NTNU
Kunnskap for en bedre verden

KANDIDAT
10013

PRØVE
TDT4117 1 Informasjonsgjenfinning

Emnekode	TDT4117
Vurderingsform	Skriftlig eksamen
Starttid	09.12.2019 08:00
Sluttid	09.12.2019 12:00
Sensurfrist	09.01.2020 22:59
PDF opprettet	05.08.2020 11:41

Skriftlig eksamen i TDT4117

Oppgave	Tittel	Oppgavetype
i	EKSAMENSOPPGAVE I FAG TDT4117	Dokument
1	Oppgave 1 (10%)	Langsvar
2	Oppgave 2 (10%)	Langsvar
3	Oppgave 3 (20%)	Langsvar
4	Oppgave 4 (10%)	Langsvar
5	Oppgave 5 (20%)	Langsvar
6	Oppgave 6 (30%)	Flervalg

1 Oppgave 1 (10%)

1. Hvorfor er ”index terms” viktig i informasjonsgjenfinningssammenheng? Hva er de viktigste kriteriene for valg av indekstermer. Forklar. (4%)
2. Tegn et blokkdiagram (med firkanter og piler) som forklarer hvordan informasjonsgjenfinningsprosessen er bygd opp. Tips: Dette er ikke tekstoperasjoner. (3%)
3. Gitt følgende utsagn:
   «*Bruken av søkefunksjonen i Netflix kan karakteriseres til å være både informasjonsgjenfinning og datagjenfinning*».
   Forklar kort hvorfor dette utsagnet er sant. (3%)

Skriv ditt svar her...

1. Index terms er et utvalg ord representative for dets dokumentet. Disse er viktige i IR fordi de tillater god estimering av relevans og "similarity". Ved å koke ned dokumentene til sine index termer kan vi effektivt se gjennom disse istedet for å søke i hele dokumentet. Ved å koke ned teksten "Nuclear holocaust is imminent" til "nuclear, holocaust, imminent" sørger vi for at den er mer relevant om man søker etter "Nuclear Holocaust" enn om man søker "Is the holocaust real?" ved å fjerne irrelevante ord som "is" og bare holder de relevante termene. De viktigste kriteriene for valg av index termer er at de er representative for innholdet av dokumentet og at de holder en klar, entydig mening

2. See attachment: The IR process is the user doing retrieval or browsing,

which extracts information from the DB. These two activities are distinct and are not done together. Retrieval is getting the best match for a given search whilst browsing is a process of floating from subject to subject. For example a search for Formula 1 is retrieval, but if the user is interested it can lead to browsing related subjects, as a user might want to read about specific cars in the race, or where the race is held which might lead to further browsing on tourism of said country. Then this can lead to retrieval, for example if the user decides to visit said country he can now do retrieval of Flights to this country.

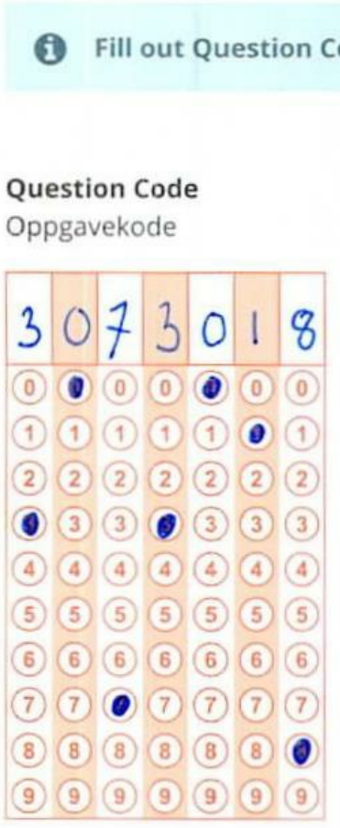
3. Det er informasjonsgjenfinning i den forstand at søket kan ha partiell match og best match. For eksempel hvis man søker "shrek 3", retrieves Shrek 2 fordi Netflix ikke har "Shrek 3" i sitt bibliotek. På den andre siden er søk på genre kun direkte match da dette ikke kan interpoleres fra innholdet i filmen. Dette er data i strukturert tabell som er manuelt innført av noen som ser filmen. Derfor om man søker "Atcion" istedenfor "Action" vil man få et søk på filmer som har tittel nært "Atcion", altså på genre vil den kun gi full matching. Mye av netflix sin data om filmene er strukturerte metadata, og søk i dette vil være datagjenfinning. Denne typen metadata er strukturert i en tabell, noe som også kjennetegner datagjenfinning.

