Course Assignment 1: Literature Review

Course: Academic Writing

A good review of related work is an essential ingredient for any research activity to understand the gap in the published results, by the others. This understanding motivates pushing forward the State-of-the-Art in the chosen research field. This course assignment is aimed at giving students a chance (and guidance) to elaborate this ingredient for their Master project in a systematic, effective, and efficient way.

The assignment is offered after the completion of the first part of the course. It deals with conducting systematic literature search and selection within the scope of the (potential) topic of a Master project. After the selection is done, the chosen literature sources are studied, read, reviewed, and analyzed with the objective to find a research gap to be attacked and the motivation to narrow this gap. The findings have to be documented in a Technical Report.

The report will be evaluated based on the following aspects that resemble the criteria used for the evaluation of review papers. These aspects are:

- Methodology (0 – 5)
- Representativeness (0 – 5)
- Relevance (0 – 5)
- Structure (0 – 10)
- Gap analysis (0 – 10)
- Writing style (0 – 10)

1 Representativeness, Relevance, and Methodology

As a pre-requisite for making this review and analysis grounded and unbiased, the collection of the literature sources needs to be:

- **Relevant**, which is containing only the sources, that, in their substance but not by the title, report the results that might be of use for, or influence the work planned in your project
- **Representative**, which is sufficiently complete, to date, for covering all the significant research and development accomplishments in the chosen field

Of course, a well-formed idea of what is planned to be done in the project needs to be elaborated to allow answering the **relevance** question in a cogent way. Therefore, iterations might be needed to refine the project idea based on the discovered literature collection of the previous iteration, which resembles the Active Learning approach [1]. This way will successively approximate your understanding of the chosen research topic to what the corresponding research community thinks of the State-of-the-Art within the topic.

Answering the question about the **representativeness** of the collected literature is not easy. One has to argue that, at least, all the mainstream related work to date is covered by the collection at hand. Here, the arguments like "… my advisor says so …" or "… I firmly believe that …" are biased and, therefore, do not fly to a sufficient extent. A more objective and mature approach to finding a representative sample of literature within a research topic has to use a methodology. Some of methodological advises are offered in the suggested basic textbook for the course [2, p.21]. The context is also discussed in the course topic 3[1].

Methodological suggestions on literature search and selection, are often informal and do not offer an instrument for the activity. To complement that, we offer several tools that might be used for selecting relevant papaers and checking if the selected collection is representative.

[1] https://cms.ucu.edu.ua/mod/resource/view.php?id=260845

1.1 Tools for Collecting Papers

Severall instruments are available publicly online. Some of those are free to use[2]. These include:

- **arHiv** advanced search (https://arxiv.org/search/advanced)
- **arHiv** sanity (lite) (https://arxiv-sanity-lite.com/about)
- **WOSviewer** (https://www.vosviewer.com/)
- **Mendeley** (https://www.mendeley.com/) including its reference manager
- **Zotero** (https://www.zotero.org/)

As one more alternative, we offer you to consider a software instrument for the collection and terminologogical analysis of the full texts of papers (please refer to the technical notes in Section 3). The instrument contains two software components in its pipeline and helps answer the relevance and representativeness questions as follows:

- **Relevance**. A user is suggested to provide a small “seed” set of research papers that are indexed by Open Alex[3]. The metadata of these papers is further fed into the Controlled Snowball Sampling tool. The tool generates the catalogue of the selected relevant publications using citation network analysis, probabilistic topic modeling, and snowball sampling. The full texts of the papers that are publicly available are downloaded for further reading and analysis. Iterations could be done via refining the “seed” set of research papers based on the result of the previous iteration.
- **Representativeness**. A user is suggested to use the collection catalogue and downloaded full texts of the publications for conducting their analysis of the completeness of the collected set of papers. Terminology saturation in the subset of papers, if observed, could be used as an objective argument in favour of the representativeness of this subset. Furthermore, the subset is smaller than the whole collected set of papers and requires less effort for reading and analysis.

The computation pipeline pictured in Fig. 1.

