Fundamentals of NLP: course projects 2024

Below is a description of the different project possibilities for the course of Fundamentals of NLP. You can choose one of the topics or propose your own project. Preferably the projects look at English texts, although you are free to experiment.

Make sure to carefully read the documentation below. Importantly**, don't stick to what we have seen in class but show that you understand, are able to combine different elements or add new elements**! Examples of things that we did not cover in class, but for which you can find lots of references (some in the extra course materials), are Aspect-Based Sentiment Analysis, Topic Analysis, and Document Clustering. If you need more processing power, it could be an option to look at some Colab-notebooks for certain parts of the project.

Universal guidelines

1. Study the data
    a. Inspect the data first!
    b. Using different structures and representations in Python
    c. Try different preprocessing/cleaning methods
2. Define and formalize the problem
    a. Do we need a formal model or not? (not necessary for more descriptive insights)
    b. What should the model do/predict?
    c. What type of model do we need
3. Research and brainstorm about possible solutions/tutorials/packages
4. Prototype on smaller subsets (!)
5. Refine and tune your model
6. Do not limit yourself to what we've done in class (!). If you simply apply a topic model or sentiment analysis, this is simply copy-pasting code. You must show your understanding of the materials and dare to go further.
7. Several datasets are publicly available, and some code will be available as well. I want to remind you that plagiarism is not accepted!
   You can reuse parts of the code, but in case you do, make sure to cite your source. Also, if you only copy code from others, I will of course take this into account for your grades.

Project planning

- Work on project during class
    - At the end of sessions if possible
    - Fourth session

- Autonomous finalization of the project: until April 12, 2024. (two weeks after the last session)
- Submission: per group
    - Preferably a document or presentation (can be a Markdown document or revealJS slides) with the results, interpretation and recommendations, and a description of the application functionalities, the challenges encountered and maybe what you would improve in the future
    - Code (well commented!)

Project evaluation

Project evaluation will be based on several items

- The analysis performed
    - Degree to which you go further/deeper than what is seen in class in terms of analysis
    - Correctness of the analysis
    - originality
    - Perform analyses that explore elements we haven't discussed in class
- Nicely representing the analysis in terms of tables/graphics
- Making a good report/app that clearly shows your analysis and recommendations
- For several datasets (e.g., from Kaggle), there are notebooks available that start to look into the data. Of course, I am aware of these notebooks and what they do, and the goal is really to move beyond these. They can be used as a starting point or as inspiration, but the goal is to develop your own skills!

Project Descriptions

1. ESG communications (from and about companies)

Environmental, Social and Governmental (ESG) communications are obliged for publicly listed companies in most parts of the world. We have a set of both official communications about ESG (annual reports and sustainability reports), and various news articles. The goal is to analyse to what extent ESG is discussed, and whether there are differences between the news articles and the "official" reports. This could indicate "greenwashing", in which the official communications from the firms do not really align with their actual behaviour. The analysis can consist of the use of pre-trained LLMs, the use of sentiment analysis, analysing whether certain topics (can be obtained through topic modelling or derived from the Sustainable Development Goals (SDG) which are also provided) are more prone to greenwashing or more reported on, etc. This is a very rich dataset, with a lot of possible directions to go to. You may also have a look at some fine-tuned LLMs already available in the ESG domain.

2. Analyse job descriptions

A lot of websites show job descriptions, or allow to upload CVs. Although these are more difficult to collect via an API publicly, companies that post their jobs on these websites can do so.

The goal of this analysis is to extract information from the next, so with a strong focus on entity recognition and the usage of regular expressions. Try to extract the job description, the location, the current company and skills.

You will get a dataset of ca 2000 job descriptions extracted from CVs. Take a closer look at research on named entity recognition and extraction. You might check the SpaCy, or discover other options for NER (e.g., openNLP, or the use of the Google natural language API). Also use regex to find relevant info.

This project is somewhat further from the course, but I will of course take this into account in the grading. You can also use the document "Skill Finder: Automated Job-Resume Matching System" to get some ideas.

3. Predictions of online engagement for tweets

The goal of this project is simple: predict which posts get more engagement, by focusing on the text of tweets.

Hints:
- Make a set of independent variables to explain engagement. You can think of sentiment and topic variables, but also other variables related to the text can be included.
- Use a machine learning method or a traditional statistical model (e.g., regression) to model this.
- Make sure you include some evaluation measure
- There is already quite some academic literature on this topic, as well as online sources. Be innovative enough here!

4. Analyze reviews in terms of content and sentiment

Analyze a set of reviews (there are a lot of review sets available online from Yelp or Amazon).

You can link the sentiment, and for instance try to find sentiment for specific elements mentioned (atmosphere, food quality, …) in the review. This can be linked to dependencies, and you might want to have a look at noun chunks (SpaCy has several options for this, see for instance: https://spacy.io/usage/linguistic-features).

This can be linked for instance to the helpfulness of the review. Moreover, other elements of the text (length, content, …) can serve as independent variables to predict helpfulness.

5. Using song lyrics to predict success

For this project, you will be provided with a dataset created by students in a previous year (who did a great job!).

The dataset contains data from Spotify (information on the music, and some index of popularity), but also the lyrics, which were scraped from Genius. By using this combination, you should try to predict a song's success (of course mostly based on the text!).

6. Analyzing donor request and chance of success

For this project, you will use a dataset of DonorChoose. This dataset includes requests (mostly from teachers) for materials that are needed. Subsequently, people can sign up and donate money to a project/request. There is a target goal of money to collect, so one can assess whether the call has been successful.

Your task is to analyze whether and how the content and/or style of a request has an influence on the success rate. Similarity to other calls can also be an interesting avenue to pursue in this setting.

7. Analyzing fake news

Fake news datasets are problematic for governments, individuals and companies. In this project, it is your task to make a model to predict whether a news article is fake or not, based on the text of the article. There is already quite some research in this area, so you should be creative and think about specific elements that can be informative. Experimenting with LLMs can also be an option, although you will probably be limited in the resources you have available.

8. Detect climate-related text

Mathias Kraus, one of the speakers at the MBD-event, has a series on LLM models designed to look at climate-related texts (https://www.chatclimate.ai/climatebert). In this project, you can work on these models and datasets. You will be provided with a dataset of news articles about a few companies. Your goal is to investigate to which extent climate is being discussed in this data. You can look at the sentiment regarding climate, specificity, … In order to do so, you can have a look at the LLM models already developed, but you can also use the datasets they used to train the models (see the website) and use these to build your own (simpler) predictive model.

9. Feel free to propose your own project!