TDT4117 - Assignment 2

Sander Lindberg

Task 1 - Relevance Feedback

1

Explain the difference between automatic local analysis and automatic global analysis.

Begge analyserer dokumenter og bygger en ”thesaurus” ut i fra de, men forskjellen er at local analysis først henter alle relevatne dokumenter basert på den opprinnelige querien, og deretter bygger en ”thesaurus” ut i fra disse. Global analysis bygger ut ifra alle dokumentene.

2

What is the purpose of relevance feedback? Explain the terms Query Expansion and Term Re-weighting. What separates the two?

Hele poenget med relevance feedback er å finne relevante dokumenter basert på brukerfeedback. En bruker gir en query, velger de relevante dokumentene og ”vi” finner dokumenter som er like de brukeren plukket ut som relevante.

Query Expansion handler om å omformulere den opprinnelige querien slik at vi forbedrer relevansen til dokumentene som returneres. *Term Re-weighting* handler om å gjøre vekten på urelevante termer mindre og større på relevante. Forskjellen mellom de to er at *Term Re-weighting* ikke utvider den opprinnelige querien.

Task 2 - Language Model

1

Explain the language model, what are the weaknesses and strengths of this model?

En ”language model” brukes for å forenkle spørringer etter ord i dokumenter. Dette brukes videre til å rangere dokumentene etter relevans. Modellen behandler hvert dokument som grunnlag for en modell. Altså to forskjellige dokumenter gir forskjellige modeller. Denne modellen inneholder sannsynlighetene for de forskjellige termene.

Generelt har vi at $\sum_{\text{term in doc}} \text{P(term)} = 1$. Positive aspekter ved denne modellen er at den er veldig intuitiv og enkel. Negative aspekter er at den for eksempel ikke tar hensyn til fraser

som "*To be or not to be*". Den vil da rangere dokumenter basert på termer, i stedet for hele frasen som brukeren mest sannsynlig mente.

2

Given the following documents and queries, build the language model according to the document collection

d_1 = *failure is the opportunity to begin again more intelligently.*

d_2 = *intelligence is the ability to adapt to change.*

d_3 = *lack of will power leads to more failure than lack of intelligence or ability*

q_1 = *failure*

q_2 = *intelligence opportunity*

q_3 = *intelligence failure*

Use MLE for estimating the unigram model and estimate the query generation probability using the Jelinek-Mercer smoothing

$$\hat{P}(t \mid M_d) = (1-\lambda)\hat{p}_{mle}(t \mid M_d) + \lambda \hat{p}_{mle}(t \mid C), \quad \lambda = 0.5 \tag{1}$$

For each query, rank the documents using the generated scores.

d_1 = 9 termer

d_2 = 8 termer

d_3 = 14 termer

Totalt = 31 termer

t/P	$\hat{P}_{mle}(t \mid M_{d1})$	$\hat{P}_{mle}(t \mid M_{d2})$	$\hat{P}_{mle}(t \mid M_{d3})$	$\hat{P}_{mle}(t \mid C)$
Intelligence	0.0	0.125	0.0714	0.0645
Failure	0.11	0.0	0.0714	0.0645
Opportunity	0.11	0.0	0.0	0.0322

Tabell 1: Sannsynligheter basert på term i query

$\hat{P}_{mle}(t \mid M_d)$ er kalkulert ved å telle forekomst av term i gitt dokument, delt på antall termer i dokumentet. F.eks sannsynligheten for intelligence i d_1 er gitt ved $\hat{P}_{mle}(\text{intelligence} \mid d_1) = \frac{0}{9} = 0.0$.

$\hat{P}_{mle}(t \mid C)$ er kalkulert ved å telle antall forekomster av termen i hele kolleksjonen av dokumenter, delt på antall termer i kolleksjonen. F.eks $\hat{P}_{mle}(\text{failure} \mid C) = \frac{2}{31} = 0.0645$

Nedenfor er en tabell som viser verdiene for $\hat{P}(t \mid M_d)$ ved bruk av formelen (1) over. Her har jeg bare ganget inn λ og lagt sammen leddene. F.eks:

$$\hat{P}(\text{intelligence} \mid M_{d1}) = (1-0.5) \cdot 0.0 + 0.5 \cdot 0.0645 = 0.5 \cdot (0.0 + 0.0645) = 0.03225$$

t/P	$\hat{P}(t \mid M_{d_1})$	$\hat{P}(t \mid M_{d_2})$	$\hat{P}(t \mid M_{d_3})$
Intelligence	0.03225	0.09475	0.06795
Failure	0.08725	0.03225	0.06795
Opportunity	0.0711	0.0161	0.0161

