BEP Final Project Plan

Project Planning

Week	Task Description
1,2	Refine the research question, conduct literature review and (if possible) gather initial datasets.
2,3*	*First circle meeting, work plan*
3,4	Develop the project plan, finalize methodology, start thinking of how to address text preprocessing (after having accessed and understood the data). Understand role of labelers/labeling process. Create GitHub repository for project.
5	*Informal meeting (Oct. 2nd), final project plan discussion.*
5,6	Preprocessing, store data in suitable format. Review Harry's work. Start looking into rule-based approaches.
7,8,9,10	Implement and test rule-based methods (e.g. RegEx, NER, POS, etc.) for extracting key metadata. If time is left, start research on suitable ML approaches.
10,11,12*	*Second circle meeting, midterm presentation*
11,12,13,14	Develop and train machine learning models (e.g. Bert and others—research needed) and compare their performance to rule-based methods.
15,16	Buffer weeks, allocate tasks depending on project progress.
16,17*	*Third circle meeting*
17,18	Get final feedback, make adjustments, evaluate cross-project generalizability and conduct final evaluations. Write the thesis, incorporate feedback and prepare for final submission.
19	Thesis submission.
20,21*	*Fourth circle meeting. Assessment meeting, final presentation (according to BEP Canvas page).*
22	To be defined.

Table 1: Project Timeline (*) overlapping activity: circle meetings

Research Questions

- Primary RQ: *How can NLP techniques be applied to effectively* extract and structure key metadata from Dutch administrative decisions (energy permits), while overcoming the challenge of incomplete labels using weak labeling techniques, and comparing the performance of different approaches, including rule-based methods and machine learning?*

  ***effectiveness** refers to the ability of NLP techniques to extract and structure metadata from administrative decisions in a way that achieves satisfactory outcomes across

multiple dimensions: accuracy, efficiency, scalability, and adaptability. An approach is considered effective if it yields high-quality results, is feasible to implement, and can be applied to various types of decisions with minimal supervision.

- Sub-questions:

    - *How can weak labeling techniques be applied to systematically categorize Dutch administrative decisions, addressing the issue of incomplete data labeling (e.g., by leveraging legal keywords from document titles)?* Possible reflection points:

    - *Which NLP techniques (e.g., rule-based, machine learning, or others)—alone or in combination—are most effective in extracting key metadata from administrative decisions in the absence of fully labeled data?*

- Points of relfection (while answering the questions):

    - Specific challenges associated with processing legal language in Dutch?

    - Generalization across different types of administrative decisions?

    - For when dealing with ML approaches: how can large language models (LLMs) be used for extracting key information from administrative decisions, and what strategies can be employed to ensure accuracy and reliability in their outputs?

    - Non-quantitative evaluation: how?

Proposed Methodology

1. Understand the project;

2. Work on research question(s) (with further refining happening as the project advances);

3. Store the data in a suitable format and understand its structure;

4. Research available tools (preprocessing techniques, rule-based methods and ML libraries) and materials (related work, current developments in the field);

5. Carry out experiments;

6. Evaluate results (after choice of on appropriate evaluation system);

7. Get feedback and incorporate it into the project;

8. Report findings.

Literature Search
So far, the literature review has focused on the papers recommended on the project page of the BEP Marketplace. As the project progresses, additional research will be conducted, particularly to address areas that prove more complex and require a deeper understanding

of the relevant context and methodologies. I have also reviewed scientific and online articles on weak labeling, which have provided insights and resources into how to go about the lack labels in the dataset at hand.

- Sansone, C. & Sperlí, G. (2022). Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106, 101967.

- Gray, M., Savelka, J., Oliver, W., & Ashley, K. (2023). Can GPT alleviate the burden of annotation? In *Legal Knowledge and Information Systems* (pp. 157–166). IOS Press.

- Zin, M. M., Nguyen, H. T., Satoh, K., Sugawara, S., & Nishino, F. (2023). Information extraction from lengthy legal contracts: Leveraging query-based summarization and GPT-3.5. In *Legal Knowledge and Information Systems* (pp. 177–186). IOS Press.

- Wolswinkel, C. J. (2024). Actieve openbaarmaking van beschikkingen. *Nederlands Juristenblad*, volume 24 (pp. 1851–1857) (in Dutch only)[1].

- Lison, P., Barnes, J., & Hubin, A. (2021). Weak Supervision Made Easy for NLP.

- Ratner, A., Liss, S., Selsky, J., & Snoek, J. (2017). Snorkel: Rapid Training Data Creation with Weak Supervision. In *Proceedings of the 2017 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2201-2211). ACM.

Website references (articles, blog posts, libraries)

- Humanloop. (2023). Why I changed my mind about weak labeling for ML. Retrieved from
  `https://humanloop.com/blog/why-i-changed-my-mind-about-weak-labeling-for-ml`

- Weak labeling Python library (1): `Skweak`, `https://github.com/NorskRegnesentral/skweak`

- Weak labeling Python library (2): `Snorkel`, `https://www.snorkel.org/`

- `SpaCy` tutorial:
  `https://youtube.com/playlistlist=PL2VXyKi-KpYvuOdPwXR-FZfmZOhjoNSUo&si=MZ2zKrwSipQZL-E1`

---

[1]Read in machine translated version