Students are NOT requested to use the offered instrumental bundle. Instead, they may use any other means to collect the relevant and representative collection of publications for their literature analysis. The arguments for proving representativeness and relevance still have to be provided in the technical report.

The methodology for literature collection should also be presented in sufficient detail that allow the understanding of the set of the sources has been collected.

2 Structure, Gap Analysis, and Writing Style

The report has to be structured in a way to cover:

- The initial motivation of the author to choose their particular research topic
- The method used for literature search and selection
- The sub-topical structure of the chosen research topics – with sub-sections reviewing the State-of-the-Art within the identified sub-topics and identifying open problems
- The refined motivation to perform the research work. The focus of this work is presented as one selected research challenge out of those identified in the review.
- The analysis of the research gap between the State-of-the-Art and the anticipated result of the research and the presentation of the attacked open research problem

[2] As the tools are free, these are offered “as-are” so please do not expect any extended documentation or tutoring service in this context.

[3] https://explore.openalex.org

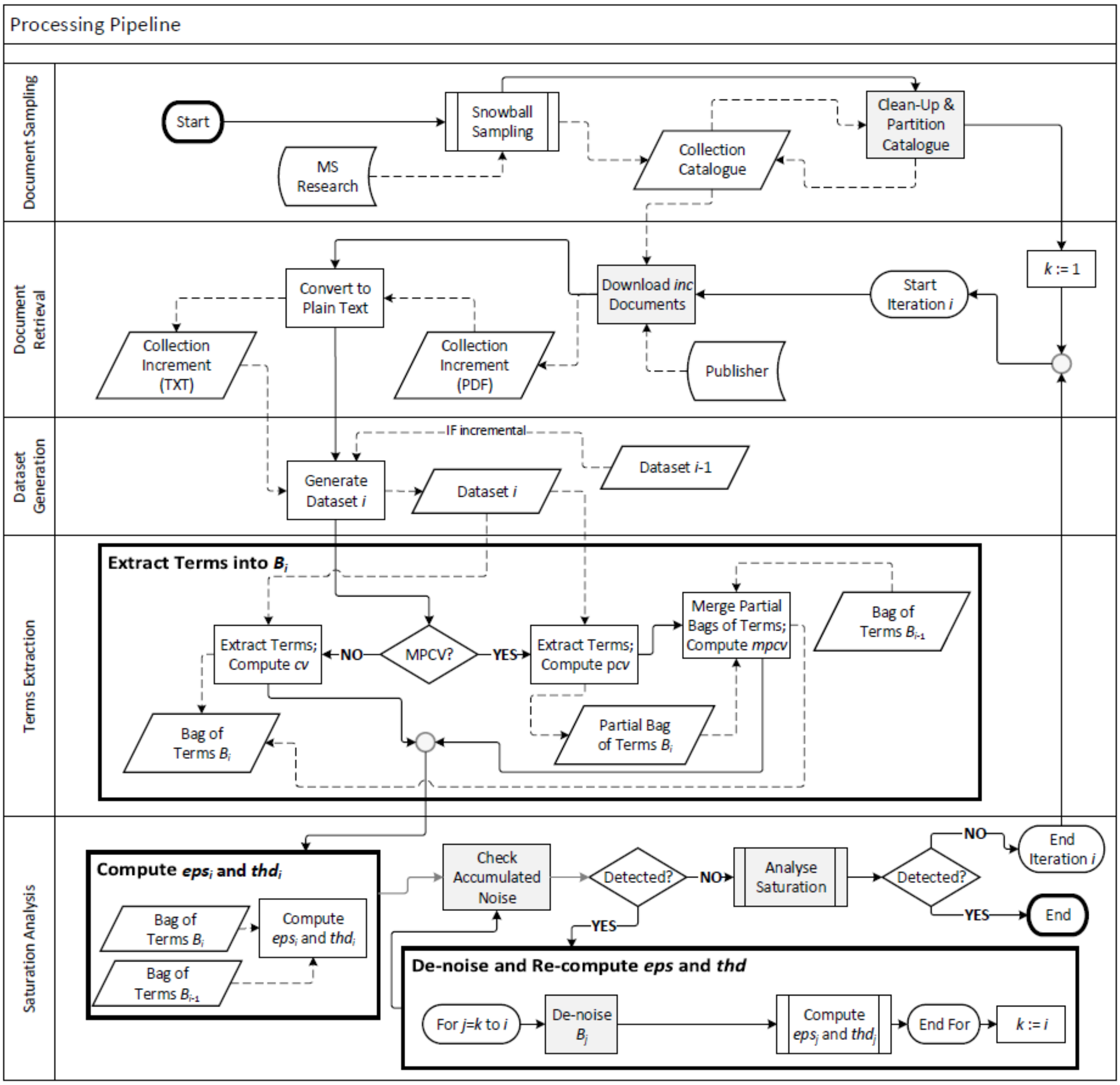

Fig. 1. The computation pipeline for controlled snowball sampling and terminology saturation analysis.

The report should also contain appropriate introduction and summary. Additional recommendations are given in course topics 2[4] and 4[5]

It is suggested that the report is written using the style recommendations discussed in the course topic 4

3 Technical Notes on the Suggested Software Tool

The tool is provided as the pipeline of python scripts that implement controlled snowball sampling and terminological saturation analysis for collecting the set of scientific publications on desired subject. The details of the approach are available in [3-8]. There is also the monograph [9] that offers all the necessary reading on the approach, including the use csases, in one source.

Requirements: To run the package you need

[4] https://cms.ucu.edu.ua/mod/resource/view.php?id=260844

[5] https://cms.ucu.edu.ua/mod/resource/view.php?id=260846

- python3
- python-poetry
- optionally: anonymous proxy to query Google Scholar

Anonymous proxy

You can use one of the proxy services listed at https://www.didsoft.com

Quick start:

1. Get the copy of this repository using

   ```
   $ git clone https://github.com/gendobr/snowball.git
   ```

