : org.apache.lucene.search.similarities*, cop. 2000-2014; McCandless et al., 2010, s. 86-88)

Termfrekvensen (TF) gir en score til dokumentet basert på antall ganger en term forekommer i dokumentet. Jo oftere en term fra et søkespørsmål forekommer i et dokument jo høyere verdi for

likhet får dokumentet. Dokumenter som har høy repetisjon av termer fra søkespørsmål antas å ha større likhet med spørsmålet og representerer derfor et bedre svar for søket. Målet kombineres med invers dokumentfrekvens (IDF) der det tas hensyn til hvor mange dokumenter som inneholder en bestemt term. En term som er brukt i mange dokumenter har lavere diskrimineringsverdi enn en term som er brukt i få dokumenter, og som dermed får en høyere verdi på invers dokumentfrekvens. Termfrekvensen tar hensyn til termens verdi i dokumentet, og invers dokumentfrekvensen tar hensyn til verdien av termen i samlingen.

Lengdenormalisering tar hensyn til dokumentets størrelse og normaliserer verdien i forhold til dette. Lange dokumenter har større sjanser for at termene repeteres flere ganger i forhold til korte dokumenter og kan derfor få høyere score på termfrekvensen. Normaliseringen gir en høyere score til termer i korte dokumenter for å kompensere for dette (Baeza-Yates & Ribeiro-Neto, 2011, s. 75-77; McCandless et al., 2010, s. 87)

DefaultSimilarity-klassen bruker i tillegg to faktorer kalt for *boost* og *coord. Boost* er en score som gis til ulike felter i et dokument avhengig av hvor viktige feltene er for gjenfinningen. Når feltene er vurdert som like relevante, kan *boost* settes likt for alle feltene. Jeg har bare et felt med innholdsbeskrivelse i mine søk, og har derfor ikke *boost* noen funksjon i mine søk. *Coord* er også knyttet til en beregning av *boost* for dokumentene, og skal ikke ha noen konsekvenser for mine søk siden dokumentene har samme *boost* i utgangspunktet. (McCandless et al., 2010, s. 87)

Vektorbaserte gjenfinningsmodeller er sammen med probabilistiske modeller mye brukt i gjenfinningseksperimenter, blant annet for *passage retrieval*. Derfor har jeg valgt denne modellen. Jeg bruker modellen for en relativt liten samling med korte innholdsbeskrivelser. I en vektorbasert modell er det viktig hvordan en term oppfattes av systemet både innenfor et dokument og i forhold til de andre termene i samlingen. Termens relevans for å beskrive et program gis av dens frekvens i dokumentet og i samlingen. En slik beregning kan bli påvirket av flere tilfeldigheter når samlingen er liten og det er få termer i samlingen. Dette kan få konsekvenser for søkeresultatet, der små variasjoner i antall termer kan ha utslag for resultatene.

4 Analyse

Resultatene fra de 16 søkene og beregning av fullstendighet og presisjon, viser at det er enkelte faktorer som har spesiell innflytelse på søkeresultatene for de ulike søkene. Noen resultater kan forklares med de metodologiske valgene, andre med NRKs indekseringspraksis og andre er en direkte konsekvens av inndelingen av programmene i innslag.
Forklaringene av resultatene er som oftest sammensatte og er en kombinasjon av de ulike faktorene som blir presentert. Flere av faktorene må derfor nevnes innenfor de overordnetde grupperingene for å kunne gjengi det sammensatte bildet av gjenfinningen.
Tallresultatene for alle søkene, både score for rangeringen gitt av systemet og verdiene for presisjon og fullstendighet, er samlet i vedlegg 4. I samme vedlegg er også teksten for søkespørsmålene og fasiten for søkene.

4.1 Konsekvenser av de metodologiske valgene

Flere av de metodologiske valgene påvirker resultatene av søkene. En analyse av søketreffene viser at spesielt valg av gjenfinningsmodell og valget av rangeringsmetoden for innslagene påvirker resultatene av søk i de to indeksene.
Den valgte gjenfinningsmodellen bruker en vektorbasert algoritme. Algoritmen sammenlikner egenskapene ved dokumentrepresentasjonen og spørsmålsrepresentasjonen. Systemet gir en score som resultat av sammenligningen til hver dokumentrepresentasjon og rangerer dem i trefflistene etter score. Score for hvert av programmene eller innslagene baseres på verdiene av termfrekvensen (TF), invers dokumentfrekvens (IDF) og lengdenormalisering. (McCandless et al., 2010, s. 87) Det er flere tilfeller i resultatene hvor man kan se effekten av bruken av denne algoritmen i trefflistene for søk i begge indeksene, og hvor den ene variabelen i algoritmen har større innflytelse på score enn de andre.

4.1.1 Lengdenormalisering

Et eksempel på hvordan lengdenormaliseringen påvirker rangeringen i trefflisten finnes i resultatet for spørsmål nr 12 ("Arne Treholt dømmes til 20 års fengsel for spionasje"). Ved søk i indeksen av

innslagene er det to relevante programmer som ikke blir funnet innenfor de første 20 treffene, mens de har posisjon nr 5 og 12 ved søk i indeksen av programmene. Disse har nøkkel "1990/03098" og "1990/03827". Programmene består av ett innslag. Programmer som består av bare ett innslag, har fått en noe lengre beskrivelse av det ene innslaget i forhold til programmer der det er mange innslag som skal beskrives. Formelen for lengdenormalisering gir korte tekster en høyere verdi enn lange tekster. (McCandless et al., 2010, s. 87) De to programmene som har samme tekst i de to indeksene, får derfor høyere score for lengdenormalisering når de er sammenliknet med tekster som er lengre i indeksen av programmene enn i indeksen av innslagene hvor sammenligningen er med korte innslag. Som konsekvens blir deres score høyere i indeksen av program enn i indeksen av innslag, og får en bedre posisjon i trefflisten i denne indeksen. Resultatet er en kombinasjon av score gitt ved lengdenormaliseringen og av indekseringspraksisen hos NRK. Ved indeksering av programmene i SIFT-basen har man ikke brukt en standard lengde for fulltekstbeskrivelsen, hverken for de enkelte innslagene eller for hele programmer. Dette gir utslag i en gjenfinningsmodell som operer med lengdenormalisering. De to programmene er som sagt rangert lavere i indeksen av innslagene, og finnes i posisjon nr 21 og nr 25. Til tross for at de ikke blir funnet innenfor de første 20 treffene som ønsket, er de likevel ikke så langt nede i trefflisten. Forskjellen skapt av indekseringspraksis og lengdenormalisering er derfor ikke så stor.

Spørsmål 87 har tekst "Kommune- og fylkestingsvalg i Norge". I første posisjon i trefflisten fra søk i indeksen av programmer plasserer systemet programmet med nøkkel "2007/29846". Programmet er ikke relevant for søkespørsmålet. Programmet har termene "kommune" og "Norge" felles med søkespørsmålet, men det handler ikke om valg. Dokumentet består av bare ett innslag. Lengdenormaliseringen gjør at det rangeres høyere opp når enheten for søket er programmet fordi den har en kort beskrivelse i forhold til andre programmer. Programmet er i stedet i posisjon nr 11 i indeksen av innslagene hvor teksten er lang sammenliknet med andre innslag, og score for lengdenormaliseringen blir lavere. Indeksen av innslag gir bedre resultat ved å plassere lenger ned i trefflisten et program som ikke er relevant for søkespørsmålet.

Et annet eksempel på lengdenormaliseringens effekt på trefflistene finnes i resultatene fra spørsmål nr 22 ("Norge blir klare for VM i fotball 1994 etter å ha slått Polen 3-0 på bortebane. Jostein Flo, Jan Åge Fjørtoft og Ronny Johnsen skåret målene"). Ved søk i begge indeksene blir det samme programmet med nøkkel "2007/22013" rangert øverst i de to trefflistene. Programmet er en årskavalkade med bare ett innslag som er relevant i forhold til søket. Årskavalkader har veldig lange fulltekstbeskrivelser fordi de inneholder langt flere innslag enn vanlige programmer. Lengdenormalisering gjør at en årskavalkade som har lang tekst, får lavere verdi i forhold til andre

kortere programmer eller samme program delt i innslag. Innslagene til en årskavalkade er ofte kortere enn andre innslag, og får derfor høyere score på lengdenormaliseringen i forhold til andre lengre innslag.
Innslaget som er relevant i dette programmet inneholder følgende tekst:

> "21. FOTBALL (e. 41'15") 930602: Norge slår England 2-0 i VM-kvalifiseringskamp. Scoringene til Øyvind Leonhardsen og Lars Bohinen (41'15"-41'50") 931013: Norge slår Polen 3-0 i landskamp (VM-kvalifiseringskamp). Scoringene til Jostein Flo, Jan Åge Fjørtoft og Ronny Johnsen. Kutt landslagstrener Egil OLSEN (mv) (Drillo) (41'50"-42'50") 931220: VM-trekningen i Las Vegas. Kutt Egil OLSEN (mv) (Drillo) (43'00"-43'30") (K795.33 - Fotball (eo))"

Teksten har termene "VM", "fotball", "Norge" og navnene på spillerne til felles med søkespørsmålet. Likheten mellom spørsmålsrepresentasjon og dokumentrepresentasjon er derfor stor, men lengden på teksten blir avgjørende for hvilken score den får i trefflisten. Lengdenormaliseringen får ikke konsekvenser for rangeringen, men forskjellen mellom verdi av score for samme program er stor i de to indeksene. Score er på 0,3654 i trefflisten for indeksen av program og 1,5 i trefflisten i indeksen av innslag.

4.1.2 Termfrekvensen

Fulltekstbeskrivelsene i datamaterialet er relativt korte beskrivelser sammenliknet med fulltekstdatabaser som inneholder hele artikler eller bøker. Når tekstene er så korte gir en enkel repetisjon av en term utslag i rangeringen gjennom termfrekvensen.
I søkespørsmål nr 12 ("Arne Treholt dømmes til 20 års fengsel for spionasje") blir programmet med nøkkel "1993/00019" plassert i posisjon nr 8 i trefflisten for søk i indeksen av innslagene, til tross for at den ikke er relevant for søket. Forklaringen finnes i en kombinasjon av verdien for termfrekvensen og invers dokumentfrekvens. I programmet er det ett innslag med tekst:

> "12. ARNE TREHOLT BENÅDET - e. 27'30" 920703: Kutt fra intervju med justisminister Kari GJESTEBY (mv), og med Arne TREHOLT (mv). (K345.02 – Straffesaker)"

Innslaget er ikke vurdert relevant i fasiten fordi det handlet om benådningen, og ikke om dommen mot Arne Treholt. Termene "Arne" og "Treholt" blir repetert to ganger i samme innslag. Da blir termfrekvensen høy i forhold til tekstens lengde. Navnet finnes sannsynligvis ikke i så mange programmer i forhold til andre mer generelle termer. Da får disse termene også en noe høyere Invers dokumentfrekvens. (McCandless et al., 2010, s. 87) Termfrekvensen og invers

dokumentfrekvens gir her en høy score til innslaget til tross for at det bare har to termer felles med søkespørsmålet.
I resultatene for søkepørsmål nr 63 ("VM i skiskyting") finnes det et annet eksempel på hvordan termfrekvensen får innflytelse på rangeringen. Programmet med nøkkel "2011/02781" blir plassert øverst i trefflistene for søk i begge indeksene. I teksten for programmet er det 3 innslag om VM i skiskyting. I alle innslagene, som har flere andre nyheter seg imellom, finnes alle termene fra søkespørsmålet repetert flere ganger. Det er ikke noen emneord lagt til programmet. Til sammen er termen "VM" repetert 4 ganger og "skiskyting" 5 ganger i hele programmet. Termfrekvensen for begge termene blir høy ved søk i indeksen av programmer. I indeksen av innslag er det ett innslag hvor ordene som er felles med søkespørsmålet er repetert to ganger i en veldig kort tekst. Teksten for innslaget er:

> "4. VM i skiskyting: fellesstart kvinner Direkte fra VM i skiskyting"

Innslaget får derfor høy score, og hele programmet blir rangert etter score fra dette innslaget i trefflisten.
Spørsmål nr 100 ("Joshua French og Tjostolv Moland blir dømt til døden i Kongo") har 5 relevante programmer i fasit. Alle blir funnet og plassert øverst i trefflisten ved søk i begge indeksene. Søk i begge de to indeksene returnerer det best mulige resultatet for presisjon og fullstendighet. En gjennomgang av innholdsbeskrivelsen av de tre første programmene viser at programmene blir rangert høyest på trefflisten fordi alle inneholder navnene på de to involverte i saken og termen "Kongo". Alle termene blir repetert flere ganger i de tre programmene. Termfrekvensen gir utslag for resultatet, og dette bidrar til stor likhet mellom spørsmålsrepresentasjon og dokumentrepresentasjon. Personnavnene bidrar i tillegg med høy invers dokumentfrekvens siden navn er ikke høy frekvente ord gjennom samlingen. Det er ingen emneord ved siden av den innholdsbeskrivende fulltekst, men resultatet for gjenfinningen er likevel det beste resultatet som kunne oppnås.

4.1.3 Invers dokumentfrekvens

Termer som finnes i veldig mange dokumenter kan ikke brukes for å skille dokumentene fra hverandre ved et søk. De er ikke spesifikke for det enkelte dokumentet, og får derfor en lavere diskrimineringsverdi ved invers dokumentfrekvens. Det finnes likevel enkelte tilfeller i

søkeresultatene hvor systemet ikke har klart å rangere dokumentene som forventet, og dokumenter som bare har generelle termer til felles med søkespørsmålet, har blitt rangert høyt i trefflisten.
Et eksempel er i treffene for søkespørsmål nr 120 ("Terrorangrepene i Norge 2011 i Regjeringskvartalet og på Utøya. 77 mennesker omkommer, mange blir skadet, og det ble store materielle ødeleggelser"). I trefflisten fra søk i indeksen av innslagene er det 3 ikke-relevante programmer som blir funnet før ett som er relevant. Det første programmet med nøkkel "2007/29309", består av bare ett innslag med følgende tekst:

> "Ukeslutt Minst tre mennesker er omkommet i bussulykke i Verdal - 24 er sendt til sykehus med skader. Kommunene her i landet bryter loven i stor stil når de gambler med innbyggernes penger, sier redaktøren i Dine penger. Røde Kors vil gi humanitær hjelp til illegale flyktninger - dette er å oppfordre til lovbrudd, sier Fremskrittspartiet."

Innslaget handler om noe helt annet enn søkespørsmålet, men metadataene og spørsmålet har følgende termer felles: "mennesker", "stor", "omkommer". Termene er ikke unike for denne hendelsen og bør få lavere IDF-vekt enn mer spesifikke termer som *named entities* (f eks Utøya eller Regjeringskvartalet). En mulig forklaring kan finnes i den begrensede størrelsen på datasamlingen og i de korte innholdsbeskrivelsene slik at systemet ikke klarer å gi riktige vekt til enkelte termer av generell karakter. En anne forklaring finnes også i valg av søkespørsmålene. Jeg bruker hele setninger fra Wikipedia. Konsekvensen er at det finnes generelle termer i søkespørsmålet. Hadde jeg valgt søk med nøkkelord ville termene "mennesker" og "stor" blitt utelukket fra søkespørsmålet.
Et lignende tilfelle finnes også med søkespørsmål nr 13 som har den korte teksten "Tsjernobyl-ulykke". Her har termen "Tsjernobyl" større betydning for å identifisere relevante innslag enn termen "ulykke". "Tsjernobyl" er et stedsnavn og er stort sett omtalt i mediene i forbindelsen med atomulykken fra 1987, mens termen "ulykken" er en allmenn term som kan brukes i forbindelse med mange flere hendelser. Dette bør fanges opp gjennom beregning av invers dokumentfrekvens.
I trefflisten fra indeksen av programmer blir to ikke-relevante programmer plassert foran et relevant treff. Programmene har nøkkel "1990/01069" og "1990/02869". I programmet med nøkkel "1990/01069" er termen "ulykke" repetert to ganger, mens i det andre programmet finnes den samme termen tre ganger. Dette vektes med termfrekvensen. Begge programmene refererer til andre ulykker enn den i Tsjernobyl. Programmet som er relevant, og som er plassert etter disse har en veldig kort innholdsbeskrivende fulltekst, og ingen emneord som følger den. Teksten til hele innslaget er:

"0004 G7 penger Tsjernobyl 0:21 Reporter: bulletin"

Termen "Tsjernobyl" blir ikke repetert i innslaget og termen "ulykke" er ikke med. Dette resulterer i en lav termfrekvens. Rangeringen i trefflisten viser at termfrekvensen får en høyere innflytelse på resultatet enn invers dokumentfrekvensen i dette tilfellet.
Et siste eksempel er i trefflisten fra søkespørsmål nr 12 ("Arne Treholt dømmes til 20 års fengsel for spionasje") hvor programmet med nøkkel "2011/08240" får bedre posisjon enn relevante programmer ved søk i indeksen av innslag. Innslaget som har felles termer med spørsmålet er

> "Etter at kunstneren Odd Nerdrum ble dømt til to års fengsel for skattesnusk denne uka, har debatten gått i alle kanaler"

Fellestermer er "års" og "fengsel". Invers dokumentfrekvensen klarer ikke å prioritere nok spesifikke termer som navnet "Arne Treholt" foran generelle termer.

4.1.4 Rangeringsmetoden for indeksen av innslag

For å rangere treffene fra indeksen av innslag har jeg valgt metoden kalt for *"Maximum of supporting spans"* av Jimmy Lin. Innslagene fra samme program samles på posisjonen til innslaget som er rangert øverst i trefflisten. Denne metoden favoriserer ifølge Lin de artiklene som bare har ett innslag som er relevant, og som er rangert høyt oppe i trefflisten. (2009, s. [5]) En del av treffene fra indeks av innslag viser denne svakheten ved metoden.
I søkespørsmål nr 33 ("Frankrikes president Jacques Chirac annonserer en "definitiv slutt" på landets atomprøvesprengninger") illustrerer programmet med nøkkel "1995/09791" problemet med den valgte metoden. Programmet har både termen "Frankrike" og "atomprøvesprengning" til felles med søkespørsmålet, og termen "Frankrike" får høyere termfrekvens fordi den finnes både i fullteksten og som emneord. Programmet blir derfor rangert relativt høyt oppe i begge trefflistene. I programmet er det to innslag om temaet, og innslagene blir delt i indeksen av innslag. Termfrekvensen blir derfor ikke så høy i indeksen av innslag siden termene ikke repeteres i hvert innslag. Programmet får heller ikke noe høyere score fordi det er to relevante innslag. Den blir plassert etter innslaget som er høyest trefflisten ifølge den valgte metoden. Resultatet er at programmet med begge innslagene blir plassert høyere i trefflisten fra søk i indeksen av programmene, enn i trefflisten fra søk i inndeksen av innslag. Forskjeller mellom posisjonene er ikke så stor: i trefflisten fra søk i indeksen av program er programmet i posisjon nr 4 og i trefflisten

fra søk i indeksen av innslag i posisjon nr 6. Søk i indeksen av innslag gir til slutt 5 ikke-relevante treff øverst i trefflisten, noe som ikke er ønskelig.
Valget av denne metoden får større konsekvenser for rangeringen ved søkespørsmål nr 120.
Det er 22 relevante programmer til dette søket. Ved søk i indeksen av programmer blir 9 av de 22 relevante programmene funnet innen de første 20 treffene. Ved søk i indeksen av innslag er 7 relevante programmer funnet innen de første 20 treffene. Det er de samme programmene som blir funnet i de to indeksene. De 2 relevante programmene som blir funnet bare ved søk i indeksen av programmer har nøkkel "2011/07894" og "2012/07277". Disse finnes i posisjon nr 60 og nr 47 ved søk i indeksen av innslagene. Begge programmene har flere innslag som handler om samme hendelse. Termene fra søkespørsmålet er i ulike innslag. Hver for seg får derfor innslagene lavere score. Termfrekvensen for termene og antall felles termer øker når innslagene er samlet i ett program. Her vises det igjen hvordan den valgte rangeringsmetoden for innslagene får problemer med å plassere programmer der flere av innslag er relevante. I dette tilfellet blir programmene plassert langt nede i trefflisten, og de ville sannsynligvis ikke blitt oppdaget av en bruker.

4.1.5 Karakteristikkene ved spørsmålene

Et av kriteriene for valg av spørsmål var at beskrivelsene fra Wikipedia skulle ha forskjellig lengde for å se på eventuelle konsekvenser for gjenfinningen i samlingen.
De tre korteste spørsmålene er nr 13 ("Tsjernobyl-ulykke"), nr 63 ("VM i skiskyting") og 87 ("Kommune- og fylkestingsvalg i Norge"). Når teksten på spørsmålene formaliseres av *QueryParser* i Lucene, blir disse til en sekvens av substantiver som ligner mer på et søk med nøkkelord enn et søk i naturlig språk (*natural language queries*).
Søkespørsmål nr 13 er allerede blitt omtalt i forbindelse med invers dokumentfrekvens. Spørsmålet består av bare to termer, hvor den ene er veldig generell og kan finnes i flere programmer. Spørsmålet ville sannsynligvis fått bedre resultater med en AND-forbindelse i forhold til OR-forbindelse som er standard for alle spørsmålene i *QueryParser*. Med AND-forbindelse ville de programmene som bare inneholder termen "ulykke" blitt utelukket i stedet for å komme foran relevante programmer.
Det samme skjer for spørsmål nr 87 hvor programmet med nøkkel "2007/29846" og som handler om norske kommuner uten å handle om valg, blir plassert øverst i trefflisten for indeksen av programmer. Eksempelet er allerede nevnt i analysen av lengdenormaliseringen. Resultatet er igjen en kombinasjon av ulike faktorer, hvor det korte spørsmålet og bruk av OR-forbindelse bidrar til at

ikke-relevante treff kommer langt oppe i trefflisten.
Det siste av de korte søkespørsmålene er nr 63. Resultatene for søket er ikke så mye knyttet til søkespørsmålet, men mest til andre forhold ved gjenfinningen. Det er to relevante programmer i fasit. Det første av de relevante programmene er rangert øverst i begge trefflistene på grunn av høy termfrekvens for begge termene fra søkespørsmålet. Det andre relevante programmet har et innslag om flere sportsbegivenheter hvor VM i skiskyting er nevnt. Termen som nærmer seg mest termene i spørsmålet er "skiskytter-VM". Termen "skiskyting" blir ikke gjenkjent av systemet som lik "skiskytter". Det er ikke noe kontrollert vokabular som justerer dette, og rotlemmatiseringen forkorter ikke termene nok. Derfor blir programmet rangert som nr 64 ved søk i indeks av programmer og som nr 52 ved søk i indeks av innslag.
Blant de lengste søkespørsmålene er søkespørsmålene nr 22 ("Norge blir klare for VM i fotball 1994 etter å ha slått Polen 3-0 på bortebane. Jostein Flo, Jan Åge Fjørtoft og Ronny Johnsen skåret målene") og nr 129 ("Flommen på Østlandet 2013: Deler av tettstedet Kvam i Gudbrandsdalen ble ødelagt, og veier og jernbaner ble stengt").
For lange søkespørsmål ser det ut som OR-forbindelsen fungerer bedre. Her er det mange termer, og man kan ikke forvente å finne alle i enhver beskrivelse. De lange søkespørsmålene ville ikke gitt noe treff med AND-forbindelse. Etter indekseringen består hver spørsmålsrepresentasjon av både generelle termer som f eks "fotball", "jernbaner", og flere *named entities* (personnavn og stedsnavn).
Til tross for mange termer og flere av disse er veldig spesifikke, får likevel ikke lange søkespørsmål noe bedre resultater enn de korte. I forbindelse med eksemplene for invers dokumentfrekvens, har jeg allerede pekt på problemer med søkespørsmål som inneholder flere generelle termer. Noen av resultatene viser at generelle og høyfrekvente termer blir vektlagt like mye som *named entities* som har lavere frekvens i samlingen. Systemet klarer ikke å differensierer dem tydelig nok gjennom IDF, og det får konsekvenser for rangeringene i trefflistene.
Et eksempel på at flere spesifikke termer ikke alltid er nok til å få det best mulige resultatet ses i søkespørsmålet 129. I begge trefflistene blir programmet med nøkkel "1996/00689" plassert øverst selv om det ikke er relevant. Programmet inneholder et innslag om et tidligere flom på Østlandet, og har flere termer felles med søkespørsmålet. Søkene er begrenset til feltet for innhold, og det tas ikke hensyn til sendedato. Dette søket ville fått noe bedre resultat ved å ta hensyn til alle feltene for registreringen.

