Lab 1 – EDAN20

The purpose of this lab was to learn about how to collect all words within a number of text documents and produce a vector containing the positions of each word in each document. Moreover, learn about indexing and thereby build a master index file containing all data from all documents, thereafter 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 indexfiles, created out of each single document.

This is an example of the out print in the master index file of the word “gömt”,:

(each integer tells the position of the word in the correspsonding document).

'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],

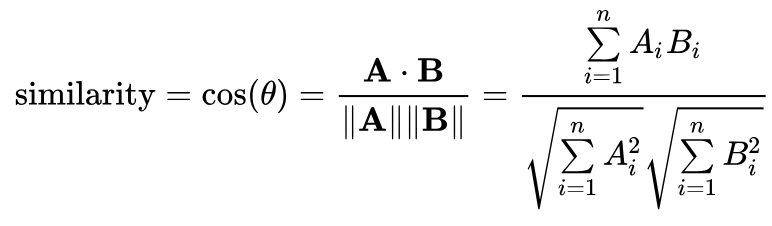
'troll.idx': [195055]},

The tf-idf values was obtained by using the master index file containing all data needed. Tf-idf is calculated by two different equations, one for tf and one for idf.

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 the cosine similarities. The cosine similarities are calculated by the following formula:



A and B corresponds to the tf-idf for the chosen word. When applying this in the code a for loop is used, which is looping through each word from one specific document, in the documents that are being compared.

for word in idtflist1:  
 sumd1times2 += idtflist1[word] \* idtflist2[word]  
 sumd1sqr += idtflist1[word] \* idtflist2[word]  
 sumd2sqr += idtflist2[word] \* idtflist2[word]  
  
return sumd1times2 / (math.sqrt(sumd1sqr) \* math.sqrt(sumd2sqr))

All files are being compared with one and another, but it turned out the documents kejsaren and troll are the moste similar ones with a value of 0.517. If the value is 1 are the texts exactly the same.