Name:

Enrolment No:

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

End Semester Examination, December 2018

Programme Name: BTech-CS-All Branches | **Semester : V**
Course Name : Information Retrieval and Search Engines | **Time : 03 hrs**
Course Code : CSEG3012 | **Max. Marks: 100**
Nos. of page(s) : 04

SECTION A

S. No.		Marks	CO
Q 1	Assuming Zipf's law with a corpus independent constant A = 0.1, what is the fewest number of most common words that together account for more than 18% of word occurrences (i.e. the minimum value of m such that at least 18% of word occurrences are one of the m most common words).	05	CO1
Q 2	How can we design a new Mean Average Precision metric to incorporate multi-grade relevance judgments, e.g., 0 - irrelevant, 1 - fair, 2 - good, 3 - excellent, 4 - perfect.	05	CO3
Q 3	How to utilize user clicks in search evaluations?	05	CO2
Q 4	You want to avoid always giving blog posts with zero comments a retrieval score of zero. Otherwise, they would never be retrieved! How could you mitigate this problem?	05	CO4

SECTION B

Q 5	A search engine supports error correction in the following way: If an error is suspected in a query term, the system provides a link labelled **"Did you mean X?"** where X is the corrected term, in addition to its normal results. Explain why it is non-trivial to implement this feature efficiently. Discuss methods for implementing this feature in a realistic setting.	10	CO1
Q 6	Consider a lexicon (dictionary) in which there are no words of length 1 or 2. Assume the probability of a word in this lexicon having length i is proportional to $1/i^2$, for i > 2. Assume that the distribution is truncated so that the longest possible word length is 25. Further, the lexicon has 50,000 words in it. Consider storing the words as one contiguous string, with a term pointer to the beginning of each word. How much space is used in total, including storage for the entire string as well as for the pointers that resolve the beginning of each word? Show a formula, and estimate the total as a number of bytes needed.	10	CO2
Q 7	We have 100 million documents containing 9 million terms.	10	CO4

	(a) How many posting entries are there using the simple Zipf's approximation? You may assume that the natural log of 9 million is 16. (b) Assume 12 bytes per postings entry and a machine with sufficient main memory to hold all data in memory. Assume a cost of 1 microsecond per cpu operation. How much time would it take to sort the postings entries in memory? Assume we use Quicksort with running time 2N ln N.		
	OR		
	We have 100 million documents containing 9 million terms. Consider the case where we have only 8GB of main memory, and we use blocks with 200 million postings entries in a block. Assuming 0.5 milliseconds per disk seek, 0.5 microseconds per byte following a seek in block transfer mode and 1 microsecond for all other operations, estimate the time to create one such block of 200 million sorted postings on disk. Also give the total time needed to create all such (initial) sorted blocks.	10	CO4
Q 8	GEneric Multimedia object INdexIng is considered as quick-and-dirty test to quickly discard bad objects. Do your agree the given statement. Justify your support and Discuss the steps followed in GEMINI approach.	10	CO3
SECTION C			
Q 9	Let us consider a scenario in which we use two crawls to estimate the frequency of duplicates on the web. Web search engines A and B each crawl a random subset of the web of the same size. Some of the pages crawled will be duplicates – exact textual copies of each other at different URLs. Assume that duplicates are distributed uniformly amongst the pages crawled by A and B. No pages have more than two copies. A indexes pages without duplicate elimination whereas B indexes only one copy of each duplicate page. If 45% of A's indexed URLs are present in B's index, while 50% of B's indexed URLs are present in A's index, what fraction of the web consists of pages that do not have a duplicate?	20	CO4
Q 10	Stack overflow wants to redesign its current search function: it is preferred if it can directly answer questions rather than simple keyword matching in all forum posts. Based on the concepts of ranking how will you design a tailored ranking system for them?	20	CO3
	OR		
	Information retrieval systems vary in the expressivity of the query languages they employ. For instance, some systems support proximity search: if two query terms are connected by the **"Next"** operator, then only those documents are retrieved where the query terms appear close together (i.e., within a certain number of words of each other). a) List and briefly describe other ways in which the syntax and the interpretation of query languages may vary. b) Describe with an example how the **"Next"** operator described above is implemented effectively in modern information retrieval systems. Your answer should include a description of the data structure(s) necessary to support it.	20	CO3