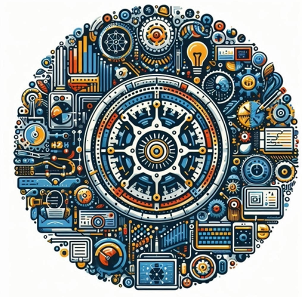
Advanced Topics on Machine Learning 2024/205



Pedro G. Ferreira (version 1 – 27/9/2024)

Objectives

The goal of this project is to explore the potential of LLMs on the analysis of data and information. Different projects are suggested and each group should choose one of the projects. Beyond what is proposed, students can also propose new functionalities to enhance the work.

P1 - Analyzing and Generating Scientific Abstracts

The *arxiv* database contains pre-prints of scientific articles that are made available before being published. Data from each paper is gathered, including name of the authors, title, abstract and other meta-data. The goal of this project is to analyze this data and develop a tool that implements some of the functionalities described below. The data is available at HuggingFace:
https://huggingface.co/datasets/gfissore/arxiv-abstracts-2021?row=25

Tasks to be performed:

1. Generate Titles from the Abstract. Evaluate the performance of the model using the appropriate measures.

2. Generate Abstracts from the Title. Evaluate the performance of the model using the appropriate measures.

3. Predict the categories based on abstract and/or title. Analyze the existing categories, create a taxonomy of categories and evaluate the predictive performance of the model. Here you can apply the cross-validation procedure to evaluate classifiers. Classification can be done considering a flat distribution of classes or a hierarchical distribution considering the previous taxonomy, where performance is evaluated at different levels of the taxonomic trees.

4. Apply other appropriate model optimization strategies.

5. Other relevant features.

P2 - Predict Ratings from Goodreads Reviews

Goodreads.com is a website where book readers can post reviews on books. The dataset to be analyzed consists in the public scrapping of the reviewers of thousands of books. The data is available at:
https://mengtingwan.github.io/data/goodreads#datasets

Since this is a very large dataset, in order to make its processing more accessible you can create a smaller dataset that is representative of different categories of books and then work with that dataset. Functions to be implemented include:

1. Create a smaller and representative dataset. Sample book reviews by genre, e.g. 5K from Children, 5K from Comics & Graphic, 5K from Fantasy & Paranormal, etc... Divide then into train/test/split folds.

2. Fine-tune LLM to predict the rating of the book. Use the appropriate techniques to create a regression model that scores the book based on the reviews text.

3. Create a classifier that classifies the book category based on the readers' reviews.

4. Identify the most informative words in task 3 and 4.

5. Other relevant features.

P3 - Building RAG Chatbots for Technical Documentation

Implement retrieval augmented generation (RAG) with *LangChain* to create a chatbot for answering questions about technical documentation. Below two examples of very long technical and legal documents are provided. You can use one of these two or suggest another relevant document with similar characteristics. This can be discussed during the first checkpoint. If the requirements to process the full document are too big, you can process only part of the document (e.g the first 50 pages).

Document 1:

The European Union Medical Device Regulation - Regulation (EU) 2017/745 (EU MDR)
https://eumdr.com/
Document:

Document 2:

Artificial Intelligence Act
European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))
Document:
https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

Other relevant documents with similar characteristics of these above.

The steps to approach the project can be as follows:

1. Split the document
2. Generate and store the embeddings
3. Create a retriever
4. Initialize the LLM and prompt template
5. Define RAG chain
6. Invoke RAG chain

Resources:
https://www.datacamp.com/tutorial/how-to-build-llm-applications-with-langchain

Guidelines

Although there is no definitive workflow to implement a project life cycle for generative AI projects, there are several steps that when performed will guide you to make informed decisions at each step of the process.

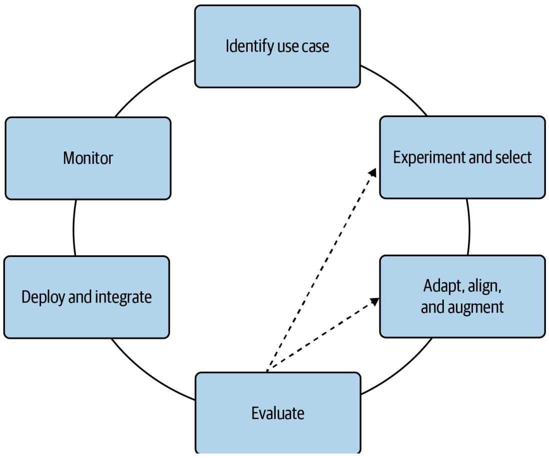


Figure 1: Framework of a generative AI project life cycle.
Image taken from the book: Generative AI on AWS by C. Freegly, A. Barth and S. Eigenbrode. O'Reilly.

