Exercise Sheet 01

Foundations of Information Retrival

Jonathan Heinz (3431767), Felix Viola (3125695)
s6jthein@uni-bonn.de, s6feviol@uni-bonn.de

April 21, 2023

1. Preprocessing

1. Only syntactical information is important and should be extracted. Formatting, punctuation and stop words do not add significant value.

2. Applied pre-processing steps: Tokenizing, linguistic token preprocessing, stop word removal, :

   - `[my, favorite, fruit, be, apple]`
   - `[what, be, your, name]`
   - `[how, excit]`
   - `[today, be, monday, we, be, not, in, house]`

3. Both approaches are used to reduce inflectional forms of words, however their approaches differ:

   Stemming: finds the common word stem of different words with the same (family of) meaning by removing the last letters. For example *horse, horse's, horses* and *horses'* get reduced to their stem *horse*

   Lemmatization: reduces words to their dictionary form (so called lemma). For example *am, are, is* get reduced to *be*.

2. Posting List, Inverted Index and Boolean Retrival Model

D1: `[My, favorite, fruit, is, apple]`

D2: `[I, like, playing, sports]`

D3: `[I, don't, enjoy, sports]`

D4: `[My, mother, is, working]`

Inverted Index:

- My → D1, D4
- favorite → D1

- fruit → D1
- is → D1, D4
- apple → D1
- I → D2, D3
- like → D2
- playing → D2
- sports → D2, D3
- don't → D3
- enjoy → D3
- mother → D4
- working → D4

Boolean queries:

- My *AND* apple ⇒ D1
- My *AND NOT* I ⇒ D1, D4
- I *AND* sports *OR* don't ⇒ D2, D3

3. Term-Document Matrix

D1: `[my, favorite, fruit, is, apple]`

D2: `[i, like, playing, sports]`

D3: `[i, don't, enjoy, sports]`

D4: `[my, mother, is, working]`

Term-Document-Matrix:

	D1	D2	D3	D4
my	1	0	1	0
favorite	1	0	0	0
fruit	1	0	0	0
is	1	0	0	1
apple	1	0	0	0
i	0	1	1	0
like	0	1	0	0
playing	0	1	0	0
sports	0	1	1	0
don't	0	0	1	0
enjoy	0	0	1	0
mother	0	0	0	1
working	0	0	0	1

4. TF*IDF

In this corpus, the term frequency $\mathrm{tf}_{t,d}$ is $1 + \log_{10}(1) = 1$ for $t \in d$ and 0 otherwise (since any term is at most once present per document).

In this corpus, the inverse document frequency depending on a terms document frequency $\mathrm{df}_t \in \{1, 2\}$ is

$$\log_{10}\left(\frac{4}{1}\right) = 0.6 \text{ or } \log_{10}\left(\frac{4}{2}\right) = 0.3$$

TF*IDF weighted matrix:

	D1	D2	D3	D4
my	0.3	0	0.3	0
favorite	0.6	0	0	0
fruit	0.6	0	0	0
is	0.3	0	0	0.3
apple	0.6	0	0	0
i	0	0.3	0.3	0
like	0	0.6	0	0
playing	0	0.6	0	0
sports	0	0.3	0.3	0
don't	0	0	0.6	0
enjoy	0	0	0.6	0
mother	0	0	0	0.6
working	0	0	0	0.6

5. Similarity Computation

With $u = (0, 0, 0, 0, 0, 0.3, 0.6, 0.6, 0.3, 0, 0, 0, 0)$ and $v = (0.3, 0, 0, 0, 0, 0.3, 0, 0, 0.3, 0.6, 0.6, 0, 0)$

cosine similarity is given by:

$$coscos(u, v) = \frac{u \times v}{||u|| \times ||v||} = \frac{0.18}{0.9439} = 0.19$$

6. Miscellaneous

- Euclidian distance measure takes the lenght of the vectors into account while cosine similarity does not. This is useful when comparing documents since the length of the vectors is unimportant.
- Stop words are very common in documents and therefor have a low TF*IDF value. Depending on the corpus this value might even go to zero, effectively gaining the same result as stop word removal.
- IDF is zero, if a term occurs exacly as often as there are documents. IDF is infite, if a term is very rare in a corpus.

Jonathan Heinz (3431767), Felix Viola (3125695)
s6jthein@uni-bonn.de, s6feviol@uni-bonn.de