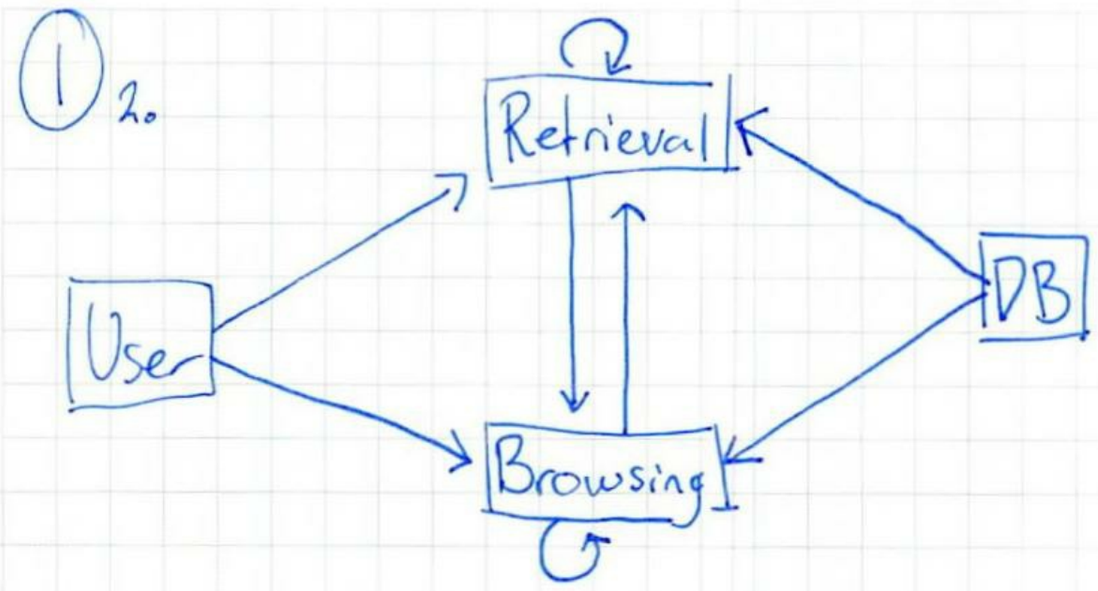
Knytte håndtegninger til denne oppgaven?
Bruk følgende kode:

3 0 7 3 0 1 8

Fill out Question Code and Test Information on every sheet. Fyll inn oppgavekode og emneinformasjon på alle skisseark.

Question Code Oppgavekode	Date Dato	Subject code Emnekode	Candidate ID KandidatID	Question nr Oppgavenr	Page number Sidetall
3073018	9.12	TDT4117	10013	1	1





2 Oppgave 2 (10%)

1. Drøft hovedforskjellene mellom multimedia og tekstgjenfinning. (2%)
2. Innen multimedia er begrepet «**features**» brukt. Hva menes med dette begrepet? Gi eksempler på tre forskjellige features som er brukt i forbindelse med bildegjenfinning. (3%)
3. Tegn opp en ***taksonomi*** (taxonomy) over multimedia datamodellen. Tips: Modellen er delt opp i flere lag. Det forventes at du gir minst et eksempel på multimedia objekttype for hvert lag. (5%)

Skriv ditt svar her...

1. I multimedia gjenfinning så har vi mye mer ustrukturert data, og data som er mye vanskeligere å indeksere på. Dette betyr at vi må definere multimediet basert på andre ting, og det er der features kommer inn. I tillegg er det mye større "dokumenter" som har mye mindre endring mellom en del til den neste(piksel til piksel, og frame til frame), dette betyr at komprimeringsmetodene spiller en stor rolle.

2. Features er "index termene" av multimediegjenfinning. De er distinktive karakteristikker ved multimediet som vi kan bruke til å til å finne similaritet/likhet mellom bilder.

Examples of features used in relation to imageretrieval: region-based binary representation of shape, colour histogram, texture matching.

