Home Assignment No. 1

Submission date: 10/04/2018

# Question 1 [25%]

Devise an algorithm to calculate the Boolean “not” on a postings list.   
Assume: there are N documents in the collection, and their IDs are successive (1, 2, 3, ..., N).

1. Explain in words and give a short example of how the algorithm works.
2. Write a pseudo code.

# Question 2 [25%]

Consider the collection made of the 3 following documents:

**Doc 1** Brian went to school with Kate.

**Doc 2** Kate goes to the cinema every Thursday.

**Doc 3** We are going to be late on Thursday.

**Stop word list:** to, with, the, on, be

Normalize and lemmatize the terms, and considering the stop-list:

1. Draw the term-document incidence matrix for this document collection.
2. recall the figure of the inverted index from lecture 1:

****

Draw the inverted index representation for this collection, as in the figure.

1. For the document collection above, what are the returned results for these queries?  
   (Note that the same tokenization algorithm should be applied on the query terms)
   1. Kate & is & going
   2. Kate & !( Thursday | school)

# Question 3 [25%]

Are skip pointers useful for queries of the form x or y? Explain and give an example.

# Question 4 [25%]

Assume a biword index.

1. Give an example of a document that will be returned for a query of New York University but is actually a false positive that should not be returned.
2. Suggest a way to filter out false positive documents.