Utkal University

MTech (CSE), 3rd Semester-2020

Natural Language Processing(3.5)

Full Marks: 70 **Time: 3 Hours**

Q1

(a) Draw the inverted index that would be built for the following document collection.
Doc 1 new home sales top forecasts
Doc 2 home sales rise in july
Doc 3 increase in home sales in july
Doc 4 july new home sales rise
Doc 5 increase in produc sales in july

(b) What is the role of Dictionary and stemmer in retrieval of documents? [7+7]

OR

(a) How should the Boolean query *x* and not *y* be handled? Why is naive evaluation of this query normally very expensive? Write out a postings merge algorithm that evaluates this query efficiently.

(b) How could a search system combine use of a positional index and use of stop words? What is the potential problem, and how could it be handled? [7+7]

Q2

(a) What is the idf of a term that occurs in every document? Compare this with the use of stop word lists.

(b) If we were to stem *poni* and *ponies* to a common stem before setting up the vector space, detail how the definitions of tf and idf should be modified. [7+7]

OR

(a) Consider the four-term query *catched in the rye* and suppose that each of the query terms has five alternative terms suggested by isolated term correction. How many possible corrected phrases must we consider if we do not trim the space of corrected phrases, but instead try all six variants for each of the terms?

(b) Compute the edit distance between Paris and India. Write down the 5 × 5 array of distances between all prefixes as computed by the algorithm. [7+7]

Q3

(a) Compute variable byte and γ codes for the postings list 7, 17, 29, 31,45,90, 110. Use gaps instead of docIDs where possible. Write binary codes in 8-bit blocks.

(b) From the following sequence of γ -coded gaps, reconstruct first the gap sequence and then the postings sequence: 1110001110101011111011011110110. [7+7]

OR

(a) Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3 in Table II. Compute the tf–idf weights for the terms car, auto, insurance, and best, for each document, using the idf values from Table I.

Table-I

term	df*t*	idf*t*
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

Table-II

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

(b) From the following sequence of γ -coded gaps, reconstruct first the gap sequence and then the postings sequence: 1110001011010101011111011011110 11. [7+7]

Q4

(a) An IR system returns eight relevant documents and ten nonrelevant documents. There are a total of twenty relevant documents in the collection. What is the precision of the system on this search, and what is its recall?

(b) The Dice coefficient of two sets is a measure of their intersection scaled by their size (giving a value in the range 0 to 1):

$$\text{Dice}(X, Y) = 2|X \cap Y|/|X| + |Y| .$$

Show that the balanced F-measure (F1) is equal to the Dice coefficient of the retrieved and relevant document sets.

(c) Consider an information need for which there are four relevant documents in the collection. Contrast two systems run on this collection. Their top ten results are judged for relevance as follows (the leftmost item is the top ranked search result):

System1 R N R N N N N N R R
System2 N R N N R R R N N N

a. What is the MAP of each system? Which has a higher MAP?
b. Does this result intuitively make sense? What does it say about what is important in getting a good MAP score?
c. What is the R-precision of each system? (Does it rank the systems the same as MAP?) [4+4+6]

OR

(a) What is a crawler? Explain how the various fundamentals step like tokenization, linguistic modules and indexer are handled by the crawler. [14]

Q5 Write short notes on the following:

(a) Page Rank Algorithm

(b) HITS Algorithm [7+7]

OR

(a) A user of a browser can, in addition to clicking a hyperlink on the page x he is currently browsing, use the back button to go back to the page from which he arrived at x. Can such a user of back buttons be modeled as a Markov chain? How would we model repeated invocations of the back button?

(b) Consider a web graph with three nodes 1, 2, and 3. The links are as follows: 1→2, 3→2, 2→1, 2→3. Write down the transition probability matrices for the surfer's walk with teleporting, for the following three values of the teleport probability: (i) $\alpha = 0$; (ii) $\alpha = 0.5$; and (iii) $\alpha = 1$. [7+7]