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 effectively utilized to extract and structure key metadata from Dutch administrative decisions (energy permits) to improve their transparency and comparability?*"

- (Possible) sub-questions:

  – *What are the specific challenges associated with processing legal language in Dutch?*
  (e.g. any difficulty related to the language, availability of other studies/ML libraries to work with).

- *Can the developed methods be generalized across different types of administrative decisions?*
  (e.g. identification of recipient, reasons, topic, date, etc. of a non-legal document).
- *How can rule-based methods and machine learning models be combined to achieve the best results?*
  (useful to compare different strategies, their advantages and disadvantages, their results).
- *What are the most relevant NLP techniques for this task?*
  (also related to previous sub-question).
- *How can these techniques be tailored to specific types of administrative decisions?*
  (this is similar to the SQ on generalization of the methods to different types of administrative decisions, but allows to explain what are the specificities of the documents we are working with).
- *What challenges arise in processing and interpreting legal or bureaucratic language?*
  (can be incorporated to the SQ on the Dutch language and its challenges).
- *How can large language models (LLMs) be used effectively for extracting key information from administrative decisions, and what strategies can be employed to ensure accuracy and reliability in their outputs?*
- *How can the results be reliably evaluated in the absence of pre-labeled data, considering both qualitative and quantitative evaluation methods?*
- *What alternative approaches (e.g., weak supervision, manual annotation, or unsupervised techniques) can be employed to assess the extracted information?*

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. This ongoing literature exploration will help refine the approach and ensure that the project stays aligned with current developments in the field.

- 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].

[1]Read in machine translated version