Information Retrieval and Data Mining

Prof. Matteo Matteucci, Ing. Luca Bondi

July, 08 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 (6 pts)**:
   With reference to the *Sequential Covering Algortihm*, answer the following
   (a) What kind of model we can learn using the Sequential Covering Algorithm? Provide its description and explain its use
   (b) Describe the Sequential Covering Algorithm
   (c) Describe the Chi-Square test for pruning and its use in the context of the Sequential Covering Algorithm
2. **Question (5 pts)**: Consider the following dataset

Trans #	A	B	C	D	E	F	G
Trans 1	1	1	1	0	0	1	0
Trans 2	1	1	1	1	1	1	1
Trans 3	1	0	1	1	0	0	1
Trans 4	1	0	1	1	1	1	1
Trans 5	1	1	1	1	0	0	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%.
   (c) What does the "Confidence anti-monotone rule" says with respect to the number of items on the right hand side of a rule?
3. **Question (8 pts)**: Consider a graph that is described by the following adjacency matrix

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

(a) Compute the in-degree centrality index of each node

(b) Compute the betweenness centrality index of node 2

(c) Compute the closeness centrality index of each node

(d) Construct the matrix A for the Seeley index and compute the first iteration (k=1) of the algorithm to compute the corresponding steady state distribution, including the normalization step.

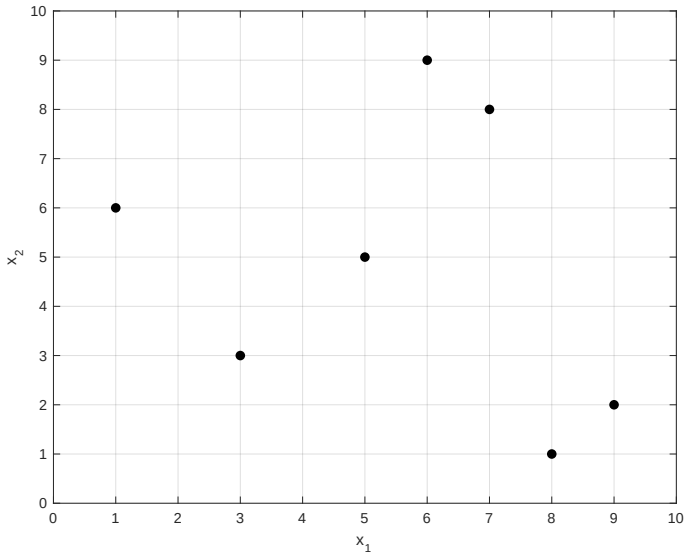
4. **Question (7 points)**
Given three rankings r_1, r_2, r_3 related to a collection of four documents $\{A, B, C, D\}$

r_1	r_2	r_3
A (0.9)	D (0.9)	A (0.8)
B (0.8)	A (0.5)	D (0.7)
D (0.7)	B (0.4)	C (0.3)
C (0.6)	C (0.1)	B (0.2)

(a) Compute the top-2 documents using the MedRank algorithm

(b) Compute the top-2 documents using the Fagin's algorithm

5. **Question (6 pts)**:
Given the following points collection



(a) Describe how to build a kd-tree index on the collection

(b) Describe how to answer a nearest neighbour query on the kd-tree index