Information Retrieval (CS 317)
Mid I Exam Fall 2016

Date: September 20, 2016　　　　**Marks: 13**　　　　**Time: 60 min.**

Name: ---

Registration No: --------------------------

Section: --------------

Attempt the examination on the question paper and write concise answers. *You can use extra sheets for rough work. Do not attach extra sheets used for rough work with the question paper. Do not fill the table titled Question/marks.*

Question	a	b	c	d	e	f	g	h	Total
Marks	/ 1	/ 2	/ 1	/ 1	/ 2	/ 1	/ 3	/ 2	/13

Part 1

a) What kind of problems can arise if we convert all text into lower case during text pre processing? Illustrate with example. [1 Point]

b) What proportion of total vocabulary (proportion of total unique words) of a novel you are expected to see if you have read 45% text of the novel. [2 Points]

c) Briefly explain steps of BSBI Indexing algorithm. [1 Point]

d) What is advantage of using positional index over bigram (biword) index? [1 Point]

e) If we have a corpus of 10 million documents, each of length 3,000 words, and a total vocabulary size of 500,000, what is the approximate maximum
  i.   size of the postings
  ii.   size of the Boolean matrix (which contains a 1 in row *i* and column *j* if word i occurs in document j and 0 otherwise)
 [2 Points]

Part 2

f) If a coin with unknown bias is flipped 10 times and it comes up heads 10 times then what is the likelihood of getting a tail in next coin flip using Laplace estimates. [1 Point]

Given the three-document corpus and a stop word list below, answer the following questions (g and h) AFTER removing stopwords.

d₁	information retrieval is process of index search retrieval
d₂	retrieval is used for evaluation of search results retrieval retrieval
d₃	evaluation in information in evaluation process search
Query	information retrieval
Stopwords	is , of, in, for, to

g) Rank documents according to their TF.IDF score. Show all calculations and fill in the table below. [3 Points]

Terms	TF			IDF	TF.IDF
	d₁	**d₂**	**d₃**		

h)   Calculate similarity of each document with the query using maximum likelihood estimate using Witten-Bell smoothing. (use three document corpus given above) [2 Points]

	Witten-Bell smoothing
d_1	
d_2	
d_3	