Web information retrieval - 2017/2018

Exam - June 18th, 2018

Time: 60 minutes

Assignment 1

Consider the collection of the following 3 textual documents (D1, D2 and D3):

- D1: data mining and social mining
- D2: social network analysis
- D3: data mining

1.1: write down the postings lists corresponding to the above documents

1.2: write down document frequency (df) for each term

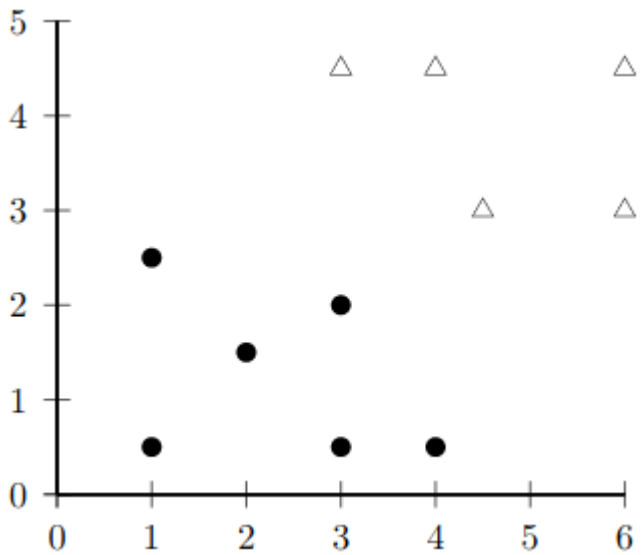
1.3: write down the term frequencies (tf) for document D1

1.4: write down the formula that relates the tf-idf weight for a term given its tf and idf

---

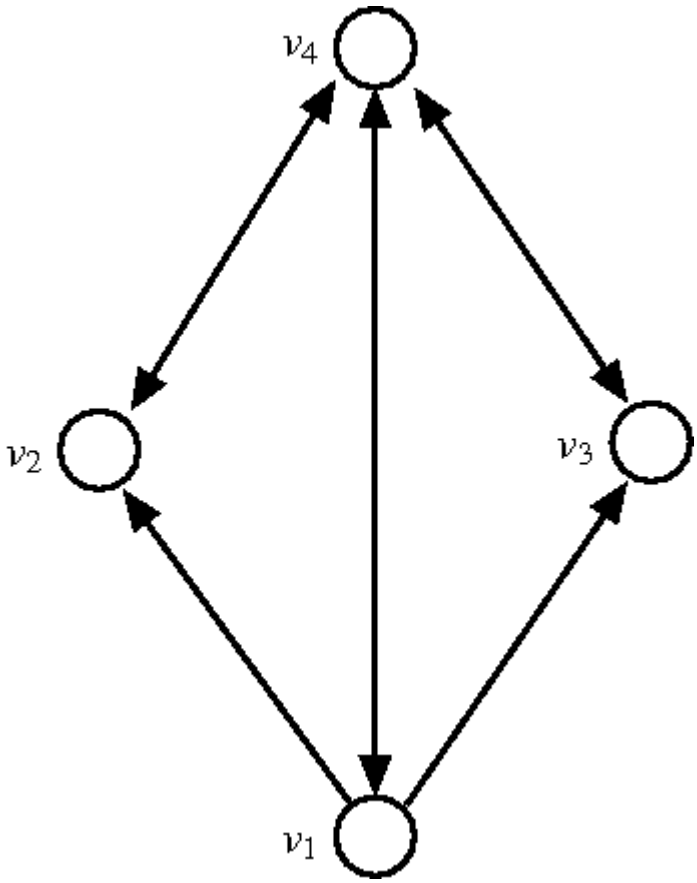
Assignment 2

In the picture below, identify the linear maximum margin (SVM) classifier for the binary problem triangle vs dot. *Draw three lines*: the two boundaries of the maximum margin and the maximum margin hyperplane. Which of the vectors are support vectors? You can solve this problem “visually” by drawing your solution into the figure.



Assignment 3

3.1. Work out the Pagerank equations to compute Personalized Pagerank vectors π_1 and π_2 for the network below, under (column) personalization vectors $\mathbf{p}_1 = (0, 1, 0, 0)^T$ and $\mathbf{p}_2 = (0, 0, 1, 0)^T$. Note that the i-th component of each personalization vector corresponds to vertex v_i in the network below. **Hint:** help yourself with symmetries whenever possible.



3.2. i) Show *how* any Personalized Pagerank vector π corresponding to personalization vector $\alpha_1\mathbf{p}_1 + \alpha_2\mathbf{p}_2$ (with $\alpha_1 + \alpha_2 = 1$) can be computed from π_1 and π_2; ii) **Bonus:** if you can, rigorously show *why* this works.