   and go into the *snowball* directory The content of the directory is

   ```
   data  - empty directory to place your data
   docs  - place to additinal documentation
   Pipfile  - list of required packages
   README.md - this file
   scripts - python code
   ```

2. Copy the directory ./docs/data into ./data/YOUR_DATA_DIRECTORY

   ```
   $ cp -R ./docs/data ./data/YOUR_DATA_DIRECTORY
   ```

3. Install all the required python packages using

   ```
   $ poetry install
   ```

4. Download the required NLTK packages

   ```
   $ poetry run python scripts/init.py
   ```

5. Find 10 - 20 seed publications in the https://explore.openalex.org.

   Each seed publication should:

   - be relevant to your search topic
   - have high citation index (however not extremely high)
   - be 7-10 years old

   These recommendations for selecting seed publications are given and explained in [4].

   Paste the publication ids in the `./data/YOUR_DATA_DIRECTORY/in-seed.csv` file. One id per row

   A publication id is the long number with the prefix “W” in the publication URL. For instance, in «https://openalex.org/W4213193956» the publication id is **W4213193956**.

6. Run one-after-one the following scripts (don't forget to change the `../data/GAN/` to `../data/YOUR_DATA_DIRECTORY/`) inside each file

- `000_download.sh` - may take several hours time to download up to 20000 baseline publications
- `001_tokenizer.sh` - performs tokenization step using NLTK tools
- `002_rarewords.sh` - detects rare words
- `003_joint_probabilities.sh` - estimates token co-occurrence probabilities
- `004_stopwords.sh` - detects stopwords
- `005_reduced_joint_probabilities.sh` - estimates token co-occurrence probabilities after the rare words and stopwords excluded
- `006_SSNMF.sh` - creates topic model in 1-2 hours
- `007_restricted_snowball.sh` - may take several hours time to download up to 20000 relevant publications
- `008_search_path_count.sh` - does search path count calculation (see. Main path analysis for explanation)
- `009_extend_items_google_scholar.sh` - the script downloads several hundreds publications from Google Scholar, so you must use a proxy to avoid the ban. The proxy address is *proxy* parameter in the configuration file. `./data/YOUR_DATA_DIRECTORY/config.ini`
- `009_extend_items_google_scholar_resume.sh` - sometimes you need to resume the previous command
- `010_download_pdfs.sh` - downloads PDF files that are available for free
- `011_export_xlsx.sh` - creates the catalogue of the selected collection of papers according to `./docs/data-requirements.txt`

The final list of publications is the file `011_exported.xlsx`. This fiename is provided in the `--outfile` parameter of the `011_export_xlsx.sh` script.