4.1.6 Fasit

Valg av programmene for fasiten er avhengig av hvordan beskrivelsene fra Wikipedia blir tolket. Som forklart i metodekapittel, blir ikke beskrivelsene av hendelsene tolket for strengt. Fasiten kan derfor inneholde programmer som har ulik grad av relevans i forhold til søkespørsmålet. Målene for gjenfinningseffektivitet tar ikke hensyn til dette, og presisjon og fullstendighet får mindre betydning for vurderingen av trefflistene i disse tilfellene. Det er i stedet interessant å finne ut hvordan systemet rangerer de relevante programmene som handler om enkelte aspekter eller konsekvenser beskrevet i søkespørsmålene. Når et program bare handler om en del av søkespørsmålet kan avstanden mellom spørsmålsrepresentasjon og dokumentrepresentasjon bli for stor, og systemet klarer ikke å rangere dem høyt i trefflistene.
Et eksempel på dette er søkespørsmålet nr 120 ("Terrorangrepene i Norge 2011 i Regjeringskvartalet og på Utøya. 77 mennesker omkommer, mange blir skadet, og det ble store materielle ødeleggelser"). Det er flere programmer fra fasiten som ikke blir funnet innenfor de første 100 treffene ved søk i de to indeksene. Programmene som ikke blir funnet har nøkkel: "2012/06962", "2012/05480", "2012/05152", "2012/04889", "2012/04669", "2012/03679", "2012/01597". Alle disse handler ikke direkte om hendelsen, men om konsekvenser av terroraksjonen. Innholdsbeskrivelsene og søkespørsmålet har for få termer til felles.
Spørsmål nr 12 («Arne Treholt dømmes til 20 års fengsel for spionasje») har også noen programmer i fasit som ikke handler om selve dommen, men om ulike aspekter ved rettssaken før dommen falt, og om andre personer involvert i saken. Selv om disse kan tolkes som litt på siden av beskrivelsen gitt i søkespørsmålet, blir de rangert langt opp i trefflistene. Dette er det motsatt av resultatene for søkespørsmål nr 120. Eksempler er programmene med nøkkel "1990/01946", "1990/01951", "1990/02507", "1990/01945". Alle programmene er rangert øverst i trefflistene fra søk i begge indeksene. En mulig forklaring kan finnes i emneordene som beskriver innslagene. Alle innslagene har de samme emneordene, og i emneordene er termen "spionasje" felles for søkespørsmålet og innslagene.
Søkespørsmål nr 120 og nr 12 viser at det er vanskelig å finne noen generelle forklaringer i treffene som er direkte konsekvenser av fasiten. Indekseringspraksisen for både den innholdsbeskrivende fullteksten og det kontrollerte vokabularet spiller en stor rolle i forhold til hvilke programmer fra fasiten som blir funnet. Det er balansen mellom uttømmenhet (*Exhaustivity*) og spesifisitet (*Spesificity*) ved indekseringen som blir avgjørende. Spesifisiteten i indekseringen gir mulighet for å finne nyanser og ulike grader av relevans i forhold til søkespørsmålet. Den skal gjøre det mulig å

skille de mest relevante fra mindre relevante programmer selv om alle handler om samme overordnede tema. Samtidig kan uttømmende indeksering dekke flere aspekter av innholdet. (Lancaster, 2003, s. 27-35)

4.2 Forhold mellom program og innslag ved søk

De innholdsbeskrivende metadataene for hvert program beskriver innslag som i enkelte programmer har et felles gjennomgående tema, men som for det meste er beskrivelser av ulike hendelser. Hendelsene har bare til felles at de har skjedd i samme tidsperiode. Et eksempel på dette er årskavalkadene hvor man oppsummerer de viktigste hendelsene for ett år i ett program. Med programmene som enhet for søket har man derfor en kombinasjon av helt uavhengige beskrivelser som kan skape problemer med gjenfinning når de er blandet.

Noen av søketreffene til spørsmål nr 27 ("Norge stemmer nei til EU-medlemskap") eksemplifiserer en av konsekvensene av å bruke programmene som enhet for søket i stedet for å bruke innslagene. I dette søket får tre relevante programmer bedre posisjon i trefflisten i forhold til søk i indeksen av innslagene, fordi i de samme programmene finnes innslag om andre temaer som bruker termer fra spørsmålet. Man får en positiv effekt, men dette er i utgangspunktet ikke ønskelig, og kan fort gi motsatt resultat. Programmene dette gjelder har nøkkel "1994/07018", "1994/28792" og "1994/27574". Alle programmene har termene "Norge" og "EU" til felles med søkespørsmålet. Termen "Norge" finnes i innslag som handler om noe annet enn EU-medlemskap. Programmet får dermed høyere score på grunn av høyere termfrekvens for denne termen. Mens søket i programmene som enheter får fordeler av termer fra ikke-relevante innslag, søket i programmene fordelt som innslag får problemer på grunn av valget av rangeringsmetode for innslagene. Hvert av programmene har flere enn ett innslag som er relevant, men rangeringsmetoden tar bare hensyn til det innslaget som har høyest score.

Et lignende eksempel på hvordan søk i indeksen av programmene får uønsket nytte av andre ikke-relevante innslag, finnes også ved søkespørsmål nr 87 ("Kommune- og fylkestingsvalg i Norge"). Programmet med nøkkel "1996/00689" er relevant. I det relevante innslaget er termene "kommunevalget" og "fylkestingsvalget". I tillegg finnes termen "Norge" i to andre innslag fra samme program. Når søket gjøres i indeksen av programmer blir derfor den plassert høyt i trefflisten, i posision nr 2. I indeksen hvor innslagene er delt hver for seg, er ikke termen "Norge" med lenger, og programmet blir rangert lengre ned i trefflisten, i posisjon nr 18.

Resultatet for søk i programmer som inneholder beskrivelser av helt forskjellige innslag, kan fort bli

det motsatte av disse eksemplene. Det kan illustreres med et hypotetisk søkespørsmål som blir stilt til datasamlingen, f eks "Valg i USA". Ett program kunne hatt ett innslag om USAs økonomi og ett annet innslag om valg i Norge. Ved å stille søkespørsmålet "Valg i USA" til systemet, ville et slik program bli vurdert som relevant. Begge termene til søkespørsmålet er med i programmet samlet og ville gitt en høy score i sammenligningen hvis man ikke har en inndeling av innslagene.
En gjennomgang av de ulike søkene viser at søk i indeksen av innslag får til litt bedre resultater for presisjon i flere av søkene, men marginene er veldig små, og det er ikke tilfellet for alle søk.
Treffene fra spørsmål nr 13 er et eksempel på noe bedre presisjon på toppen av trefflisten. Søkespørsmålet har 5 mulige relevante treff i fasit. Alle 5 blir rangert øverst i trefflisten ved søk i indeksen av innslag. I trefflisten fra indeksen av program er de 4 første relevante, mens det siste relevante programmet finnes som treff nr 7. Presisjonen er noe bedre for indeksen av innslag, men forskjellene er marginale.
Treffene fra søkespørsmål nr 12 viser det motsatt av søkespørsmål nr 13. Søk i indeksen av programmer gir bedre resultater enn søk i indeksen av innslag. Det er 12 mulige relevante programmer i fasit. Ved søk i indeksen av programmer finner systemet alle de 12 relevante programmene og plasserer dem øverst i trefflisten. Dette gir best mulig resultat for presisjon og fullstendighet. Indeksen av innslag har samme resultat tom treff nr 7. Etter 7 treff finnes det en del ikke-relevante programmer i indeksen av innslag, og systemet finner 10 av de 12 relevante treffene innenfor de første 20 treff. Både fullstendighet og presisjon er dårligere. Indekseringen etter innslag bidrar ikke til å forbedre effektiviteten i gjenfinningen i dette tilfellet.
Tre av søkespørsmålene gir i tillegg helt like resultater på presisjon i de to indeksene. Det er søkespørsmål nr 63, 100 og 121.
Forskjellene mellom søkene i de to indeksene er enda mindre når det gjelder resultatene for fullstendigheten. 10 av de 16 søkespørsmålene finner like mange relevante treff innenfor de første 20 treffene i de to indeksene. Det er søkene nr 13, 20, 22, 27, 54, 63, 100, 118, 121 og 129. Forskjellene mellom de resterende søkene er minimale. Det er en forskjell på ett eller to programmer som ikke blir funnet i den ene eller den andre indeksen.
Den oppnådde presisjonen på toppen av trefflisten blir viktigere for å evaluere forskjellen mellom søk i programmer og i innslag, når fullstendigheten er såpass lik ved søk i de to indeksene. Samtidig bekrefter dette de marginale forskjellene.
Tendensen for at gjenfinningseffektiviteten er såvidt bedre i indeksen av innslag bekreftes av resultatene fra beregningene av *R-precision*, P@n og MAP.

4.2.1 *R-precision*, P@n og MAP.

R-precision

R-precision er beregnet for resultatet av hvert enkelt søkespørsmål i begge indeksene og resultatene vises i fig. 4.

Spørsmåls-nummer	***R-precision* for indeks av programmer**	***R-precision* for indeks av innslag**
12	1,000	0,833
13	0,800	1,000
20	0,500	1,000
22	0,200	0,200
24	0,550	0,600
27	0,444	0,472
32	0,500	0,500
33	0,000	0,000
54	0,375	0,375
63	0,000	0,500
87	0,286	0,333
100	1,000	1,000
118	0,000	0,500
120	0,455	0,318
121	1,000	1,000
129	0,000	0,000

Fig.4. *R-precision* for hvert av søkene i indeks av programmer og indeks av innslag.

R-precision for hvert spørsmål bekrefter de små forskjellene mellom de to indeksene gjennom de ulike søkespørsmålene. Innenfor disse små marginene viser tabellen noe bedre resultater på presisjon for indeksen av innslag. 6 av de 16 søkene har helt like verdier på *R-precision*, 7 søk har noe høyere verdier i indeksen av innslag, og 2 i indeksen av programmer.
Ulempen med dette målet er at når det totale antallet relevante dokumenter er høy, sier målet veldig lite om hvor god systemet er til plassere de relevante dokumentene øverst på trefflisten. I og med at det er for få relevante programmer i fasiten for de ulike søkene, er dette ikke et stort problem i denne sammenheng. 6 av de 16 søk har fra 1 til 3 programmer i fasit, og det er bare tre søk som har flere enn 20 mulige relevante programmene i fasit. Målet viser derfor til representative resultater for

den oppnådde presisjonen i de to indeksene for de enkelte spørsmålene.

Hendelsene for søkespørsmålene var bevisst valgt fra ulike år for å kunne søke i metadataene skapt i ulike tidsperioder. Målet var å se om det var tydelige forandringer i indekseringspraksis som kunne avdekkes ved søk i samlingen. Resultatene for *R-precision* viser ikke noen forskjeller i indekseringspraksis over tid som har hatt konsekvenser for resultatene. Dette illustreres tydelig i tabellene til fig. 5 hvor *R-precision* øker og minsker fra et søk til det andre uten å ha en generell tendens fra det første til det siste søket. Samme resultat vises for søk i begge indeksene.

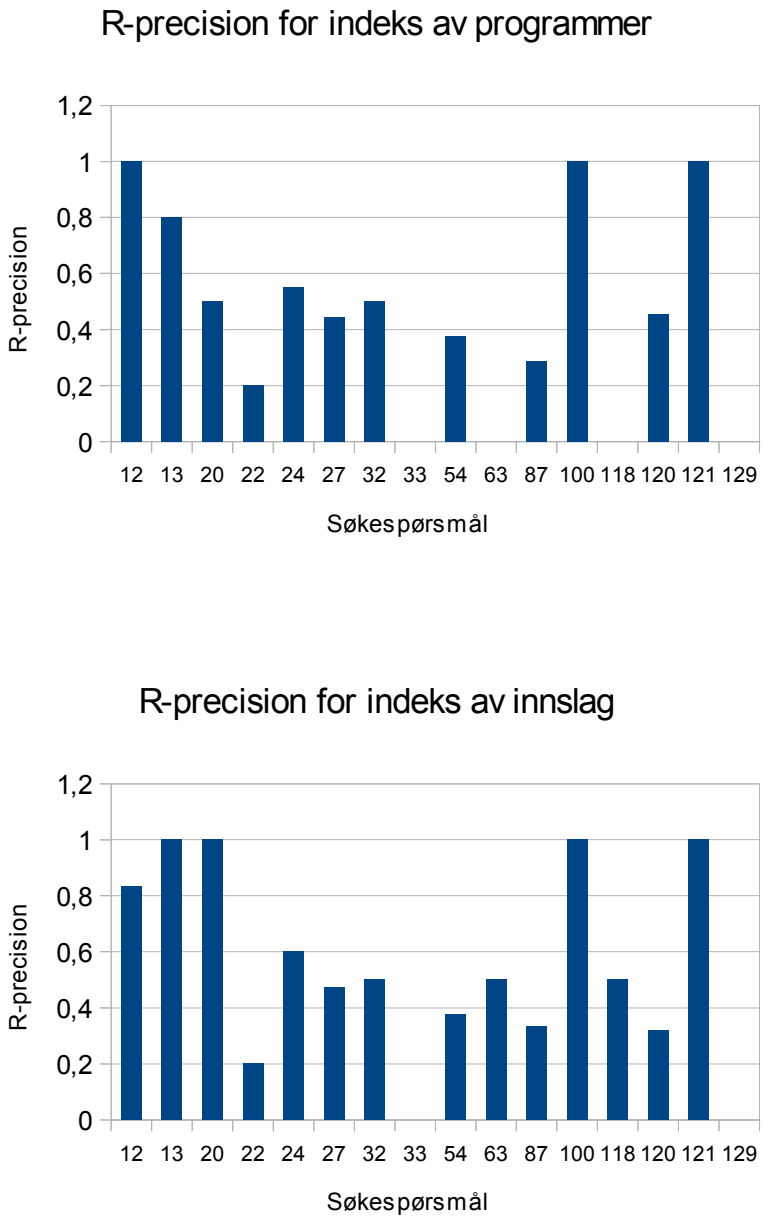


Fig.5. Diagrammer med verdiene for *R-precision*.

P@n

P@n er regnet for de vanlige verdiene ved 5, 10 og 20 treff og i tillegg ved 3 treff. Resultatene vises i fig. 6.

P@n	Indeks av programmer	Indeks av innslag
P@3	2,2140	2,2496
P@5	1,6686	1,6712
P@10*	0,6649	0,6303
P@20*	0,4838	0,4741

Fig.6. P@n i indeks av programmer og indeks av innslag

* Verdien av P@10 og P@20 er beregnet på de 7 søk som har flere enn 5 programmer i fasit

P@3 og P@5 viser hvor godt systemet er til å finne fram til relevante programmer helt øverst på trefflisten. Resultatene for beregning av P@n bekrefter de vurderingene av søkeresultatene som er gjort i analysen så langt. Forskjellene mellom søkene i de to indeksene er marginale, men tendensen er at søk i indeks av innslag gir litt bedre resultater for presisjon på toppen av trefflisten. Verdiene for P@3 og P@5 er såvidt høyere for indeksen av innslag.
Lenger ned i trefflisten er det indeksen av programmer som blir bedre. Verdiene av P@10 og P@20 er ikke beregnet for de søkene som har 5 eller færre relevante programmer slik som forklart i § 3.4.

MAP

MAP er en enkel verdi for gjennomsnittlige presisjon for alle søkene i hver av indeksene. Resultatet er i fig. 7. MAP tar hensyn til flere presisjonsverdier enn P@n som stopper ved *n'*te treffet. Som forklart i § 3.4 er det likevel satt en grense for hvilke presisjonsverdier som er med. Der de relevante programmene er funnet etter posisjon nr 100, er deres presisjon satt til 0. Sannsynligheten for at brukerne kommer til å se opp til treff 100 er lav, og presisjonsverdien er sannsynligvis nærmest null så langt nede i trefflisten.

	Indeks av programmer	Indeks av innslag
MAP	0,5961	0,6343

Fig.7. MAP i indeks av programmer og indeks av innslag

Resultatet for MAP viser igjen at systemet gir nærmest like resultater ved søk i de to indeksene, men at søk i indeksen av innslag får litt høyere verdi.

4.3 Indekseringspraksis

Indekseringspraksisen har direkte konsekvenser for resultatene av gjenfinningseffektivitet. Her skal jeg se på indekseringen av den innholdsbeskrivende fullteksten og det kontrollerte vokabularet, og deres funksjon for gjenfinningen i de to indeksene. Analysen gjøres både ved å bruke søkeresultatene fra de 16 søkespørsmålene og med en generell beskrivelse av indekseringen utenom gjenfinningsforsøket.

4.3.1 Innholdsbeskrivende fulltekst

Ett av kriteriene for valg av søkespørsmålene var at disse skulle dekke hele tidsperioden for datamaterialet. Antagelsen er at det kunne være variasjoner i hvordan den manuelle indekseringen ble gjort gjennom en lang tidsperiode. I tillegg kan selve programmet ha forandret seg en del i løpet av denne tiden. Spesielt for den innholdsbeskrivende fullteksten har man ikke de samme overordnede reglene for spesifisitet, grundighet og bruk av termer som finnes for et kontrollert vokabular, og som kan bidra til konsistensen i dokumentrepresentasjonen. (Lancaster, 2003, s. 68-77)

For å se variasjoner over tid i indekseringen er det registreringsdatoen til metadataene som er viktigere enn akkurat når hendelsen skjedde. I datamaterialet er det ikke fullt samsvar mellom når en hendelse skjedde, når den ble omtalt ("opptaksdato" og "sendedato") og registreringsdato ("Reg-dato"), spesielt for de første årene. Registreringsdato finnes ikke for hendelsene fra 70-tallet. Det første søkespørsmålet (nr 12) referer f eks til en sak fra 1986, men noen av de relevante programmene er ifølge systemet registrert i 1988. Til tross for at det er noen forskjeller mellom sendedato og registreringsdato, er det fortsatt over 20 år forskjell mellom registreringen av innslag som refererer til det første spørsmålene og den siste som er fra 2013. Ikke noen av programmene ble registrert lang tid etter at de ble sendt.

En sammenligning av presisjon og fullstendighet gjennom årene for de ulike spørsmålene viser ikke noen tydelige forskjeller og forandringer mellom de eldste registrerte hendelsene og de nyeste. (Vedlegg 4) Det er hverken en tydelig forbedring eller forverring av presisjon og fullstendighet

innenfor de første 20 treffene gjennom årene uansett om man ser på indeksen av programmer eller av innslag. Dette bekreftes også av verdiene av *R-precision* for hvert søk. I diagrammet over *R-precision* øker og minsker den samlet verdien for presisjon uten å følge et mønster fra de første til de siste søkene. (Fig. 7)

Hvis det har vært noen forskjeller i indekseringen over tid har disse ikke fått noen synlige konsekvenser for resultatene for de valgte spørsmålene. En analyse av den innholdsbeskrivende fullteksten for de enkelte søkeresultatene peker likevel på en del mangelfull og inkonsekvent indeksering i den innholdsbeskrivende fullteksten som har konsekvenser for søkeresultatene. Dette kan illustreres med enkelte eksempler.

Noen innslag blir beskrevet med veldig få ord i fullteksten, andre har langt flere termer. Konsekvensene for gjenfinningen blir tydelig i trefflistene for søkespørsmål nr 22 ("Norge blir klare for VM i fotball 1994 etter å ha slått Polen 3-0 på bortebane. Jostein Flo, Jan Åge Fjørtoft og Ronny Johnsen skåret målene"). Programmet med nøkkel "1994/11025" står i fasit, og har følgende innslaget som beskriver hendelsen:

> "0017 Fotball VM RSPN 3:55 Reporter: bredeli"

Fulltekstbeskrivelsen er veldig kort i forhold til innslaget i et annet relevant program ("2007/22013") som er sendt året i forveien:

> "21. FOTBALL (e. 41'15") 930602: Norge slår England 2-0 i VM-kvalifiseringskamp. Scoringene til Øyvind Leonhardsen og Lars Bohinen (41'15"-41'50") 931013: Norge slår Polen 3-0 i landskamp (VM-kvalifiseringskamp). Scoringene til Jostein Flo, Jan Åge Fjørtoft og Ronny Johnsen. Kutt landslagstrener Egil OLSEN (mv) (Drillo) (41'50"-42'50") 931220: VM-trekningen i Las Vegas. Kutt Egil OLSEN (mv) (Drillo) (43'00"-43'30") (K795.33 - Fotball (eo))"

Dette siste programmet er rangert øverst i trefflistene fordi likheten mellom termene i spørsmålet og i innslaget er stor. Disse har både de generelle termene "VM", "fotball" og "Norge" og navnene på fotballspillerne til felles. Innslaget fra året etter er i stedet såpass mangelfullt at systemet ikke klarer å rangere dette foran andre treff. I den ene beskrivelsen blir det gjengitt resultatet fra spesifikke fotballkamper og fotballspillernes navn, i den andre får man bare gjengitt det overordnede emnet for innslaget. Hva innslaget spesifikt handler om er vanskelig å si. Tema kan ha vært enkelte kamper i VM, men ikke nødvendigvis Norge sine kamper, eller andre forhold rundt VM, som åpningen eller avslutningen av VM eller mottakelsen av VM i det landet den holdes. Dette gjør det også vanskelig å vurdere om innslaget er relevant i forhold til søkespørsmålet.

