TDT4117 Information Retrieval
Assignment 2

Erling Storaker Moen

October 2018

1 Relevance Feedback

1.1

Automatisk lokalt søk bruker dokumentene fra søkeresultatet til å utvide søket. Automatisk globalt søk bruker i motsetning alle dokumentene i samlingen til søket.

1.2

Relevanse-tilbakemelding er en måte å få tilbakemelding fra brukere på et søkeresultat. Deretter kan man bruke dette til å formulere det bedre og få et bedre søkeresultat ved neste spørring.

Spørringsutvidelse er å utvide brukerens spørring for å finne flere relevante dokumenter. Dette kan gjøres ved f.eks. å legge til noen synonymer.

Term Re-weighting øker eller senker vekten på termen i dokumenter for å gjøre resultatene mer relevante.

2 Language Model

2.1

Language modellen definerer en språkmodell for hvert dokument i samlingen. Modellen bruker termene i dokumentet til å kalkulere sannsynligheten for at et gitt dokument genererer en gitt spørring.

+ enkel, intuitiv for brukerne
- vanskelig å implementere relevansetilbakemelding sammen med LM.

2.2

d1 = Hurricane Irma caused major disasters in Florida.
d2 = Jose was almost the same category as Irma, also headed towards Florida.
d3 = The third hurricane was named Katia.

q1 = hurricane
q2 = hurricane florida
q3 = hurricane katrina

d1= 7 termer
d2 = 12 termer
d3 = 6 termer

Totalt = 25 termer

P^(t|Md) = (1)p^mle(t|Md) + p^mle(t|C), = 0.5
P^(t|Md) = 0,5 * (forekomster i dok/antall t i dok) + 0,5 * (forekomster i samlingen/antall t i samlingen)

q1 = hurricane :
P(q1,d1) = 0.5 * (1/7) + 0.5 * (2/25) = 0.115
P(q1,d2) = 0.5 * (0/12) + 0.5 * (2/25) = 0.04
P(q1,d3) = 0.5 * (1/6) + 0.5 * (2/25) = 0.123
d3 > d1 > d2

q2 = hurricane florida:
P(q2, d1) = 0.115 * (0.5 * (1/7) + 0.5 * (2/25)) = 0.012
P(q2, d2) = 0.04 * (0.5 * (1/12) + 0.5 * (2/25))) = 0.00326
P(q2,d3) = 0.123 * (0.5 * (0/6) + 0.5 * (2/25)) = 0.00492
d1 > d3 > d2

q3 = hurricane katrina:
P(q3, d1) = 0.115 * (0.5 * (0/7) + 0.5 * (1/25)) = 0.0023
P(q3, d2) = 0.04 * (0.5 * (0/12) + 0.5 * (1/25)) = 0.0008
P(q3,d3) = 0.123 * (0.5 * (1/6) + 0.5 * (1/25)) = 0.01271
d3 > d1 > d2

2.3

«Smoothing» brukes for å fjerne 0-verdier og for å jevne ut verdiene. Jelinek-Mercer metoden bruker en lambda verdi mellom 0 og 1. I denne øvingen har vi brukt en verdi på 0.5, som vil si at frekvensen i dokumentet og i samlingen, vektes like mye.

3 Evaluation of IR Models

3.1

Precicion er andelen av dokumentene som er relevante i et søkeresultat.

$p = \frac{|R \cap A|}{|A|}$

hvor R er settet med relevante dokumenter og A er settet med søkeresultatene

p er da en variabel som avhenger av antall relevante dokumenter og antall dokumenter totalt.

Recall er andelen av alle relevante dokumenter som blir hentet ut av søket.

$r = \frac{|R \cap A|}{|R|}$.

Relasjonen mellom disse et at de er omvendt proporsjonale. Derfor må man prioritere hvilken egenskap som er viktigst for søket. Til websøk er høy precicion egnet, mens recall brukes gjerne i forskning eller medisin hvor det er viktig å få alle relevante dokumenter med

3.2

R = 14, 2, 42, 13, 300, 5, 33, 41, 8, 10, 96, 67
A = 83, 2, 76, 42, 5, 9, 23, 33, 96, 101

Rank Level	**Recall**	**Precision**
1	0	0
2	8.33	50
3	8.33	33.3
4	16.7	50
5	25	60
6	25	50
7	25	42.9
8	33.3	50
9	41.7	55.6
10	41.7	50

4 Interpolated Precision

4.1

Interpolated Precision tar hensyn til at gjennomsnittlig, gjennom mange søk, vil precision synke når recall øker. Interpolated Precision er en måte å flate ut Precision-Recall grafen på, og å finne precision verdier der det ellers ville vært umulig. For å finne i-precision = høyeste precision til høyere recall verdi.

4.2