

TDT4117 - Øving 2

Ole-Christer Selvik, Håkon Løvdal og Kristoffer Andreas Dalby

Oktober 2013

Oppgave 1

Deloppgave a

SKRIVER INN MASSE TULL HER I MORGEN

Deloppgave b

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

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

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

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

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

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

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

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

Videre setter vi opp språkmodell for samlingen:²

$$\lambda = 0.5$$

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

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

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

DERP