Dette eksempelet viser noe av forskjellen i uttømmenhet og i spesifisitet for indekseringen av to

beslektede innslag til tross for at det bare er ett år mellom de to registreringene. Lancaster hevder at fullteksten kan bidra til spesifisitet i beskrivelsen siden det er flere termer man kan få treff på, mens det kontrollerte vokabularet kan bidra med overordnede og konseptuelle termer som oppsummerer en hel fulltekst. (Lancaster, 2003, s. 270) I dette eksempelet ligner den ene innholdsbeskrivende fullteksten på Lancasters beskrivelse av et kontrollert vokabular. Innslaget inneholder bare to termer som er en overodnet beskrivelse av innslaget. Det andre innslaget har i stedet en innholdsbeskrivende fulltekst som består av langt flere termer og er mer detaljert. Denne inkonsekvensen i indekseringen påvirker søkeresultatet, der det ene programmet er rangert i begge trefflistene i posisjon nr 1, og det andre er i posisjon nr 22 i indeksen av innslagene og posisjon nr 63 i indeksen av programmene.

Et annet eksempel på forskjellene i indekseringspraksis vises av resultatene for søkespørsmål nr 24 ("Vinter-OL 1994 arrangeres på Lillehammer"). 20 programmer er vurdert som relevante i forhold til søkespørsmålet. Ved søk i indeksen av innslag får man en treffliste med 12 av de 20 relevante programmene innenfor de 20 første treffene. Ved søk i indeksen av programmer plasserer systemet 11 av de 20 relevante innenfor de første 20 treffene. Det er de samme relevante programmene som blir rangert omtrent likt i de to trefflistene. Det ene programmet som ikke kommer i søket fra indeksen av programmer er i trefflisten plassert som nr 21. Det er derfor ikke store forskjeller mellom de to trefflistene. De resterende 8 treffene som var relevante kommer noe lenger ned på trefflisten. Bare programmet med nøkkel "1994/04544" blir ikke rangert blant de første 100 treff av systemet i noen av indeksene. Det relevante innslaget har følgende innholdsbeskrivende fulltekst:

> "0018 Paralympics Reporter: Tom Eriksen RNYA 3:10"

Systemet klarer selvfølgelig ikke å finne fram til dette når ingen av termene fra søkespørsmålet finnes i innholdsbeskrivelsen. Det er to aspekter her som er interessante, innslagets relevans i forhold søkespørsmålet og spesifisitetsnivået ved innholdsbeskrivelsen. Hvis søkespørsmålet tolkes som et søk om alt som har med OL på Lillehammer å gjøre, er også Paralympics en del av dette, og det ville ha vært ønskelig med mer generelle termer som beskriver innslaget enn bare den mest spesifikke termen. Emneordene "OL" og "Lillehammer" er brukt i flere andre innslag og kobler de spesifikke innholdsbeskrivelser fra fullteksten sammen, men de er utelatt her. Termen er ikke en del av de kontrollerte emneordene, men er den eneste fulltekstbeskrivelsen til innslaget. Den viser likevel hvordan høy grad av spesifisitet fører til dårlig fullstendighet. Hjortsæter skriver om dette i rapporten "Emneordskatalogisering". Det skal brukes "det snevrest mulige, det meste spesielle emneordet som uttrykker dokumentets innhold". (Hjortsæter, 2005, s. 23) Dette fører til økt

presisjon på bekostning av fullstendigheten, noe som skjer i eksempelet.
Et annet relevant program ("1993/21480") for samme søket ble indeksert bare noen få måneder tidligere. Programmet inneholder tre innslag om OL. I alle innslagene brukes både fulltekst, emneord og klassifikasjonsnummer. Innslagene blir beskrevet med følgende termer:

> "0016 Frivillige medarbeidere til OL øver på jobbene DOPP 3:10 Alle må registreres og få legitimasjonskort - reportasje fra registrering. Intervju med John LILLEBRÅTEN (mv) (dial. Vågå), løypekryssmann. (K796.06 - OL Olympiske leker (eo) + K948.25 - Lillehammer (eo)) Reporter: Eriksen, Tom Rune 0017 Prinsesse Märtha Louise tente OL-ilden i Morgedal i dag RNYA 3:10 Kutt med Eivind STRONDLI (mv), som gir tyrifakkelen til prinsessen. Han har laget fakkelen, er skiveteran. Prinsesse Märtha Lousie tenner fakkelen (mye blitzing), og overrekker fakkelen til Åshild Lofthus. (K796.06 - OL Olympiske leker (eo)) Reporter: Boyesen + Eriksen - 0018 Reportasje fra Morgedal sentrum, der fakkelstaffeten RNYA 2:00 starter. Intervju med Dag Åsmund LARSON (mv). Reporter: Andersen"

Forskjellen i indekseringspraksis er stor selv i en kort tidsperiode, både når det gjelder innholdsbeskrivende fulltekst og bruk av emneordene. Det er høy spesifisitet i den innholdsbeskrivende fullteksten, mens det kontrollerte vokabularet kobler de spesifikke innslagene sammen med det overordnede temaet for innslagene.
Et siste relevant program knyttet til samme søk blir indeksert som treff nr 100 ved søk i indeksen av programmer. Programmet med nøkkel "1997/02229" og det relevante innslaget følgende innholdsbeskrivende fulltekst uten emneord:

> "0011 Kutt fra Lilleham-OL RNYA 0:10 Reporter:"

Beskrivelsen er igjen mangelfullt med i tillegg noen skrivefeil i teksten. Når den blir sammenliknet av systemet med innslag som både inneholder kontrollerte emneord og utfyllende fulltekstbeskrivelser, taper slike innslag på grunn av inkonsekvens i indekseringspraksis. Systemet kan ikke justerer inkonsekvensen i beskrivelsen, og dette får konsekvenser for trefflistene.

4.3.2 Det kontrollerte vokabularet

En gjennomgang av søkeresultatene viser at det kontrollerte vokabularet hovedsakelig har to funksjoner for gjenfinningen når den blir sett i sammenheng med den innholdsbeskrivende fullteksten og med den valgte gjenfinningsmodellen. Enkelte av søkeresultatene viser at det kontrollerte vokabularet forbedrer beskrivelsen ved å utvide vokabularet som allerede finnes i den innholdsbeskrivende fullteksten. I andre tilfeller bidrar det kontrollerte vokabularet til en høyere

termfrekvens for enkelte termer som allerede finnes i den innholdsbeskrivende fullteksten. Problemet med indekseringen av det kontrollerte vokabularet er mangelen på konsistens. Ikke alle innslag blir beskrevet med emneord eller tagger. I tillegg er det ulike grader av grundighet ved indekseringen siden ikke alle aspekter ved et emne blir dekket ved alle de lignende innslagene. Dette får konsekvenser for rangeringene av programmene i trefflistene.

Et eksempel er søkespørsmål nr 24 ("Vinter-OL 1994 arrangeres på Lillehammer"). I programmene som er fasit for spørsmålet, er det forskjeller i bruken av det kontrollerte vokabularet. De fleste av programmene som er rangert øverst i trefflistene har både termene OL og Lillehammer i den innholdsbeskrivende fullteksten og i det kontrollerte vokabularet. Dette gir høyere score til programmene på grunn av termfrekvensen. Flere av de relevante programmene som ikke er rangert blant de første 20 treffene, har i stedet bare fått emneordet "OL", og ikke "Lillehammer". Det er ulik grundighet i beskrivelsen av programmer indeksert i løpet av en kort periode. Eksempler er programmene med nøkkel "1993/20002", "1994/01361", "1994/02906", "1994/03282". For enkelte er mangelen på emneord kompensert med at termen Lillehammer finnes i den innholdsbeskrivende fullteksten. Programmene med nøklene "1993/20002", "1994/01361", "1994/02906" har termen verken i fullteksten eller blant emneordene. De faller derfor langt nede i trefflistene. Ved søk i indeksen av programmer står de i posisjon nr 98, 97 og 85, og ved søk i indeksen av innslag i posisjon nr 90, 73 og 57.

Søkespørsmål 33 ("Frankrikes president Jacques Chirac annonserer en "definitiv slutt" på landets atomprøvesprengninger") er et annet eksempel på inkonsekvent indekseringspraksis ved registreringen av det kontrollerte vokabularet. Spørsmålet har to programmer som er relevante. Det relevante programmet med nøkkel "1995/09791" har termene "Frankrike" og "atomprøvesprengning" felles med søkespørsmålet, og "Frankrike" er både i den innholdsbeskrivende fullteksten og i emneordene. Termfrekvensen for termen øker, og programmet får høyere score. Det andre relevante programmet "1995/12919" har bare termen "atomprøvesprengning" felles med søkespørsmålet. Termen "Frankrike" finnes ikke som emneord og i fullteksten brukes termen "fransk". Rotlemmatiseringen som gjøres av systemet, hjelper heller ikke til å koble termene "Frankrike" og "fransk", og programmet er rangert som nr 45 ved søk i indeksen av programmene.

Søkespørsmålet nr 120 ("Terrorangrepene i Norge 2011 i Regjeringskvartalet og på Utøya. 77 mennesker omkommer, mange blir skadet, og det ble store materielle ødeleggelser") får ikke gode resultater for presisjon og fullstendighet. Ved søk i indeksen av programmer blir bare 9 av de 22 relevante programmene funnet innen de første 20 treffene. Ved søk i indeksen av innslag er 7

relevante funnet innen de første 20 treffene. I trefflisten fra søk i indeksen av innslag er det heller ikke noen av de relevante innslagene plassert øverst, og forklaringen finnes i valget av rangeringsmetoden. Resultatet for søk i indeksen av programmer er noe bedre, men fortsatt ikke så optimalt som for andre søk. Hvis man ser på bruk av tagger for de ulike programmene, er det bare to av de relevante programmene som har fått tagger, og begge programmene blir rangert blant de første treffene. Det er programmene med nøkkel "2011/07539" i posisjon nr 2 og "2011/08240" i posisjon nr 5. Disse får en fordel i fohold til mange andre programmer hvor termene bare finnes i den innholdsbeskrivende fullteksten. Programmet i posisjon nr 1 ("2011/07801") er også relevant, og er rangert øverst uten å ha tagger fordi den har en veldig lang innholdsbeskrivende fulltekst med flere innslag om temaet og derfor høy repetisjon av termene.

4.3.3 Indekseringspraksis og karakteristikker ved metadataene

Resultatene fra søkene viser karakteristikkene ved de registrerte metadataene, og hvordan disse påvirker gjenfinningen. Evalueringen ved bruk av mål for gjenfinningseffektivitet baseres på de metadataene som finnes og fasit bygges fra de samme metadataene. For å få bedre forståelse av den manuelle indekseringen av dataene og eventuelle konsekvenser for gjenfinningen, har jeg valgt å presentere noen tall om mengde og typer metadata som viser sider av indekseringspraksis som ikke vises ved en analyse av gjenfinningen, men som likevel har konsekvenser for søkeresultatene. Noen av tallene som beskriver den innholdsbeskrivende fullteksten og det kontrollerte vokabularet i samlingen er oppsummert i fig. 8.

	Antall programmer	**Antall innslag**
Har ikke innholdsbeskrivelse i feltet for innhold	591 (35%)	591 (10%)
Har bare innholdsbeskrivende fulltekst i feltet for innhold	414 (25%)	3516 (59%)
Har bare kontrollert vokabular i feltet for innhold	0	0
Har både innholdsbeskrivende fulltekst og kontrollert vokabular i feltet for innhold	661 (40%)	1776 (30%)
Totalt antall	1666	5883

Fig.8. Bruk av innholdsbeskrivende fulltekst og kontrollert vokabular ved den manuelle indekseringen

Ved den manuelle inndelingen av dataene i innslagene ble jeg oppmerksom på at ikke alle innholdsfeltene har en beskrivelse av innholdet. 14 programmer har helt tomt innholdsfelt og 577 har en tekst som ikke er en innholdsbeskrivelse. Disse programmene har i stedet en referanse til andre systemer hvor data er bevart, eller inneholder bare tittelen for programmet. Til sammen er det 591 programmer, dvs. 35% av de registrerte "Ukeslutt"-programmene, som ikke har en innholdsbeskrivelse som kan brukes for gjenfinningen av innhold.

Programmene med en tekst som ikke referer til innhold er alle fra perioden 2001-2010. At det store flertallet er samlet i en tiårsperiode viser forandringer i indekseringspraksisen for akkurat disse ti årene. Dette hadde ikke noen synlige konsekvenser for de 4 søkene jeg hadde fra tiårsperioden (søkespørsmål nr 54, 63, 87 og 100). Blant kriteriene for valg av søkespørsmål var at det skulle finnes metadata i samlingen om hendelsene beskrevet i spørsmålene. Det skulle være mulig å finne en fasit og beregne gjenfinningseffektivitet basert på fasiten. Hva de programmene som ikke er beskrevet handlet om, kan ikke finnes ut ved søk i samlingen. Man kan bare anta at det ville vært enda flere mulige relevante svar til søkespørsmålene hvis alle "Ukeslutt"-programmene hadde fått en innholdsbeskrivelse.

Fig. 8 viser videre hvor mange av programmene som har et kontrollert vokabular, dvs. emneord eller tagger, i innholdsbeskrivelsen. Det er ingen programmer som er beskrevet med bare det kontrollerte vokabularet, enten er det både innholdsbeskrivende fulltekst og kontrollerte termer, eller det er bare innholdsbeskrivende fulltekst. Dette er et signal på at det er den innholdsbeskrivende fullteksten som er prioritert i indekseringen.

Emneordene og taggene finnes i 40% av programmene. Prosenten av innslag som har emneord og tagger i beskrivelsen er i stedet på 30%. Forskjellen i dekningsgraden mellom programmene og innslagene tydet på at ikke alle innslagene er like prioritert i beskrivelsen innenfor samme program. I tillegg bekrefter disse tallene prioriteringen av den innholdsbeskrivende fullteksten i indekseringen i forhold til det kontrollerte vokabularet. Konsekvensen er at den innholdsbeskrivende fullteksten får større betydningen for gjenfinningen siden den dekker en større del av beskrivelsen for programmene.

Et annet viktig forhold mellom de to typene metadata er å se hvor mange av termene fra det kontrollerte vokabularet som allerede finnes i den innholdsbeskrivende fullteksten. Dette har konsekvenser for hvilken funksjon termene får for gjenfinningen. I tillegg har det vært en forandring i indekseringspraksisen fom 2012 hvor blant annet emneordene ble erstattet med tagger. Derfor er det her delt mellom emneord og tagger selv om begge er betraktet som en del av det

kontrollerte vokabularet. Tallene presenteres i fig. 9.

	Emneord (1977-2011)	**Tagger (2012, 2013)**	**Kontrollerte termer samlet**
Antall termer som ikke finnes i innholdsbeskrivende fulltekst	1817 (59%)	841 (62%)	2658 (60%)
Antall termer som finnes i innholdsbeskrivende fulltekst	1278 (41%)	522 (38?)	1800 (40%)
Totalt antall termer	3095 (100%)	1363 (100%)	4458 (100%)

Fig.9. Emneord og tagger: deres forhold til den innholdsbeskrivende fullteksten

Slik som forklart i metodedelen (§ 3.7.1), er emneordene noe tilpasset for å kunne sammenlikne dem med den innholdsbeskrivende fullteksten. Termene blir til enkeltstående termer og stoppordene er fjernet.

Sammenligningen av emneordene og taggene med fullteksten viser at 40% av det totale antallet emneord og tagger allerede finnes i fullteksten. Det er ikke betydelige forskjeller mellom taggene og emneordene når det gjelder repetisjon av termene fra den innholdsbeskrivende fullteksten. 38% av taggene finnes allerede i innholdsbeskrivende fullteksten for samme innslag, mens 41% av emneordene er en repetisjon av termer fra fullteksten.

Emneord eller tagger som er repetisjon av termer fra fullteksten utvider ikke treffmulighetene for de programmene det gjelder. Det kontrollerte vokabularet bidrar derfor i begrenset form til å oppnå enten noe bedre spesifisitet eller uttømmenhet ved å utvide vokabularet til den innholdsbeskrivende fullteksten. (Foskett, 1997, s. 23)

Emneordene og taggene har likevel en funksjon i en vektorbasert gjenfinningsmodell. Termfrekvensen øker når disse finnes flere ganger i teksten. Bruk av emneord og tagger med termer som allerede er i fullteksten får dermed konsekvenser for rangeringen i trefflistene.

I fig. 9 kan man også se på eventuelle forskjeller i registreringen av emneordene og taggene. Taggene dekker metadata for bare to år, 2012 og 2013, mens det resterende materialet fra 1977 til 2011 har emneord. Sammenligningen blir noe begrenset fordi dataene dekker svart forskjellige perioder. Til tross for at det er registrert tagger i bare 2 år, utgjør disse over en tredjedel av antall emneord som dekker over 30 år med metadata. Tendensen etter de to første årene med registrering av tagger viser at disse blir prioritert høyere enn man har gjort med registrering av emneord. Tallet

må likevel tas med forsiktighet. Det er store variasjoner i datasamlingen gjennom årene for om det ble brukt emneord eller ikke i beskrivelsen. Tallene er heller ikke noe garanti for at tendensen fortsetter.

Tallene for bruken av emneord og tagger i forhold til fullteksten viser at det kontrollerte vokabularet dekker en liten del av samlingen. (Fig. 8) Programmer som har emneord og tagger får fordel av det kontrollert vokabularet både når dette er en utvidelse av fullteksten og når det er en repetisjon av termer på grunn av bruk av termfrekvensen i gjenfinningsmodellen. Konsekvensen er at relativt få programmer får en fordel i gjenfinningen. Det er derfor heller ikke ønskelig med en vekting av det kontrollerte vokabularet ved indekseringen, eller søk som er rettet spesifikk på termer i det kontrollerte vokabularet. Da vektlegges de programmene som allerede har et fortrinn, mens andre programmer som kan være like relevante blir nedprioritert på grunn av mangelfull beskrivelse.

Disse tallene refererer til ett enkelt radioprogram fra SIFT-basen. Programmet "Ukeslutt" kan ikke regnes alene som representativt for alle de andre programmene fra basen. Det er til sammen 1666 registreringen av "Ukeslutt", som er en veldig liten andel av det totale antall registreringer i hele SIFT-basen på ca 750.000 poster. Disse tallene kan derfor ikke generaliseres til å gjelde for hele SIFT-basen, men tegner likevel et bilde av denne delen av samlingen.

5 Diskusjon

Oppgaven har hatt som formål å sammenlikne gjenfinningen mellom en samling hvor metadataene om ulike temaer er blandet i indekseringen, i forhold til samme samlingen hvor temaene er indeksert hver for seg. Problemstillingen var følgende:

Hvordan påvirkes gjenfinningen ved søk i metadata hvor ulike temaer er indeksert samlet sammenliknet med et søk hvor de er semantisk inndelt?

For å besvare på problemstillingen har jeg to forskningsspørsmål:

1. Hvordan påvirkes gjenfinningen i innholdsmetadata til NRKs radioprogram "Ukeslutt" med programmene som søkeenhet i forhold til innslagene?

2. Hvilken betydning har det kontrollerte vokabularet i forhold til den innholdsbeskrivende fullteksten for gjenfinningen av innslagene og programmene?

Jeg skal her svare på problemstillingen ved å koble resultatene fra analysen med de to konkrete forskningsspørsmålene, og med tidligere forskning.

5.1 Gjenfinningen i indeksen av programmer og i indeksen av innslag for radioprogrammet "Ukeslutt"

Analysen av søkeresultatene viser at disse er en kombinasjon av flere faktorer som må ses i sammenheng. Resultatene ved søk i indeksen av programmer og i indeksen av innslag kan ikke ses isolert fra de metodologiske valgene tatt for å utføre gjenfinningsforsøket. Valget av gjenfinningsmodell, av rangeringsmetode og valg knyttet til utformingen av søkespørsmål og fasit, er interessante for å forklare resultatene. Beregning av de ulike målene for presisjon og fullstendighet gjør det mulig å sammenlikne søkene i de to indeksene og oppsummere felles tendenser for gjenfinningen.

5.1.1 De metodologiske valgene

Gjenfinningsmodellen operer med en formel for beregning av score. Score brukes av systemet for rangeringen av trefflistene i de to indeksene. Lengdenormalisering, termfrekvens og invers termfrekvens er faktorer i formelen. Analysen viser hvordan disse faktorene påvirker gjenfinningen i de to indeksene. Enkelte av eksemplene fra søkeresultatene tar for seg hvordan ikke-relevante innslag blir funnet foran andre som er relevante. I andre tilfeller er funnene av relevante innslag interessante i forhold til de ulike valgene for modellen. Kombinasjonene av ulike faktorer gjør det vanskelig å generalisere fra den kvalitative analysen og trekke konklusjoner som skal gjelde for hele samlingen.

Analysen trekker også fram enkelte karakteristikker ved metadataene som må ses i sammenheng med den valgte gjenfinningsmodellen. Den innholdsbeskrivende fullteksten i samlingen består av en tekst som ikke har en fast lengde hverken når programmene eller innslagene er enheten for søket. I tillegg består datasamlingen av relativt få programmer med få termer i innholdsbeskrivelsen for hvert av innslagene.

Programmer med innholdsbeskrivelser av forskjellig lengde kombinert med lengdenormaliseringen gir forskjellige resultater i rangeringene i de to trefflistene. Dette får konsekvenser for gjenfinningseffektiviteten. Programmer som bare består av ett innslag, har f eks samme tekstlengde i indeks av programmer som i indeks av innslag. Ved lengdenormaliseringen blir disse sammenliknet med svært forskjellige tekster i de to indeksene. Disse programmene får høyere verdi for lengdenormaliseringen når de blir sammenliknet med andre og lengre programmer, og derfor høyere score ved søk i indeksen av programmene. Hvis programmene er relevante blir gjenfinningseffektiviteten bedre når det søkes i indeks av programmer. Er de ikke relevante er det indeks av innslag som får nytte av dette fordi programmene ikke blir rangert så langt oppe når de blir sammenliknet med kortere tekster.

Fulltekstbeskrivelsene i datamaterialet inneholder relativt få termer sammenliknet med andre fulltekstdatabaser. Når tekstene er så korte gir en enkel repetisjon av en term stort utslag på termfrekvensen, og innslaget blir rangert høyere opp i trefflistene. Søk i indeksen av programmer der flere innslag er relevante, kan få større uttelling på termfrekvensen, og derfor gi bedre resultater enn søk i indeksen av innslag.

Kombinasjonen av invers dokumentfrekvens og en relativt liten samling metadata, gjør at verdien for invers dokumentfrekvensen påvirkes av en del tilfeldigheter. For at systemet skal kunne identifisere termer med lav diskrimineringsverdi, er det nødvendig med et stort antall termer i

metadata. Når antall termer er begrenset, gir få repetisjoner utslag på termens diskrimineringsverdi, og det får konsekvenser for resultatene av søket.

