R.M.

Library
Set

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR	
	Signature of the Invigilator

Please fill up carefully the boxes provided below

EXAMINATION (End Semester)												SEMESTER (Spring)	
Roll Number										Section		Name	
Subject Number	C	S	6	0	0	9	2	Subject Name	Information Retrieval				
Name of the Department / Center of the Student													

Instructions and Guidelines to Students Appearing in the Examination

1. Ensure that you have occupied the seat as per the examination schedule.
2. Ensure that you do not have a mobile phone or a similar gadget with you even in switched off mode. Note that loose papers, notes, books should not be in your possession, even if those are irrelevant to the paper you are writing.
3. Data book, codes or any other materials are allowed only under the instruction of the paper-setter.
4. Use of instrument box, pencil box and non-programmable calculator is allowed during the examination. However, exchange of these items is not permitted.
5. Additional sheets, graph papers and relevant tables will be provided on request.
6. Write on both sides of the answer script and do not tear off any page. Report to the invigilator if the answer script has torn page(s).
7. Show the identity card whenever asked for by the invigilator. It is your responsibility to ensure that your attendance is recorded by the invigilator.
8. You may leave the examination hall for wash room or for drinking water, but not before one hour after the commencement of the examination. Record your absence from the examination hall in the register provided. Smoking and consumption of any kind of beverages is not allowed inside the examination hall.
9. After the completion of the examination, do not leave the seat until the invigilator collects the answer script.
10. During the examination, either inside the examination hall or outside the examination hall, gathering information from any kind of sources or any such attempts, exchange or helping in exchange of information with others or any such attempts will be treated as adopting 'unfair means'. Do not adopt 'unfair means' and do not indulge in unseemly behavior as well.

Violation of any of the instructions may lead to disciplinary action of varied nature.

To be filled in by the examiner											
Question(s)	1-20	21-23	24-29	30	31	32	33	34	35	36	**Total**
Marks											

Marks obtained (in words)	Signature of Examiner	Signature of Scrutineer

Information Retrieval -- CS60092

Maximum Marks: 80

This question paper has 6 pages and 36 questions. Answers are to be written in the question paper itself. Please remember to return the question paper.

Questions 1 to 20 contain ONE mark each. For each correct answer you score ONE mark; for each wrong answer you lose HALF mark. Tick only the appropriate answer.

1. The initial state of random walk in PageRank algorithm determines the estimation quality of its stationary distribution.
   (a) Yes
   (b) No
   (c) Depends on the nature of the chosen initial state
2. For an N X N square positive matrix, the number of distinct eigenvalues is
   (a) Exactly N
   (b) Strictly less than N
   (c) Less than or equal to N
3. For square positive matrices, in case of matrix-vector products it is the small eigenvalues and the corresponding eigenvectors that are
   (a) extremely important
   (b) not so important
   (c) occasionally important
4. If C is the term document matrix then we can construct a symmetric diagonal decomposition of
   (a) C
   (b) CC^{-1}
   (c) CC^{T}
5. For a rank-k approximation, the error in approximation can be approximated by
   (a) the singular value at index k+1
   (b) the singular value at index k
   (c) sum of singular values at indices k and k+1
6. Vector space models are very efficient in handling
   (a) Synonymy related problems
   (b) Polysemy related problems
   (c) Classification problems
7. The most important cost of universal crawling is
   (a) Storage
   (b) Processing
   (c) Bandwidth
8. Topical locality

(a) Degrades very slow
(b) Degrades very fast
(c) Does not degrade

9. Topical crawlers rely on
   (a) Co-citation more
   (b) Co-reference more
   (c) Both co-citation and co-reference equally
10. Cue phrases and location markers can be used as is for generating
   (a) Extractive summaries
   (b) Abstractive summaries
   (c) Both extractive and abstractive summaries
11. In degree centrality based document summarization the similarity measure used is
   (a) tf-modified
   (b) idf-modified
   (c) tf-idf-modified
12. In LexRank/TextRank, one is interested to compute the prestige of
   (a) a word
   (b) a sentence
   (c) a noun phrase
13. In c-LexRank the order of processing is
   (a) First cluster then PageRank
   (b) First PageRank the cluster
   (c) PageRank and cluster simultaneously
14. A new lexical chain needs to be necessarily inserted in the existing list of chins if
   (a) a new word could not be inserted anywhere in the existing chains
   (b) a new word could be inserted in multiple chains and there is a confusion
   (c) a new sentence is added to the text
15. Sentences representative of strong chains are those that has the
   (a) Multiple occurrence of the keyword from the chain
   (b) First occurrence of the keyword from the chain
   (c) Last occurrence of the chain
16. In Wordnet based summarization approach extracting collocations is
   (a) Insensitive to stop word removal
   (b) Sensitive to stop word removal
   (c) Only occasionally sensitive to stop word removal
17. In Wordnet based summarization, after the sentence selection phase, each row of the matrix M represents
   (a) Overall meaning captured by a sentence
   (b) Important synsets
   (c) Share of meaning captured by a sentence out of the meaning of the whole text

18. In Wordnet based summarization, top eigenvectors of the matrix M refers to the
    (a) Important meaning dimensions of the text
    (b) Important sentences to be inserted into the summary
    (c) Important sentence-meaning pairs
19. Undefined references should be
    (a) Always included in a summary
    (b) Never included in a summary
    (c) Occasionally included in a summary as per need
20. The ROUGE-N measure is recall centric because it considers
    (a) all n-gram matches between the candidate and the reference summaries in the numerator
    (b) all n-grams from all reference summaries in the denominator
    (c) Both (a) and (b)

Questions 21 to 23 contain FOUR marks each. For each correct blank entry you score TWO marks; for each wrong entry you lose ONE mark.

	d_1	d_2	d_3	d_4	d_5
romeo	1	0	1	0	0
juliet	1	1	0	0	0
happy	0	1	0	0	0
dagger	0	1	1	0	0
live	0	0	0	1	0
die	0	0	1	1	0
free	0	0	0	1	0
new-hampshire	0	0	0	1	1

Figure 1. Term-document matrix

21. Consider the term-document matrix in the Figure 1. The singular values after rank 2 approximation are _____ and _____.
22. Consider the term-document matrix in the Figure 1. The rank 2 latent space co-ordinates of d_5 are _____ and _____.

