Data Engineering - Mini Project (Version 0.2)

1 Introduction and Use Case

In this mini-project, imagine a research funding agency asks you to evaluate, compare, and benchmark Hamburg's universities and universities of applied science computer science departments.

Your dataset should enable an analysis of 1) publication metrics (publication per person, citations per year per institute), 2) co-author networks inside, outside, and in-between universities, 3) funded projects, and 4) geographic features of publication venues and others. Use at least 2 data sources. You can come up with more metrics that you as a student find interesting.

2 Task (7 points)

Prepare a dataset for statistical analysis for all computer science researchers from Hamburg's academic research institutes!

In your project, document the data engineering as source code and documentation. The documentation should be a brief report (A4 letter, 11 pt font size, single spacing, 2cm borders). The document should also describe the distribution of tasks in your team (if any). Present the results briefly at the beginning of your oral exam.

Cover the following dimensions in your documentation:

- Which data sources and formats did you use? Describe them.
- Did you use any annotations or did you do any manual or automatic labeling? Why not?
- Document the data quality and metadata. Describe briefly your observations.
- Which string metrics did you need and why?
- Document whether you provide any embeddings and if which, and if not, why not. Where could embeddings be used in such a scenario?
- Did you do any outlier detection? If so, which? In not, why not?
- Document any ethical issues that could arise with your data.
- Explain how your dataset is now better compared to each of the single resources (raw data) and how it helps to fulfill the analysis task! Give us five meaningful insights (text, graphic, map) from your dataset.
- Give us a brief feedback text on the project.

If you answer all points, the project counts instantly as 51% of the points for the course.

3 Data Sources

Among others, you can use the following resources:

- Semantic Scholar Resources for the Global Research Community: `https://www.semanticscholar.org/resources`
- DBLP: `https://lod-cloud.net/datasets?search=dblp`, `https://dblp.org/xml/`
- OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts `https://arxiv.org/abs/2205.01833` or `https://openalex.org/`
- Datasets listed on **https://nfdi-search.nliwod.org/**

Do not hesitate to use more data sources!