VIT
Vellore Institute of Technology

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Continuous Assessment Test – I - Winter Semester 2019-2020

Programme Name & Branch: B. Tech. (CSE)

Course Name & Code: CSE3024 Web Mining

Class Number: VL2019205001934, VL2019205001939 Slot: G2+TG2

Faculty Name: Prof. S. M. Satapathy, Exam Duration: 90 Minutes Maximum Marks:50
Prof. Anbarsi M.

ANSWER ALL THE QUESTIONS

S.No.	Question	Course Outcome (CO)
1.	Consider the following document corpus DOC1: italy ~~is~~ world champion 2006 DOC2: germany ~~and~~ italy played ~~each~~ ~~other~~ ~~in~~ ~~the~~ semifinal DOC3: germany ~~was~~ ~~in~~ ~~the~~ semifinal 2006 DOC4: germany won ~~the~~ semifinal ~~in~~ italy 1990 a) Assume that the following terms are stop words: is, and, in, the, was, each, other. Considering a vector space model, rank the documents according to the query **"italy semifinal"** using Cosine and Euclidean measures. **(15 marks)** b) Construct a trie for the corpus shown above. **(05 marks)**	CO2
2	a) How can the owner of a web site design a "spider trap"? Is it possible for a crawler to distinguish between a spider trap and a legitimate web site with 100% certainty? Explain. Describe one method that a crawler might use to avoid spider traps. What are the drawbacks of this method? **(5 marks)** b) Explain why it is important for a crawler to detect whether two pages that it has downloaded are "near duplicates". Give two reasons that a crawler would want to record the URLs of all the near duplicate pages it has downloaded, rather than discard them. How does the existence of near duplicate pages affect the computation of PageRank? **(5 marks)**	CO2
3	a) 20 documents are retrieved on the basis of query Q. Assume that the precision is 0.40 and the recall is 0.25 for this retrieval. What is the total number of relevant documents in the collection for query Q? **(2.5 marks)**	CO2

	b) A spam filter checks 1000 emails. 465 of them are spam. The spam filter tags 386 emails as spam, but only 322 of them really are. What is the performance of the spam filter in Precision, Recall? After some learning, the spam filter checks again the 1000 emails. Now it tags 267 as spam and 255 of them really are. Does the spam filter perform better now? **(2.5marks)**	CO2
	c) **State TRUE or FALSE** **(5 marks)** i. Stemming increases the size of the vocabulary. ii. In the bag of words model, the exact ordering of terms within the document is both significant and relevant to processing. iii. Precision in an information retrieval system refers to the fraction of relevant documents in the collection that were returned by the system. iv. Stop words are used to delimit the word segments in a sentence. v. The IDF value depends on the term frequency and the document frequency.	CO1
4	a) What are the different issues for implementation of a crawler? Describe it in detail with suitable example. **(5 marks)** b) Decode the Golomb encoded sequence of numbers 011100001010 with b=10. Encode the same decoded sequence of numbers with b=5. **(5 marks)**	CO1