Beregningen av invers dokumentfrekvens er avhengig av samlingens størrelse, og noen av resultatene knyttet til IDF, bør kunne løses ved bruk av samme gjenfinningsmodell for hele NRKs base. Med en mye større base, er antall termer høyere og beregningene for termenes diskrimineringsverdi kan justeres av systemet. IDF er en samlingsfaktor og påvirker søket i de to indeksene likt i gjenfinningsforsøket siden de inneholder de samme metatadaene. (Salvesen, 1994, s. 35-37) Termfrekvensen er i stedet knyttet til beskrivelsene av de enkelte programmene, og viser til forskjellige resultater ved søk i de to indeksene. Karakteristikkene ved termfrekvensen kan av samme grunn ikke forbedres med en utvidelse av samlingen fra "Ukeslutt" programmer til hele NRKs base, siden den er koblet til det enkelte programmets beskrivelse.

Rangeringsmetoden for indeksen av innslag er et annet av de metodologiske valgene som påvirker analysen. Den er bare knyttet til den ene indeksen, og er viktig for å forklare eventuelle forskjeller mellom resultatene av søk i de to indeksene.

Trefflistene viser at den valgte rangeringsmetoden for innslagene er problematisk når det er flere innslag fra samme program som er relevante. Disse kommer ikke høyt nok opp i rangeringslistene fordi de blir rangert etter det innslaget som har høyest score. Søk mot indeksen av programmer gir derfor bedre gjenfinningseffektivitet i disse tilfellene. Termfrevensen bidrar også til dette fordi i programmer med flere innslag med samme tema, blir termene fra søkespørsmålet repetert flere ganger i løpet av programmet. Ulempen for denne rangeringsmetoden er utpekt som et mulig problem også i tidligere forskning av *passage retrieval*. Denne metode har likevel gitt noe bedre resultater i forhold til andre metoder i gjenfinningsforsøket fra Jimmy Lin. (Lin, 2009, s. [8])

Differensieringen mellom lange og korte søkespørsmål fører ikke til tydelige variasjoner i gjenfinningseffektiviteten. De lange søkespørsmålene har flere termer, og flere av disse er veldig spesifikke. Dette kunne ha ført til bedre presisjon i trefflistene og bedre gjenfinningseffektivitet, men de søkene som er valgt for gjenfinningsforsøket gir ikke noe bedre samlet presisjon med lange søkespørsmål enn korte.

De korte søkespørsmålene gir noen problemer med rangeringen av relevante og ikke-relevante treff på grunn av kombinasjonen av få termer i søkespørsmålet og OR-forbindelsen mellom termene. De korte søkespørsmålene har få termer. Når noen av termene i tillegg er generelle og brukt i flere innslag, blir en del ikke-relevante treff rangert høyere enn de relevante. Dette skjer ved søk i begge indeksene.

5.1.2 Presisjon og fullstendighet ved indeks av innslag og av programmer

Kvantitative gjenfinningsforsøk utføres med langt større samlinger enn det jeg bruker, flere søk og flere relevante dokumenter i fasit. De baserer analysen utelukkende på de samlede gjennomsnittsverdiene for presisjon og fullstendighet. I denne oppgaven er presisjon og fullstendighet brukt for å sammenlikne søkene i de to indeksene, men de kan likevel ikke brukes til å beskrive hele NRKs metadatasamling.

En gjennomgang av verdiene for presisjon for de ulike søkene viser at forskjellene mellom indeksene er marginale. Tendensen er likevel at bruk av indeksen av innslag gir noe bedre resultater for presisjon i flere av søkene på toppen av trefflisten. Dette vises spesielt med verdiene for P@3 og P@5, mens resultatene snur seg for P@10 og P@20. *R-precision* og MAP bekrefter de små forskjellene, og at det samlet sett er noe bedre verdier for indeksen av innslag.

Dette samsvarer med funnene gjort av Kaskiel og Zobel, som får gjennomgående bedre verdier for gjennomsnittpresisjon ved søk i ulike typer *passages* enn for søk i hele dokumenter. Samtidig har noen av deres resultater ikke statistisk signifikans fordi forskjellene er for små. (Kaszkiel & Zobel, 2001, s. 352-353)

Den kvalitative analysen viser i tillegg at søk i indeksen av programmer kan bli problematisk når score for enkelte relevante innslag påvirkes av termer som beskriver andre innslag. Eksemplene vist i analysen referer til programmer som er relevante, men det samme fenomenet kan finnes også for ikke-relevante programmer. Disse kan få bedre posisjon i trefflisten på grunn av termer som beskriver forskjellige innslag og som er felles med søkespørsmålet. Problemet unngås med en indeks av innslag.

Verdiene for fullstendighet viser enda større likhet for søk i de to indeksen enn presisjonsverdiene, hvis man tar for seg de første 20 treffene for hvert søk. 10 av de 16 søkespørsmålene gir like mange relevante treff i begge indeksene. De kvantitative gjenfinningsforsøkene som er omtalt innledningsvis i oppgaven, har langt flere dokumenter i fasit og mye større samlinger, og får andre resultater for fullstendighet i trefflistene som ikke er sammenlignbare med mine. Dataene er ofte hentet fra TREC-eksperimentene hvor fasiten er skapt automatisk gjennom *pooling method.* (Chowdhury, 2010, s. 315-316)

Når det er så stor likhet i fullstendigheten mellom indeksene i min analyse, blir presisjonen på toppen av trefflisten viktigere for å evaluere gjenfinningseffektiviteten. Samtidig bekrefter fullstendigheten at forskjellene mellom indeksene er marginale.

5.2 Betydningen av kontrollert vokabular og innholdsbeskrivende fulltekst for gjenfinningen

Analysen av indekseringspraksisen viser ikke noen tydelige endringer gjennom den tidsperioden metadataene er registrert. Det er hverken en tydelig forbedring eller forverring av presisjon og fullstendighet innenfor de første 20 treffene gjennom årene, uansett om man ser på indeksen av programmer eller av innslag. Resultatene for *R-precision* viser heller ikke noen forskjeller i presisjonsverdiene over tid som kan tyde på en forandring av indekseringspraksis.

Det er likevel flere eksempler i analysen på mangelfull og inkonsekvent indeksering som bidrar til dårligere gjenfinning i datasamlingen. Det er forskjeller i den innholdsbeskrivende fullteksten fra program til program sendt innenfor en kort tidsperiode, og fra innslag til innslag som tilhører samme program. Inkonsekvensen kan ikke rettes opp av en inndeling av indeksen i innslag. Termene er de samme i de to indeksene.

Den innholdsbeskrivende fullteksten er prioritert i den manuelle indekseringen i forhold til det kontrollerte vokabularet. Det er bare 30% av de indekserte programmene som har emneord eller tagger ved siden av fullteksten.

Når emneord og tagger brukes i indekseringen er det ulike grader av uttømmenhet for lignende innslag. Ikke alle aspekter ved et emne blir dekket ved beskrivelsen av samme tema. Dette får konsekvenser for rangeringene av programmene i trefflistene i begge indeksene.

Det kontrollerte vokabularet har hovedsakelig to funksjoner i gjenfinningssammenheng. Emneord og tagger som er en repetisjon av termer fra den innholdsbeskrivende fullteksten, bidrar til høyere termfrekvens for gjenfinningen og derfor høyere rangering i trefflistene. Dette gjelder ca 40% av termene fra det kontrollerte vokabularet. Resten av termene som ikke finnes fra før i den innholdsbeskrivende fullteksten, bidrar i stedet med å utvide beskrivelsen og spesifisitet i beskrivelsen. Sannsynligheten for likhet med søkespørsmålene blir høyere i dette tilfellet.

Programmer som har et kontrollert vokabular i tillegg til den innholdsbeskrivende fullteksten, har derfor en fordel av vokabularet uansett om termene finnes fra før i den innholdsbeskrivende fullteksten eller ikke. Dette er en konsekvens av den vektorbaserte gjenfinningsmodellen som bruker termfrekvensen for beregning av rangeringen i trefflistene.

Flere eksempler viser at den innholdsbeskrivende fullteksten og det kontrollerte vokabularet har funksjoner som ligner de Lancaster peker på som fordeler og ulemper for fulltekster og kontrollerte vokabularer. (Lancaster, 2003, s. 270) Flere av innslagene beskrives med en lengre fulltekst som

øker spesifisiteten, mens det kontrollerte vokabularet gir konsistens i representasjonen ved å bruke de samme termene for den overordnede emnene i beskrivelsen og gjør at innslag om samme tema kan finnes gjennom de samme emneordene eller taggene. Problemet er at denne type indeksering ikke er gjort konsekvent, og det finnes flere unntak hvor enten den innholdsbeskrivende fullteksten ligner mer på et kontrollert vokabular med bare få overordnet termer i beskrivelsen, eller hvor det kontrollerte vokabularet ikke finnes i det hele tatt.

Sammenligningen av emneordene og taggene viser at det har vært en endring i bruken av disse. Ved overgangen til tagger de siste to årene, er det brukt et langt større vokabular enn tidligere.

Taggene er ikke registrert i direkte forbindelse med de innslagene de referer til, men samlet for ett helt program enten i starten eller på slutten av registreringen. Denne indekseringspraksisen, som bare unntaksvis er brukt for emneordene, kan gi problemer for emnesøk. Nærhetsoperatorer kunne vært en mulig løsning for søk av semantiske enheter uten at det skal være en inndeling etter innslag på forhånd. Muligheten blir borte når beskrivelsene av de ulike semantiske enhetene er blandet. Innholdsbeskrivende fulltekst og kontrollert vokabular som referer til samme innslag er delt, og det er vanskelig å rekonstruere hvilke innslag de ulike taggene refererer til.

6 Konklusjon

I denne oppgaven har jeg utført et gjenfinningsforsøk med metadata fra radioprogrammet "Ukeslutt", og analysert resultatene av forsøket. Jeg har sammenlignet gjenfinningen i en samling av metadata med blandet innhold, og samme samling med temaene indeksert hver for seg. Formålet var å undersøke om gjenfinningseffektiviteten kunne forbedres med en semantisk inndeling av metadataene.

Gjenfinningsevalueringen er basert på få søk og på en liten metadatasamling. Det er derfor ikke mulig å trekke konklusjoner som kan gjelde for hele NRKs metadatasamling, men den kvalitative analysen viser likevel noen generelle trekk.

Fra analysen av søkeresultatene kan jeg konkludere at det er mulig å finne fram til de enkelte innslagene eller temaene også i indeksen der innholdet er blandet og søkeenhetene er programmene. Forskjellene i målene for gjenfinningseffektivitet for de to indeksene er marginale og flere av de relevante programmene blir funnet i begge. Inndeligen i innslag bidrar til noe bedre presisjon, men bare helt øverst i trefflistene, og resultatene for fullstendighet er nærmest like. Søk i samlingen delt etter innslag forbedrer ikke gjenfinningen såpass mye at det er verdt å dele metadataene fra NRKs metadatasamling manuelt.

Den kvalitative analysen av søketreffene viser at enkelte av de metodologiske valgene har direkte konsekvenser for resultatene, og påvirker søket i indeksen av programmer og indeksen av innslag ulikt. Generelt bør valg av gjenfinningsmodell ta hensyn til karakteristikkene ved indekseringspraksisen for samlingen. Dette bekrefter at systemet og valgene knyttet til den automatiske indekseringen og gjenfinningen må ses i sammenheng.

De forskjellige søkespørsmålene gir ingen systematiske forskjeller i gjenfinningen i de to indeksene, men enkelte karakteristikker ved søkespørsmålene og hvordan disse blir håndtert av systemet påvirker likevel gjenfinningen generelt, uavhengig om det er reelle brukere eller ikke som velger dem. På denne måten bekrefter oppgaven betydningen av at gjenfinningsevalueringen tar hensyn til faktorer utenfor gjenfinningssystemet, slik som det holistiske perspektivet trekker fram.

Til slutt avdekker analysen en del mangelfull og inkonsekvent indeksering som gir dårligere gjenfinning. Det er spesielt det kontrollerte vokabularet som er nedprioritert i forhold til den innholdsbeskrivende fullteksten. Innslag som er beskrevet med et kontrollert vokabular i tillegg til den innholdsbeskrivende fullteksten, kobles med innslag fra andre programmer som har det samme overordnede temaet. I tillegg får disse innslagene en fordel ved at det brukes en vektorbasert

gjenfinningsmodell, spesielt ved beregningen av termfrekvensen. Den innholdsbeskrivende fullteksten øker spesifisiteten i indekseringen. Problemet er at dette er gjort lite konsekvent gjennom samlingen, og enkelte av innslagene blir derfor ikke funnet igjen.

Det ville absolutt vært interessant å kunne utvide gjenfinningsforsøket til en større metadasamling som er mer representativ for flere av NRKs metadata. På denne måten kunne det vært mulig å generalisere resultatene til å gjelde for hele samlingen, og målene for gjenfinningseffektivitet ville fått en større betydning i analysen. Til slutt ville eventuelle søkelogger fra NRK vært til hjelp i gjennfinningen slik at gjenfinningsforsøket kunne baseres på reelle brukerbehov og uttrykt med de termene som faktisk brukes i virkeligheten.

7 Litteraturliste

Baeza-Yates, R. & Ribeiro-Neto, B. (2011). *Modern information retrieval: the concepts and technology behind search*. Harlow: Addison Wesley.

Beall, J. (2008). The Weaknesses of Full-Text Searching. *The Journal of Academic Librarianship, 34*(5), 438-444. doi:http://dx.doi.org/10.1016/j.acalib.2008.06.007

Belkin, N. J. & Croft, W. B. (1987). Retrieval techniques. *Annual review of information science and technology, 22*, 109-145.

Belkin, N. J., Oddy, R. N. & Brooks, H. M. (1982). Ask for Information Retrieval : Part II. Results of a design study. *Journal of Documentation, 38*(3), 145-164. doi:doi:10.1108/eb026726

Bendersky, M. & Kurland, O. (2010). Utilizing passage-based language models for ad hoc document retrieval. *Information Retrieval, 13*(2), 157-187. doi:10.1007/s10791-009-9118-8

Büttcher, S., Clarke, C. L. A. & Cormack, G. V. (2010). *Information retrieval: implementing and evaluating search engines*. Cambridge, Mass.: MIT Press.

Callan, J. P. (1994). *Passage-level evidence in document retrieval*. Paper presentert på: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Dublin, Ireland

Chowdhury, G. G. (2010). *Introduction to modern information retrieval*. London: Facet.

Class DefaultSimilarity : org.apache.lucene.search.similarities. (cop. 2000-2014). Hentet 28.06.2014 fra http://lucene.apache.org/core/4_9_0/core/org/apache/lucene/search/similarities/DefaultSimilarity.html

Class NorwegianAnalyzer : org.apache.lucene.analysis.no. (cop. 2000-2014). Hentet 28.06.2014 fra http://lucene.apache.org/core/4_9_0/analyzers-common/org/apache/lucene/analysis/no/NorwegianAnalyzer.html

Class QueryParser : org.apache.lucene.queryparser.classic. (cop. 2000-2014). Hentet 20.08.2014 fra http://lucene.apache.org/core/4_9_0/queryparser/org/apache/lucene/queryparser/classic/QueryParser.html

Class Similarity : org.apache.lucene.search.similarities. (cop. 2000-2014). Hentet 28.06.2014 fra http://lucene.apache.org/core/4_9_0/core/org/apache/lucene/search/similarities/Similarity.html

Class StopwordAnalyzerBase : org.apache.lucene.analysis..util. (cop. 2000-2014). Hentet 19.06.2014 fra http://lucene.apache.org/core/4_9_0/analyzers-common/org/apache/lucene/analysis/util/StopwordAnalyzerBase.html

Craig, W. & Miles, E. (2013). *Finding information in books: characteristics of full-text searches in a collection of 10 million books*. Paper presentert på: Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries. Montreal, Quebec, Canada

Fernández, R., Losada, D. & Azzopardi, L. (2011). Extending the language modeling framework for sentence retrieval to include local context. *Information Retrieval, 14*(4), 355-389. doi:10.1007/s10791-010-9146-4

Fernández, R. T. & Losada, D. E. (2012). Effective sentence retrieval based on query-independent evidence. *Information Processing & Management, 48*(6), 1203-1229. doi:http://dx.doi.org/10.1016/j.ipm.2012.01.007

Foskett, A. C. (1997). *The subject approach to information* (5th ed. utg.). London: Library Association Publishing.

Fuhr, N. (2005). *Advances in XML information retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004 : revised selected papers*. Berlin: Springer.

Fuhr, N. (2008). *Focused access to XML documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007, Dagstuhl Castle, Germany, December 17-19, 2007 : revised selected papers* (Bind 4862). Berlin: Springer.

Hearst, M. A. & Plaunt, C. (1993). *Subtopic structuring for full-length document access*. Paper presentert på: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. Pittsburgh, Pennsylvania, USA

Hemminger, B. M., Saelim, B., Sullivan, P. F. & Vision, T. J. (2007). Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts. *Journal of the American Society for Information Science and Technology, 58*(14), 2341-2352. doi:10.1002/asi.20708

Hjortsæter, E. (2005). *Emneordskatalogisering: innholdsanalyse, emnerepresentasjon og lagring* (Bind nr 2). Oslo: Høgskolen i Oslo.

Information retrieval. Hentet 20.09.2014 fra http://en.wikipedia.org/wiki/Information_retrieval

Ingwersen, P. & Järvelin, K. (2005). *The turn: integration of information seeking and retrieval in context*. Dordrecht: Springer.

Johannessen, A. & Tufte, P. J. (2002). *Introduksjon til samfunnsvitenskapelig metode*. Oslo: Abstrakt.

Kaszkiel, M. & Zobel, J. (2001). Effective ranking with arbitrary passages. *Journal of the American Society for Information Science and Technology, 52*(4), 344-364. doi:10.1002/1532-2890(2000)9999:9999<::AID-ASI1075>3.0.CO;2-#

Kelly, D. & Ruthven, I. (2011). *Interactive information seeking, behaviour and retrieval*. London: Facet.

Krikon, E. & Kurland, O. (2011). A study of the integration of passage-, document-, and cluster-based information for re-ranking search results. *Information Retrieval, 14*(6), 593-616. doi:10.1007/s10791-011-9168-6

Lancaster, F. W. (2003). *Indexing and abstracting in theory and practice*. London: Facet Publishing.

Lesk, M. (2005). *Understanding digital libraries* (2nd ed. utg.). Amsterdam: Elsevier.

Lin, J. (2009). Is searching full text more effective than searching abstracts? *BMC Bioinformatics,*

10, 1-15. doi:10.1186/1471-2105-10-46

McCandless, M., Hatcher, E. & Gospodnetić, O. (2010). *Lucene in action*. Greenwich: Manning.

Mooers, C. N. (1951). Zatocoding applied to mechanical organization of knowledge. *American Documentation, 2*(1), 20-32. doi:10.1002/asi.5090020107

NISO. (2004). Understanding metadata. Hentet fra http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Porter, M. F. (1979). An algorithm for suffix stripping. I K. Sparck Jones, & P. Willett (Red.), *Readings in Information Retrieval* (s. 589). San Francisco: Morgan Kaufmann publ.

Ranganathan, S. R. (2006). *The five laws of library science*. Hentet fra http://arizona.openrepository.com/arizona/handle/10150/105454

Ringdal, K. (2007). *Enhet og mangfold: samfunnsvitenskapelig forskning og kvantitativ metode*. Bergen: Fagbokforl.

Ruthven, I. (2008). Interactive Information Retrieval. *Annual review of information science and technology, 42*, 43-91.

Salton, G. & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Salvesen, G. (1994). *Termvekting i automatisk gjenfinning*. [S.l.]: [G. Salvesen].

Saracevic, T. (1997). Extension and Application of the Stratified Model of Information Retrieval Interaction. *Proceedings of the Annual Meeting of the American Society for Information Science, 34*, 313-327.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology, 58*(13), 1915-1933. doi:10.1002/asi.20682

Sparck Jones, K. (2000). Further reflections on TREC. *Information Processing & Management,*

36(1), 37-85. doi:http://dx.doi.org/10.1016/S0306-4573(99)00044-8

Swanson, D. R. (1988). Historical note: Information Retrieval and the future of an illusion. I K. Sparck Jones, & P. Willett (Red.), *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann Publ.

Søbak, V. D. B. (2013). *Desentralisert indekseringspraksis: en studie av det semikontrollerte vokabularet i NRK*. Oslo: Høgskolen i Oslo og Akershus.

TREC tracks. (2002). Hentet 10.10.2014 fra http://trec.nist.gov/tracks.html

Wade, C. & Allan, J. (2005). Passage Retrieval and Evaluation. *CIIR Technical Report*, [8] s. Hentet fra http://ciir-publications.cs.umass.edu/pub/web/getpdf.php?id=541

Wallace, D. P. & Van Fleet, C. (2012). *Knowledge into action: research and evaluation in library and information science*. Santa Barbara, Calif.: Libraries Unlimited.

