Big Data Algorithms
Lab-3

Due: 30.01.2022

Please, collect web pages data before laboratories!!!

---

1. (6pts) **Basic page rank**
   This exercise consists of two parts:

   - create simplified web crawler and collect some web pages data (for this exercise you are interested in links connecting us to other pages),
   - use collected data to prepare ranking of the most important pages, using page rank as a measure.

   You can crawl any pages, but I suggest using Wikipedia web page for this exercise and stick only to internal pages (links you can find there are in nice format */wiki/topic*), so it will simplify the crawler code. Lets assume that number of crawled pages will be an input parameter to the application and the application itself will work in multiple modes:

   - crawl mode - will take number of pages to crawl as an input and output files with crawled web pages information (and exclude any of Wikipedia side pages, like files pages, authors pages etc. - such links are usually in form *File:sth*, *Template:sth* - once you collect first set of links, you should recognize it),
   - page rank mode - will use collected data to create file with web pages and their rankings, additionally printing the results sorted according to page importance (this mode should take an input parameter for number of matrix multiplications used at page rank calculation; this phase should be optimized to take the sparse nature of links matrix into account),
   - link analysis mode - application should be able to print information like, which pages have links to the chosen page, how many such links were there, average number of links per page etc.

   Hint 1 For this exercise calculate page rank only for crawled pages (assuming you will use Wikipedia page, which has a quite large number of links per page, for 100 crawled pages you will probably have several thousands of not visited links, making page rank calculation longer and probably not feasible on personal computers).
   Hint 2 You can assume that all pages has out-links - for Wikipedia this should be always true anyway, and you can remove all spider trap pages before calculating page rank.

2. (3pts) **Spider trap prevention**
   Add some pages without out links to your input for page rank - you can use collected, non-crawled links for it. Check how it impacts the page rank calculation. Implement a page rank version which prevents the spider traps. Try to add more complex spider trap - ex. hard coding the set of pages which will link to each another in a cycle (in a way that would block the crawler from exiting such trap, but still enabling him some movements)

3. (3pts + 3 bonus pts*) **Search engine**
   Use previous exercises to create simple search engine, which for some input word will print the 5 most important web pages containing that word. Additionally, you can implement extensions to the search engine, which will improve the quality of the search results (ex. topic sensitive page rank, increasing result importance when some word is more frequently appearing there, some topic matching etc.) - each extension will be worth 1 bonus point, up to 3 pts.