**The LNM Institute of Information Technology**

**End Semester Exam-2015**

**Information Retrieval and Web Search(IRWS**)

**Max Marks: 40 Time: 3 Hr.**

**Note: All questions are compulsory.**

**Q1** a) Precision and recall are computed as follows: P = TP / ( TP + FP) and R = TP / (TP + FN) Name and define the three values TP, FP, FN appearing here. Also discuss F-measure.

b) Suppose 200 college students are to be classified based on height. In actuality, there are 60 tall students and 140 who are not tall. A classification technique classifies 115 students as tall and 85 as not tall.  Draw the confusion matrix and calculate the precision and recall as well as F-measure. **[3+4]**

**Q2**. Consider the following pages and the set of web pages that they link to:

* Page A points to page E.
* Page B points to pages D and E.
* Page C points to pages D and E.

Consider running the HITS (Hubs and Authorities) algorithm on this subgraph of pages. Simulate the algorithm for three iterations. Show the authority and hub scores for each page twice for each iteration, both before and after normalization, order the elements in the vectors in the sequence: A, B, C, D, E.  **[6]**

**Q3.** a)How is “soft” clustering different from traditional “hard” clustering?

b)Consider the problem of clustering the following documents using complete-link hierarchical agglomerative clustering.

• Doc1: alpha alpha gamma

• Doc2: alpha beta

• Doc3: alpha gamma

• Doc4: alpha alpha

Show all similarity calculations needed to cluster the documents, and the final cluster hierarchy. Apply normalization and tf-idf weights with cosine similarity. **[2+6]**

**Q4.** The PageRank algorithm uses a model of a “random surfer” to calculate the importance or validity of a page. Describe how the random surfer can be modelled as an ergodic Markov chain, and how this leads to the PageRank values being calculated as a principal left eigenvector of the transition probability matrix. **[6]**

**Q5. a)** Why is naive Bayes text categorization typically implemented using logarithms of probabilities rather than probabilities themselves?

**b)** State theassumptions in naïve Bays classification. Use the below table data for training your classifier using naïve Bays (for **play** Tennis or **don’t play**) and classify an unseen sample **X = <rain, hot, high, false>** in one of the class. **[2+6]**



**Q6**. Describe the opportunities and challenges in building XML based information retrieval system. Draw the DOM tree for the XML document given below. **[5]**

<play>

<author>Shakespeare</author>

<title>Macbeth</title>

<act number=“I”>

<scene number=“”vii”>

<title>Macbeth’s castle</title>

<verse>Will I with wine

…</verse>

</scene>

</act>

</play>