

TDT4117 - Øving 2

Ole-Christer Selvik, Håkon Løvdal og Kristoffer Andreas Dalby

Oktober 2013

Oppgave 1

Deloppgave a

Først la oss si hovedformålet med språkmodellen: Ideen med språkmodellen er å rangerere dokumenter etter sannsynligheten for at dokumentet inneholder/vil kunne generere søkestrengen/spørringen.

Dette fungerer ved at modellen setter en sannsynlighet for en gitt sekvens med ord (spørring) ved hjelp av sannsynlighetsfordeling. Gitt et dokument, vil vi sette opp en språkmodell for dette dokumentet. Dette vil si at hver term i dokumentet gis en vekt basert på sannsynligheten for at det søkes etter den termen i dokumentet. Videre, ved en samling med dokumenter, og en spørring Q , vil dokumentene i samlingen rangeres basert på sannsynligheten for at de vil generere termene i spørringen. Det finnes flere former for språkmodellen, men den vanligste er unigram-modellen.

Fordeler med språkmodellen

- Effektiv og rask i IR-oppgaver.
- Enkel og lett å bruke.
- Stort sett gode resultater på spørringer

Ulemper med språkmodellen

- Enkle, unigram-modeller som vanskeliggjør å inkludere brukerens tilbakemeldinger eller preferanser.
- Vanskelig å velge gode sannsynligheter for termer.
- Forutsetter at dokumentet at søkestreng omhandler samme tema/er av samme type.
- Må benytte smoothing-teknikker istedenfor tf_{idf} -vekter.

Deloppgave b

For å lettere å forstå hvor de ulike termvektene i dokumentet kommer ifra, setter vi opp en språkmodell for hvert dokument:

$P(q | M_d)$ gir oss sannsynligheten for at spørringen q er i et enkelt dokument (ikke tatt hensyn til en samling med dokumenter).

$$q = \{\text{NTNU, campus}\}$$

$$M_1 = \{\text{NTNU is a university in Trondheim}\}$$

$$= \{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\}$$

$$P(q | M_1) = \frac{1}{6}$$

$$M_2 = \{\text{Gløshaugen is a Campus at NTNU, Øya is another campus.}\}^1$$

$$= \{\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}\}$$

$$P(q | M_2) = \frac{1}{10}$$

Videre setter vi opp språkmodell for samlingen:² Dette vil si at vi regner ut sannsynligheten for at et dokumentet i samlingen vil generere strengen i spørringen. Med dette kan vi rangere dokumentene.

$$\lambda = 0.5$$

$$\begin{aligned} P(q, d_1) &= ((1-0.5)\frac{2}{16}) + (0.5(\frac{1}{6})) \times ((1-0.5)\frac{2}{16}) + (0.5(\frac{0}{6})) \\ &= \frac{7}{48} \times \frac{1}{16} = \frac{7}{768} = 0.00911 \end{aligned}$$

$$\begin{aligned} P(q, d_2) &= ((1-0.5)\frac{2}{16}) + (0.5(\frac{1}{10})) \times ((1-0.5)\frac{2}{16}) + (0.5(\frac{2}{10})) \\ &= \frac{9}{80} \times \frac{13}{80} = \frac{117}{6400} = 0.01823 \end{aligned}$$

Som vi ser av dette er $d_1 < d_2$, som tilsier at d_2 vil rangeres som det beste dokumentet.

¹Vi forutsetter at språkmodellen ikke tar hensyn til store og små bokstaver.

²Formelen vi benytter til å regne ut er gitt i oppgaveteksten til oppgave 1b.

Oppgave 2 - Interpolated Precision

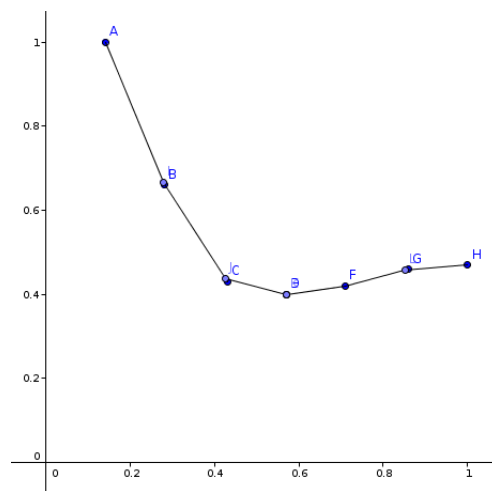
Deloppgave a

Interpolated precision er noe som brukes når man ønsker å lage en mer meningsfylt Recall-Precision graf. Den fungerer ved at man setter en presisjon for et intervall med recall istede for å ha en presisjon for alle recall verdier, noe som gjør at man får en svært hakkete graf. Dette fungerer godt fordi brukeren sannsynligvis vil sjekke ut flere relevante dokumenter og da er det grei å kjøre den høyeste gitte presisjons-verdien for et recall-intervall.

Deloppgave b

Gitt de følgende dokumentene $d = \{2, 6, 72, 10, 84, 15, 103, 66, 37, 45, 12, 201, 33, 94, 22\}$ og de relevante dokumentene $r = \{2, 72, 103, 201, 22, 45, 33\}$. Vi var usikre på hva som mentes med de ti første dokumentene så vi har besvart oppgaven på grunnlag av alle dokumentene.

d	Releant	Recall	Precision
2	REL	$\frac{1}{7}=0.14$	$\frac{1}{1}=1.0$
6			
72	REL	$\frac{2}{7}=0.28$	$\frac{2}{3}=0.66$
10			
84			
15			
103	REL	$\frac{3}{7}=0.43$	$\frac{3}{7}=0.43$
66			
37			
45	REL	$\frac{4}{7}=0.57$	$\frac{4}{10}=0.40$
12			
201	REL	$\frac{5}{7}=0.71$	$\frac{5}{12}=0.42$
33	REL	$\frac{6}{7}=0.86$	$\frac{6}{13}=0.46$
94			
22	REL	$\frac{7}{7}=1.0$	$\frac{7}{15}=0.47$



Figur 1: Graf som representerer dataene fra oppgaven. Y-aksen representerer Precision og X-aksen representerer Recall

Oppgave 3 - Relevance Feedback

Deloppgave a

Meningen med Relevance Feedback er å bruke data fra en tidligere spørring for å gi bedre presisjon av relevans på den neste spørringen. Det er vanlig å tilegne seg informasjon om den forrige spørringen var relevant for brukeren og bruke denne informasjonen.

Query Expansion

Query Expansion er en operasjon som gjerne brukes for å få flere relevante resultater på en gitt spørring. Det fungerer ved at logikken som tar imot spørringen kjører spørringen på flere forskjellige måter med forskjellige variasjoner av spørringen som ble motatt. Dette er svært vanlig å gjøre i søkemotorer. Endringene som er vanlig å gjøre på spørringen er for eksempel å endre ord i strengen til synonymer med samme betydning slik at man får et bredere spekter med resultater. Andre endringer kan være og rette skrivefeil og prøve forskjellige bøyninger av ordet.

Term Reweighting

Term Reweighting går ut på å modifisere en spørring basert på resultatene den forrige spørringen fikk. Man justerer ikke selve stringen i spørringen, men man tar heller og veier alle termene i spørringen på nytt basert på forekomster av termene i resultatene på søket.

Hva er forskjellen?

Forskjellen på term reweighting og query expansion er at term reweighting ikke endrer termene i den originale stringen, men kun veier dem på nytt. Query expansion baserer seg på å endre stringen så den skal få flere relevante treff.