Tabell 2: $\hat{P}(t \mid M_d)$

Bruker nå formelen

$$\hat{P}(Q \mid M_d) = \prod_{t \in Q} \hat{P}(t \mid M_d) \tag{2}$$

for q_1, q_2 og q_3, ved hjelp av Tabell 2. Eksempel på utregning med q_2 og d_1:

$$\begin{aligned} \hat{P}(q_2 \mid M_{d_1}) &= \hat{P}(\text{intelligence} \mid d_1) * \hat{P}(\text{opportunity} \mid d_1) \\ &= 0.03225 * 0.0711 \\ &\approx 0.0023 \end{aligned}$$

Verdiene er oppsummert i Tabell 3

q/d	d_1	d_2	d_3
q_1	0.08725	0.03225	0.06795
q_2	0.0023	0.001525	0.0011
q_3	0.0028	0.00304	0.0046

Tabell 3: $\hat{P}(Q \mid M_d)$

Utifra tabellen ser vi at rangeringen for de forskjellige queriene er som følger:

$$q_1 = d_1 > d_3 > d_2$$

$$q_2 = d_1 > d_2 > d_3$$

$$q_3 = d_3 > d_2 > d_1$$

3

Explain what smoothing means and how it affects retrieval scores. Describe your answer using a query from the previous subtask

Smoothing gjøes for å få vekk null-verdier. Det gjøres ofte ved å legge til en λ gjerne lik 1 eller $\frac{1}{2}$. Med *Jelinek-Mercer smoothing* som er brukt i forrige deloppgave gjøres det ved å legge til sannsynligheten for at termen er i hele kolleksjonen ganget med lambda. Hvis termen ikke finnes i kolleksjonen i det hele tatt, vil vi fortsatt få en null verdi, men det var ikke tilfelle i denne oppgaven.

F.eks q_3 og d_2 ville vi fått:

$$\prod_{t \in Q} \hat{P}(t \mid M_{d2}) = 0.0125 * 0.0$$

$$= 0.0$$

uten denne smoothingen.

Task 3 - Evaluation of IR Systems

1

Explain the terms Precision and Recall, including their formulas. Describe how differently these metrics can evaluate the retrieval quality of an IR system.

Precision er definert som $\frac{\text{Relevant og Hentet}}{\text{Hentet}}$, mens recall er definert som $\frac{\text{Relevant og Hentet}}{\text{Totalt relevante}}$.

Precision kan alstå tenkes på som ”hvor nyttig er søkeresultatene?”, mens recall kan tenkes på som ”hvor komplett er søkeresultatene?”.

Formlene er utledet fra tabellen:

	Relevant	Ikke-relevant
Hentet	True positive (TP)	False positive (FP)
Ikke-hentet	False positive (FN)	True negatives (TN)

Fra tabellen ser vi precision = $\frac{\text{TP}}{\text{TP+FP}}$ og recall = $\frac{\text{TP}}{\text{TP+FN}}$

2

Given the following set of relevant documents Rel = $\{82, 21, 45, 271, 72, 300, 94, 56, 88, 150\}$**, and the set of retrieved documents ret** = $\{91, 21, 45, 56, 82, 221, 72, 215\}$**, provide a table with the calculated precision and recall at each level.**

DocID	Relevant?	Precision	Recall
91	Nei	$\frac{0}{1}$	$\frac{0}{10}$
21	Ja	$\frac{1}{2}$	$\frac{1}{10}$
45	Ja	$\frac{2}{3}$	$\frac{2}{10}$
56	Ja	$\frac{3}{4}$	$\frac{3}{10}$
82	Ja	$\frac{4}{5}$	$\frac{4}{10}$
221	Nei	$\frac{4}{6}$	$\frac{4}{10}$
72	Ja	$\frac{5}{7}$	$\frac{5}{10}$
215	Nei	$\frac{5}{8}$	$\frac{5}{10}$

Tabell 4: Racall og precision

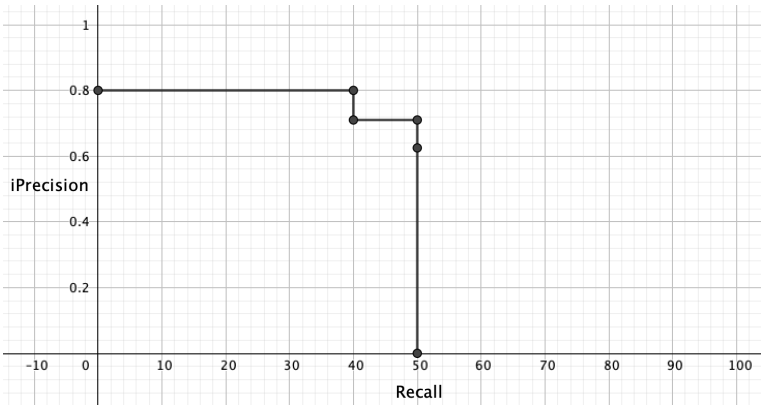
Task 4 - Interpolated Precision

1

What is interpolated precision?

Interpolated precision er gjennomsnittlig precision ved fikserte verdier av recall.

2



Figur 1: Interpolated Precision graph