DPRPy 2024/2025

Homework assignment no. 4 (max. = 40 p.)

Maximum grade: 40 p.

Homework should be sent via the `MS Teams` platform. You should include all files containing solutions to tasks as one `.zip` archive named i.e. `Last-name_First-name_HA_3.zip`.

Note: Your solutions may be in one or multiple Jupyter Notebook files, knitr + Markdown files, `.py` or `.R` scripts, etc. Just make sure that each file is clearly named.

1 Tasks description

Scientometric analysis is widely used for the estimation of research output. In this task we will take a look at the data fo research papers published via the ArXiv - a free distribution service and an open-access archive for nearly 2.4 million scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics.

Data that interest us is available at: https://www.kaggle.com/datasets/Cornell-University/arxiv?resource=download

This file contains an entry for each paper, containing:

- `id`: ArXiv ID (can be used to access the paper, see below)
- `submitter`: Who submitted the paper
- `authors`: Authors of the paper
- `title`: Title of the paper
- `comments`: Additional info, such as number of pages and figures
- `journal-ref`: Information about the journal the paper was published in
- `doi`: https://www.doi.org
- `abstract`: The abstract of the paper
- `categories`: Categories / tags in the ArXiv system
- `versions`: A version history

2 Task 1 [20 p.]

Based on abstracts and assigned to each paper category (variable `abstract` and variable `categories`) perform the following analyses.

1. What is the most commonly used category in papers in ArXiv? Prepare visualization of the distribution of categories (e.g. prepare barplot presenting the number of occurrences of each category). [5 p.]
2. For one category of your choice, prepare a word cloud based on the available abstracts. [10 p.]
3. For 5 top most commonly used categories find top 3 words used in abastracts. [5 p.]

3 Task 2 [20 p.]

Based on available data create a graph that shows some kind of relationship between authors.

For example, let $G = (V, E)$, where the set of node consist of all authors, i.e. $V = u_1, ..., u_n$ and u_i denotes i-th author. Moreover, E denotes the set of all edges and $\{u_i u_j\} \in E$ if there is an interaction between authors. You may chose the type of interaction you are interested in. For example, the interaction can be viewed as:

- u_i and u_j collaborated on a paper i.e. they are co-authors of a paper;
- u_i and u_j have papers from the same category;
- u_i and u_j published in the same journal;
- etc.

Create a visualization of created graph. Note that in order to do so, you may need to filter the nodes / edges that are of low degree / weight, experiment with different layouts.