A3: Structure of wikipedia links

Go to https://zenodo.org/record/2539424 and fetch the file:

https://zenodo.org/record/2539424/files/enwiki.wikilink_graph.2004-03-01.csv.gz?download=1

Investigate the graph:

- Dead ends
- Distribution of in-degrees
- Distribution of out-degrees of nodes
- Implement the page rank algorithm from slide 18
- Implement direct (sparse) matrix multiplication
- Compare results
- *Is this graph strongly connected?*

Data preprocessing (prep)

The data has the following layout:

page_id_from	page_title_from	page_id_to	page_title_to
12	Anarchism	34568	16th century
12	Anarchism	35416	1793
…	…	…	…

- □ Extract only the 1st and the 3rd column
- □ Convert page_id's into consecutive integers, in such a way that you can return back to the original numbering
- □ Both columns should use the same coding!
- □ Save the prepared data on HD

Exploratory data analysis (eda)

- Dead ends: find nodes with no outgoing edges. How many have you found?
- Distribution of in-degrees: for every node compute the number of incoming edges
- Distribution of out-degrees: for every node compute the number of outgoing edges
- Make nice & informative plots of both distributions
- What is the average out-degree and the average in-degree of the graph?

Estimate RAM requirements: (eda)

1. How much RAM would you need to store the transition matrix M and the initial vector v in RAM? Assume double precision (64 bits per number).

2. The same question assuming that you store M in a sparse matrix (in RAM)?

3. The same question, assuming that you use data structures as described on slide 17.

embed your answers in the notebook eda.ipynb

Implement PageRank algorithm (sparse)

1. Store both M (as a sparse matrix) and v (in RAM).
2. Run 25 iterations of the “classical” update rule from slide 10, with Beta=0.8.
3. Plot the MSE of the differences (25 numbers): v – Mv
4. Assume that your computer has 1GB RAM and the average out-degree of a graph G is 15.

   What is the maximal number of nodes of G such that your algorithm could be executed on your computer?

What to deliver?

Four Jupyter notebooks:

- prep.ipynb
- eda.ipynb
- sparse.ipynb
- PageRank.ipynb

that correspond to all the subtasks.

Your notebooks should read/write files from/to the same directory as your notebooks. We will test your programs in a directory which contains wikilink_graph.2004-03-01.csv