	Course:	Information Retrieval and Text Mining	Course Code:	CS567
	Program :	MS (Computer Science)	Semester:	Fall 2016
	Duration:	80	Total Marks:	23
	Paper Date:	October 18, 2016	Weight:	20
	Section:	CS	Pages:	6
	Exam:	Midterm		

Instructions/Notes: 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.

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

Question	1	2	3	4	5	6	7	8	9	Total
Marks	/ 2	/ 1	/ 2	/ 2	/ 1	/ 3	/ 3	/ 5	/4	/23

Part 1

Q1) a) If the length of two postings lists are *x* and *y*, then what is the tightest upper bound on the running time of merging the postings lists in an OR query (using merge algorithm studied in class)? [1 Point]

b) Consider the answer you obtained for merging 2 lists in part (a). Can we do better (Can we improve this running time)? If yes how? [1 Point]

Q2) What do we need to store in index in order to answer proximity queries ? [1 Point]

Q3) 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]

Q4) 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

Q5) 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 (Q 6 and Q7) AFTER removing stopwords.

d_1	information retrieval is process of index search retrieval
d_2	retrieval is used for evaluation of search results retrieval retrieval
d_3	evaluation in information in evaluation process search
Query	information retrieval
Stopwords	is , of, in, for, to

Q6) Rank documents according to their TF.IDF score. Show all calculations.
[3 Points]

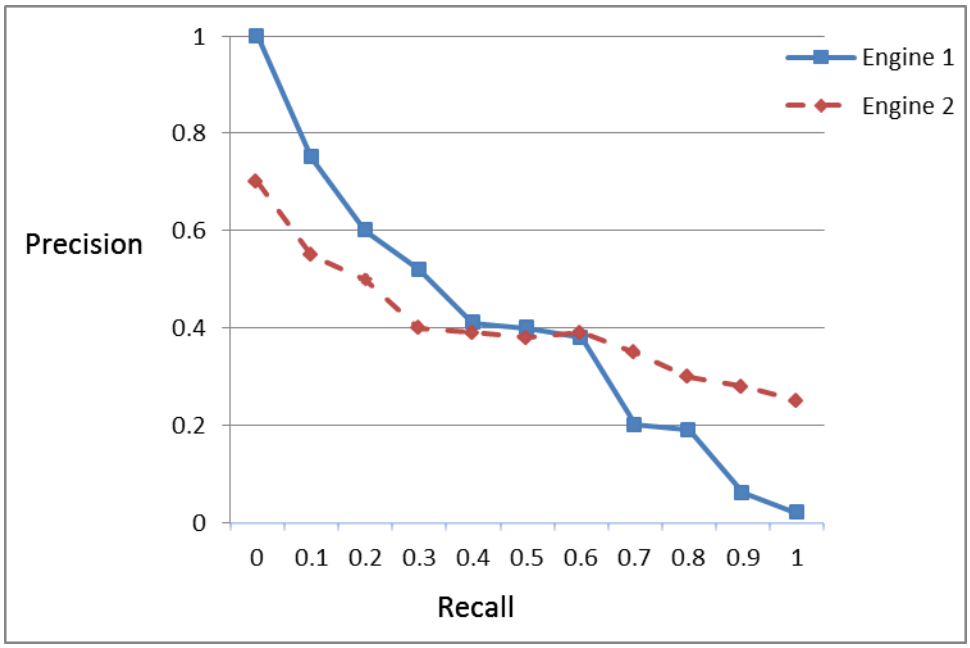
'

Q7) Calculate similarity of each document with the query using maximum likelihood estimate using Witten-Bell smoothing. (use three document corpus given above) [3 Points]

Part 3

Q8) a) What is F-measure value, MAP, and R-Precision of following ranked list of documents? Suppose total number of relevant documents is 10. Leftmost document is top ranked document. [3 Points]

R R N R N N N N R R



b) The figure above depicts interpolated precision-recall curves for two search engines that index research articles. There is no difference between the engines except in how they score documents. Imagine you're a scientist looking for all published work on some topic. You don't want to miss any citation. Which engine would you prefer and why? [2 Points]

Q9) a) In Normalized Discounted Cumulative Gain (NDCG), we normalize the Discounted Cumulative Gain (DCG) for each topic with a normalizer. What is this normalizer? Why do we need to do this normalization step? Why do we not need to do this normalization step for the Mean Average Precision (MAP)? [2 Points]

b)The table below shows the final ranked list of results for an IR search together with their human-rated relevance grades. Assume the table contains all documents with non-zero relevance. Compute the values of the DCG evaluation metrics for each value of n and add them to the table. [2 Points]

n	Documents	Relevance	DCG
1	D1	4	
2	D2	1	
3	D3	0	
4	D4	3	
5	D5	5	