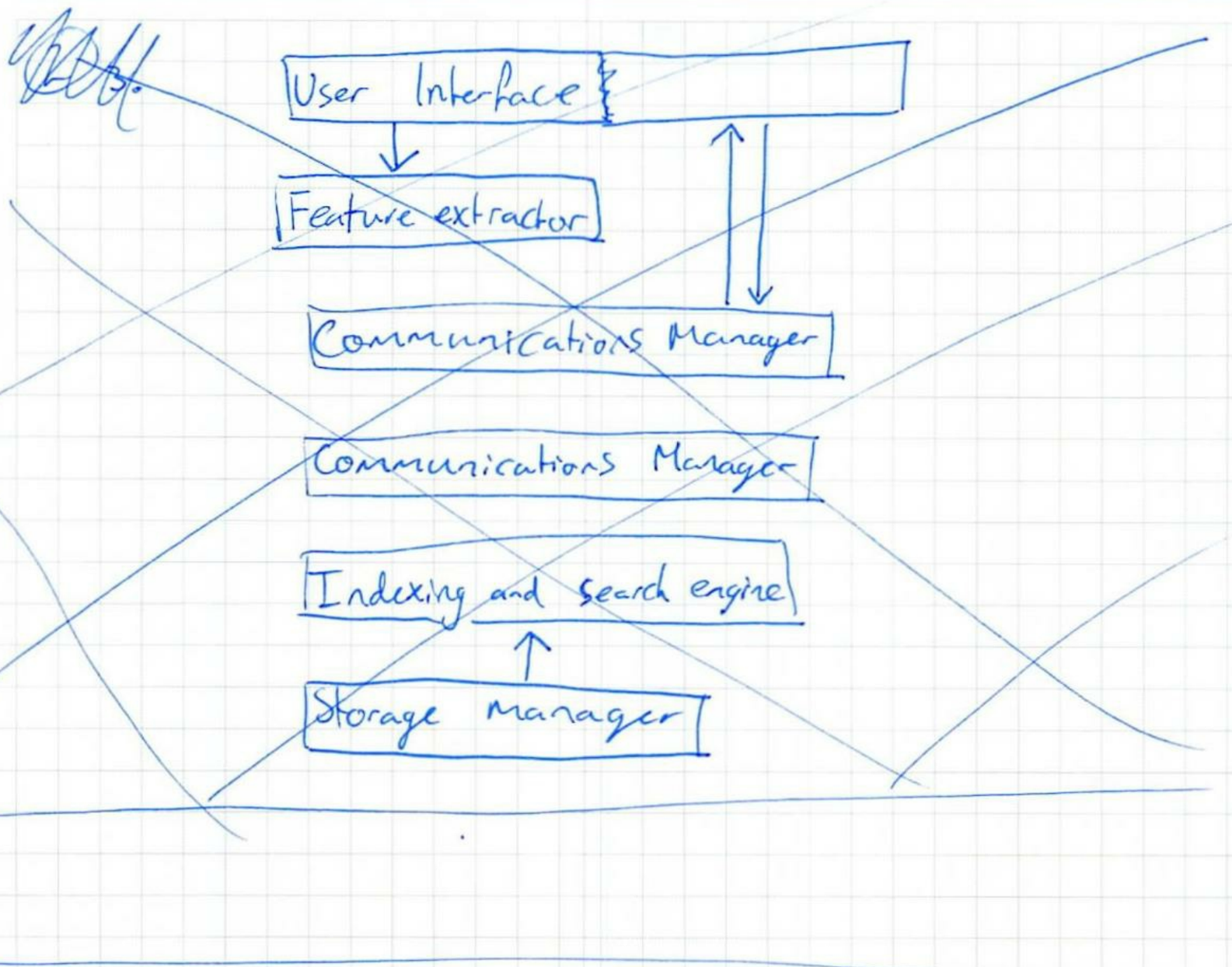
3. se vedlegg

Knytte håndtegninger til denne oppgaven?
Bruk følgende kode:

5 2 1 2 7 4 0

Fill out Question Code and Test Information on every sheet. Fyll inn oppgavekode og emneinformasjon på alle skisseark.

Question Code Oppgavekode	Date Dato	Subject code Emnekode	Candidate ID KandidatID	Question nr Oppgavenr	Page number Sidetall
5 2 1 2 7 4 0	9.12	TDT4117	10013	2	2



3 Oppgave 3 (20%)

1. Gitt følgende tekst:
   «*Competition enforcers on both sides of the Atlantic are now looking into how dominant tech companies use and monetise data*».

