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 : CSEG393 | **Max. Marks: 100**
Nos. of page(s) : 04

SECTION A

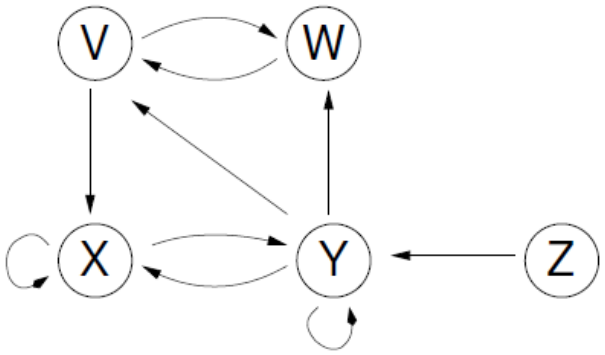
S. No.		Marks	CO
Q 1	Bing wants to become a better-personalized search engine. Combining the concepts and techniques, give your concrete suggestions of where and how an IR system can be personalized. Cover at least three components in a typical retrieval system, e.g., query processing module, ranking functions, and feedback modeling.	05	CO1
Q 2	Demonstrate Zipf's law and Heap's Law for modelling Natural Language. In a given corpus of Spanish documents, the frequency of the most frequent word is 1,270,873. Then what is the estimated frequency for the second most frequent word in this corpus?	05	CO3
Q 3	Compare controlled-vocabulary indexing and free-text indexing.	05	CO2
Q 4	Differentiate between concept of Query and Information Need. Differentiate by using concept of Relevance feedback and Information Gap between them.	05	CO4

SECTION B

Q 5	Given the following four documents in our archive, where the first two documents are stored in machine one and the second two documents are stored in machine two, draw the procedure of inverted index construction, and the resulting inverted index. new home sales top forecasts home sales rise in July increase in home sales in July	10	CO4

July new home sales rise

Q 6 Illustrate the importance of Page Rank in Information Retrieval. Given the following hyperlink structure among five web pages, demonstrate the step by step construction of the transition matrix for PageRank with dumping factor d= 0.8.



10 CO1

Q 7 Why canonical tree is used in the Huffman Coding? Encode the given string ***"That house has a garden. The garden has many flowers. The flowers are beautiful"*** using the standard Huffman coding. 10 CO2

OR

The following table shows the output of an information retrieval system on two queries. You can assume that there are no relevant documents in ranks lower than the top 10 ranks shown. Calculate Average precision for these two queries, using an appropriate interpolation method if necessary, and sketch it in a precision-recall graph.

Rank	Q1	Q2
1	X	–
2	–	X
3	–	–
4	X	X
5	X	–
6	–	X
7	–	X
8	–	–
9	–	X
10	–	X

CO3

Q 8 Construct B tree and B+ tree for 1, 4, 7, 10, 17, 21, 31, 25, 19, 20, 28, 42 with n=4. 10 CO4

SECTION-C

Q 9 You work for a large company where there are many meetings, both of internal staff 20 CO4

and between staff and external clients. Meetings are recorded in formal minutes. The company's files of minutes are large, and the material has to be kept for many years since it may be necessary to check back on decisions taken early in large projects.

You are asked to design a retrieval system so that company staff can locate minutes on a particular topic. Because of the legal implications that past discussions and decisions may have, the company is particularly concerned that the new retrieval system will be reliable and effective. Outline the design of your system, indicating the particular features it will have that are intended to meet the company's requirements (you can assume that minutes are always clearly dated and have explicit lists of participants).

The company is willing to allow the installation of a pilot system so your approach can be evaluated under realistic conditions. Describe, in detail, your design for the evaluation: what data, operational conditions and aspects of your system would you consider, and why? What performance measures would you apply, and why?

Q 10 Consider the following hypothetical information retrieval scenario. Suppose it has been found at Edinburgh Royal Infirmary that due to equipment malfunction, the results of blood tests taken on 2013-12-04 are unreliable for diabetic patients. The hospital would like to contact all diabetic patients who had any kind of blood test on that day, to repeat the test. The hospital uses an information retrieval system to identify these patients. Suppose the collection of patients' medical records contains 10000 documents, 150 of which are relevant to the above query. The system returns 250 documents,125 of which are relevant to the query.

(a) Calculate the precision and recall for this system, showing the details of your calculations.

(b) Based on your results from (a), explain what the two measures mean for this scenario. How well would you say that the hospital's information IR system works?

20 CO2

(c) According to the precision-recall tradeoff, what will likely happen if an IR system is tuned to aim for 100% recall?

(d) For the given scenario, which measure do you think is more important, precision or recall? Why? Given your answer, what value would you give to the weighting factor α when calculating the F-score measure for the hospital's IR system?

OR

Explain Precision and Recall in detail. If for a collection of 50 retrieved documents, it is found that only 20 documents are relevant. What is the probability that the document retrieved after firing a query will be non-relevant? What will be the probability that relevant document is never retrieved? CO1

If the relevancy is calculated on searching the term “ASSIGN” and out of 20 relevant documents, it is found that 5 documents are found to contain the term then find out how much important is the term “ASSIGN” to the searcher.

