Statistical Methods of Language Technology

{ruppert,yimam}@informatik.uni-hamburg.de

Summerterm 2019

Practice class session 10 19.06. 2019

Problem 10.1 PageRank

Compute the pageranks (damping factor = 0.8) for nodes in the graph in Figure 1. Initialize with a pagerank value of 0.2.

a) Use the recursive formula.

b) Use matrix multiplication.

c) (Extra) Use the undirected version of this graph, apply either of the two methods.

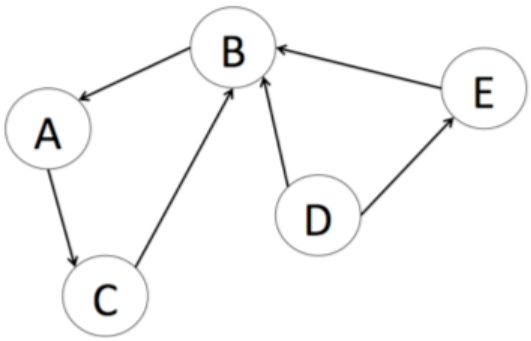

Figure 1:

Problem 10.2 MCL Graph Clustering

Cluster the graph (nodes A..F) given in Figure 2 with Markov Chain Clustering, k=2.
What are the clusters?

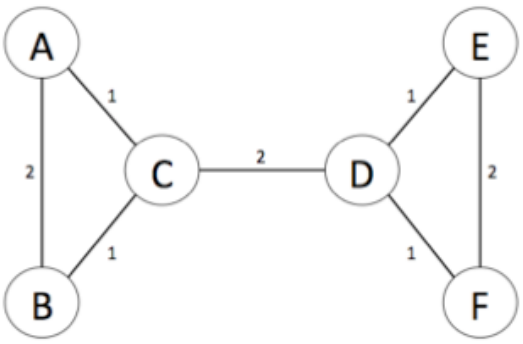

Figure 2:

Problem 10.3 Summarization

Consider the following sentences:

1. Bears eat men in the dark.
2. When it is dark, bears eat men and women.
3. Women eat men.
4. Dark bears eat only big men.

a) compute the TF*IDF feature vectors for all these sentences, only using open class words (no stopwords). Assume lowercase for all words. Stopword list: in, the, is, it, and, only, when

b) compute the graph of these sentences by calculating cosine similarity of their vectors. Apply a threshold of 0.1

Recap: $$cosine\ similiarity(Sx, Sy) = \frac{\sum_{i=1}^{n} Sx_i Sy_i}{\sqrt{\sum_{i=1}^{n} Sx_i^2} \times \sqrt{\sum_{i=1}^{n} Sy_i^2}}$$

c) (Extra) What is the ranking of the sentences after applying PageRank?