Roll No: ----------

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

End Semester Examination, December 2017

Program: B.Tech. (CSE) All IBM Branches — **Semester – V**
Subject (Course): Information Retrieval and Search Engines — **Max. Marks : 100**
Course Code: CSEG393 — **Duration : 3 Hrs**
No. of page/s: 3

Section-A **[4 x 5 = 20 Marks]**

Q1. Explain the difference between concept of a query and an information need.

Q2. Name the technique, which substitute the words with their respective stems. Mention the different method for the same technique

Q3. How does Zipf's law ensure effective inverted index compression?

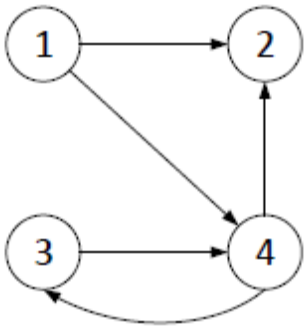
Q4. Generate the Huffman tree for the given data

character	Frequency
a	5
b	9
c	12
d	13
e	16
f	45

Q5. Discuss the steps used in text processing for indexing.

Section-B **[10 x 4 = 40 Marks]**

Q6.Given the following hyperlink structure among four web pages, demonstrate the step by step construction of the transition matrix for PageRank with dumping factor d= 0.1. Write down the intermediate results.



Q7. 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.
Q8. Bing wants to become a better personalized search engine. Combining the concepts and techniques that we have learned in this semester, please give your concrete suggestions of where and how an IR system can be personalized. You need to cover at least three components in a typical retrieval system, e.g., query processing module, ranking functions, and feedback modeling.
Q9. We have an indexed collection of one million documents that includes the following terms:

Term	# docs
computing	300,000
networks	200,000
computer	100,000
files	100,000
system	100,000
client	80,000
programs	80,000
transfer	50,000
agents	40,000
p2p	20,000
applications	10,000

a). Compute the similarity between the following documents D1 and D2 using tf-idf weights and the cosine measure:
D1 = "p2p programs help users sharing files, applications, other pro-grams, etc. in computer networks"
D2 = "p2p networks contain programs, applications, and also files"

b). Assume we are using the cosine measure and tf-idf weights to compute document similarity. Give a document containing two different terms exactly that achieves maximum similarity with the following document "p2p networks contain programs, applications, and also files".

Compute this similarity and justify that it is indeed maximum among documents with two terms.

Section-C **[20 x 2 = 40 Marks]**

Q10. 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 2017-11-24 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?
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?

Q11. Write a short note on following (Any four):
a) Digital Library
b) Spatial Access Method
c) HITS algorithm
d) Compression model and its types
e) Information retrieval models (any three)

Roll No: ----------

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

End Semester Examination, December 2017

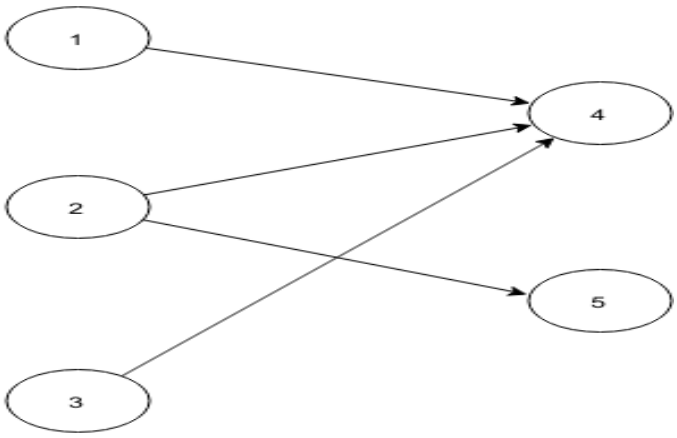
Program: B.Tech. (CSE) All IBM Branches | **Semester – V**
Subject (Course): Information Retrieval and Search Engines | **Max. Marks : 100**
Course Code: CSEG393 | **Duration : 3 Hrs**
No. of page/s:2

Section-A **[5 x 4 = 20 Marks]**

Q1.Explain the web search techniques in detail.
Q2. Differentiate between Zipf's law and Heap's law.
Q3. Discuss the priciples that are of special interest to human computer interface in information access system.
Q4. What is collaborative and content filtering? Explain with the help of example.
Q5. Determine the Euclidean and Manhattan distance between the vector (3,2,4,6,5) and (6,8,3,5,7).

Section-B **[10 x 4 = 40 Marks]**

Q6. Calculate the best hub and authority for k=3, assume u=1.



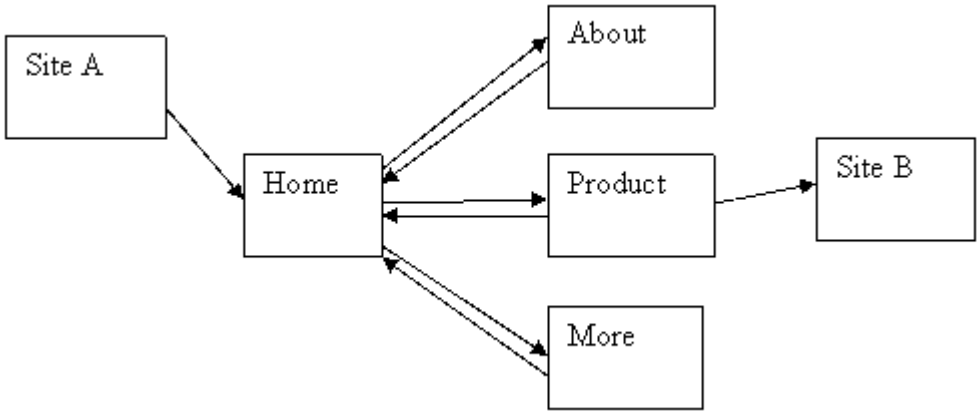
Q7. Explain the crawler-indexer architecture.
Q8. Discuss the two elements of inverted index. Given the following four documents in our archive, where the first two documents are stored in machine one and the second two documents are stored in machine two, draw the resulting inverted index.

a) new home sales top forecasts
b) home sales rise in july
c) increase in home sales in july
d) july new home sales rise

Q9. Why canonical tree is used in the Huffman Coding. Encoded the given string " for each rose, a rose is a rose " using the standard Huffman coding.

Section-C **[20 x 2 = 40 Marks]**

Q10. Explain the concept of PageRank and its relevance for web search. Given the following structure among six web pages, demonstrate the stepwise transaction matrix for page rank with dumping factor d=0.85.



Q11. Write a short note on following (Any four):

f) Extended Boolean Model
g) GEMINI algorithm
h) Web Crawler
i) Meta Search Engine
j) Digital Library Architecture