Assignment 4

Course: *Big Data*
Due date: *April 22nd, 2024*

Assignment

As we do not have a Hadoop cluster the focus will be on the implementation of the solution. You will not run your programs on large datasets. The aim of this homework is for you to get the mindset of the MapReduce paradigm.

Write the data preprocessing stages and map reduce procedures using Python and MRJob library for the **PageRank** algorithm (`https://en.wikipedia.org/wiki/PageRank`).

How to implement PageRank with map reduce is described in `https://michaelnielsen.org/blog/using-mapreduce-to-compute-pagerank/`. You should watch these videos:

- `https://www.youtube.com/watch?v=u8HtO7Gd5q0`
- `https://www.youtube.com/watch?v=kxzSFocwMT0`
- `https://www.youtube.com/watch?v=9e3geIYFOF4&t`
- `https://www.youtube.com/watch?v=_Wc9OkMKS3g&t`.

There is plenty of material about PageRang on the internet.

Test it on different datasets.

There are a lot of datasets suitable for testing PageRank available on the internet. I advice you to:

- Test your implementation on the the Florentine families dataset: `http://www.casos.cs.cmu.edu/computational_tools/datasets/external/padgett/index2.html`
- Find your own "large" dataset to test your program.

Submit a Jupyter notebook with your code and "report".

Your Jupyter notebook should contain: problem description, short description of the solution, description of the selected dataset, and conclusions.