Master in Smart Data Science — Project Proposal

Comparison of Large Language Models (LLMs) and a local NLP method : a case study on the citizen consultation “How can we protect and restore biodiversity together?” (Make.org)

Supervisor : Anne-Cécile GAY
Organization : Independent Data Scientist, in collaboration with Make.org
Contact : annececile.gay@gmail.com

---

1. Context

Make.org is a large-scale citizen consultation platform that enables citizens to submit proposals and vote on issues of general interest.
One of its consultations focused on the question : **“How can we protect and restore biodiversity together?”.**

The goal of Make.org is to identify areas of consensus from thousands of citizen contributions in order to co-construct concrete and applicable solutions.

In this project, students will analyze the citizen proposals to identify the main themes expressed. Three Natural Language Processing (NLP) approaches will be compared :

- **Word2Vec** (local model trained on the corpus),
- **LLM used as an embedding generator**,
- **LLM queried directly in a question–answering mode** to extract themes.

---

2. Data and research question

Students will have access to :

- 5,550 citizen proposals written in French,
- 1.8 million associated votes (for / against / neutral).

Main research question:
Which major themes emerge from the biodiversity consultation, and how do they differ depending on the method used (Word2Vec, LLM embeddings, or direct LLM interrogation) ?

3. Proposed Methodology

The project will be conducted in Python and may be structured in the following steps :

1. **Exploratory Data Analysis**

- Examination of the corpus (vocabulary, distribution of votes, etc.).

2. **Topic Modeling with Word2Vec**

- Training Word2Vec to obtain embeddings of the proposals,
- Projecting words/proposals into a vector space,
- Applying clustering methods to identify themes.

3. **Topic Modeling with an LLM**

- Using a pre-trained language model (e.g. via Hugging Face),
- Extracting embeddings and clustering them to identify themes,
- *Optional (if time allows):* directly querying the LLM to generate a categorization.

4. **Comparison of Approaches**

- Cross-comparison of the clusters/themes obtained,
- Discussion of the strengths and limitations of a local vs. large-scale vs. generative approach.

4. Resources and Bibliography

- Make.org: https://make.org/
- Word2Vec (foundational paper): https://arxiv.org/pdf/1301.3781.pdf
- Python Implementation (Gensim) https://radimrehurek.com/gensim/models/word2vec.html

Note: Oral exchanges between the supervisor and students will be conducted in French. Written exchanges will be conducted in English.