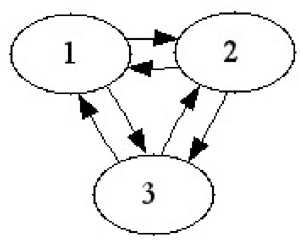
Cerca i Anàlisi d'Informació (CAI)
Grau en Ciència i Enginyeria de Dades, UPC

Session 4: Web search

Exercise List, Fall 2019

Basic comprehension questions.
Check that you can answer them before proceeding.

1. Compute by eye the PageRank of all nodes of the following graph:



2. True or False: depth-first search is more polite than breadth-first search crawling.
3. True or false: the Pagerank of a web page depends on the query.
4. True or false: the PageRank algorithm does not take into account the content of a web page.

Exercise 1

Study from the slides the block diagram of a web searcher system.
OK. Now without looking at it, make sure you can explain:

- The main parts of the system, and why they are necessary.
- How the information collected by the crawlers is processed, and where it goes.
- What needs to be done by the links collected.
- What needs to be done with the text collected.

- What needs to be done with the queries received from the user.
- How everything fits together.

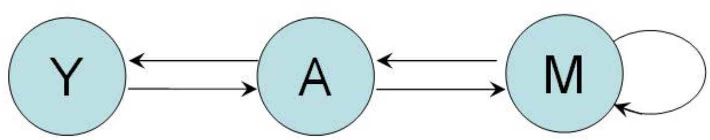
Exercise 2

Consider a small web with three pages, A, B, and C, where A has links to B and C, B has a link to C, and C a link to B.

1. Give the initial equations for this system (no damping), the associated transition matrix, and the resulting node PageRank values.
2. Now give the Google matrix using damping factor 0.85, the associated system of equations for PageRank, and the resulting node PageRank values.

Exercise 3

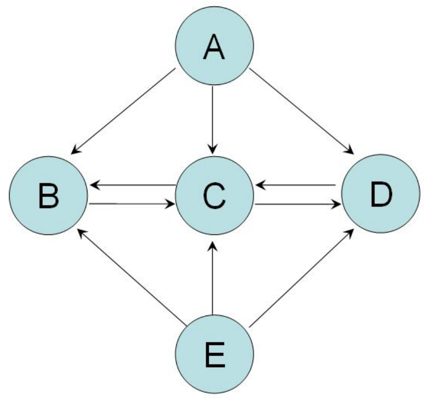
Consider the following miniature web:



Compute the PageRank equations with no damping, and the PageRank of each node. Repeat with damping factor 0.85.

Exercise 4

Consider the following miniature web:



1. Tell the PageRank values of A and E as a function of the damping factor. There is a simple argument for this. Don't go to the system of equations yet.
2. Justify that B and D have the same PageRank, no matter the damping factor.
3. Fix the damping factor to 0.9. Give the the Google matrix and the associated PageRank equations. Then compute the PageRank of each node.

Exercise 5

Give an example of a strongly connected graph with three nodes such that 1) each node has exactly two edges arriving into it 2) not all three nodes have the same page rank.

Set up the page equations for the graph you give, solve the system, and check by direct substitution that the solution you give satisfies the equations.

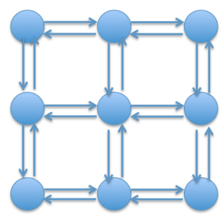
Exercise 6

Let G be the Google matrix of a web. We know that the vector of page ranks of the nodes is the one that satisfies $G^T P = P$, where ^{T} denotes the transpose of a matrix. Argue that if we do the computation without transposing G, that is, if we compute the vector S such that $GS = S$, there is always a trivial solution, independent from the web graph. Of course, we only consider vectors whose components' sum is 1. Suggestion: start working out a simple example, without damping factor, to see what goes on.

Exercise 7

Consider a graph with N^2 nodes aligned in a grid, where each node has links to its 4 closest neighbors, with the exception of nodes at the borders and the corners, which have links to their 3 and 2 closest neighbors, respectively. The figure gives the example for $N = 3$.

Compute the Pagerank of each node using as a function of N using damping factor $\lambda = 1$ first. Then generalize to damping factors $0 < \lambda < 1$.



.

Suggestion: Do it by hand for $N = 3$ and $N = 4$, try to generalize the solution for N and check it. Or write a program to solve it symbolically. Or numerically. Or something. You are a data scientist now.