University of Ioannina,
Computer Science & Engineering Department

Data Mining – CS 059
Fall 2013

Assignment 3

This is the second part of Assignment 3. The deadline for the assignment is **January 24**. You can submit electronically, or bring the assignment to my office. Electronic submission should be done either via turn-in, or via email. The details for the turn-in are on the web page of the course.

Question 1

Prove that for an undirected graph the stationary distribution of a random walk is proportional to the degree of the nodes. If P is the transition matrix of the random walk, and π is the stationary distribution for which $\pi = \pi \cdot P$, show that for node i the probability π_i is proportional to d_i where d_i is the number of edges incident on node i.

Question 2

For this question you will implement the HITS algorithm on a bipartite graph. On the Material page of the class you are given a file containing tab-separated pairs of user ids and venue names, where the user with that id left a tip on that specific venue on Foursquare. Consider the venues as authorities and the users as hubs, and use the HITS algorithm to compute the hub and authority weights. (Attention: The authority weight is computed only for the venues and the hub weight only for the users). Sort the venues in decreasing order of their authority weight.

Hand in your code, and the sorted list of venues. Inspect the top-20 venues and try to explain why they appear at the top.

Question 3

You are given a document D consisting of n phrases. The document refers to d different **concepts** $c_1, c_2, \dots, c_d$. The concepts correspond to words or phrases that appear as Wikipedia lemmas. Identifying the concepts in the document has already been done by a different process, and for each sentence we know which concepts appear in it. A concept can appear in multiple sentences.

We want to create a summary of the document using exactly K sentences of the document, such that as many of the d concepts of the document appear (at least once) in the summary.

- Show that there is a greedy algorithm for the problem that has a constant approximation ratio with the optimal algorithm that constructs the summary that contains the maximum number of concepts.
- What happens if we want to cover all of the d conepts?