LNMIIT
The LNM Institute of Information Technology

The LNM Institute of Information Technology

Department of Computer Science and Engineering

Information Retrieval (IR)
End Semester Exam

Time: 3 Hour **Date: 29/11/2018** **Max. Marks: 40**

Instructions: 1) Look through the whole exam and answer the questions that you find easiest first.
2) If necessary, you may make assumptions that are reasonable, and if you do make an assumption, state it clearly.
3) You may use a calculator.

Q1. Draw the Decision tree for the data sample given in the below table: **[5 + 1]**

income	age	student	Credit_rating	Buys_computer
high	<=30	no	fair	No
high	<=30	no	excellent	No
high	31…40	no	fair	Yes
medium	>40	no	fair	Yes
low	>40	yes	fair	Yes
low	>40	yes	excellent	No
low	31…40	yes	excellent	Yes
medium	<=30	no	fair	No
low	<=30	yes	fair	Yes
medium	>40	yes	fair	Yes
medium	<=30	yes	excellent	Yes
medium	31…40	no	excellent	Yes
high	31…40	yes	fair	Yes
medium	>40	no	excellent	No

Using the drawn decision table with above training data Classify the test sample **X= {(age<=30, Income=medium, Student=yes , Credit_rating=Fair) as buys_computer "Yes" or "No".**

Q2. Describe various text preprocessing methods used in Information Retrieval. **[5]**

Q3. What is **Laplace smoothing in Naïve Bayes classification**? There are two classes **"ham" and "spam"**. Use naïve bays method and the training data given below:

Training Data

Document	Terms	Class label
D1	"good"	ham
D2	" very good"	ham
D3	"bad"	spam
D4	"very bad"	spam
D5	"very bad, very bad"	spam

Test document D6: "good ? bad! Very bad", classify this document as **ham** or **spam**. **[1+5]**

Q4. Derive the confusion matrix for the following clusters and calculate the **purity, Rand Index, NMI and F-Measure.**

Cluster1= **{R, R, R, G, G, B}** Cluster2= **{G, G, G, G, B}** Cluster3 = **{R, R, B, B, B, B}** **[2+1+2+3+2]**

Q5. **[1 mark]**

5.1. Social network analysis consists of
- a. Numerical analysis
- b. Content analysis
- c. Structure analysis
- d. None of the above

5.2. Which of the following are the quality control measures tried in crowdsourcing? **[1 mark]**
- a. Assessment of the requester's behavior
- b. Assessment of the crowd
- c. Identification of the right answer from the crowd
- d. Identification of the right request from the requester

5.3. Assume that you have a group that generated 300 tweets and the number of tweets in the largest polarity cluster and the largest emotion cluster are 50 and 100 respectively. Calculate the abstraction and expression scores. Also give the probabilities of this group to be a crowd, herd, mob and gang. **[3 marks]**

Q6. **Write short notes on any two with suitable example:** **[2X4]**
- a) **Latent Semantic Indexing.**
- b) **Recommender System.**
- c) **Hierarchical clustering**
- d) **Any popular social media (only technical aspect)**

{Best of Luck}