SRN | | | | | | | | | | | | |

PES UNIVERSITY

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE15CS412

Dec 2018: END SEMESTER ASSESSMENT (ESA) B.TECH. VII Semester

UE15CS412- Algorithms for Information Retrieval

Time: 3HrsAnswer all questions preferably in the same orderMax Marks: 100

Clearlystate any assumptions made if question is ambiguous.Write short answers for question 5.

1	a)	Let D be a document in a text collection. Suppose we add a copy of D to the collection. How would this affect the IDF values of all the words in the collection? Why?	4
	b)	Given a Boolean conjunctive query "term-a AND term-b", write an algorithm to show how the posting lists containing the two terms are intersected, assuming the posting lists have skip pointers.	8
	c)	Determine the similarity of vector space model between the given query and the documents. Use tf*idf for weights where tf is raw term frequency and no normalization for both document and query. (ntn.ntn). Use dot-product for similarity and rank the documents. Q: "gold silver truck" D1: "Shipment of gold damaged in a fire" D2: "Delivery of silver arrived in a silver truck" D3: "Shipment of gold arrived in a truck"	8
2.	a)	Write the formula for Zipf's Law (define your terms). Give one practical example of its use (i.e., a situation where it would be useful). What are the implications of Zipf's law for IR?	4
	b)	Outline the implementation of spellchecker developed by Peter Norvig. Mention 2 ways of further improving this model.	6
	c)	What is the future of search. Discuss in terms of new possible features or new search applications	4
	d)	For each of the scenarios below, choose the most appropriate evaluation metric to judge a ranked list. Justify your choice. a) A student searching for the CS412 course page. b) A customer who wants to buy a laptop case from an online store. The search results have different sizes, looks, ratings, and prices. The customer likes or dislikes them in varying degrees. c) A lawyer who is working on finding related documents from the Claims cases of Supreme. Court for her upcoming trial. She has a staff to look through many documents if necessary.	6
3.	a)	A metasearch engine (or aggregator) is a search tool that uses multiple other search engine's data to produce its own results from the Internet. Metasearch engines take input from a user and simultaneously send out queries to third party search engines for results. Devise a metasearch algorithm that ranks web pages based on the results of other search engines	4
	b)	Given the following collection (one doc per line), what is the revised query based on Rocchio relevance feedback algorithm. good movie trailer shown trailer with good actor unseen movie Assume the following : (a) dictionary consists of just three words : movie, trailer and good (b) document is represented as a vector of raw tf scores..	8

SRN | | | | | | | | | | | | |

		(c) user judges the first 2 documents relevant for the query "movie trailer". (d) balancing weights in Rochhio algorithms is Alpha=+1 for original query, Beta=+1 for relevant documents and Gamma= - 0.5 for non-relevant documents	
	c)	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 you have learned in this course, can you design a tailored ranking system for them? Hint: you can discuss what components we can reuse from a generic search engine, and what components we need to redesign to accommodate documents from Stack-Overflow (e.g., consider its discussion structure, program snippet in post, content, tags, vote count, authors' reputation, etc.).	8
4.	a)	Given the following 3 documents and their classes, classify the test document into one of the two classes using KNN. Represent each document as a vector of raw tf, use cosine similarity for finding the nearest neighbors, and assume k=3. "Iraq election" => world news "French executive injured" => world news "Chief executive smiles" => business "Krispy Kreme executive resigns" => business Classify the document "executive gets married"	8
	b)`	Cluster to following documents using K-means with K=2 and cosine similarity. D1: "come lilliput come" D2: "come walking" D3: "walking lilliput" D4: "lilliputlilliput" Assume D1 and D3 are chosen as initial seeds. Use tf (no idf). Show the clusters and their centroids for each iteration. The algorithm should converge after 2 iterations'.	8
	c)	Mention different methods of text summarization and briefly explain one of them.	4
5.	a)	In the context of a tf-idf weighting why might you decide to store raw term frequencies in the postings, rather than the final tf-idf weighting for a term?	2
	b)	What is the primary advantage of "R-precision" over the "Precision at a fixed cutoff" metric?	2
	c)	For efficient ranking, in what order of query terms would you search in the inverted index. Why	2
	d)	Why cosine similarity is preferred over Euclidian distance in Vector Space Models?.	2
	e)	Convert 100 to variable byte and gamma encoding.	2
	f)	If you wanted to search for s*ng in a permuterm wildcard index, what key(s) would one do the lookup on?	2
	g)	You decide to form a start-up company being inspired by this course. What kind of product or service (related to IR) would your company offer that you think will be useful in the real world	2
	h)	Breadth-first crawling helps us avoid duplicated visit of web sites. Explain whether this statement is true or false	2
	i)	List 4 important characteristics/features of a web crawler.	2
	j)	You are in charge of Customer Relationship Management. Give two examples of how IR techniques/tools can be employed to improve your customer relationships.	2