7. You than continue with the Automated Term Extraction and Saturation Analysis steps. Please refer to [9] for the details.

- `012_ate_pdf2txt.sh` - extract plain texts from PDF files
- `013_ate_clear_txt.sh` - clear extracted texts
- `014_ate_generate_datasets.sh` - join extracted texts in the sequence of datasets
- `015_ate_get_terms.sh` - extract terms
- `016_ate_clear_terms.sh` - remove trash terms (list of trash terms is the file `./data/YOUR_DATA_DIRECTORY/ate_stopwords.csv`)
- `017_ate_saturation.sh` - does terminological saturation analysis

References

1. Settles, B.: Active Learning Literature Survey. Computer Sciences Technical Report 1648 (2009). Available at: https://minds.wisconsin.edu/bitstream/handle/1793/60660/TR1648.pdf
   - **openalex:** https://explore.openalex.org/W2903158431
2. Zobel, J.: Writing for Computer Science. Third Edition. Springer London Heidelberg New York Dordrech (2014)
   - **openalex:** https://explore.openalex.org/W10254817
3. Dobrovolskyi, H., Keberle, N., Todoriko, O. (2017). Probabilistic Topic Modelling for Controlled Snowball Sampling in Citation Network Collection. In: Różewski, P., Lange, C. (eds) Knowledge Engineering and Semantic Web. KESW 2017. Communications in Computer and Information Science, vol 786. Springer, Cham. https://doi.org/10.1007/978-3-319-69548-8_7
   - **openalex:** https://explore.openalex.org/works/W2766806542
4. Dobrovolskyi, Hennadii, and Nataliya Keberle. "Collecting the Seminal Scientific Abstracts with Topic Modelling, Snowball Sampling and Citation Analysis." ICTERI. 2018.

   - **openalex:** https://explore.openalex.org/works/W2899429816

5. Dobrovolskyi, H., & Keberle, N. (2018, May). On convergence of controlled snowball sampling for scientific abstracts collection. In International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications (pp. 18-42). Springer, Cham.
   - **openalex:** https://explore.openalex.org/works/W2912283770
6. Kosa, V., Chaves-Fraga, D., Dobrovolskyi, H., Fedorenko, E., & Ermolayev, V. (2019). Optimizing Automated Term Extraction for Terminological Saturation Measurement. ICTERI, 1, 1-16.
   - **openalex:** https://explore.openalex.org/works/W2954824683
7. Kosa, V., Chaves-Fraga, D., Dobrovolskyi, H., & Ermolayev, V. (2019, June). Optimized term extraction method based on computing merged partial C-values. In International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications (pp. 24-49). Springer, Cham.
   - **openalex:** https://explore.openalex.org/works/W3000160792
8. Dobrovolskyi, H., & Keberle, N. (2020). Obtaining the Minimal Terminologically Saturated Document Set with Controlled Snowball Sampling. In ICTERI (pp. 87-101).
   - **openalex:** https://explore.openalex.org/works/W3108422417
9. Kosa, V., Ermolayev, V.: Terminology Saturation: Detection, Measurement, and Use. Cognitive Science and Technology, Springer – Nature (2022)
   - **openalex:** https://explore.openalex.org/W4213193956
10. Lecy, J. D., & Beatty, K. E. (2012). Representative literature reviews using constrained snowball sampling and citation network analysis. Available at SSRN 1992601
    - **openalex:** https://explore.openalex.org/W1499565482