   Gjør de antakelsene du finner nødvendige og svar på følgende spørsmål:

   a. Kjør **leksikal analyse** og **fjerning av stoppord** etter å ha forklart hva disse går ut på. Hvilket resultat får man? (3%)
   b. Tegn opp et «**vocabulary trie**» basert på teksten over. (5%)
   c. Bruk teksten over til å forklare prinsippet bak **inverterte filer/invertert indeks** (inverted files/inverted index). (6%)

2. Anta følgende tekst:
   «*EU antitrust regulators say they are investigating Google's data collection*».

   Gitt videre følgende hashkoder for *eu, antitrust, regualtors, say, investigating, google, data, collection*:
   f(eu) = 100001
   f(antitrust) =100010
   f(regualtors) = 100011
   f(say) = 100110
   f(investigating) = 100111
   f(google) = 101110
   f(data) = 101111
   f(collection) = 110010

   Vi skal bruke metoden ***signaturfil*** til å indeksere teksten vår. Forklar hvordan du vil gå frem. Velg en blokkstørrelse på 3 og gjør ellers de antakelsene du finner nødvendig for å løse oppgaven. (6%)

Skriv ditt svar her...

1. Lexical analysis: handling all characters into indexable types. This means removing punctuation, turning digits into characters, removing hyphens, setting uniform case for all characters, and so on.

Stopword removal: Removing unimportant words, words which do not play any part in the meaning of the document, e.g. articles and conjunctions.

After Lexical analysis and Stopword removal:

competition enforcers sides atlantic looking dominant tech companies monetise data

b) see attachment

c) Inverted files consists of a vocabulary and occurrances. The occurrances point to the place of the occurance of a word. If i were to search for the "*atlantic*" in this inverted file, i would lookup in vocabulary and see the occurance of it in 44. To create an inverted file of

the given document, we can use the vocabulary trie and add a term list with occurrances. This would be a complete inverted file. The main feature of this is that we an use a direct access to the occurrance and get what we need, including if we need a proximity search.

2: We first devide into block of similar size(+/- 1). To get the signature file we OR the hashed index terms of each block. Each block is now a binary string which we can lookup if a word is contained in the block, barring any false drops.

First we divide into blocks:

Block 1: *EU antitrust regulators*

Block 2: *say they are*

Block 3: *investigating Google's*

Block 4*: data collection*

The we OR the hash of the index terms in each block:

Block 1: *100001* OR *100010* OR *100011 = 100011*

Block 2: *100110 = 100110*

Block 3: *100111* OR *101110 = 101111*

Block 4*: 101111* OR *110010 = 111111*

We have now made the signature file.

We can clearly see the pitfall of Signature files in block 4, as we can now falsely believe that Block 4 contains any of the terms as it is 1 in each corresponding bit to all terms.