Wikipedia : 20. århundre. Hentet 09.09.2014 fra http://no.wikipedia.org/wiki/20._%C3%A5rhundre

8 Vedlegg

Vedlegg 1 – Søkespørsmålene

Vedlegg 2 – Program "LagIndex"

Vedlegg 3 – Program "MittSøk"

Vedlegg 4 – Søkeresultater, fasit og verdier av presisjon og fullstendighet

Vedlegg 1 – Søkespørsmålene

Spørsmåls-nummer	Tekst	År
1	Astronomer oppdager ringene rundt Uranus.	1977
2	Camp David-avtalen ble underskrevet av Israel og Egypt.	1978
3	Margaret Thatcher blir Storbritannias statsminister.	1979
4	Vietnamesiske styrker inntar Kambodsjas hovedstad Phnom Penh og overtar makten etter Røde Khmer	1979
5	Boligplattformen Alexander L. Kielland kantrer. 123 mennesker mister live	1980
6	Iran–Irak-krigen starter ved at Irak angriper Iran.	1980
7	Polske soldater skyter mot demonstranter under uroligheter i Warszawa og andre polske byer.	1981
8	Israels invasjon av Libanon starter med flyangrep	1982
9	Falklandskrigen bryter ut, når en argentinsk styrke på 8 000 mann invaderer de britiske Falklandsøyene uten forvarsel.	1983
10	Musikalen Annie blir fremført for siste gang etter 2 377 forestillinger på Broadway.	1984
11	Sør-Afrika får sin første regjering med ikke-hvite medlemmer.	1985
12	Arne Treholt dømmes til 20 års fengsel for spionasje.	1986
13	Tsjernobyl-ulykke	1987
14	Odelstinget behandler lov om Sametinget og andre samiske rettsforhold.	1988
15	I Sovjetunionen velges Mikhail Gorbatsjov til president.	1989
16	Den sovjetiske tilbaketrekningen fra Afghanistan fullføres	1990
17	Nelson Mandela, politisk fange i 27 år, blir løslatt fra Victor Verster-fengselet i Cape Town, Sør-Afrika.	1991
18	Gulfkrigen: Den amerikanske kongressen gir grønt lys for bruk av militære styrker for å drive Irak ut av Kuwait.	1992
19	TV2 har sin første tv-sending.	1993
20	Rosemarie Køhn blir Norges første kvinnelige biskop	1993
21	Forleggeren William Nygaard ble livstruende skadet etter å ha blitt skutt i et attentat begått av islamistiske fundamentalister utenfor sin bolig på Slemdal i Oslo i forbindelse med fatwaen mot Salman Rushdie	1993
22	Norge blir klare for VM i fotball 1994 etter å ha slått Polen 3-0 på bortebane. Jostein Flo, Jan Åge Fjørtoft og Ronny Johnsen skåret målene	1993
23	648 kosovoalbanere som har gått i kirkeasyl i Norge blir lovet ny behandling av asylsøknadene dersom de forlater kirkene	1994
24	Vinter-OL 1994 arrangeres på Lillehammer	1994
25	Parlamentsvalget i Italia ender med seier for Silvio Berlusconi og høyerealliansen Polo delle Libertà - Polo del Buon Governo	1994
26	MF "Estonia" sank utenfor Utö. 852 omkom i forliset	1994
27	Norge stemmer nei til EU-medlemskap	1995
28	Et gassangrep tar livet av 12 personer i Tokyo. Ytterligere 1 300 blir skadet	1995
29	Terrorangrep i Oklahoma City med 168 døde og enorme ødeleggelser	1995
30	Srebrenica-massakren finner sted	1995

31	Den israelske statsministeren Yitzhak Rabin blir snikmyrdet av en ekstrem israeler	1996
32	Amerikanske fredsbevarende styrker rykker inn i Bosnia-Hercegovina som et resultat av Dayton-avtalen	1996
33	Frankrikes president Jacques Chirac annonserer en "definitiv slutt" på landets atomprøvesprengninger	1996
34	Tyskland blir Europamestere i fotball, de vinner finalen på Wembley med 2–1 e.e.o mot Tsjekkoslovakia	1996
35	Thorbjørn Jaglands regjering overtok etter Gro Harlem Brundtlands tredje regjering	1997
36	Russland trekker sine styrker ut av Tsjetsjenia	1997
37	104 personer omkommer da et SilkAir-fly styrter i Indonesia	1997
38	Bjørn Dæhlie, Norge, vinner verdenscupen i langrenn for menn	1997
39	Ottawa-avtalen mot bruk av landminer undertegnes av 122 stater	1998
40	Flere tornadoer i det sentrale Florida tar livet av 42 mennesker og gjør mye skade	1998
41	Et jordskjelv som måler 6,9 på Richters skala rammer det sørøstlige Iran	1998
42	Lewinsky-skandalen: En føderal dommer bestemte at Secret Service-agenter kan måtte vitne i saken	1998
43	De amerikanske ambassadene i Kenya og Tanzania blir ødelagt i to koordinerte terrorangrep. Al-Qaida får skylden for angrepet, som tar livet av 200 mennesker og sårer mer enn 4000	1999
44	NATO går til krig mot Jugoslavia	1999
45	Krigen i Kosovo: Kosovos grense blir stengt av serbiske styrker for å forhindre at albanere skal få forlate landet	1999
46	En 7,4-jordskjelvet slår Izmit, Tyrkia, drepte mer enn 17 000 og skadet 44 000	1999
47	Boris Jeltsin går av som russisk president, Vladimir Putin blir utnevnt til ny russisk president	2000
48	Åsta-ulykken, to motgående tog kolliderte på Rørosbanen, 19 personer omkom	2000
49	Kjell Magne Bondeviks første regjering søker avskjed etter å ha stilt kabinettspørsmål i "gasskraftsaken" og fått flertall mot seg i Stortinget. Søknaden innvilges 17. mars og Jens Stoltenbergs første regjering blir utnevnt til ny regjering	2000
50	Den russiske atomubåten "K-141 Kursk" synker etter en eksplosjon ombord i Barentshavet. Alle 118 ombord omkom. Først fire dager senere fikk Norge beskjed om ulykken	2000
51	De 27. olympiske sommerleker blir arrangert i Sydney i Australia	2001
52	Holmlia-drapet: 15 år gamle Benjamin Hermansen blir drept	2001
53	Schengen-traktaten iverksettes i Norge	2001
54	USA og andre allierte styrker angriper Afghanistan og Taliban	2001
55	Linate-ulykken, et SAS McDonnell Douglas MD-87 kolliderer med et lite forretningsfly under avgang med tett tåke i Milano	2002
56	Euroen blir introdusert	2002
57	Nederland blir det første landet i verden som legaliserer dødshjelp	2002
58	Rundt 50 væpnede tsjetsjenske terrorister stormer Dubrovka-teateret i Moskva og tar omkring 800 mennesker som gisler	2002
59	EU annonserer at ti nye medlemmer (Estland, Latvia, Litauen, Polen, Tsjekkia, Slovakia, Ungarn, Slovenia, Malta og Kypros) skal inn i unionen fra 1. mai 2004	2003
60	Mennesker i 600 byer over hele verden (anslagsvis 10–15 millioner) protesterer imot en amerikanskledet krig imot Irak	2003
61	Irak-krigen: USAs president George W. Bush gir i en TV-tale Iraks Saddam Hussein 48 timer på	2003

	å overgi seg	
62	Verdens helseorganisasjon erklærer at SARS-epidemien er under kontroll	2003
63	VM i skiskyting	2004
64	Frankrike vedtar forbud mot bruk av religiøse klesplagg (eks. hodetørkle for islamske jenter) eller symboler (eks. krusifiks for kristne) i landets skoler	2004
65	Omkring 1000 mennesker omkommer i et jordskjelv i Marokko	2004
66	En rapport fra FN slår fast at Irak ikke har hatt noen masseødeleggelsesvåpen "av betydning" siden 1994, i motsetning til hva mange har trodd	2004
67	198 mennesker omkommer og ca. 1400 blir skadet i en serie terrorangrep i Spanias hovedstad Madrid. Flere bomber blir detonert med få minutters mellomrom på tog og togstajoner	2004
68	Amerikansk fangemishandling i Abu Ghraib-fengslet avsløres 28. april	2004
69	I Beslan i Ossetia, Russland tar tsjetjenske terrorister omkring 1 300 mennesker, de fleste barn, som gisler i en barneskole. De truer med å drepe gislene om deres krav ikke oppfylles	2004
70	Oransjerevolusjonen: I Ukraina utropes Viktor Janukovytsj som vinner av landets presidentvalg. Det framkommer sterke påstander om valgfusk	2004
71	VM på skøyter, allround arrangeres på Hamar	2004
72	USA rammes av storm, 16 omkommer	2005
73	Direktoratet for naturforvaltning gir tillatelse til lisensjakt på 5 ulver i Hedmark, til tross for stor motstand i befolkningen og massiv internasjonal kritikk	2005
74	Kyoto-avtalen trer i kraft etter å ha blitt ratifisert av 127 land	2005
75	Israelsk politi og militære begynner å fjerne jødiske bosettere fra Gaza-stripen med makt	2005
76	I Belfast, Nord-Irland blir 50 politimenn og mange sivile skadet i protestantiske opptøyer	2005
77	Ekstremværet Kristin: Rekordstor nedbør forårsaker flom, ras og ødeleggelser i Sogn og Fjordane, Hordaland og Rogaland. To omkommer og flere blir husløse etter jordras i Bergen	2005
78	Fugleinfluensa–viruset H5N1 påvises i Romania	2005
79	Den russiske tråleren "Elektron", som var tatt i arrest av Kystvakten for grove brudd på fiskeribestemmelsene, rømmer til russisk farvann mens to norske fiskeriinspektører holdes tilbake om bord	2005
80	Voldsomme ungdomsopptøyer i den innvandrer-dominerte bydelen Clichy-sous-Bois i Paris, Frankrike. Opptøyene fortsetter i flere uker og sprer seg etterhvert til hele landet	2005
81	En ny Al-Qaida-video på Internett oppfordrer muslimer til å angripe Norge, Danmark og Frankrike for å ha trykt karikaturer av Muhammed	2006
82	Ski-VM 2011 tildeles Oslo under kongressen til det internasjonale skiforbundet (FIS) i Faro i Portugal	2006
83	22 ryttere som skulle delta i Tour de France 2006 ble kastet ut etter en stor dopingskandale der det har kommet frem en liste med 58 navn over syklister som skal være involvert i organisert doping, hvorav 22 av navnene er syklister som skal delta i Tour de France. Forhåndsfavorittene Jan Ullrich og Ivan Basso er blant dem som ikke får delta. Aleksandr Vinokurovs lag Astaná-Würth får har så mange involvert at hele laget trekker seg	2006
84	FNs sikkerhetsråd vedtok enstemmig en resolusjonen hvor de fordømmer Nord-Koreas prøvesprengning av atomvåpen tidligere i uken	2006
85	Ingunn Yssen publiserer sitt oppsigelsesbrev i Verdens Gang. Dette ble starten på Kontroversen rundt Gerd-Liv Valla	2007
86	Statsminister Jens Stoltenberg sier at Norge skal redusere sine CO2-utslipp med 30 prosent innen 2020	2007
87	Kommune- og fylkestingsvalg i Norge	2007

88	11. september var det seks år siden terrorangrepet på World Trade Center	2007
89	Valg i Russland: Vladimir Putin sitt parti "Det forente Russsland" fikk over 60 % av stemmene i valget på ny nasjonalforsamling. Opposisjonen hevder at det offisielle valgresultatet er bygget på omfattende valgfusk	2007
90	Terrorangrepet mot Serena Hotell: Syv mennesker ble drept, inklusiv Dagbladets journalist Carsten Thomassen, og flere skadd i angrepet, deriblant en UD-ansatt	2008
91	Krisen på Gaza: Palestinske militanter sprenge åpninger i grensemuren mot Egypt, og titusenvis av palestinere krysser over i jakt på mat og andre forsyninger	2008
92	Fidel Castro kunngjorde at han går av på permanent basis som Cubas president	2008
93	Offisiell åpning av Svalbard globale frøhvelv	2008
94	Nye overgrepsbilder fra det amerikanske Abu Ghraib-fengslet i Irak offentliggjøres av Philip Zimbardo, en av verdens mest anerkjente og berømte psykologer	2008
95	Oljeprisen når rekordhøyden 108,21 dollar fatet på den amerikanske råvarebørsen New York Mercantile Exchange (Nymex)	2008
96	Samtlige partier på Stortinget ble enige om et forlik på grunnlovsendring som betyr at kirken selv (fra 2012) vil overta ansvaret med å utnevne biskoper, samt at Grunnlovens paragraf 2, 4 og 16 om kirken oppheves og erstattes med nye paragrafer	2008
97	Det norske Veritas legger frem sin rapport om Tromsøs OL-søknad. Rapporten viser at OL vil bli 9,5 milliarder dyrere enn Tromsø 2018 hadde beregnet	2008
98	Presidentvalg i USA	2008
99	Opptøyer i Oslo, som følge av konflikten mellom Israel og Gaza. Pro-israelske demonstranter blir angrepet av anti-israelske demonstranter, og opptøyer bryter ut. Politiet svarer med tåregass	2009
100	Joshua French og Tjostolv Moland blir dømt til døden i Kongo	2009
101	Høstutstillingen blir avholdt i Oslo	2009
102	Jens Stoltenberg og den rødgrønne regjeringen fortsetter etter seieren i stortingsvalget	2009
103	Folkerepublikken Kina feirer sitt 60-årsjubiluem. Republikken ble opprettet 1949 da kommunistpartiet gikk seirende ut av den kinesiske borgerkrig	2009
104	USAs president Barack Obama tildeles Nobels Fredspris for 2009	2009
105	Kraftig uvær og flom i Storbritannia og Irland	2009
106	Fns klimakonferanse 2009 arrangeres i København	2009
107	Synlig oppstilling av tobakksvarer blir forbudt i Norge	2010
108	Minst 75 mennesker omkommer som følge av jordskred og oversvømmelser i den brasilianske delstaten Rio de Janeiro. Oversvømmelsene fører også til at landets eneste atomkraftverk må stenges	2010
109	De nominerte til Den 82. Oscar-utdelingen blir annonsert	2010
110	Streik i norske kommuner. Streiken blir en av de mest omfattende på tiår, og nesten 50 000 er i streik da partene blir enige	2010
111	Verdens helseorganisasjon erklærer at svineinfluensaen ikke lenger er en pandemi	2010
112	President Obama erklærer at krigen i Irak er over	2010
113	Norges håndball jenter vinner EM	2010
114	Minst 15 839 mennesker omkom under jordskjelvet og den påfølgende tsunamien ved T?hoku i Japan	2011
115	Bryllupet mellom prins William av Wales, hertug av Cambridge og Catherine, finner sted i Westminster Abbey i London	2011
116	Osama bin Laden, grunnleggeren av terroristnettverket al-Qaida, er drept av amerikanske	2011

	spesialstyrker i Pakistan	
117	Mai–juni: E.coli-utbrudd med 50 døde, hovedsakelig i Tyskland	2011
118	Gjenopptakelseskommisjonen bestemmer at Treholt-saken ikke skal gjenopptas	2011
119	Hurtigruten: Minutt for minutt sendes på NRK2	2011
120	Terrorangrepene i Norge 2011 i Regjeringskvartalet og på Utøya. 77 mennesker omkommer, mange blir skadet, og det ble store materielle ødeleggelser	2011
121	Borgerkrigen i Libya: Under slaget om Tripoli inntar opprørerne det meste av Tripoli og Gaddafis bunker	2011
122	Steve Jobs trekker seg som toppsjef for Apple	2011
123	Rettssaken mot Anders Behring Breivik etter terrorangrepene i Norge 2011 startet	2012
124	EM i fotball 2012 arrangeres i Polen og Ukraina	2012
125	Sommer-OL 2012 arrangeres i London, England	2012
126	Stavanger nye konserthus ble åpnet	2012
127	I et intervju med Oprah Winfrey innrømte Lance Armstrong at han dopet seg systematisk gjennom sin karriere som proffsyklist	2013
128	Findus kunngjorde at store partier frossen ferdigmat deklarert med innhold av storfekjøtt i virkeligheten inneholdt hestekjøtt	2013
129	Flommen på Østlandet 2013: Deler av tettstedet Kvam i Gudbrandsdalen ble ødelagt, og veier og jernbaner ble stengt	2013
130	Magnus Carlsen ble verdensmester i sjakk	2013

Vedlegg 2 – Program "LagIndex"

```
import java.io.File;
import java.io.IOException;
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.ResultSet;
import java.sql.SQLException;
import java.sql.Statement;
import org.apache.lucene.analysis.Analyzer;
import org.apache.lucene.analysis.Analyzer.TokenStreamComponents;
import org.apache.lucene.document.DateTools;
import org.apache.lucene.document.DateTools.Resolution;
import org.apache.lucene.document.Document;
import org.apache.lucene.document.Field;
import org.apache.lucene.document.TextField;
import org.apache.lucene.index.IndexWriter;
import org.apache.lucene.index.IndexWriterConfig;
import org.apache.lucene.index.IndexWriterConfig.OpenMode;
import org.apache.lucene.index.Term;
import org.apache.lucene.store.FSDirectory;
import org.apache.lucene.util.Version;
import org.apache.lucene.analysis.no.NorwegianAnalyzer;

public class LagIndex {

       public void LagIndexMetode() throws SQLException, IOException {

           Connection connection = getDBConnection();
           Analyzer analyzer = new NorwegianAnalyzer(Version.LUCENE_3_6);

              File INDEX_DIR = new File
       ("C:/Users/Elisa/Documents/Masteroppgave/IndexFraBearbeidetDataGammelVersion");

         IndexWriterConfig config = new IndexWriterConfig(Version.LUCENE_3_6, analyzer);
          config.setOpenMode(OpenMode.CREATE);
          IndexWriter writer = new IndexWriter(FSDirectory.open(INDEX_DIR), config);

          String query = "SELECT keys, sendedato, tittel, innhold, klassifikasjon,
             tagger FROM bearbeidet_data";
          Statement statement = connection.createStatement();
          ResultSet result = statement.executeQuery(query);

          while (result.next()) {
           String keys = result.getString("keys");
           String innhold = result.getString("innhold");
           String sendedato = result.getString("sendedato");
           String tittel = result.getString("tittel");
           String klassifikasjon = result.getString("klassifikasjon");
           String tagger = result.getString("tagger");
           String samletinnhold = innhold + " " + klassifikasjon + " " + tagger;
           System.out.println(keys + sendedato + tittel + innhold + klassifikasjon +
             tagger);
         Document document = new Document();
         document.add(new TextField("keys", result.getString("keys"), Field.Store.YES));
         document.add(new TextField("sendedato", result.getString("sendedato"),
             Field.Store.YES));
```

```
                document.add(new Field("tittel", result.getString("tittel"),
Field.Store.NO, Field.Index.ANALYZED));
                document.add(new Field("innhold", result.getString("innhold"),
Field.Store.NO, Field.Index.ANALYZED));
                document.add(new Field("klassifikasjon",
result.getString("klassifikasjon"), Field.Store.NO, Field.Index.ANALYZED));
                document.add(new Field("tagger", result.getString("tagger"),
Field.Store.NO, Field.Index.ANALYZED));
                document.add(new Field("samletinnhold", samletinnhold, Field.Store.NO,
Field.Index.ANALYZED));
                writer.addDocument(document);
                writer.commit();
              }
           writer.close();
        }


        private Connection getDBConnection() {
              String DB_DRIVER = "oracle.jdbc.driver.OracleDriver";
              String DB_CONNECTION = "jdbc:oracle:thin:@localhost:1521:XE";
              String DB_USER = "Elisa";
              String DB_PASSWORD = "Elisa";
              Connection dbConnection = null;

              try {
                    Class.forName(DB_DRIVER);
              } catch (ClassNotFoundException e) {
                    System.out.println(e.getMessage());
              }

              try {
                    dbConnection = DriverManager.getConnection(
                    DB_CONNECTION, DB_USER,DB_PASSWORD);
                    return dbConnection;
              } catch (SQLException e) {
                    System.out.println(e.getMessage());
              }

              return dbConnection;
}
}
```

Vedlegg 3 – Program "MittSøk"

```
import java.io.File;
import java.io.IOException;
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.PreparedStatement;
import java.sql.ResultSet;
import java.sql.SQLException;
import java.sql.Statement;
import java.text.ParseException;
import java.util.ArrayList;

import org.apache.lucene.analysis.Analyzer;
import org.apache.lucene.analysis.no.NorwegianAnalyzer;
import org.apache.lucene.analysis.standard.StandardAnalyzer;
import org.apache.lucene.queryparser.classic.MultiFieldQueryParser;
import org.apache.lucene.queryparser.classic.QueryParser;
import org.apache.lucene.search.*;
import org.apache.lucene.*;
import org.apache.lucene.document.Document;
import org.apache.lucene.document.Field;
import org.apache.lucene.document.TextField;
import org.apache.lucene.index.CorruptIndexException;
import org.apache.lucene.index.DirectoryReader;
import org.apache.lucene.index.IndexReader;
import org.apache.lucene.search.Explanation;
import org.apache.lucene.search.IndexSearcher;
import org.apache.lucene.search.Query;
import org.apache.lucene.search.ScoreDoc;
import org.apache.lucene.search.TopDocs;
import org.apache.lucene.search.similarities.DefaultSimilarity;
import org.apache.lucene.search.similarities.Similarity;
import org.apache.lucene.store.Directory;
import org.apache.lucene.store.FSDirectory;
import org.apache.lucene.util.QueryBuilder;
import org.apache.lucene.util.Version;


public class MittSok {

	static String sql = "SELECT nummer,tekst FROM spoersmaal WHERE NUMMER in (select
		spoersmaal from fasit)";
	static Connection conn;

	public static void main(String args[])
		throws Exception
	{
	 File IndexFraProgram = new File
	("C:/Users/Elisa/Documents/Masteroppgave/IndexFraProgram");
	 File IndexFraBearbeidetData = new File
	("C:/Users/Elisa/Documents/Masteroppgave/IndexFraBearbeidetData");

	  MittSok ms = new MittSok();
	  conn=getDBConnection();

	  String deleteSQL="delete from RESULTAT";
```

```
Statement stmt=conn.createStatement();
stmt.executeUpdate(deleteSQL);
stmt.close();

  Statement statement = conn.createStatement();
     ResultSet result = statement.executeQuery(sql);

     while (result.next()) {
        int nummer = result.getInt("NUMMER");
        String tekst = result.getString("TEKST");
        ms.testQueryParser2(tekst, IndexFraProgram, nummer, "innhold");
        ms.testQueryParser2(tekst, IndexFraBearbeidetData, nummer,
              "samletinnhold");
        }
}

public void testQueryParser2(String q, File r, int n, String feltnavn)
        throws CorruptIndexException, IOException, ParseException,
org.apache.lucene.queryparser.classic.ParseException, SQLException    {

  Directory dir = FSDirectory.open(r);
  String tabellnavn = r.getName();
  IndexReader reader = DirectoryReader.open(dir);
  IndexSearcher searcher = new IndexSearcher(reader);

  Analyzer analyzer = new NorwegianAnalyzer(Version.LUCENE_4_9);

  QueryParser parser = new QueryParser(Version.LUCENE_4_9,feltnavn,analyzer);
  Query query = parser.parse(q);

  System.out.println("----------");
  TopDocs docs = searcher.search(query, 100);

  int treffnummer = 0;

  for (ScoreDoc match : docs.scoreDocs) {
       treffnummer ++;
       System.out.println("----------");
       Document doc = searcher.doc(match.doc);
       String keys=doc.get("keys");

      System.out.println(match.doc);
      Float score =match.score;

 String pinsertTablesql = "INSERT INTO RESULTAT (INDEKS, SPOERSMAAL, KEYS, SCORE,
      TREFFNUMMER) VALUES (?,?,?,?,?)";
 PreparedStatement pstmt=conn.prepareStatement(pinsertTablesql);

              pstmt.setString(1, tabellnavn);
                    pstmt.setInt(2, n);
                    pstmt.setString(3, keys);
                    pstmt.setFloat(4, score);
                    pstmt.setInt(5, treffnummer);
                    pstmt.executeQuery();
              pstmt.close();
          }
    }
    private static Connection getDBConnection() {
```

