DEC - 2018

Total No. of Questions : 12]

SEAT No. :

P3961

[Total No. of Pages : 2

[5462]-686
M.E. (Computer Engineering)
INFORMATION RETRIEVAL
(2017 Pattern) (Semester - III) (610102)

Time : 3 Hours] *[Max. Marks : 50*

Instructions to the candidates :

1) *Answer Q.1 or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q.7 or Q.8, Q.9 or Q.10 and Q.11 or Q.12*
2) *Neat diagrams must be drawn necessary.*
3) *Figures to the right indicate full marks.*
4) *Assume suitable data, if necessary.*

Q1) a) Write down the entries in the permuterm index dictionary that are generated by the term *mama*. **[4]**

b) If $|si|$ denotes the length of string si, show that the edit distance between s1 and s2 is never more than $\max\{|s1|, |s2|\}$. **[5]**

OR

Q2) a) Explain k-gram indexes for wildcard queries with suitable example. **[5]**

b) Find two phonetically similar proper nouns whose soundex codes are different. **[4]**

Q3) How would you create the dictionary in blocked sort-based indexing on the fly to avoid an extra pass through the data? **[8]**

OR

Q4) State the statistical properties of terms in information retrieval and explain it with suitable example. **[8]**

Q5) Consider the postings list (4, 10, 11, 12, 15, 62, 63, 265, 268, 270, 400) with a corresponding list of gaps (4, 6, 1, 1, 3, 47, 41, 202, 3, 2, 130). Assume that the length of the postings list is stored separately so the system knows when a postings list is complete.Using variable byte encoding: **[8]**

a) What is the largest gap you can encode in 1 byte?

b) What is the largest gap you can encode in 2 byte?

c) How many bytes will the above posting list require under this encoding? (Count only space for encoding the sequence of numbers.)

OR

P.T.O.

Q6) a) When using weighted zone scoring, is it necessary for all zones to use the same Boolean match function? **[4]**

b) In above question 6(a) with weights g1 = 0.2, g2 = 0.31 and g3 = 0.49 what are all the distinct score values a document may get? **[4]**

Q7) a) Why is the IDF of a term always finite? **[4]**

b) What is the IDF of a term that occurs in every document? Compare this with the use of stop word lists. **[4]**

OR

Q8) If we were to stem jealous and jealousy to a common stem before setting up the vector space, detail how the definitions of TF and IDF should be modified. **[8]**

Q9) a) Write down all the structural terms occuring in the XML document in below fig. 1.0 **[5]**

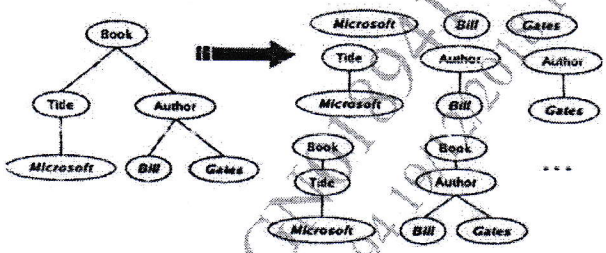

Figure 1.0

b) Differentiate language modeling versus other approaches in IR. **[4]**

OR

*Q10)*Enlist the different language models and explain it with suitable examples. **[9]**

*Q11)*Explain the Naive Bayes text classification algorithm. **[8]**

OR

*Q12)*Consider making a language model from the following training text: *the martian has landed on the latin pop sensation ricky martin* **[8]**

a) Under aMLE-estimated unigram probability model, what are P(the) and P(martian)?

b) Under the aMLE-estimated bigram model, what are P(sensation/pop) and P(pop/the)?