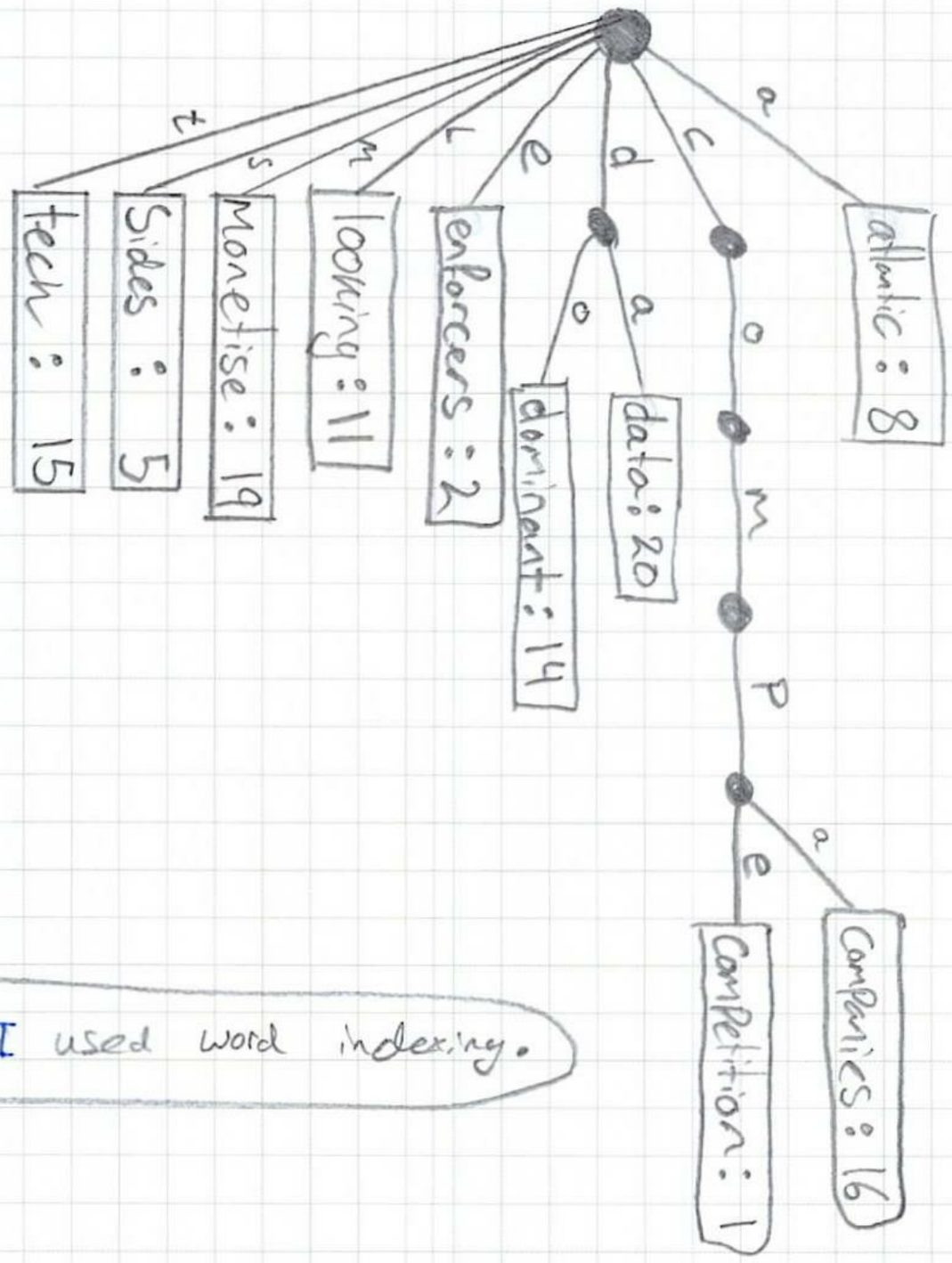
Knytte håndtegninger til denne oppgaven?

3 8 0 3 4 4 3

Bruk følgende kode:

Fill out Question Code and Test Information on every sheet. Fyll inn oppgavekode og emneinformasjon på alle skisseark.

Question Code Oppgavekode	Date Dato	Subject code Emnekode	Candidate ID KandidatID	Question nr Oppgavenr	Page number Sidetall
3803443	9.12	TDT4117	10013	3	3



PS: For simplicity, I used word indexing.

4 Oppgave 4 (10%)

1. Bruk hovedprinsippet bak modellene til å sammenlikne språkmodellen (language model) og vektorbasert (vector space model) similaritetsmodell. Hvilken modell ville du foretrekke dersom du skulle lage et tekstgjenfinningssystem selv. Tips: Fokuser på styrke og svakheter med hver av modellene til å hjelpe med forklaringen din. (6%)
2. Hva er hovedforskjellen mellom sannsynlighetsmodellen (probabilistic model) og språkmodellen (language model). (4%)

Skriv ditt svar her...

1. Whilst the language model concerns itself with finding relevance based on probability that the generated model of a document will produce the query, the vector space model compares similarity of occurrances of keywords and represents the similarity of the query and a document based on the angle between the two in a vector space based on the occurrances on index words. The key difference is that the language model estimates the relevance, whilst the VSM estimates the similarity of a document and the query.

They both allow for ranking and partial matching. VSM uses term frequency and inverse document frequency, whilst language model only uses term frequency and not IDF in the standard model.Whilst VSM is quite computationally simple and intuitive, the language model can become computationally slow when the index term list is long and we have many documents, especially if one implements smoothing which is important in allowing for partiap matching. VSM does however have the drawback of assuming that the relevance of a document is independent of the relevance of other documents.

Personally i would choose language model if i were to implement my own IR system. Since I am making it the collection is probably small, and the documents are not that long. This means that the computational drawback is not as heavy, and the unnatural assumption of independant relevance is a far larger drawback than any of the language model.