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 : CSEG393 | **Max. Marks: 100**
Nos. of page(s) : 04

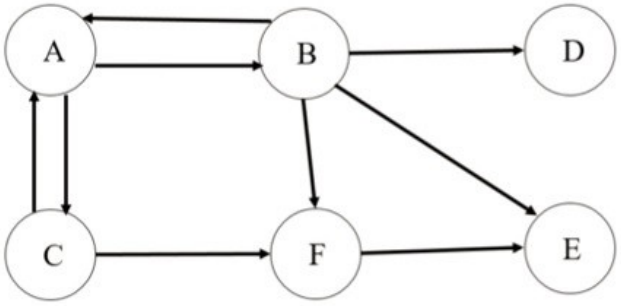
SECTION A

S. No.		Marks	CO
Q 1	Twitter is building its next generation of tweet search engine. Engineers are arguing that classical IR techniques are sufficient to build it so that there is no need for them to work overtime. As Twitter's chief research scientist in IR, do you agree with them? If not, what are the major technical challenges? What has to be innovated and what has to be adapted in such a system?	05	CO1
Q 2	Walmart lab is conducting an important research project to improve walmart.com's product search effectiveness. Multiple teams claim their algorithms are the best and should be deployed. As the manager of this project, what would be your judging criteria? How should you make a reasonable decision accordingly?	05	CO3
Q 3	Compare free-text-indexing and controlled-vocabulary indexing.	05	CO2
Q 4	Illustrate Web Crawler Indexer Architecture.	05	CO4

SECTION B

Q 5	Given the following four documents in our archive, where the first two documents are stored in machine one and the second two documents are stored in machine two, draw the procedure of inverted index construction, and the resulting inverted index. new home sales top forecasts home sales rise in July increase in home sales in July July new home sales rise	10	CO4

Q 6 Illustrate Page Rank in context with Information Retrieval. Given the following hyperlink structure among six web pages, demonstrate the step by step construction of the transition matrix for PageRank with dumping factor d= 0.8.



10 CO1

Q 7 Why canonical tree is used in the Huffman Coding? Encode the given string ***"for each rose a rose is a rose."*** using the standard Huffman coding. 10 CO2

OR

The figure below shows the output of two information retrieval systems on the same two queries in a competitive evaluation. The top 15 ranks are shown. Crosses correspond to a document which has been judged relevant by a human judge; dashes correspond to irrelevant documents. CO3

System 1		
Rank	Q1	Q2
1	–	X
2	X	–
3	X	–
4	X	–
5	–	–
6	–	–
7	–	–
8	X	–
9	X	–
10	X	–
11	X	–
12	–	–
13	–	X
14	–	X
15	X	–

System 2		
Rank	Q1	Q2
1	X	X
2	X	–
3	X	–
4	–	X
5	X	X
6	X	–
7	–	–
8	–	–
9	–	–
10	–	–
11	X	–
12	X	–
13	–	–
14	–	–
15	X	–

(a) Explain the following evaluation metrics and give results for query Q1 for both systems.

(i) Precision at rank 10

(ii) Recall at precision 0.5

(b) The metrics in part (a) above are not adequate measures of system performance for arbitrary queries. Why not? What other disadvantages do these metrics have?

(c) Give the formula for mean average precision (MAP), and illustrate the metric by calculating System 1's MAP.

(d) For each system, draw a precision-recall curve. Explain how you arrived at your result.

Q 8 Perform insertion in B+ tree for 3, 5, 7, 13, 12, 21, 34, 25, 16, 20, 28, 43 with n=4. Also, show the deletion for 5, 7, 21, 25, 20, and 43. 10 CO4

SECTION-C

Q 9 You work for a company that takes news stories from all over the world and provides reports on specified topics to customers, for example on political developments in the new Republic of Rumbaza during 1997. Your company's staff use a retrieval system to extract the material on which they base their reports from the company's very extensive archive of stories. The retrieval system is a 20-year-old makeshift and is to be scrapped. You are asked to design the new retrieval system.

(a) Give a detailed description of the retrieval devices the new system will offer, explaining why they are available and how they will work. What facilities will the user have for modifying his or her search specification in response to system output? 20 CO4
(b) What do you regard as the most difficult problem to be tackled, and why?

[Assume that the stories in the file are texts between 1 and 100 sentences long, with Date and source headers in a standardized form.]

OR

Given the following 5 documents and 1 queries, rank the documents for each of the queries by the Boolean and vector space model. Suppose that log(x) denotes log10(x). Note that idf_i is given by log (N/n_i) CO2

Documents

Document #	Document
1	Peas pudding hot, peas pudding cold
2	Peas pudding in the pot,
3	Nine days old.
4	Peas pudding, peas pudding,
5	Eat the lot.

Queries

Query #	Queries
1	hot pudding
2	Pudding
3	Eat

Consider the following equation for query terms for vector space model.

$w_{iq} = (0.5 + 0.5 * [freq_{iq} / max(freq_{lq})]) * log10 (N/n_i)$ when $freq_{iq} > 0$

$= 0$ otherwise

Q 10	Given the query Q=" good computer software programmers", compute the similarity between Q and the following documents, if we use tf-idf weights for the document, binary weights for the query, and the cosine measure. Determine their relative ranking: D1 = programmers write computer software code D2 = most software has bugs, but good software has less bugs than bad software D3 = some bugs can be found only by executing the software, not by examining the source code D4 = good programmers write good computer software code.	20	CO1