Identify Use Case.

Define the scope of your project, the goals and the specific task(s) you plan to address with your gen AI application. Define the inputs and outputs, the characteristics of the data, the requirements and the possible major challenges.

Experiment and select.

Decide if exists a foundational model that is suitable for your application needs. If so, it is recommended that you start with the simplest model and test various possible models that your hardware can handle.

Adapt, align and augment.

While generative models are trained on vast amounts of data and possess enormous amount of information and knowledge, they often need to be adapted and augmented with specific knowledge that helps solve your specific problem. Select and test the most appropriate techniques for data augmentation and model alignment.

Evaluate.

To achieve models with satisfactory performance you will probably need to iterate the development process several times. It is important that you establish well-defined evaluation metrics and benchmarks to understand the progresses that result from your optimizations. Tasks of classification and regression can benefit from the traditional machine learning evaluation best practices (e.g. division in train/test/validation datasets and cross-validation).

Deploy and integrate.

Once you have well-tuned and aligned your generative model, it is time to deploy and create a prototype solution for inference. While the development of a tool is not the goal here, the creation of a minimal user-friendly interface or easy-to-use application is an aspect to be valued.

Monitor.

The monitorization of the application goes beyond the scope of this project. Nevertheless, you can discuss in the report how you would setup proper metrics and monitoring systems to follow-up all the components of your generative AI application.

Deliverables

For this project you should deliver the following elements:

- Report.
- Notebook: All code documented. Description of the datasets used.
- Video Presentation.
- Slides of the presentation and checkpoints. This should be submitted as pdf file.

Notebook

A fully operational Jupyter notebook with the selected experiments as clear and concise as possible. Described how the data was processed. Try to structure the document in different part highlighting all the steps of the process. It is recommended a creation of a github webpage with all the documentation that supports the project, including additional scripts that were developed.

Report

This should be submitted as pdf file. Document with no more than 6 pages and letter of maximum size 11pt. Avoid output dumps. Recall that the report is going to be evaluated by your very busy professors and that they will focus on the main points of the assignment. Always highlight your best results. Please note:

– The objectives for each experiment and plots should be clear so that the reader understands why it is worth to read a particular part.

– The conclusions should be a short high-level account of what was observed.

– It is not necessary to describe the technical details of the methods (unless requested, but you should know their concepts and how they work). It is more important to point out the differences in the methods and the reasons for the results in terms of methods characteristics.

Presentation

A **4 minutes video** (or link to a video), **per element of the group** with a recorded presentation of the respective part of the work. The presentations of the group, when combined, describe the whole of the group's work.

Slides

The project slides presentation and the slides at checkpoints to be submitted at different dates.

Evaluation

This assignment is worth the values described in Sigarra, according to the course you are following. Components:

Report 30%

- Narrative 10%
- Writing style 10%
- Presentation 10% (includes the checkpoints and the final presentation).

Technical 70%

- Diversity of the results for the experiments 20%
- Correctness 30%
- Challenge performance 10%
- Conclusions 10%

You are free to propose new ideas and functionalities to implement. However, this should be discussed with the responsible of the course during the classes and checkpoints. New contributions will be evaluated in the item "Diversity of the results for the experiments".

Groups

Assignments are submitted by **groups of 3 students**. Different elements may have different grades. Other group sizes will not be considered.

It is advisable that the students from the same group perform overlapping work and only after that, exchange ideas with each other. Group work is important for learning from other people.

The projects should be distributed uniformly according to the number of groups and the number the three types of projects, i.e. 5 groups will perform P1, 5 groups P2 and 5 groups P3.

Submissions

Formal final deadline is **3 of November 2024**, to be submitted in Moodle, and only in Moodle. Submissions after that date will be multiplied by a monotonously decreasing factor that starts in 1.

Checkpoints:
In the classes of the 7th to 21st of October there will be a checkpoint. Each group should present an update with the status of the project. You will have around 5 minutes for this presentation, where you can show your current results and list your main difficulties.

After each checkpoint and for the project deadline you have to submit the slides on Moodle. Only one submission per group is required, however you should indicate if any element of the group did not participate or contributed differently from the other elements.

Ethical principles

When submitting, students commit themselves to follow strong ethical principles. All the work must be done by the elements of the group alone. All members of the group will be involved with the whole of the work. All the materials used and consulted must be credited in the work.