Exercise on Python and PageRank

Mauro Sozio
Marie Al Ghossein (TA), Maroua Bahri (TA)
Arnaud Guerquin (TA), Pierre-Alexandre Murena (TA)
name.lastname@telecom-paristech.fr

In this lab session, we are going to learn and practice with Python and the PageRank algorithm. We suggest to use IPython/Jupyter to edit and run your code, however any other editors or IDE can be used. We recommend those who are not familiar with Python to check the tutorial on Python on the Web page of the course and proceed to Section 1. After that, move to Section 2. Those who are familiar with Python can go directly to Section 2.

1 Learning Python

This section contains a few suggestions in order to practice with the basics of Python:

- construct a list of integers L. Then build another list M with the same size of L that contains the square of the elements in L.
- define a function f that receives in input a list of integers and returns a new list containing only even integers.
- write a few lines in Python that read a list of integers from a file and store them into a list L. Then run f on L.

2 Exercise on Python and PageRank

This exercise consists of implementing the PageRank algorithm in Python. It consists of the following steps:

1. Implement the PageRank algorithm in Python. The algorithm receives in input a directed graph G which is represented as a list of lines of the kind "ij'' denoting that there is an edge between node i and j. The output is the PageRank vector for G. In this step, we can assume that there are no dead ends in G. The matrix M_G should be represented using a sparse matrix representation, i.e. only non-zero entries should be represented. This allows to deal with very large graphs. The product between the matrix M and the vector r should also be done assuming M is sparse. For this exercise we can assume $\beta = 0.8$ and $\epsilon = 0.1$. For testing, the PageRank vector shown in Figure 2, is $\{\frac{2}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}\}$ with no random jumps ($\beta = 1$).

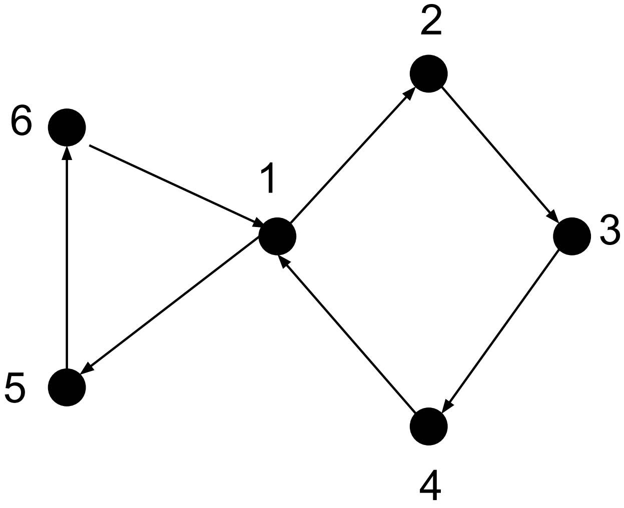

Figure 1: A simple web graph.

2. Construct the Web graph from the webpages extracted from Wikipedia which you can find at `https://sites.google.com/site/maurosozio/teaching/data-mining18`. To do so, for each page, extract all links using the *findall()* method of the regular expression module *re* (`https://docs.python.org/2/library/re.html`) and add an edge from page i to page j if there is a link on page i to page j. Beware of duplicate link and selfloop. Tip: all links are preceded by 'a href="'. After having constructed the graph, run the PageRank algorithm on that.

3. Implement in Python an algorithm that given an directed graph G in input, it removes dead-ends iteratively until no dead-ends are left.

4. Design a heuristic that: 1) removes the deadends from an input graph G, obtaining a graph H; 2) compute the PageRank vector for the nodes in H; 3) compute the pagerank vector for the nodes in $G \setminus H$ by starting from the PageRank scores of the nodes in H.