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

Figure 2: Distances between US cities in miles

23. With reference to Figure 2, the last two clusters that would merge in single-linkage clustering are _____ and _____.

Questions 24 to 29 contain THREE marks each. For each correct blank entry you get ONE AND HALF marks; for each wrong entry you lose ONE mark.

24. For universal crawlers, the most important policy issues are _____. For preferential crawlers, the PageRank priority is usually computed based on pages _____.
25. In graph-based summarization, setting weak thresholds on edges mean _____ while strong thresholds mean _____.
26. In Wordnet based summarization, the traversal for sub-graph construction is done for a (small) fixed depth to avoid _____. In the PCA phase, the different eigenvectors of the matrix M required to rank sentences are chosen proportional to _____.
27. One advantage of using ROUGE-LCS is that it does not require _____ matches but _____ matches.
28. In *k*-NN classification, decision boundaries are constructed _____ based on the _____ hypothesis.
29. In the context of Rocchio classification, the expression for w = _____ and b = _____.

30. Consider the following two documents. [5 marks]

    D1: Peru Votes in Presidential Election as Fujimori Leads Polls

    D2: Fujimori daughter leads exit polls in presidential election

    Given the query, "Peru Election Fujimori", use the mixture model to compute the relevance scores for each of the documents. You should use lambda = ½. You can lowercase all the words but do not remove the stop words.

    Relevance score of D1: _____ Relevance score of D2: _____

31. Let $A(t,v)$ denote the lexical association between terms t and v, computed as per PMI measure. Express how you would use this to incorporate query expansion while computing $P(q|M_d)$ as per the language modeling, where M_d corresponds to the document language model $P(.|M_d)$. Let V denote all the terms in the vocabulary. [3 marks]

    $P(q|M_d)$ =

32. Consider the following sentences (after removing stop-words) along with the class they belong to – Dependency Parsing (DP) and Word Sense Disambiguation (WSD)
    S1: Syntactic Parser Combination Improved Dependency Analysis (DP)
    S2: Sentence Diagram Generation Using Dependency Parsing (DP)
    S3: Dependency Parsing Short Dependency Relations Unlabeled Data (DP)

S4: Word Sense Disambiguation Using Label Propagation Based Semi Supervised Learning (WSD)

Use these sentences to build an NB classifier and find the scores for the sentence, "S5: Using Dependency analysis improved sense disambiguation" to belong to these classes. [5 marks]

P(DP|S5) = _____ P(WSD|S5) = _____

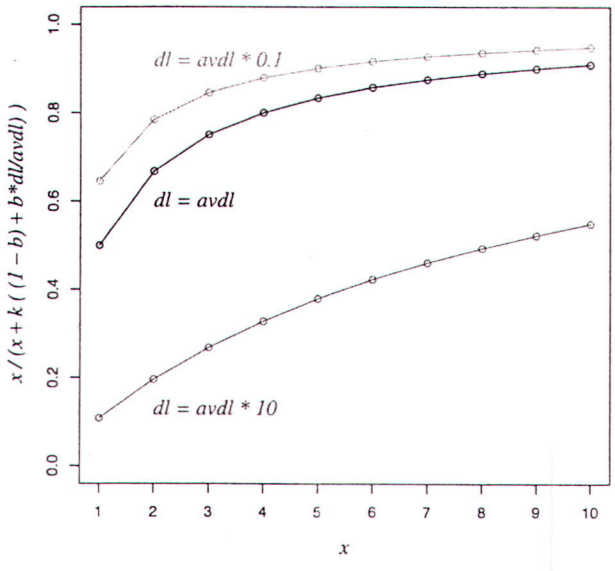
33. In a task of text-classification, assume that you want to categorize various documents into 3 classes, 'politics', 'sports' and 'nature'. The table below shows 12 such documents, their true labels (Actual) and the label assigned by your classifier (Predicted). [3+3 marks]

Doc	Actual	Predicted	Doc	Actual	Predicted
1	Politics	Sports	2	Sports	Sports
3	Politics	Politics	4	Sports	Sports
5	Politics	Nature	6	Sports	Politics
7	Nature	Nature	8	Nature	Politics
9	Sports	Nature	10	Politics	Politics
11	Nature	Sports	12	Nature	Nature

Construct the confusion matrix.

Macro-averaged precision: _____ Micro-averaged precision: _____

34. The following figure describes how the BM25 indexing behaves as a function of term frequency (x) for different document lengths. Find the parameters k and b. [2 marks]



k = _____ b = _____

35. For a given query, an IR system returns 10 documents, out of which only the documents at odd positions are relevant. Also, documents at position 1,3 have a relevance score of "3", 5,7 have a relevance score of "1" and 9 has a relevance score of "2". Assume that the query has 5 relevant documents. Answer the following. [1+1+2+2 marks]

    Precision @ 5 = _____ Recall@5 = _____

    Average Precision = _____ NDCG@5 = _____

36. Consider that the posting list of a term contains the following docIds

    57, 148, 512 …

    The corresponding representation using variable byte encoding would be [3 marks]

    = _____

Rough Work