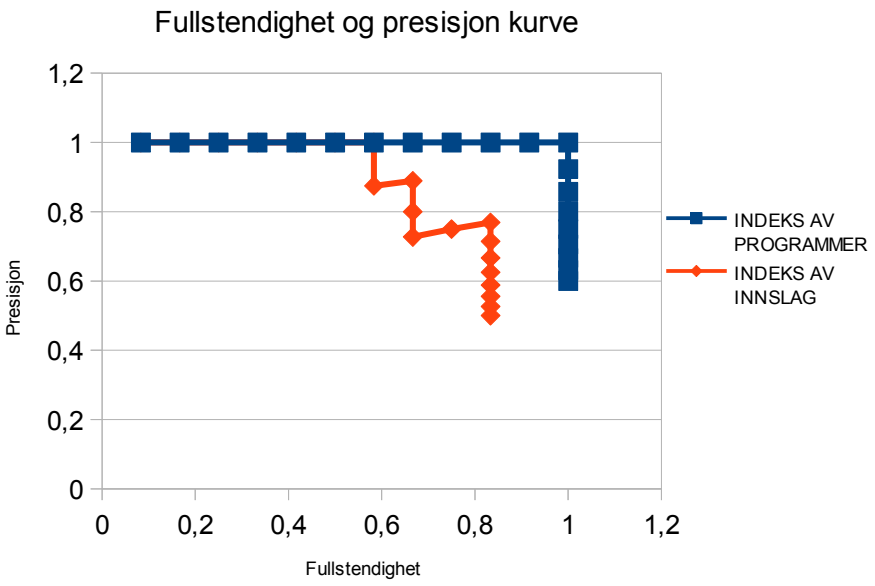
```
		String DB_DRIVER = "oracle.jdbc.driver.OracleDriver";
		String DB_CONNECTION = "jdbc:oracle:thin:@localhost:1521:XE";
		String DB_USER = "ELISA";
		String DB_PASSWORD = "ELISA";
		Connection dbConnection = null;

		try {
			Class.forName(DB_DRIVER);
		} catch (ClassNotFoundException e) {
			System.out.println(e.getMessage());
		}

		try {
			dbConnection = DriverManager.getConnection(
			DB_CONNECTION, DB_USER,DB_PASSWORD);
			return dbConnection;
		} catch (SQLException e) {
			System.out.println(e.getMessage());
		}

		return dbConnection;

	}
}
```

Vedlegg 4 – Søkeresultater, fasit og verdier av presisjon og fullstendighet

Søkespørsmål nr 12: "Arne Treholt dømmes til 20 års fengsel for spionasje"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,0833333333	1	0,35293654	"1990/01945"
2	0,1666666667	1	0,3019698	"1995/15384"
3	0,25	1	0,28743753	"1990/01951"
4	0,3333333333	1	0,27290896	"1990/05390"
5	0,4166666667	1	0,2670674	"1990/03098"
6	0,5	1	0,24820068	"1990/02507"
7	0,5833333333	1	0,23418726	"1990/01946"
8	0,6666666667	1	0,23067723	"1990/02370"
9	0,75	1	0,17782886	"1990/01949"
10	0,8333333333	1	0,14313594	"1990/02595"
11	0,9166666667	1	0,13855423	"1990/04168"
12	1	1	0,10942845	"1990/03827"
13	1	0,9230769231	0,10749766	"1998/03003"
14	1	0,8571428571	0,09857012	"1993/01663"
15	1	0,8	0,09769074	"2006/05271"
16	1	0,75	0,097317986	"1995/02110"
17	1	0,7058823529	0,08793077	"2002/00399"
18	1	0,6666666667	0,074300334	"1997/03424"
19	1	0,6315789474	0,07402482	"2011/08240"
20	1	0,6	0,07392759	"1990/05537"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,0833333333	1	0,4564197	"1990/01945"
2	0,1666666667	1	0,37986615	"1995/15384"
3	0,25	1	0,37445888	"1990/01951"
4	0,3333333333	1	0,34789562	"1990/05390"
5	0,4166666667	1	0,32099354	"1990/02507"
6	0,5	1	0,30082947	"1990/01946"
7	0,5833333333	1	0,2979362	"1990/02370"
8	0,5833333333	0,875	0,27236664	"1993/00019"
9	0,6666666667	0,8888888889	0,22917618	"1990/01949"
10	0,6666666667	0,8	0,21623364	"2011/08240"
11	0,6666666667	0,7272727273	0,18971619	"1998/00123"
12	0,75	0,75	0,18181153	"1990/02595"
13	0,8333333333	0,7692307692	0,1806958	"1990/04168"
14	0,8333333333	0,7142857143	0,18014216	"2010/00717"
15	0,8333333333	0,6666666667	0,1756416	"1997/03366"
16	0,8333333333	0,625	0,16325757	"2013/05565"
17	0,8333333333	0,5882352941	0,15889107	"2009/03666"
18	0,8333333333	0,5555555556	0,15563808	"2004/12339"
19	0,8333333333	0,5263157895	0,15526551	"1995/02110"
20	0,8333333333	0,5	0,15526551	"1996/04395"

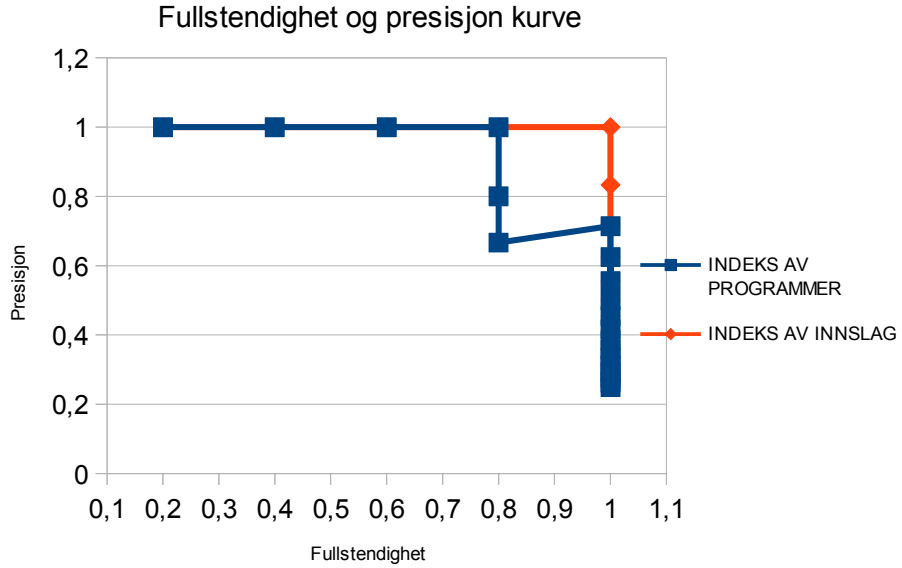


Fasit representert ved nøkkel
"1990/02507"
"1990/03827"
"1990/02595"
"1990/05390"
"1990/01946"
"1990/01951"
"1990/02370"
"1990/03098"
"1990/01949"
"1990/01945"
"1990/04168"
"1995/15384"

Søkespørsmål nr 13: "Tsjernobyl-ulykke"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,2	1	1,1377596	"1990/03273"
2	0,4	1	0,48293537	"1990/06210"
3	0,6	1	0,4394932	"2011/03020"
4	0,8	1	0,38071686	"1990/03033"
5	0,8	0,8	0,24368604	"1990/01069"
6	0,8	0,6666666667	0,17406145	"1990/02869"
7	1	0,7142857143	0,17074344	"1994/11025"
8	1	0,5555555556	0,16079108	"1993/09660"
9	1	0,625	0,16079108	"1994/30563"
10	1	0,5	0,1406922	"1995/16312"
11	1	0,4545454545	0,12184302	"1994/20571"
12	1	0,4166666667	0,12059331	"1990/00651"
13	1	0,3846153846	0,12059331	"1994/21599"
14	1	0,3333333333	0,11369646	"1994/30571"
15	1	0,3571428571	0,11369646	"2005/02387"
16	1	0,3125	0,104436874	"2005/14420"
17	1	0,2941176471	0,10049443	"1990/03877"
18	1	0,2631578947	0,09948441	"1994/05685"
19	1	0,2777777778	0,09948441	"2011/07217"
20	1	0,25	0,08527235	"1996/12929"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,2	1	1,6323869	"1990/03033"
2	0,4	1	1,5287168	"2011/03020"
3	0,6	1	1,4002635	"1990/03273"
4	0,8	1	0,9936964	"1994/11025"
5	1	1	0,56211954	"1990/06210"
6	1	0,8333333333	0,45397988	"2005/02387"
7	1	0,7142857143	0,42801633	"2004/16682"
8	1	0,625	0,42801633	"1996/10230"
9	1	0,5555555556	0,37451428	"1994/30563"
10	1	0,5	0,32433888	"1990/01069"
11	1	0,4545454545	0,32101226	"2010/03612"
12	1	0,3846153846	0,32101226	"1995/00528"
13	1	0,4166666667	0,32101226	"2004/10592"
14	1	0,3571428571	0,32101226	"2010/08450"
15	1	0,2941176471	0,30265325	"2011/07217"
16	1	0,3333333333	0,30265325	"1994/30571"
17	1	0,3125	0,30265325	"1998/11865"
18	1	0,2631578947	0,2675102	"2009/10603"
19	1	0,2777777778	0,2675102	"2010/09858"
21	1	0,25	0,2648216	"1994_05685"

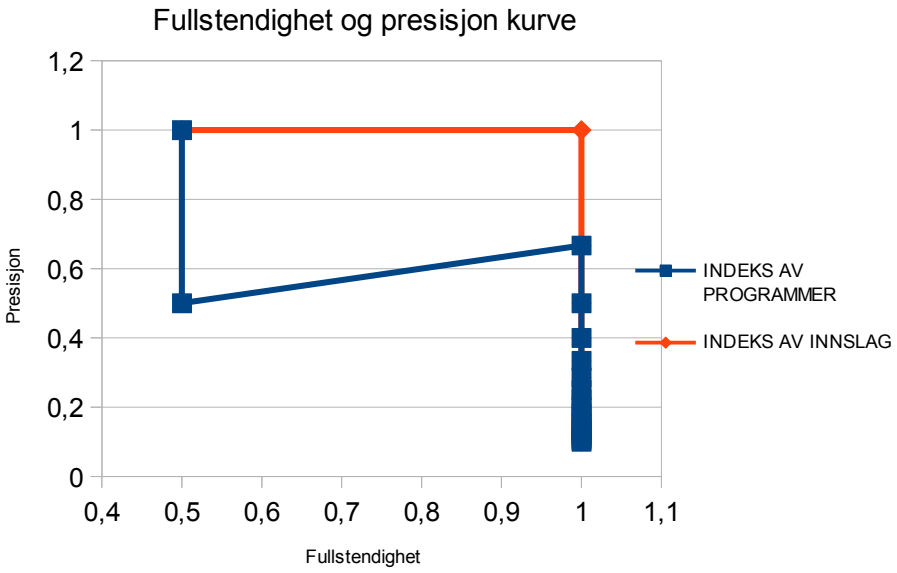


Fasit representert ved nøkkel
"1990/03273"
"1990/06210"
"1990/03033"
"1994/11025"
"2011/03020"

Søkespørsmål nr 20: "Rosemarie Køhn blir Norges første kvinnelige biskop"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,5	1	1,5272776	"1993/09893"
2	0,5	0,5	0,574028	"1990/06731"
3	1	0,6666666667	0,5265242	"2007/22013"
4	1	0,5	0,4842434	"1998/00123"
5	1	0,4	0,44310987	"1998/12154"
6	1	0,3333333333	0,44053888	"1998/11272"
7	1	0,2857142857	0,42244267	"1995/13082"
8	1	0,25	0,37823442	"1996/00689"
9	1	0,2222222222	0,37159115	"1999/01799"
10	1	0,2	0,35301438	"1997/16975"
11	1	0,1818181818	0,22816595	"1991/00742"
12	1	0,1666666667	0,12083614	"1994/30563"
13	1	0,1538461538	0,1184348	"2010/04086"
14	1	0,1428571429	0,11342796	"2005/02387"
15	1	0,1333333333	0,111603014	"1995/12340"
16	1	0,125	0,10479747	"2005/01334"
17	1	0,1176470588	0,10371578	"1990/04229"
18	1	0,1111111111	0,09188932	"1997/04806"
19	1	0,1052631579	0,09057605	"2009/02935"
20	1	0,1	0,08708269	"2011/09862"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,5	1	2,684929	"2007/22013"
2	1	1	2,5944486	"1993/09893"
3	1	0,6666666667	1,3281894	"1998/00123"
4	1	0,5	1,1827842	"1998/12154"
6	1	0,4	0,90874255	"1996/00689"
7	1	0,3333333333	0,87740725	"1998/11272"
8	1	0,2857142857	0,7571945	"1995/13082"
9	1	0,25	0,7101656	"1997/16975"
10	1	0,2222222222	0,6695506	"1990/06731"
11	1	0,2	0,6629866	"1999/01799"
12	1	0,1818181818	0,3148595	"1991/00742"
13	1	0,1666666667	0,30307347	"1994/30563"
14	1	0,1538461538	0,21090475	"1995/12340"
15	1	0,1428571429	0,19041528	"2010/07410"
16	1	0,1333333333	0,17879042	"2005/02387"
17	1	0,125	0,15575576	"2003/15056"
18	1	0,1176470588	0,15492736	"2010/04086"
19	1	0,1111111111	0,15457137	"1990/04229"
20	1	0,1052631579	0,15219595	"2009/10300"
21	1	0,1	0,14752342	"2009/03455"

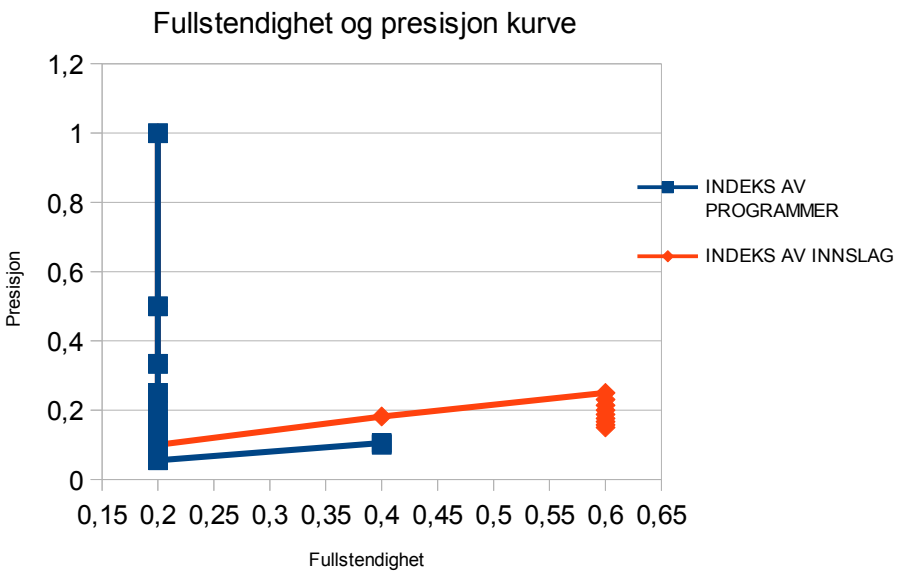


Fasit representert ved nøkkel
"1993/09893"
"2007/22013"

Søkespørsmål nr 22: "Norge blir klare for VM i fotball 1994 etter å ha slått Polen 3-0 på bortebane. Jostein Flo, Jan Åge Fjørtoft og Ronny Johnsen skåret målene"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,2	1	0,3654697	"2007/22013"
2	0,2	0,5	0,14533356	"1996/08071"
3	0,2	0,3333333333	0,13360934	"1998/00123"
4	0,2	0,25	0,09549727	"1999/00994"
5	0,2	0,2	0,089114554	"1998/06513"
6	0,2	0,1666666667	0,082684495	"1995/08259"
7	0,2	0,1428571429	0,07780095	"1990/03827"
8	0,2	0,125	0,06711304	"1998/08313"
9	0,2	0,1111111111	0,06602014	"1990/03033"
10	0,2	0,1	0,06341656	"2010/07177"
11	0,2	0,0909090909	0,06303295	"2012/05152"
12	0,2	0,0833333333	0,0624238	"1993/21153"
13	0,2	0,0769230769	0,059138563	"1996/09911"
14	0,2	0,0714285714	0,057104383	"1998/06852"
15	0,2	0,0666666667	0,056531873	"2007/02789"
16	0,2	0,0625	0,05622172	"2005/02387"
17	0,2	0,0588235294	0,053862136	"1997/09155"
18	0,2	0,0555555556	0,053638987	"2009/02124"
19	0,4	0,1052631579	0,053603828	"1993/23215"
20	0,4	0,1	0,053297102	"2012/08250"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,2	1	1,5170828	"2007/22013"
2	0,2	0,5	0,34981108	"1996/08071"
3	0,2	0,3333333333	0,25970086	"1999/00994"
5	0,2	0,25	0,15544228	"1998/07162"
6	0,2	0,2	0,14105077	"1998/08313"
7	0,2	0,1666666667	0,13816135	"2012/05152"
8	0,2	0,1428571429	0,114734136	"2005/02387"
9	0,2	0,125	0,11214852	"1995/08259"
10	0,2	0,1111111111	0,097078174	"1994/21044"
11	0,2	0,1	0,09668044	"1990/03827"
12	0,4	0,1818181818	0,096333325	"1993/23215"
13	0,6	0,25	0,094173126	"1994/17196"
14	0,6	0,2307692308	0,087960444	"2010/07177"
15	0,6	0,2142857143	0,08611139	"1995/09199"
16	0,6	0,2	0,08584591	"1993/21153"
17	0,6	0,1875	0,08514036	"1998/06513"
18	0,6	0,1764705882	0,08283045	"1998/08148"
19	0,6	0,1666666667	0,08129281	"2010/06459"
20	0,6	0,1578947368	0,071446106	"1997/09155"
21	0,6	0,15	0,069099315	"2009/10300"

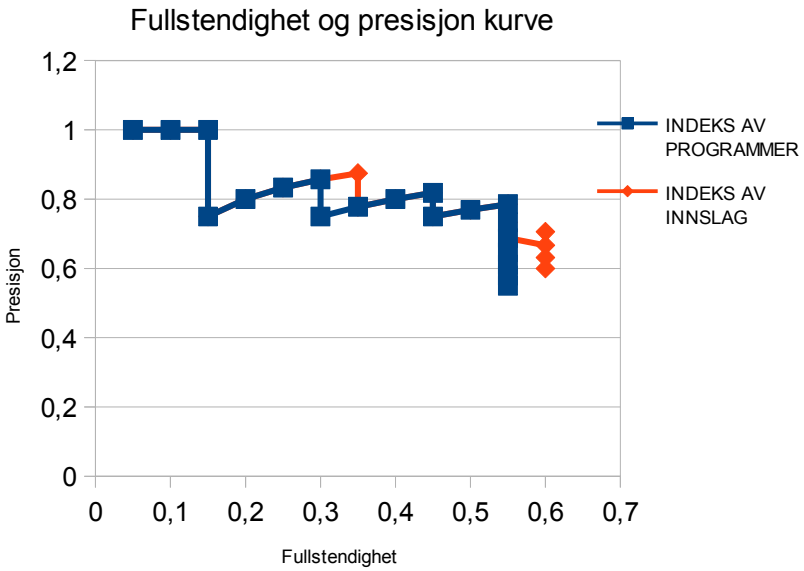


Fasit representert ved nøkkel
"1994/11025"
"1994/08824"
"2007/22013"
"1993/23215"
"1994/17196"

Søkespørsmål nr 24: "Vinter-OL 1994 arrangeres på Lillehammer"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,05	1	1,6268231	"1990/05536"
2	0,1	1	1,0338819	"1990/07061"
3	0,15	1	0,9209908	"1990/02181"
4	0,15	0,75	0,45009094	"1998/01872"
5	0,2	0,8	0,28157157	"1990/05571"
6	0,25	0,8333333333	0,24113461	"1990/05106"
7	0,3	0,8571428571	0,21197401	"1993/09903"
8	0,3	0,75	0,15797037	"1992/00044"
9	0,35	0,7777777778	0,13010426	"1994/05863"
10	0,4	0,8	0,120397195	"1994/02886"
11	0,45	0,8181818182	0,113304876	"2010/03795"
12	0,45	0,75	0,09197868	"1999/01799"
13	0,5	0,7692307692	0,09029789	"1994/00786"
14	0,55	0,7857142857	0,08962883	"1993/21480"
15	0,55	0,6875	0,08936153	"1995/12653"
16	0,55	0,7333333333	0,08936153	"2009/11378"
17	0,55	0,6470588235	0,08870747	"1993/00199"
18	0,55	0,6111111111	0,08227749	"1998/02444"
19	0,55	0,5789473684	0,081252545	"1997/12578"
20	0,55	0,55	0,08011865	"2009/08733"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,05	1	2,0052938	"1990/05536"
2	0,1	1	1,1325577	"1990/02181"
3	0,15	1	1,0803827	"1990/07061"
4	0,15	0,75	0,84570843	"1998/01872"
5	0,2	0,8	0,508518	"1993/09903"
7	0,25	0,8333333333	0,35957652	"2010/03795"
8	0,3	0,8571428571	0,35742438	"1990/05571"
9	0,35	0,875	0,35596263	"1994/02886"
10	0,35	0,7777777778	0,32084438	"1992/00044"
11	0,4	0,8	0,3060948	"1990/05106"
12	0,45	0,8181818182	0,28358907	"1993/21480"
13	0,45	0,75	0,2766042	"1997/12578"
14	0,5	0,7692307692	0,27582934	"1994/05863"
15	0,55	0,7857142857	0,2674576	"1994/00786"
16	0,55	0,7333333333	0,26419717	"1998/02444"
18	0,55	0,6875	0,254259	"2009/08733"
19	0,6	0,6666666667	0,22687125	"2009/11378"
20	0,6	0,7058823529	0,22687125	"1994/04091"
21	0,6	0,6315789474	0,22627465	"1998/06513"
22	0,6	0,6	0,2089419	"1999/01799"

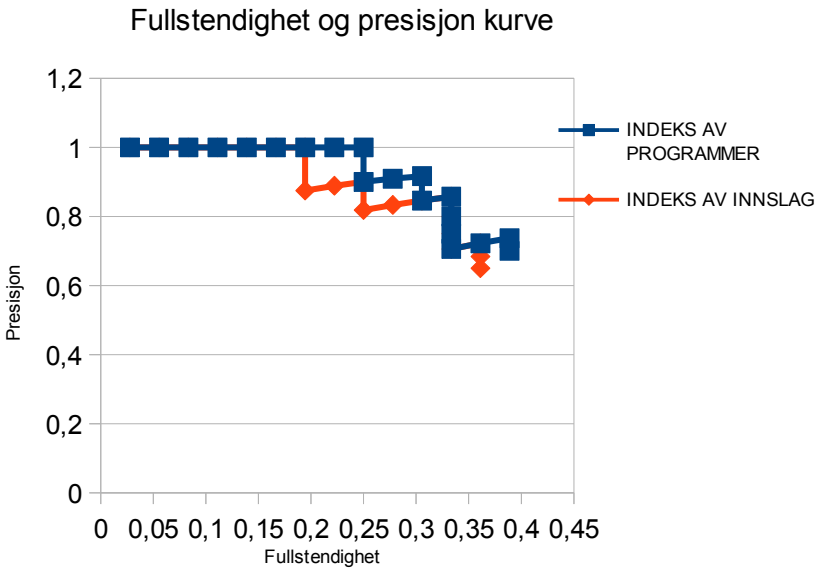


Fasit representert ved nøkkel
"1990/05106"
"1990/02181"
"1990/05571"
"1990/07061"
"1993/09903"
"1993/21480"
"1994/00786"
"1994/05863"
"1994/04544"
"1995/00528"
"1997/02229"
"1994/03343"
"1994/03282"
"1993/20002"
"1994/02886"
"1994/01361"
"1994/02906"
"1994/04091"
"2010/03795"
"1990/05536"

