1128
M.E. (Computer Science and Engineering)
Third Semester
Elective – V
CS-8304: Information Retrieval

Time allowed: 3 Hours   **Max. Marks: 50**

NOTE: Attempt five questions in all, including Question No. I which is compulsory and selecting two questions from each Section.

x-x-x

1   **(i)** What is stemming? How it differs from lemmatization   02

   (ii) What are skip pointers?   02

   (iii) What do you mean by ranking of documents with respect to a query?   02

   (iv) What is the purpose of robot.txt file in crawling process?   02

   (v) What do you understand by document zone?   02

Section – A

2   **a)** Consider the following posting list for two terms:

Term 1 : [4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180]   05

Term 2 : [47]

Calculate number of comparisons required to perform two-word query if the posting list is stored with skip pointers with skip length of 'p/5'.   'p' is length of posting.   05

b) What is bag of words model? Explain with an example.

3   **a)** What is Euclidean distance? Why it is bad to use Euclidean distance for calculating vector proximity?   05

b) What is inverse document frequency (idf)? Do idf has some impact on one term queries as compare to document frequency?   05

4   **a)** What do you mean by tolerant retrieval? Explain how the wild card queries can be processed.   05

b) What so you mean by case folding? Explain with example.   05

Section – B

5   Draw and explain the architecture for a crawler.   10

6   **a)** What is cosine similarity? How it used to calculate similarity between two documents?   05

b) Some time rare terms in a collection are more informative than the frequent terms. How rarity of a term can be accounted for ranking the documents?   05

7   **a)** How does a k-Nearest Neighbor technique make predictions about new data points? How does a distance-weighted k-Nearest Neighbor technique differ from a standard k-Nearest Neighbor technique?   05

b) What is cloaking? How it used to fool the spiders?   05

x-x-x