# Theoretical homework #6, TTTV 2017

Deborah Lambregts (11318643) & Bram Otten (10992456)
Group G, Douwe van der Wal
May 21, 2017

## Exercise 1

(a)

| | Doc1 | Doc2 | Doc3 | q |
|---|---|---|---|---|
| Dam | 1 | 2 | 1 | 1 |
| Dutch | 1 | 0 | 0 | 0 |
| Damhotel | 0 | 1 | 0 | 0 |
| Amsterdam | 0 | 2 | 1 | 1 |
| Centraal | 0 | 1 | 0 | 0 |
| Jam | 0 | 0 | 1 | 0 |
| US | 0 | 0 | 1 | 0 |
| Melkweg | 0 | 0 | 1 | 0 |
| Netherlands | 0 | 0 | 1 | 1 |

(b) Query q = What is the Dam in Amsterdam, Netherlands?

Dot product similarity:
$\vec{Doc}1 * \vec{q} = 1*1+1*0+0*0+0*1+0*0+0*0+0*0+0*0+0*1 = 1$
$\vec{Doc}2 * \vec{q} = 2*1+0*0+1*0+2*1+1*0+0*0+0*0+0*0+0*1 = 2+2 = 4$
$\vec{Doc}3 * \vec{q} = 1*1+0*0+0*0+1*1+0*0+1*0+1*0+1*0+1*1 = 1+1+1 = 3$

Cosine similarity:
Doc1: $\frac{\vec{Doc}1*\vec{q}}{\sqrt{2}\times\sqrt{3}} = \frac{1}{\sqrt{2}\times\sqrt{3}} \approx 0.408 \rightarrow 65.91^\circ$

Doc2: $\frac{\vec{Doc}2*\vec{q}}{\sqrt{10}\times\sqrt{3}} = \frac{4}{\sqrt{10}\times\sqrt{3}} \approx 0.730 \rightarrow 43.09^\circ$

Doc3: $\frac{\vec{Doc}3*\vec{q}}{\sqrt{6}\times\sqrt{3}} = \frac{3}{\sqrt{6}\times\sqrt{3}} \approx 0.707 \rightarrow 45^\circ$

Intuitively, doc1 is most relevant to the query because it actually answers the question. When using just the dot product, doc2 seems most relevant to the query (score 4 above scores 3 and 1). According to the results of cosine similarity, doc2 seems the most relevant document too (0.730 above scores 0.408 and 0.707). So, both methods give the intuitively 'wrong' document and there is no substantial difference between the outcomes. Both methods rank the relevance as (high to low) doc2 - doc3 - doc1.

# Exercise 2

(a) True positives: 1 (doc 1).
False positives 2 (docs 2 and 3).
True negatives: 8.
False negatives: 4.

(b) Precision $= \frac{1}{3}$
Recall $= \frac{1}{5}$
F-score $= \frac{2\times\frac{1}{3}\times\frac{1}{5}}{\frac{1}{3}+\frac{1}{5}} = \frac{1}{4}$

(c) Dam in doc 3:
tf = 1
idf $= \log\frac{N}{n_t} = \log\frac{15}{10} \approx 0.176$
tf-idf $= tf_t \times idf_t = 1 \times 0.176 \approx 0.176$

Melkweg in doc 3:
tf = 1
idf $= \log\frac{15}{2} \approx 0.875$
tf-idf $= 1 \times 0.875 \approx 0.875$

# Exercise 3

$P(I| < s >)$ and such are given. The main data is only for bigrams though, so we'll make the convenient assumption n = 2 for (a) and (b).

(a) *I want Chinese food* with unsmoothed n-grams
$P = P(I| < s >) \times P(want|I) \times P(chinese|want) \times (food|chinese) \times (< /s > |food)$
$P = 0.25 * 0.33 * 0.0065 * 0.52 * 0.68 = 0.00019$

(b) *I want Chinese Chinese food* with unsmoothed n-grams
$P = P(I| < s >) \times P(want|I) \times P(chinese|want) \times P(chinese|chinese) \times (food|chinese) \times (< /s > |food)$
(a) * P(chinese|chinese).
$P = 0.25 * 0.33 * 0.0065 * 0.52 * 0.68 * 0 = 0$

(c) *I want Chinese food* with Laplace-smoothed bigrams
$P^*_{LaPlace}(W_n|W_{n-1}) = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V}\ levertfiguur4.6op.$
P = 0.25 * 0.21 * 0.0029 * 0.052 * 0.68 = 5.38 * $10^-6$

(d) *I want Chinese Chinese food* with Laplace-smoothed bigrams
(c) * P(chinese|chinese).
$P = 0.25 * 0.21 * 0.0029 * 0.00062 * 0.052 * 0.68 = 3.34 * 10^-9$