LUND UNIVERSITY, INSTITUTION FOR ELECTRICAL AND INFORMATION TECHNOLOGY

EDAN20 - Lab1

Gustav Fahlén elt14gfa@student.lu.se

October 3, 2018

The purpose of this lab was to learn how to collect all words within a number of text files and produce a vector containing the positions of each word in each file. Moreover, represent each word in each document with their associated tf-idf value, to tell which words are considered more important than an other. These values are than used in cosine-similarities to notify which documents are most similar to one and another.

The master index file was created from many smaller index files, and each index file is created from each text file. Below is an example of the out print of the master index file of the word “gömt”:

```
'gömt': {'bannlyst.idx': [41033, 249276],
        'gosta.idx': [161617, 191886, 231828, 301396, 624031, 636328, 700662],
        'herrgard.idx': [113255],
        'jerusalem.idx': [148778],
        'kejsaren.idx': [154063],
        'marbacka.idx': [151164],
        'nils.idx': [293601, 743942, 1008398],
        'osynliga.idx': [145071, 180012, 387151, 612228, 647068, 867570],
        'troll.idx': [195055]},
```

Figur 0.1: Out print from the ’Gömt’ in the master index file. Each integer tells the position of the word in the corresponding text file.

Tf-idf is calculated by multiplying two different equations, one for tf and one for idf. The necessary paramters to calculate tf and idf are obtained from the mater index file.

Tf = (nbr of times a word is found in a specific document) / (total nbr of words)

Idf = log10(nbr of documents / nbr of documents containing the word)

The Tf-Idf values are than used to calculate cosine similarities, which are calculated by the following formula:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figur 0.2: Cosine similarity - used to compare how similair different texts are

A and B corresponds to the tf-idf value for the chosen word. When applying this to the code a for loop is used, which is looping through each word from one specific file and comparing it with with the word in the other files.

```
for word in idtflist1:
    sumd1times2 += idtflist1[word] * idtflist2[word]
    sumd1sqr += idtflist1[word] * idtflist1[word]
for word in idtflist2:
    sumd2sqr += idtflist2[word] * idtflist2[word]

return sumd1times2 / (math.sqrt(sumd1sqr) * math.sqrt(sumd2sqr))
```

Figur 0.3: Cosine similarity - in python

All files are being compared with one and another, but it turned out that 'kejsaren.idx' and 'troll.idx' are the most similar ones with a similarity value of 0.088. If the value is 1 they are exactly the same.