2. The main difference is the way the probability is calculated. The probabilistic model calculates the probability that a document is relevant based on a initial relevance estimate, however the language model calculates the probability of relevance based on the chance of the query being produced by the language model of the document.

Knytte håndtegninger til denne oppgaven?
Bruk følgende kode:

0 7 4 3 1 6 5

5 Oppgave 5 (20%)

1. Hva menes med *Mean Average Precision (MAP*)? (2%)
2. Gitt at du får 20 returnerte resultater fra en spørring, som basert på en enkel evaluering har følgende relevante treff (numrene angir plassering i resultatlista): 1, 3, 4, 7, 8 10, 15. Anta videre at det er i alt 8 relevante dokumenter for denne spørringen.
   a. Hva er *precision* og *recall* for denne spørringen? (2%)
   b. Hva er *harmonic mean/f-measure* for denne spørringen? (2%)
   c. Lag en tabell som viser *precision- og recall-punkt (points)* for denne spørringen. (4%)
   d. Hva blir *R-precision*? (2%)
   e. Tegn opp grafen som viser de interpolerte verdiene av precisions. (8%)

Skriv ditt svar her...

1. MAP er et evalueringsmål som regner ut snittet av snittene av presisjoner hver gang et relevant dokument blir gjenfunnet. Dette er et mål over flere dokumenter og er en slags tilnærming ar arealet under presicion-recall kurven, der jo større, jo bedre.

2.

Presicion = |Relevant ∩ retrieved| / |retrieved| = 7/20 = 0.35

Recall= |Relevant ∩ retrieved|/|Relevant| = 7/8= 0.875

b)F-measure er den harmoniske gjennomsnitt av Precision of recall. Harmonisk snitt fungerer slik at det straffer stor varians mellom parameterne, slik at høy precision og lav recall er dårligere enn middels av begge. I dette tilfellet har vi:

F-measure = 2 / ((1/R) + (1/P)) = 2 / ((8/7) + (20/7)) = 1/2

c)

DocNo.	Relevant?	Precision	Recall
1	R	1	1/8
2			
3	R	2/3	2/8
4	R	3/4	3/8
5			
6			

7	R	4/7	4/8
8	R	5/8	5/8
9			
10	R	6/10	6/8
11			
12			
13			
14			
15	R	7/15	7/8
16			
17			
18			
19			
20			

d) R-presicion er presisjonen ved det R-te dokumentet der R er antallet relevante dokument. R-precision i dette tilfellet er presisjon ved det åttende dokumentet. Presisjonen er da 5/8 = 0.625