Søkespørsmål nr 27: "Norge stemmer nei til EU-medlemskap"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,0277777778	1	0,8285652	"1995/00627"
2	0,0555555556	1	0,5168254	"1995/00635"
3	0,0833333333	1	0,46354145	"1994/29909"
4	0,1111111111	1	0,3777543	"1994/30203"
5	0,1388888889	1	0,30128232	"1995/02450"
6	0,1666666667	1	0,24397013	"1994/09284"
7	0,1944444444	1	0,24170528	"1993/00019"
8	0,2222222222	1	0,2192525	"1994/08700"
9	0,25	1	0,2192525	"1994/28214"
10	0,25	0,9	0,20052388	"1994/25502"
11	0,2777777778	0,9090909091	0,19591385	"1993/08034"
12	0,3055555556	0,9166666667	0,18291108	"1994/16332"
13	0,3055555556	0,8461538462	0,17750597	"1999/00994"
14	0,3333333333	0,8571428571	0,17233129	"1994/16056"
15	0,3333333333	0,8	0,16830233	"2012/08250"
16	0,3333333333	0,75	0,16647236	"1998/15806"
17	0,3333333333	0,7058823529	0,15980951	"1992/01558"
18	0,3611111111	0,7222222222	0,15548162	"1994/29419"
19	0,3888888889	0,7368421053	0,15439512	"1995/00528"
20	0,3888888889	0,7	0,15394431	"1995/15925"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,0277777778	1	1,1107191	"1995/00627"
3	0,0555555556	1	0,73537654	"1995/02450"
4	0,0833333333	1	0,6759853	"1994/09284"
5	0,1111111111	1	0,6664347	"1995/00635"
6	0,1388888889	1	0,5680629	"1994/08700"
7	0,1666666667	1	0,50275594	"1994/30203"
8	0,1944444444	1	0,47076774	"1995/00528"
9	0,1944444444	0,875	0,46036246	"1998/15806"
10	0,2222222222	0,8888888889	0,44406927	"1994/16056"
11	0,25	0,9	0,42604718	"1994/28214"
12	0,25	0,8181818182	0,41366535	"1995/15925"
13	0,2777777778	0,8333333333	0,408776	"1994/2990"
14	0,3055555556	0,8461538462	0,3665529	"1994/17196"
16	0,3333333333	0,8571428571	0,3330715	"1993/00019"
17	0,3333333333	0,8	0,33104354	"1995/06804"
18	0,3333333333	0,75	0,32614923	"2010/08909"
19	0,3333333333	0,7058823529	0,31422248	"2012/08250"
20	0,3611111111	0,7222222222	0,30947506	"1994/29419"
21	0,3611111111	0,6842105263	0,28790656	"1998/14097"
22	0,3611111111	0,65	0,255438	"1994/25502"

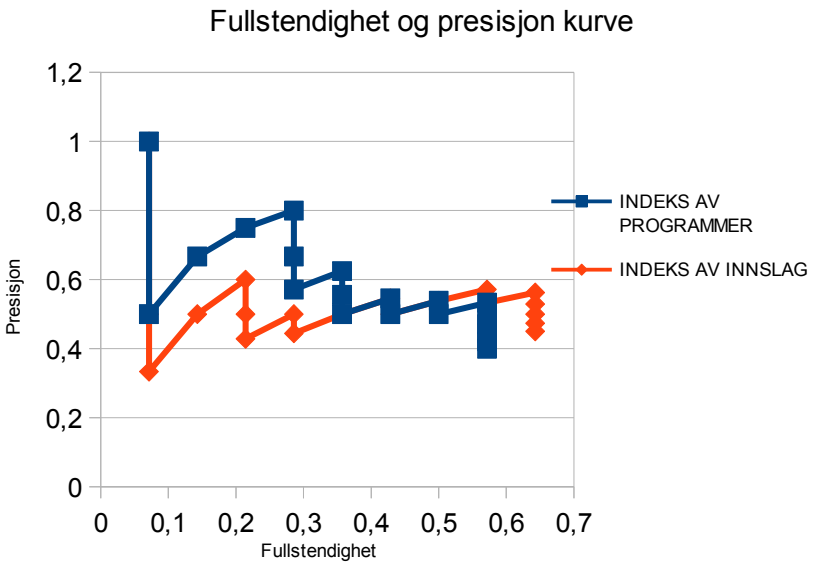


Fasit representert ved nøkkel
"1995/00627"
"1995/00528"
"1994/29130"
"1994/28792"
"1994/28166"
"1994/17637"
"1994/30203"
"1994/29419"
"1994/09691"
"1994/27027"
"1994/29909"
"1994/07133"
"1994/07018"
"1993/00019"
"1993/08034"
"1994/00247"
"1994/03282"
"1994/03343"
"1994/04544"
"1994/05387"
"1994_05685"
"1994/06062"
"1994/07647"
"1994/08700"
"1994/09546"
"1994/09284"
"1994/16056"
"1994/16332"
"1994/16428"
"1994/16472"
"1994/17196"
"1994/17455"
"1994/27574"
"1994/28214"
"1995/00635"
"1995/02450"

Søkespørsmål nr 32: "Amerikanske fredsbevarende styrker rykker inn i Bosnia-Hercegovina som et resultat av Dayton-avtalen"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,0714285714	1	0,12142402	"1995/00528"
2	0,0714285714	0,5	0,09905303	"1998/15467"
3	0,1428571429	0,6666666667	0,095665134	"1992/01273"
4	0,2142857143	0,75	0,090070345	"1995/06804"
5	0,2857142857	0,8	0,070926696	"1992/01565"
6	0,2857142857	0,6666666667	0,062027127	"2009/06651"
7	0,2857142857	0,5714285714	0,059667945	"2011/05923"
8	0,3571428571	0,625	0,055761773	"2007/22013"
9	0,3571428571	0,5555555556	0,049994137	"1990/07768"
10	0,3571428571	0,5	0,04582091	"1998/06250"
11	0,4285714286	0,5454545455	0,04393702	"1996/13240"
12	0,4285714286	0,5	0,04320037	"2008/16961"
13	0,5	0,5384615385	0,0418535	"1995/15351"
14	0,5	0,5	0,03999531	"1992/00999"
15	0,5714285714	0,5333333333	0,039336454	"1995/09907"
16	0,5714285714	0,5	0,038766954	"1999/01482"
17	0,5714285714	0,4705882353	0,037800323	"1999/05645"
18	0,5714285714	0,4444444444	0,037714053	"1998/14097"
19	0,5714285714	0,4	0,035344765	"1994/07133"
20	0,5714285714	0,4210526316	0,035344765	"1996/05299"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,0714285714	1	0,1853366	"1995/06804"
2	0,0714285714	0,5	0,15413451	"2009/06651"
3	0,0714285714	0,3333333333	0,15347865	"1998/15467"
4	0,1428571429	0,5	0,14401945	"2007/22013"
5	0,2142857143	0,6	0,12347096	"1996/13240"
6	0,2142857143	0,5	0,12260792	"1998/06250"
7	0,2142857143	0,4285714286	0,12247734	"1999/03016"
8	0,2857142857	0,5	0,12006709	"1992/01565"
9	0,2857142857	0,4444444444	0,115595855	"2008/16961"
10	0,4285714286	0,5454545455	0,11521556	"1992/01273"
11	0,3571428571	0,5	0,11521556	"1995/15351"
12	0,4285714286	0,5	0,111595176	"2011/05923"
13	0,5714285714	0,5714285714	0,10896782	"1995/00528"
14	0,5	0,5384615385	0,10896782	"1995/09907"
15	0,5714285714	0,5333333333	0,10114638	"1999/05645"
16	0,6428571429	0,5625	0,095346846	"1995/09791"
18	0,6428571429	0,5	0,08058585	"1994/07133"
19	0,6428571429	0,5294117647	0,08058585	"1996/05299"
21	0,6428571429	0,4736842105	0,05978851	"1990/07768"
22	0,6428571429	0,45	0,05978851	"1999/01482"



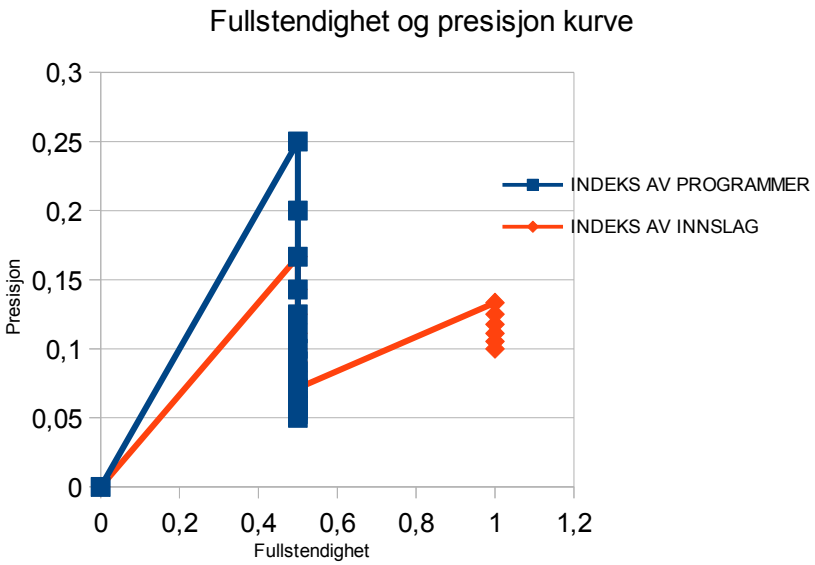
Fasit representert ved nøkkel
"1995/09791"
"1995/09907"
"1996/04395"
"1994/31520"
"1992/01273"
"1992/01565"
"1995/15351"
"2007/22013"
"1995/09356"
"1995/06804"
"1994/27027"
"1996/13240"
"1994/29909"
"1995/00528"

Søkespørsmål nr 33: "Frankrikes president Jacques Chirac annonserer en "definitiv slutt" på landets atomprøvesprengninger"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0	0	0,12886712	"1999/02919"
2	0	0	0,12558317	"2011/07410"
3	0	0	0,10917649	"1995/06804"
4	0,5	0,25	0,09096883	"1995/09791"
5	0,5	0,2	0,078297436	"2010/08450"
6	0,5	0,1666666667	0,0667638	"2007/28589"
7	0,5	0,1428571429	0,062245954	"1994/16428"
8	0,5	0,125	0,05799582	"2011/00411"
9	0,5	0,1111111111	0,054528203	"2008/00546"
10	0,5	0,1	0,04996735	"2009/04986"
11	0,5	0,0909090909	0,048228085	"2011/03020"
12	0,5	0,0833333333	0,045588173	"1992/00016"
13	0,5	0,0769230769	0,045531966	"1998/06852"
14	0,5	0,0714285714	0,044886783	"1998/10979"
15	0,5	0,0625	0,043240923	"2011/01673"
16	0,5	0,0666666667	0,043240923	"2010/07410"
17	0,5	0,0588235294	0,04228644	"1997/11632"
18	0,5	0,0555555556	0,040472	"1995/14389"
19	0,5	0,0526315789	0,037822615	"1998/04824"
20	0,5	0,05	0,037761472	"1994/21184"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0	0	0,32896802	"2011/07410"
2	0	0	0,24615876	"1999/02919"
3	0	0	0,21824114	"2008/00546"
4	0	0	0,21386121	"1995/06804"
5	0	0	0,19323984	"2010/08450"
6	0,5	0,1666666667	0,16374068	"1995/09791"
7	0,5	0,1428571429	0,15165414	"1994/16428"
8	0,5	0,125	0,14397392	"1998/06852"
9	0,5	0,1111111111	0,12520932	"2009/04986"
10	0,5	0,1	0,12516977	"2007/28589"
11	0,5	0,0909090909	0,12055357	"2011/03020"
12	0,5	0,0833333333	0,08485162	"2011/00411"
13	0,5	0,0769230769	0,078386106	"2012/07893"
14	0,5	0,0714285714	0,07670785	"1997/15990"
15	1	0,1333333333	0,07218955	"1995/12919"
16	1	0,125	0,0678813	"1994/16332"
17	1	0,1176470588	0,06711937	"1997/11632"
18	1	0,1111111111	0,06633115	"1992/00016"
19	1	0,1052631579	0,06600995	"2005/01894"
20	1	0,1	0,06331277	"2010/07410"

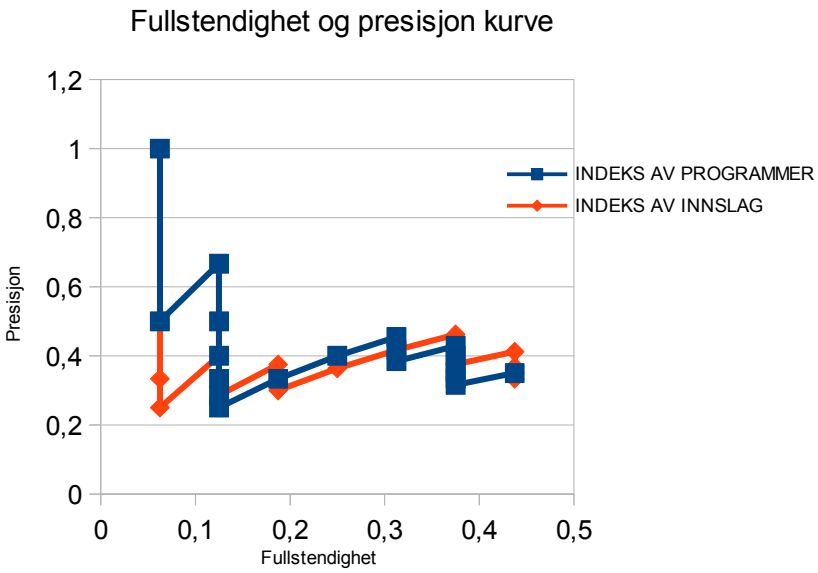
Fasit representert ved nøkkel
"1995/12919"
"1995/09791"



Søkespørsmål nr 54: "USA og andre allierte styrker angriper Afghanistan og Taliban"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,0625	1	0,3780884	"2001/04866"
2	0,0625	0,5	0,30940503	"2001/04649"
3	0,125	0,6666666667	0,1795678	"2009/04192"
4	0,125	0,5	0,11217745	"1992/01123"
5	0,125	0,4	0,11206303	"1997/13226"
6	0,125	0,3333333333	0,10321065	"2004/12339"
7	0,125	0,2857142857	0,09913873	"2013/07003"
8	0,125	0,25	0,09607947	"1998/02150"
9	0,1875	0,3333333333	0,09471834	"2012/07644"
10	0,25	0,4	0,0764403	"2002/00399"
11	0,3125	0,4545454545	0,07359561	"2010/06980"
12	0,3125	0,4166666667	0,06826699	*"2007/14407"
13	0,3125	0,3846153846	0,06488588	"1999/01482"
14	0,375	0,4285714286	0,05811089	"2009/13074"
15	0,375	0,4	0,056843117	"1999/11156"
16	0,375	0,375	0,05647516	"2007/20657"
17	0,375	0,3529411765	0,052039955	"1990/03033"
18	0,375	0,3333333333	0,050032053	"1990/01416"
19	0,375	0,3157894737	0,049760647	"1990/00127"
20	0,4375	0,35	0,048586372	"2010/09169"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,0625	1	0,45943743	"2001/04866"
2	0,0625	0,5	0,3917573	"2001/04649"
3	0,0625	0,3333333333	0,30182254	"2013/07003"
4	0,0625	0,25	0,26174772	"1998/02150"
5	0,125	0,4	0,19577162	"2012/07644"
6	0,125	0,3333333333	0,18580593	"2007/14407"
7	0,125	0,2857142857	0,16134152	"1997/13226"
8	0,1875	0,375	0,14720386	"2010/09169"
9	0,1875	0,3333333333	0,14414282	"1992/01123"
10	0,1875	0,3	0,14414282	"2003/15056"
12	0,25	0,3636363636	0,14166783	"2009/04192"
13	0,3125	0,4166666667	0,13825686	"2009/13074"
14	0,375	0,4615384615	0,13271852	"2010/08600"
15	0,375	0,4285714286	0,12170064	"2009/11378"
16	0,375	0,4	0,12028981	"2007/20657"
18	0,375	0,375	0,1153049	"2010/02484"
19	0,4375	0,4117647059	0,105499275	"2002/00399"
20	0,4375	0,35	0,10089179	"1993/13411"
21	0,4375	0,3333333333	0,10089179	"2013/07278"
22	0,4375	0,3684210526	0,10089179	"1997/09682"



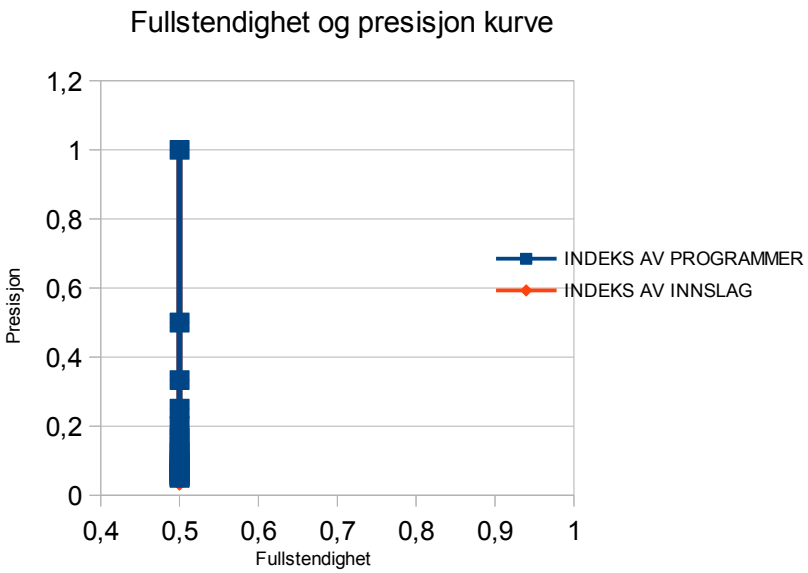
Fasit representert ved nøkkel
"2008/00237"
"2007/19685"
"2007/28876"
"2007/28450"
"2009/13074"
"2012/07644"
"2010/08600"
"2009/09682"
"2001/04866"
"2002/00399"
"2009/04192"
"2011/03659"
"2010/07410"
"2010/06980"
"2010/09169"
"2010/01237"

Søkespørsmål nr 63: "VM i skiskyting"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,5	1	1,1643611	"2011/02781"
2	0,5	0,5	0,20914552	"1993/10504"
3	0,5	0,3333333333	0,18075201	"2010/01742"
4	0,5	0,25	0,15093777	"1995/10577"
5	0,5	0,1666666667	0,14942078	"2011/02148"
6	0,5	0,2	0,14942078	"2011/02356"
7	0,5	0,1428571429	0,14788821	"1997/03366"
8	0,5	0,1111111111	0,12324017	"1994/08824"
9	0,5	0,125	0,12324017	"2002/06375"
10	0,5	0,1	0,12200155	"1999/11089"
11	0,5	0,0909090909	0,12075022	"2011/01870"
12	0,5	0,0833333333	0,12075022	"2010/07177"
13	0,5	0,0769230769	0,116915904	"1998/06513"
14	0,5	0,0714285714	0,11297001	"1998/02444"
15	0,5	0,0666666667	0,10672913	"1998/08313"
16	0,5	0,0588235294	0,10457276	"1998/06852"
17	0,5	0,0625	0,10457276	"2002/04898"
18	0,5	0,0526315789	0,09859214	"1994/10217"
19	0,5	0,05	0,09859214	"1995/09199"
20	0,5	0,0555555556	0,09859214	"2005/02783"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,5	1	4,100512	"2011/02781"
4	0,5	0,5	0,5006103	"1997/03366"
5	0,5	0,3333333333	0,4424812	"1994/27027"
6	0,5	0,2	0,4424812	"1998/08615"
7	0,5	0,1666666667	0,4424812	"1998/07817"
8	0,5	0,25	0,4424812	"1999/11089"
9	0,5	0,125	0,40290752	"1998/02444"
10	0,5	0,1428571429	0,40290752	"2010/01742"
11	0,5	0,0909090909	0,38319993	"2011/02356"
12	0,5	0,1111111111	0,38319993	"1995/10577"
13	0,5	0,1	0,38319993	"1998/06513"
15	0,5	0,0769230769	0,3754577	"1995/09199"
16	0,5	0,0833333333	0,3754577	"2005/02783"
17	0,5	0,0666666667	0,35398495	"1996/04709"
19	0,5	0,0625	0,35398495	"1998/08148"
20	0,5	0,0714285714	0,35398495	"1999/11474"
21	0,5	0,0555555556	0,31288144	"1990/03033"
22	0,5	0,0526315789	0,31288144	"1997/10975"
23	0,5	0,0588235294	0,31288144	"1997/09155"
24	0,5	0,0416666667	0,30973685	"2010/07410"

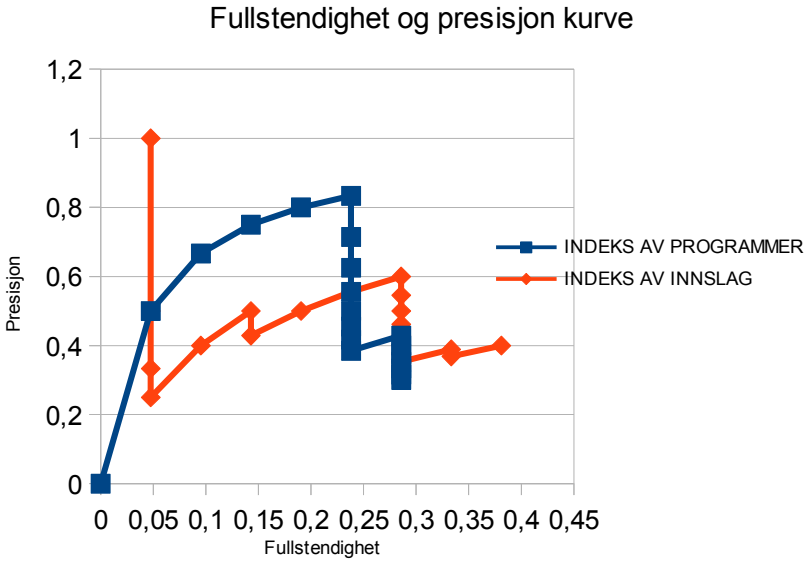
Fasit representert ved nøkkel
"2011/02781"
"2005/03697"



