DMAssignment_1

Parth Patel

February 3, 2016

Problem 1.

a :

The most important advantage of tf-idf is that it reduces the length of any big corpus by a substantial amount by removing the frequently occuring stop words like 'the', 'so' etc. The simple tf-idf formula $x_{ij} = \frac{m_{ij}}{m_i}.log\frac{n}{n_j}$ makes it really easy to compute. The term frequency part of the formula also takes care of normalization, in a sense that, all the documents in picture might not have the same length and hence it is higly likely that the number of times a particular word occurs in a document with greater length is larger than the number of occurences of the same word in a document with comparative shorter length.

b :

One way of rephrasing this question is to mention the disadvantages of 'IDF' concept. The IDF gives greater importance to the least/rarely occuring words. Thus the significance of each word can be considered to have an inversve proportional relation with the number of documents that has the same word. But there might me scenerios that a particular word or set of words represent the topic of a given doucment, having multiple occurences in the same document. Such words can be of great importance in cases of text categorical classification, but IDF completely ignores these kind of words and hence might prove ineffective for such classification problems.

c:

If the term occurs in only one document, according to the formula

$x_{ij} = \frac{m_{ij}}{m_i}.log\frac{n}{n_j}$

the value of n_j becomes one which makes the log factor huge (depending on the size n i.e the total number of documents), but the term frequency value will be a non zero value for just that one document 'i', thus eliminating this unique word 'j' from the comparision log.

If the term occurs in every document, the given transformation clearly eliminates the word, since $\frac{n}{n_j}$ becomes one and log(1) is always zero.