e) se vedlegg

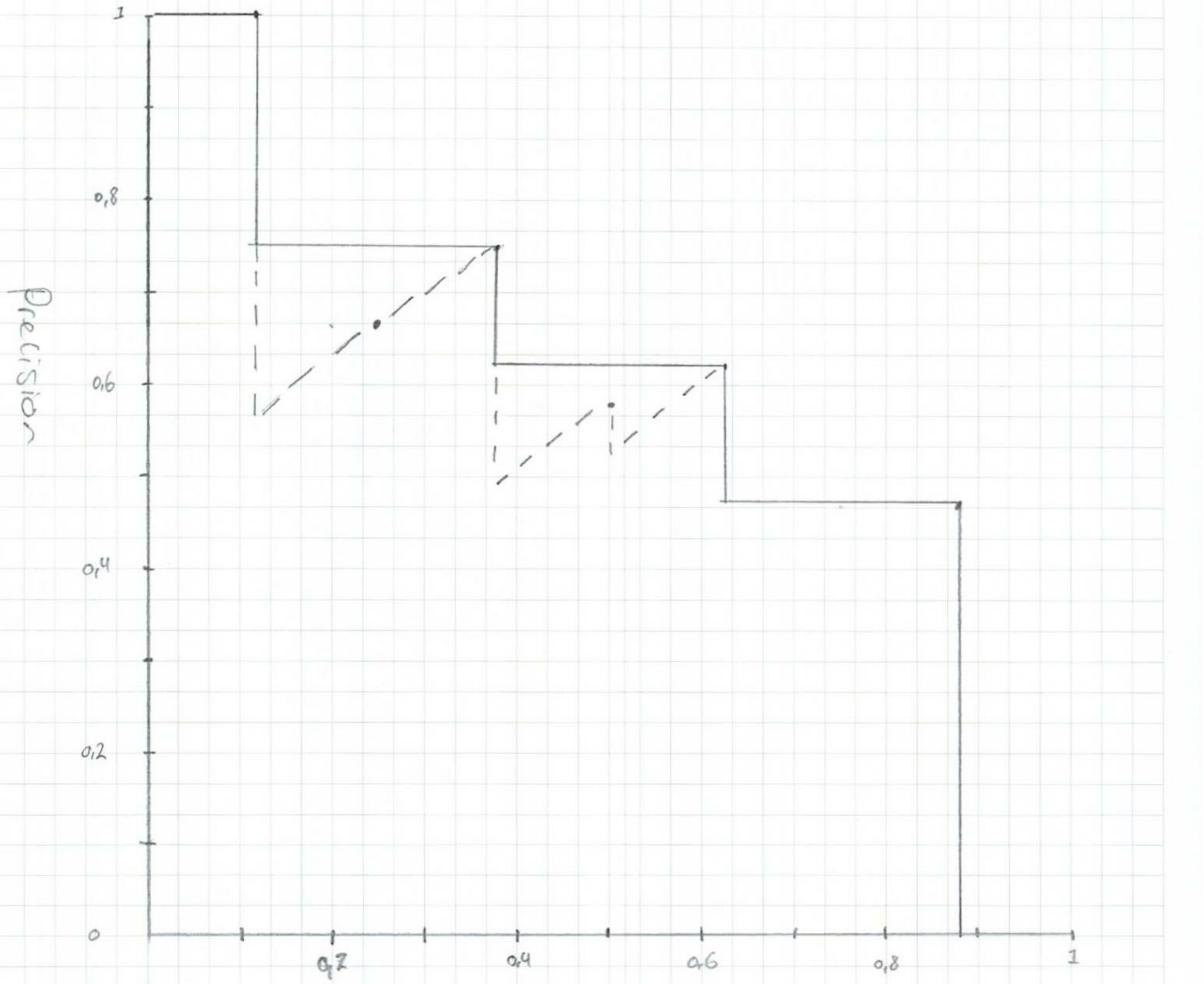
Knytte håndtegninger til denne oppgaven?
Bruk følgende kode:

8 4 2 2 4 0 2

Fill out Question Code and Test Information on every sheet. Fyll inn oppgavekode og emneinformasjon på alle skisseark.

Question Code Oppgavekode	Date Dato	Subject code Emnekode	Candidate ID KandidatID	Question nr Oppgavenr	Page number Sidetall
8 4 2 2 4 0 2	9.12	TDT4117	10013	5	4

5e



6 Oppgave 6 (30%)

Velg et av alternative i hver deloppgave. Dersom du mener flere av utsagnene i hver oppgave er riktig, velg det du mener det mest riktige.

Velg ett alternativ

- Fargehistogram kan brukes til å finne similaritet/likhet mellom bilder og dermed kan det brukes som features.
- Fargehistogram gir statistisk informasjon om piksler i et bilde og derfor er det godt egnet til å lage en gjenfinningsvennlig komprimeringsmetode.
- Fargehistogram kan ikke brukes i forbindelse med gjenfinning av bilder fordi det bare gir informasjon om pikselfordeling i bilder.
- Fargehistogram kan ikke brukes til å finne similaritet/likhet mellom bilder og dermed kan det heller ikke brukes som features.

Velg ett alternativ

- Websøkesystemer bruker ”stemming” fordi stemming bidrar generelt til økt kapasitet til å lagre websidene lokalt.
- Websøkesystemer bruker ikke ”stemming” fordi stemming ikke passer til web-søk generelt fordi selv om det bidrar til økt recall bidrar det ikke nødvendigvis til økt precision.
- Websøkesystemer bruker ”stemming” fordi selv om det koster erkjenner man at man får økt recall og er stemming derfor veldig viktig.
- Websøkesystemer bruker ”stemming” fordi stemming bidrar generelt til økt hastighet for web crawlere.

