Information Retrieval & Text Mining

Marius Popescu

popescunmarius@gmail.com

2024 - 2025

Project 2

Retrieval Augmented Generation (RAG) for Romanian

The Task

Improve an already existing information retrieval system. Your program should receive a query and return a sorted list of documents that are relevant for that query.

Your system should only take into consideration documents from a given local path folder.

Instruments

Create an ensemble between an Information Retrieval system and a Romanian LLM locally employed.

You can use your own system from the first project, a system based on another library or you can write one from scratch

Quality Control

Who gives better answers?

1. Your IR system?
2. The LLM?
3. The ensemble?

Try different prompts for better results

Deliverables

- Code
- Research Report:
  - Should be structured like a short research paper and have at least 2 pages
  - Document everything you tried, including the list of files, queries and prompt engineering (could be in an appendix)
  - Document obtained results. Can be manually extracted and/or based on average precision for certainty thresholds
  - Should cite at least 2 papers published in or after 2020

Alternatives

If you are already working on an alternative project using Language Models you might get our approval to use it instead of this project.

Deadline: week 15

January 26th 23:59