(b) Consider a document-term matrix, where tfij is the frequency of the i th word (term)

in the j th document and m is the number of documents. Consider the variable

transformation that is defined by

(1) tf’ij = tfij ∗ log m/dfi

where dfi is the number of documents in which the i th term appears, which is known as the document frequency of the term. This transformation is knows as the inverse document frequency transformation.

i. (4 points) What is the effect of this transformation if a term occurs in one document? In every document?

ii. (3 points) What might be the purpose of this transformation?

Ans:

i)

1)

If a term occurs in only one document then the final transformation will be product of log(total no of document) and the frequency of word in that document.

For example,If a word “Carolina” occurs 4 times in only one document which has 1000 words and out of 100,000 document then after the transformation the weight will be assigned to Carolina in following way:

tfij = 4 / 1000 = 0.004

log (m/dfij) = log(100000/1) = log(105) = 5

Therefore tf’ij will be a product of 5 and 0.004 which is 0.02.

With assessing tf’ij of other words in a document we can check how important the term ‘Carolina’ is to the single document and whether it is part of words which can give idea about the paragraph in few words.

2) If a term occurs in all documents then the final transformation will be zero.

For example,if a term ‘an’ occurs in all the documents then m=dfij ,which will lead to log of 1,therefore the zero weight.

ii)

The main idea behind inverse document frequency transformation is to assess the significance of a word in a single document out of collection of documents.

An intuitive idea behind assessing the significance of the word in a document would be to collect the frequency of the each word in a document and word with most occurrence would be the most significant.

However, the significance depends heavily upon the language and the subject of documents.For example ‘a’,’an’,’the’ are generally most significant words in a document written in an English language ,therefore these words rank higher in the frequency chart but if the same set of words are measured against all the documents then we can deduce that they occur in almost all the documents and therefore they are not significant or rare enough to represent a collective idea behind a document.

These words, though being frequent in a specific document, don’t really separate that document from the others. Hence we have an idf terms which assesses if a word is *rare* (and hence important) while considering all the documents together.