Velg ett alternativ

- Sentraliserte web-søkemotorer er søkemotorer med ”Harvest”-arkitektur som består av en server og flere crawlere.
- Søkemotorer med ”Harvest”-arkitektur er en variant av sentralisert web-søkemotorarkitektur.
- Søkemotorer med ”Harvest”-arkitektur er en variant av distribuert web-søkemotorarkitektur. .
- Søkemotorer med crawlere har samme arkitektur som de med «brokers» og «gatherers»

Velg ett alternativ

- Thesaurus-bygging er naturlig del i automatisk global analyse (automatic global analysis), og bruker hele dokumentsamlingen til å gjøre dette.
- Thesaurus-bygging er naturlig del i automatisk lokal analyse (automatic local analysis), og bruker hele dokumentsamlingen til å gjøre dette.
- Thesaurus-bygging er naturlig del i både automatisk lokal analyse (automatic local analysis) og automatisk global analyse (automatic global analysis), og begge bruker hele dokumentsamlingen til å gjøre dette.
- Thesaurus-bygging er naturlig del i automatisk global analyse (automatic global analysis), og bruker de returnerte dokumentene fra et søk til å gjøre dette.

Velg ett alternativ

- Både «Language Model" og «Okapi BM25» bruker sannsynlighet for relevans til å rangere resultater fra en spørring.
- De største likhetene mellom «Language Model" og «Okapi BM25» er hvordan TF og IDF blir brukt til å estimere sannsynlighet.
- Den største likheten mellom «Language Model" og «Okapi BM25» er at ingen av dem bruker TF eller IDF til å estimere sannsynlighet.
- Den største forskjellen mellom «Language Model" og «Okapi BM25» er måten sannsynligheten blir beregnet på.

Velg ett alternativ

- "Vocabulary Trie" og "Suffix Trie" er to begrep som ikke har noe med indeksering å gjøre men en tre basert tekstkomprimering.
- "Vocabulary Trie" og "Suffix Trie" er to begrep om som beskriver to forskjellige indekskomprimeringsmetoder.
- "Vocabulary Trie" og "Suffix Trie" er to begrep som brukes i to forskjellige indekseringsmetoder.
- "Vocabulary Trie" og "Suffix Trie" er to begrep som brukes i forbindelse med en og samme type indekseringsmetode.

Velg ett alternativ

- Stemming har generelt positive påvirkninger på recall, mens fjerning av stoppord har positive påvirkninger på precision.
- Både fjerning av stoppord og stemming har negative påvirkninger på Recall.
- Hverken fjerning av stoppord eller stemming har negative påvirkninger på precision.
- Både fjerning av stoppord og stemming har generelt negative påvirkninger på precision.

Velg ett alternativ

- MRR (Mean Reciprocal Rank) er veldig godt egnet til evaluere systemer der man mest er opptatt av å finne relevante resultater i en topp-k (feks. topp-10) resultatliste.
- MRR (Multimedia Retrieval Ranking) er godt egnet som rangeringsmetode for bilder.
- MRR (Machine-based Result Ranking) er en vektorbasert metode for rangering.
- MRR (Mean Reciprocal Rank) er en annen variant av MAP (Mean Average Precision).

Velg ett alternativ

- User Relevance Feedback (URF) bruker brukerens tilbakemelding kombinert med feks. Rochio's standard metode til å bestemme en forbedret spørring.
- User Relevance Feedback (URF) bruker brukerens tilbakemelding til å bestemme hastigheten på returnering av søkeresultater.
- User Relevance Feedback (URF) er ofte brukt til å redefinere spørringer slik at man får økt søkehastighet.
- User Relevance Feedback (URF) er sterkt avhengig av Rochio's standard metode alene for å produsere gode søkeresultater.

Velg ett alternativ

- HITS og PageRank bruker ikke de samme prinsippene for websøk. Mens HITS bruker linkinformasjon fra hele dokumentsamlingen, bruker PageRank nøkkelordvekt (index term weights) som basis for rangering av søkeresultatene.
- HITS og Page Rank bruker ikke de samme prinsippene for websøk. Mens PageRank bruker linkinformasjon fra hele dokumentsamlingen, bruker HITS nøkkelordvekt (index term weights) som basis for rangering av søkeresultatene.
- HITS og PageRank gjør akkurat de samme nyttene for websøk, men bruker forskjellige prinsipp for rangering. Mens PageRank bruker de returnerte søkeresultatene, bruker HITS hele dokumentsamlingen.
- HITS og PageRank gjør akkurat de samme nyttene for websøk, men bruker forskjellige prinsipp for rangering. Mens PageRank bruker hele samlingen av websider, bruker HITS de returnerte søkeresultatene.

Knytte håndtegninger til denne oppgaven?
Bruk følgende kode:

2 3 2 6 0 5 1