VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Continuous Assessment Test - I, AUG- 2018

Course Code	: CSE3024	**Duration**	: 90 Minutes.
Course Name	: Web Mining	**Max. Marks**	: 50
Slot	: F2		

Answer ALL(5 x 10 = 50 Marks)

1. The Web has many unique characteristics, which make mining useful information and knowledge a fascinating and challenging task. Justify the statement with suitable example. (10)

2. Prepare a document object model (**DOM**) tree from following HTML page. Explain focused Crawler in detail with a neat diagram. Also mention how "soft strategy" differs from "hard-strategy". (3+7)

```
<html>
  <head>
    <title>Here comes the DOM</title>
  </head>
  <body>
    <h2>Document Object Model</h2>
    <img align="right" alt="dom pict" src="dom.png">
    <p>
      This is a simple
      <code>HTML</code>
      page to illustrate the
      <a href="http://www.w3.org/DOM/">DOM</a>
    </p>
  </body>
</html>
```

3. a) Encode the decimal numbers 9, 10, and 11 using Elias Gamma, Elias Delta and Golomb (b=3 and b=10).

   b) *Using b= 10* decode this Golomb encoded sequence 11010011000011110. Also find the sequence of integers in each step. (5+5)

4. For the following table find the

   a)TF coordinate matrix **b)** IDF for all terms **c)** TF-IDF MATRIX for all documents(ignore any normalization) **d)** Calculate cosine similarity and Euclidian distance with given query vector q=(0,0,0,0.871,0.371) and rank the documents. (3+1+2+4)

	Term1	Term2	Term3	Term4	Term5
D1	7	2	24	30	5
D2	0	40	10	3	0
D3	0	20	30	40	30
D4	2	3	5	1	0
D5	4	13	3	33	49
D6	100	12	2	54	1

5. Write short notes on

   a) Spider trap **b)** Inverted index **c)** bag-of-words

   d) link-cluster conjecture **e)** Topical Locality and Cues (2X5)