Søkespørsmål nr 87: "Kommune- og fylkestingsvalg i Norge"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0	0	0,3288036	"2007/29846"
2	0,0476190476	0,5	0,26986623	"1996/00689"
3	0,0952380952	0,6666666667	0,23387209	"1995/10233"
4	0,1428571429	0,75	0,22990215	"1996/08655"
5	0,1904761905	0,8	0,18585303	"1995/11285"
6	0,2380952381	0,8333333333	0,17824171	"1995/11792"
7	0,2380952381	0,7142857143	0,17015903	"2006/05271"
8	0,2380952381	0,625	0,15816417	"2011/11921"
9	0,2380952381	0,5555555556	0,14255722	"1995/15925"
10	0,2380952381	0,4545454545	0,14255722	"2011/03446"
11	0,2380952381	0,5	0,14255722	"2011/09312"
12	0,2380952381	0,4166666667	0,13949955	"1998/08846"
13	0,2380952381	0,3846153846	0,13949955	"1996/10539"
14	0,2857142857	0,4285714286	0,13839366	"1995/07028"
15	0,2857142857	0,4	0,13152145	"1995/07341"
16	0,2857142857	0,3333333333	0,13152145	"2008/12207"
17	0,2857142857	0,375	0,13152145	"2009/03455"
18	0,2857142857	0,3529411765	0,13152145	"2009/09682"
19	0,2857142857	0,3	0,13109839	"1993/01227"
20	0,2857142857	0,3157894737	0,13109839	"2007/29309"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,0476190476	1	0,5526202	"1996/08655"
2	0,0476190476	0,5	0,44468287	"1998/08846"
3	0,0476190476	0,3333333333	0,39034653	"2013/04618"
4	0,0476190476	0,25	0,35935804	"2011/09312"
5	0,0952380952	0,4	0,34560764	"1995/10233"
6	0,1428571429	0,5	0,31443828	"1995/07028"
7	0,1428571429	0,4285714286	0,29879683	"2010/03612"
8	0,2380952381	0,5555555556	0,28800637	"1995/11285"
9	0,1904761905	0,5	0,28800637	"1999/11089"
10	0,2857142857	0,6	0,28511176	"1995/11792"
11	0,2857142857	0,5	0,2760167	"2007/29846"
12	0,2857142857	0,5454545455	0,2760167	"2007/29309"
13	0,2857142857	0,4615384615	0,26951852	"1995/08259"
14	0,2857142857	0,4285714286	0,24396658	"2007/21802"
15	0,2857142857	0,375	0,24151461	"2007/00375"
16	0,2857142857	0,4	0,24151461	"2013/05326"
17	0,2857142857	0,3529411765	0,23393036	"2006/05271"
18	0,3333333333	0,3888888889	0,23040509	"1996/00689"
20	0,3333333333	0,3684210526	0,2213497	"2011/11921"
22	0,380952381	0,4	0,17280382	"1995/11631"

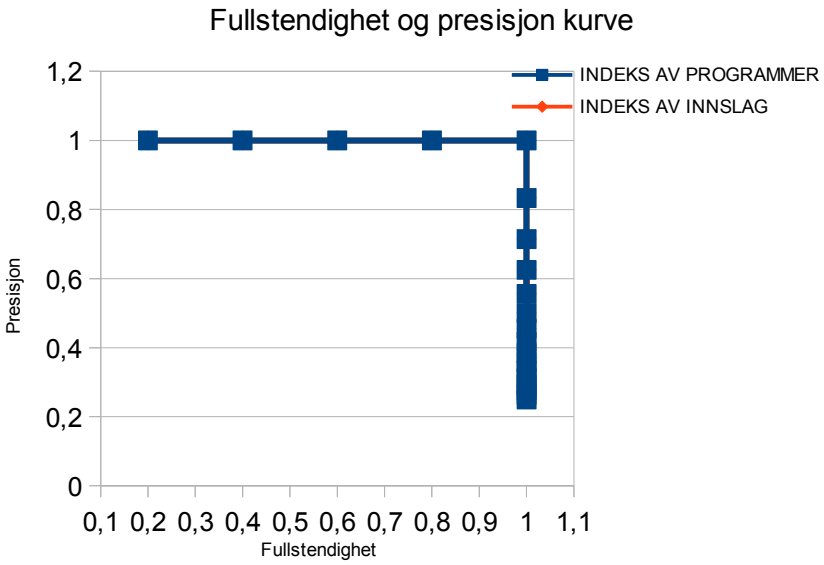


Fasit representert ved nøkkel
"1996/08655"
"1996/00689"
"1995/10577"
"1995/10922"
"1995/11285"
"1995/11631"
"1995/11792"
"1999/10799"
"1999/11089"
"1999/11770"
"1999/12170"
"2007/18092"
"2007/18928"
"2007/22778"
"2011/08114"
"2011/08240"
"2011/08338"
"2011/08918"
"2011/09110"
"1995/07028"
"1995/10233"

Søkespørsmål nr 100: "Joshua French og Tjostolv Moland blir dømt til døden i Kongo"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,2	1	1,0611838	"2009/10603"
2	0,6	1	0,78565454	"2009/09359"
3	0,4	1	0,78565454	"2009/12144"
4	0,8	1	0,5734888	"2010/09438"
5	1	1	0,26682165	"2009/13074"
6	1	0,8333333333	0,08238567	"1994/25864"
7	1	0,7142857143	0,080343045	"1992/00577"
8	1	0,625	0,073069885	"1993/10501"
9	1	0,5555555556	0,06255011	*"1996/00689"
10	1	0,5	0,05356203	"2009/03233"
11	1	0,4545454545	0,04975879	"2002/00399"
12	1	0,4166666667	0,046996545	"1998/00123"
13	1	0,3846153846	0,045411006	"1999/08869"
14	1	0,3333333333	0,044773348	"1999/02370"
15	1	0,3571428571	0,044773348	"1998/07817"
16	1	0,3125	0,04281391	"2009/05374"
17	1	0,2941176471	0,038630847	"1997/16364"
18	1	0,2777777778	0,037462167	"1996/09614"
19	1	0,2631578947	0,03150032	"1995/10577"
20	1	0,25	0,031410594	"2009/10882"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,2	1	2,211095	"2009/12144"
2	0,4	1	1,8514369	"2009/09359"
3	0,6	1	1,4943302	"2009/10603"
4	0,8	1	0,94368577	"2010/09438"
5	1	1	0,78739697	"2009/13074"
6	1	0,8333333333	0,16256651	"1999/08869"
7	1	0,7142857143	0,118790545	"1995/10577"
8	1	0,5555555556	0,11495188	"2009/03233"
9	1	0,625	0,11495188	"1996/09614"
10	1	0,5	0,11093921	"1994/25864"
11	1	0,4545454545	0,102938116	"1997/16364"
12	1	0,4166666667	0,10014215	"1996/00689"
13	1	0,3846153846	0,08851399	"1993/10501"
14	1	0,3571428571	0,08494819	"1999/06606"
15	1	0,3333333333	0,081765726	"1994/02906"
16	1	0,3125	0,07347062	"2009/10882"
19	1	0,2941176471	0,060762286	"2002/00399"
20	1	0,2777777778	0,060067434	"1992/00577"
21	1	0,2631578947	0,057817094	"1998/03352"
22	1	0,25	0,051951572	"1997/07112"



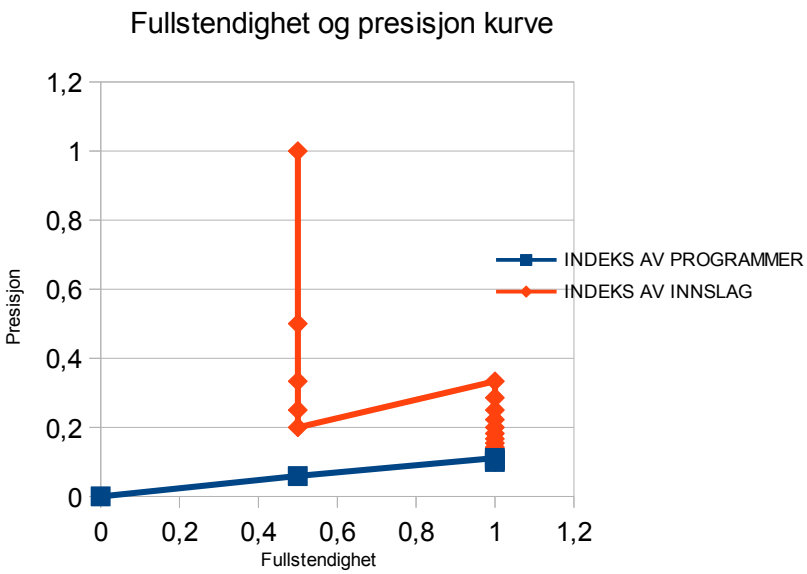
Fasit representert ved nøkkel
"2009/13074"
"2009/09359"
"2009/10603"
"2010/09438"
"2009/12144"

Søkespørsmål nr 118: "Gjenopptakelseskommisjonen bestemmer at Treholt-saken ikke skal gjenopptas"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0	0	0,13291982	"1990/04168"
2	0	0	0,1328621	"1995/15384"
3	0	0	0,12039004	"1990/02370"
4	0	0	0,11741921	"1990/01945"
5	0	0	0,10638183	"1990/01946"
6	0	0	0,10078822	"2012/08048"
7	0	0	0,10007825	"1990/02507"
8	0	0	0,09993549	"1994/24639"
9	0	0	0,095020026	"1993/00019"
10	0	0	0,0890451	"1998/10979"
11	0	0	0,08820268	"1990/01951"
12	0	0	0,07910785	"1990/03827"
13	0	0	0,075490326	"1990/01949"
14	0	0	0,074709356	"1990/02595"
15	0	0	0,06870738	"1990/04749"
16	0	0	0,06870738	"1990/06110"
17	0,5	0,0588235294	0,067320466	"2010/09169"
18	1	0,1111111111	0,06379417	"2010/09623"
19	1	0,1052631579	0,061997667	"1996/09968"
20	1	0,1	0,051530533	"1990/00713"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,5	1	0,24505955	"2010/09169"
2	0,5	0,5	0,23926984	"2012/08048"
3	0,5	0,3333333333	0,17690608	"1995/15384"
4	0,5	0,25	0,17426871	"2007/18928"
5	0,5	0,2	0,17293426	"1990/04168"
6	1	0,3333333333	0,16007224	"2010/09623"
7	1	0,2857142857	0,15767539	"1990/02370"
8	1	0,25	0,15313426	"1990/01945"
9	1	0,2222222222	0,14236736	"1996/09968"
10	1	0,2	0,13861388	"1990/01946"
11	1	0,1538461538	0,13070153	"1993/00019"
12	1	0,1818181818	0,13070153	"2012/05152"
13	1	0,1666666667	0,13070153	"2012/04237"
14	1	0,1428571429	0,1295995	"1994/24639"
15	1	0,1333333333	0,12941015	"1990/02507"
16	1	0,125	0,115970574	"1990/01951"
17	1	0,1176470588	0,108917944	"2010/04954"
18	1	0,1111111111	0,10644401	"1990/03827"
20	1	0,1052631579	0,10439473	"1998/10979"
21	1	0,1	0,09912607	"1990/01949"

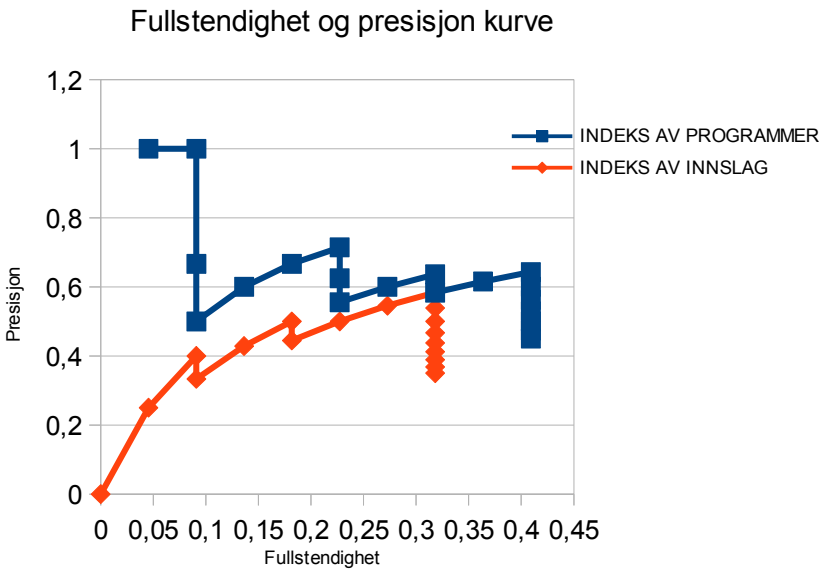
Fasit representert ved nøkkel
"2010/09169"
"2010/09623"



Søkespørsmål nr 120: "Terrorangrepene i Norge 2011 i Regjeringskvartalet og på Utøya. 77 mennesker omkommer, mange blir skadet, og det ble store materielle ødeleggelser"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,0454545455	1	0,2747465	"2011/07801"
2	0,0909090909	1	0,25098473	"2011/07539"
3	0,0909090909	0,6666666667	0,23402986	"2007/29309"
4	0,0909090909	0,5	0,20765606	"2011/11921"
5	0,1363636364	0,6	0,16276203	"2011/08240"
6	0,1818181818	0,6666666667	0,14641094	"2012/06053"
7	0,2272727273	0,7142857143	0,14481293	"2011/08114"
8	0,2272727273	0,625	0,10252866	"2011/03020"
9	0,2272727273	0,5555555556	0,08205387	"2011/09862"
10	0,2727272727	0,6	0,07453239	"2011/07894"
11	0,3181818182	0,6363636364	0,06991422	"2012/07277"
12	0,3181818182	0,5833333333	0,0685833	"2013/08925"
13	0,3636363636	0,6153846154	0,06356817	"2012/03858"
14	0,4090909091	0,6428571429	0,0626228	"2012/05224"
15	0,4090909091	0,6	0,051867235	"2010/07410"
16	0,4090909091	0,5625	0,04983913	"2007/02789"
17	0,4090909091	0,5294117647	0,045321904	"1990/01146"
18	0,4090909091	0,5	0,044660166	"2011/02781"
19	0,4090909091	0,4736842105	0,044012167	"2011/12147"
20	0,4090909091	0,45	0,042831928	"2011/06824"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0	0	0,3180756	"2007/29309"
2	0	0	0,2685238	"2013/08925"
3	0	0	0,25803944	"2011/11921"
4	0,0454545455	0,25	0,23346421	"2011/08240"
5	0,0909090909	0,4	0,19431451	"2012/03858"
6	0,0909090909	0,3333333333	0,19397669	"2011/03020"
7	0,1363636364	0,4285714286	0,18031675	"2012/06053"
8	0,1818181818	0,5	0,1456165	"2011/07801"
9	0,1818181818	0,4444444444	0,11787966	"2007/02789"
10	0,2272727273	0,5	0,11681708	"2011/08114"
12	0,2727272727	0,5454545455	0,113615975	"2012/05224"
13	0,3181818182	0,5833333333	0,10433984	"2011/07539"
15	0,3181818182	0,5384615385	0,09692969	"2013/07878"
17	0,3181818182	0,5	0,08631002	"1993/18890"
19	0,3181818182	0,4666666667	0,07787806	"2011/09862"
20	0,3181818182	0,4375	0,07630333	"2010/03845"
23	0,3181818182	0,4117647059	0,06969505	"2010/06174"
24	0,3181818182	0,3888888889	0,06866165	"1990/01146"
27	0,3181818182	0,3684210526	0,06115186	"2009/13946"
28	0,3181818182	0,35	0,06053801	"2011/04923"

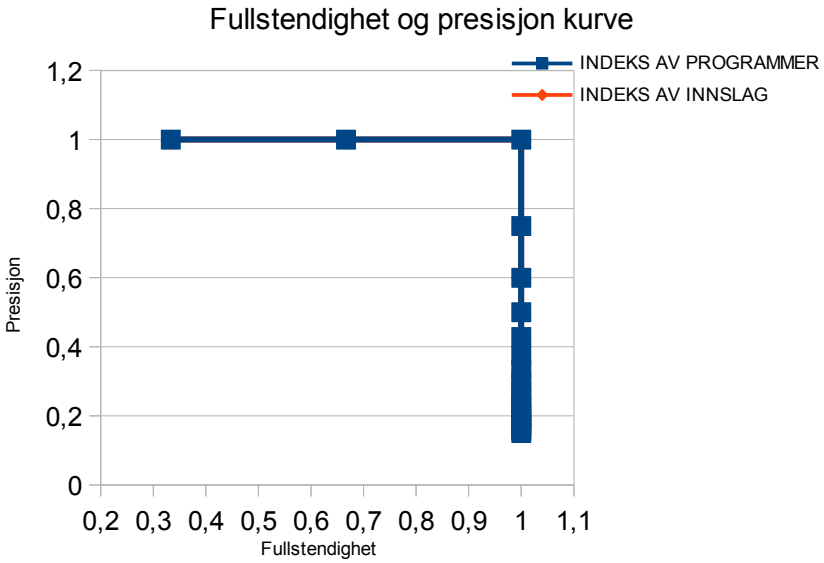


Fasit representert ved nøkkel
"2012/03679"
"2012/03858"
"2012/03987"
"2012/01597"
"2011/09312"
"2011/09110"
"2011/08240"
"2011/08114"
"2011/07894"
"2011/07801"
"2011/07539"
"2012/04237"
"2013/01573"
"2012/07277"
"2012/06962"
"2012/06053"
*"2012/05480"
"2012/05224"
"2012/05152"
"2012/04889"
"2012/04334"
"2012/04669"

Søkespørsmål nr 121: "Borgerkrigen i Libya: Under slaget om Tripoli inntar opprørerne det meste av Tripoli og Gaddafis bunker"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,3333333333	1	0,1576817	"2011/08338"
2	0,6666666667	1	0,09480086	"2011/10627"
3	1	1	0,04884853	"2011/02148"
4	1	0,75	0,040651817	"1994/16428"
5	1	0,6	0,038835686	"2010/10766"
6	1	0,5	0,027913446	"1992/01273"
7	1	0,4285714286	0,027913446	"2000/00280"
8	1	0,375	0,024124878	"2012/08756"
9	1	0,3333333333	0,023229612	"2007/20956"
10	1	0,3	0,019731358	"2011/02356"
11	1	0,2727272727	0,017058864	"2009/01152"
12	1	0,25	0,015558598	"1999/01482"
13	1	0,2307692308	0,014621884	"2012/07893"
14	1	0,2142857143	0,014621884	"2009/00637"
15	1	0,2	0,014378027	"2007/14407"
16	1	0,1875	0,014075228	"2003/15056"
17	1	0,1764705882	0,013434763	"2004/19362"
18	1	0,1666666667	0,013034952	"1993/01227"
19	1	0,1578947368	0,012184904	"1997/04542"
20	1	0,15	0,012184904	"2007/18928"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0,3333333333	1	0,37443182	"2011/08338"
2	0,6666666667	1	0,13223776	"2011/10627"
3	1	1	0,13223776	"2011/02148"
4	1	0,75	0,07140385	"1994/16428"
5	1	0,6	0,046956588	"2013/04383"
6	1	0,5	0,044271097	"2010/10766"
7	1	0,4285714286	0,040836222	"2008/00237"
8	1	0,375	0,038816378	"2007/20956"
9	1	0,3333333333	0,03305944	"1992/01273"
10	1	0,3	0,03305944	"2000/00280"
11	1	0,2727272727	0,032668978	"2009/01152"
12	1	0,25	0,028585358	"2013/0718"
13	1	0,2307692308	0,027391344	"1997/05344"
14	1	0,2142857143	0,024501733	"2009/13347"
15	1	0,2	0,024501733	"2010/07788"
16	1	0,1764705882	0,021138763	"2007/14407"
17	1	0,1875	0,021138763	"2011/10889"
19	1	0,1578947368	0,020418111	"2012/08756"
20	1	0,1666666667	0,020418111	"2003/15056"
21	1	0,15	0,020418111	"2011/02356"



Fasit representert ved nøkkel
"2011/10627"
"2011/02148"
"2011/08338"

Søkespørsmål nr 129: "Flommen på Østlandet 2013: Deler av tettstedet Kvam i Gudbrandsdalen ble ødelagt, og veier og jernbaner ble stengt"

INDEKS AV PROGRAMMER				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0	0	0,05810025	"1996/00689"
2	0	0	0,052300517	"2007/11935"
3	1	0,3333333333	0,05205017	"2013/04383"
4	1	0,25	0,047571093	"2011/06673"
5	1	0,2	0,03824349	"1990/02179"
6	1	0,1666666667	0,037859123	"1990/05946"
7	1	0,1428571429	0,037276153	"2008/00546"
8	1	0,125	0,031950988	"1995/10233"
9	1	0,1111111111	0,030674746	"1990/07506"
10	1	0,1	0,029301086	"1994/00786"
11	1	0,0909090909	0,023100035	"2007/28876"
12	1	0,0833333333	0,018618284	"1992/01275"
13	1	0,0769230769	0,015975494	"2007/24220"
14	1	0,0714285714	0,014838745	"2010/08235"
15	1	0,0666666667	0,014757476	"2008/12207"
16	1	0,0625	0,014051556	"1995/00285"
17	1	0,0588235294	0,013990955	"1996/09968"
18	1	0,0555555556	0,013990955	"2012/08756"
19	1	0,05	0,013104682	"1998/08148"
20	1	0,0526315789	0,013104682	"2009/12729"

INDEKS AV INNSLAG				
Treffnummer	Fullstendighet	Presisjon	Score	Nøkkel
1	0	0	0,17527737	"1996/00689"
2	1	0,5	0,17138731	"2013/04383"
4	1	0,3333333333	0,13321432	"2008/00546"
5	1	0,25	0,12895958	"1995/10233"
6	1	0,2	0,1136982	"2011/06673"
7	1	0,1666666667	0,09929535	"1994/00786"
9	1	0,1428571429	0,07127715	"2007/28876"
11	1	0,125	0,05702172	"2013/00289"
12	1	0,1111111111	0,050400555	"2007/11935"
13	1	0,1	0,047526684	"2010/04086"
15	1	0,0909090909	0,044624176	"1990/02179"
16	1	0,0833333333	0,04417568	"1990/05946"
17	1	0,0769230769	0,04417568	"1998/00123"
18	1	0,0666666667	0,04412325	"2008/12207"
19	1	0,0714285714	0,04412325	"2010/03134"
20	1	0,0625	0,0417729	"2013/05326"
21	1	0,0588235294	0,03960557	"1998/08148"
22	1	0,0555555556	0,03786487	"2010/03845"
23	1	0,0526315789	0,037497528	"2012/08756"
25	1	0,05	0,03645481	"2005/14420"

Fasit representert ved nøkkel
"2013/04383"