Data Mining

Project 2

Due: 21/12/2015, 23:59.

Instructions

You must hand in the homeworks electronically and before the due date and time.

Handing in: You must hand in the homeworks by the due date and time by an email to `aris@dis.uniroma1.it` that will contain as attachment (not links!) a .zip or .tar.gz file with all your answers and subject
`[Algorithmic Methods for Data Mining] Project 2`
After you submit, you will receive an acknowledgement email that your project has been received and at what date and time. If you have not received an acknowledgement email within 2 days after you submit then contact the instructor.

The solutions for the theoretical exercises must contain your answers either typed up or hand written clearly and scanned.

The solutions for the programming assignments must contain the source code, instructions to run it, the MongoDB data, and the output.

For information about collaboration, and about being late check the web page.

In this homework we will practice our skills for downloading data from Twitter, storing them into `MongoDB`, and performing some graph operations.

1. We start by performing some graph operations. Download the Enron dataset from
   `http://snap.stanford.edu/data/email-Enron.html`.
2. Use the `NetworkX` package to perform the following simple operations:
   (a) Compute and plot the degree distribution.
   (b) Compute and the degree, closeness, betweeness, and PageRank centralities
   (c) Compute the clustering coefficient of some of the nodes, and the clustering coefficient of the graph.
   (d) Compute the connected components of the graph.
   (e) Perform the k-core decomposition of the largest component.
3. Now we will create our own graph.
   (a) First download at least 10,000 tweets, related to some popular keyword (e.g., immigration), using the search API.
   (b) Obtain the followers of the users who tweeted. Note that you need to wait quite some time for this part.
   (c) Store the information of the tweets and of the graph into `MongoDB`.
   (d) Create the graph of the users and perform the operations you did in the first part.