Information Retrieval (CS60092)

Mid-Semester Examination

Maximum Marks 45

This question paper has 4 pages and 5 questions.

Note: There are no clarifications. In case of doubt, you can take a valid assumption, state that properly and continue.

Please answer all subparts for a given question at one place only.

In case of multiple attempts, cross any attempt that you do not want to be graded for.

Question 1

Consider the following terms and their document frequency values in brackets

w_1 (n_1), w_2 ($2n_1$), w_3 ($3n_1$), w_4 ($4n_1$), w_5 ($5n_1$), w_6 ($6n_1$)

Assume the size of vocabulary to be N >>n_1

Approximately how many operations will the following queries take in the worst case using the best method of processing these? Also, what will be a good upper bound on the number of relevant documents retrieved. **[2+2+2 = 6]**

(a) w_1 AND (NOT w_3)

(b) (w_1 OR w_3) AND (w_2 OR w_6) AND (w_4 AND w_5)

(c) (w_1 AND w_5) OR (w_3 AND w_6)

Question 2

Answer the following questions. **[2+2+4+3 = 11]**

a) Explain how permuterm index would be used to match the wildcard query *ntr* with the words, "intro" and "intra" while avoiding "train" (which can be permuted to "intra").
b) Consider a misspelled query term, "heathrw". How would an IR system produce the candidate terms within an edit distance of 1? What would be the complexity? Let *V* be the size of the term voabulary, and the dictionary be stored as a balanced binary tree.
c) Suppose you want to use n-gram overlap method for spelling correction using a Jaccard overlap threshold of *k (0<k<1)*. Explain how you would modify the standard posting merge method.
   i). Would you use binary merges as in standard posting merge or a multi-way merge?
   ii). Provide a brief sketch of your algorithm.

d) Draw a balanced binary tree to store the following words in the dictionary: {"apple", "and","ear","eat"}. Please use the same format for the binary tree as discussed in the class and the book.

Question 3

For the query "auto insurance", find the relevance score for the document "car premium insurance" using the ***smart notation Lnc.Lnc*** *(uppercase in L is just for clarity)* for vector space model. Assume that the corpus contains 1 million vocabulary terms, and the document frequency for {auto, insurance, premium, car} is {1000, 100, 100, 10000}, respectively. Show the computation steps clearly. **[5]**

Question 4

Assume that your collection contains 100 documents out of which only 5 documents are relevant for two very similar queries, Q_1 and Q_2. Also, each of these 5 documents have the same relevance (assume the relevance to be 1 for each of these 5 documents). Suppose your system is only able to retrieve 5 documents for both the queries. The manual judgements for the documents retrieved at top 5 ranks are shown in the table below. R denotes a relevant document as per relevance judgement, and NR denotes a non-relevant document.

Rank	Q_1	Q_2
1	R	NR
2	R	R
3	NR	R
4	R	NR
5	NR	NR

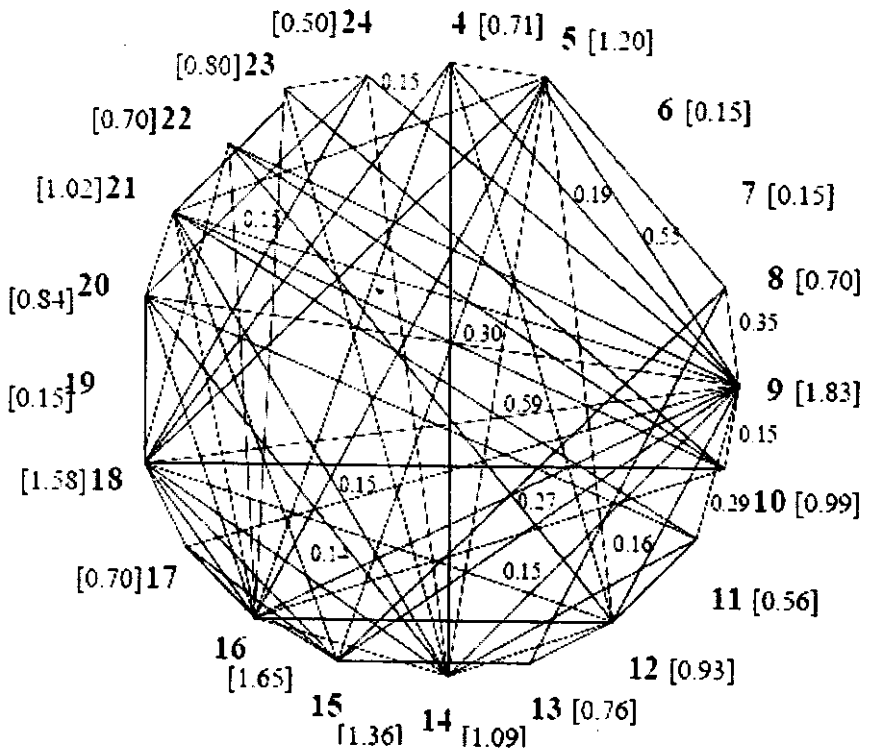
Compute P@5, R@5, MAP and NDCG@5 of your system. **[1.5+1.5+2.5+2.5 = 8]**

Question 5

a) Consider the following 24 sentences from which an extractive summary needs to be generated. Outline the basic steps of how a similarity graph can be constructed from these. In particular, how are the nodes and edges defined in this network? What is the total number of possible edges in this graph? **[1+2+1=4]**

3: BC-HurricaineGilbert, 09-11 339
4: BC-Hurricaine Gilbert, 0348
5: Hurricaine Gilbert heads toward Dominican Coast
6: By Ruddy Gonzalez 7: Associated Press Writer
8: Santo Domingo, Dominican Republic (AP)
9: Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas
10: The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
11: "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday.
12: Cabral said residents of the province of Barahona should closely follow Gilbert's movement.
13: An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
14. Tropical storm Gilbert formed in the eastern Carribean and strenghtened into a hurricaine Saturday night.
15: The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
16: The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.
17. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
18: Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.
19: There were no reports on casualties.
20: San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
21: On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast.
22: Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.
23: Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
24: The first, Debby

b) Very briefly outline the basic idea of the TextRank algorithm. If we run the TextRank algorithm on the graph so constructed we get the TextRank values as presented in the square brackets in the figure below. Based on this, if we have to generate a 100 word summary, write down the whole summary. **[2+5=7]**



c) Suppose there are two experts who have written the following two summaries form these 24 sentences. To which among these is the TextRank summary more close? Justify your answer (qualitative justification should be okay). **[2+2=4]**

Manual abstract I

Hurricane Gilbert is moving toward the Dominican Republic, where the residents of the south coast, especially the Barahona Province, have been alerted to prepare for heavy rains, and high wind and seas. Tropical storm Gilbert formed in the eastern Carribean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with winds of 75 mph. Flooding is expected in Puerto Rico and in the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and down-graded to a tropical storm.

Manual abstract II

Tropical storm Gilbert in the eastern Carribean strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday to be about 140 miles south of Puerto Rico and 200 miles southeast of Santo Domingo. It is moving westward at 15 mph with a broad area of cloudiness and heavy weather with sustained winds of 75 mph gusting to 92 mph. The Dominican Republic's Civil Defense alerted that country's heavily populated south coast and the National Weather Service in San Juan, Puerto Rico issued a flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.