Information Retrieval and Data Mining

Prof. Matteo Matteucci, Ing. Luca Bondi

February, 05 2016

Very Important Notes

- Answers to questions 1, 2, and 3 should be delivered on a different sheet with respect to 4 and 5
- If you need a calculator this should not be to any extent programmable or network connected

1. **Question (8 pts)**: With reference to the Decision Tree model answer the following
   (a) What is a decision tree? Provide its description and explain its use
   (b) Describe the training algorithm for decision trees
   (c) Describe the *sub-tree rising*, *sub-tree replacement*, and *rule-based pruning* techniques for decision trees? What these techniques are useful for?
   (d) Describe the method used by C4.5 to compute the upper bound for the error on new data at a decision tree node given the corresponding node error on the training data.
2. **Question (5 pts)**: Consider the following dataset

TID	A	B	C	D	E
T1	1	1	1	0	0
T2	1	1	1	1	1
T3	1	0	1	1	0
T4	1	0	1	1	1
T5	1	1	1	1	0

   (a) Apply the a-priori algorithm to it and extract all the frequent itemset having support greater or equal to 50%.
   (b) Then take (one of) the largest itemset and extract at least one rule, if it exists, with confidence higher than 40%.
3. **Question (6 pts)**: Let consider a graph represented by the following adiacency matrix

$$E = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

   (a) Draw the corresponding graph
   (b) Compute the between centrality for each node
   (c) Is the graph ergodic? why should you care about that?

(d) Can you apply Pagerank algorithm to it? If not, why? If yes, how?

(e) Considering the graph as a Markov network. Write the corresponding transition probability matrix P and describe how could we compute the average time required to get in the absorbing state by randomly moving? (You are not require to compute it)

4. **Question (8 pts)**:

Consider the following tokenized documents collection

d_1 = "Sheldon—likes—comics"
d_2 = "Amy—likes—Sheldon—"
d_3 = "Howard—Leonard—friends"
d_4 = "Amy—Leonard—Sheldon—friends"
d_5 = "Sheldon—Leonard—comics—friends"

(a) Aswer the query q_1 = "(Amy OR Leonard) AND NOT comics" using a Boolean model

(b) Rank the documents with respect to queries q_2 = "Sheldon—comics" and q_3 = "Amy—Leonard—friends" using a Vector Space model. Consider term frequencies equal to the number of occurrences of the term in each document and use Cosine Similarity as distance metric.

5. **Question (5 pts)**: Explain what is rank aggregation. Describe and compare Borda's and Condorcet's algorithms with a small example.