Capstone Project

Task Description

- **Your task**: Carry out a small but complete social media analytics project. Describe a real or fictitious scenario for your project, explaining its goal and purpose. The project should have its main focus on techniques covered in the course (e.g. web scraping, network analysis, text mining, sentiment analysis, text or token classification, LLM applications, retrieval augmented generation, topic modeling). You should NOT cover all techniques, but focus on those techniques that are useful for your project, i.e. for the problem or question that you are addressing. I want to inspire your creativity and give room for different kinds of projects (machine learning, exploratory, LLMs, ...). If you are unsure whether your project idea is suitable, then ask me.
- **Language**: Allowed languages are English and German.
- **Data**: Your project must be based on a real-world data set. You can either use an existing data set or engineer your own data set via APIs, web scraping, and possibly annotating the data. The data context can be social media platforms or other forms of text data such as news articles, product reviews, etc. If you intend to carry out a text classification task, keep in mind that you need a data set that comes with labels (which is not the standard case with text data), or spend some time on annotating data yourself.
- **Focus**: If your project has a significant data engineering part (data acquisition, cleaning, annotation, etc.), then a compact analytical part is sufficient. If your project is based on an existing data set, then I expect a larger scope and/or depth on the analytical side (in-depth evaluation of multiple models and parameter choices).
- **Documentation**: Include all relevant steps of your project into a Jupyter notebook and guide the reader through the project in your own words. Make sure that the reader can follow your data processing steps and your thought process. In particular, carefully evaluate the results and explain the main insighs and learnings. It may make sense to split your submission into two Notebooks, e.g. if you have both a data engineering and an analytical part.
- **Resources**: You may use all the code from the lectures. Copying and adapting from other sources is allowed in small quantities. Copying code in large quantities will be treated as intent to deceive and result in a score of zero points. Cite all relevant resources on which your project is based or from which you draw inspiration.
- **Chat GPT usage**: You are encouraged to use AI assistance to learn about concepts and approaches, write better code or similar tasks. However, YOU are the author and must therefore 100% know (and be able to explain) what you are doing and why. If you use AI assistance, you must clearly state this in your submission.
- **Submissions**: Submit all that is needed to fully reproduce your work (notebook, scripts, data, etc.) on Moodle, or submit a link to a public GitHub repository. If you are using some API, you should not provide API keys, but rather a script that can be used to reproduce the data acquisition process. The raw data used in the project must be included in the submission.

Continuous Work and Feedback

- In the beginning of the semester, it can be difficult to make a decision on the project topic, because you may not yet have a complete understanding of the topics covered towards the end of the semester. However, you should start thinking about potential project ideas early on. And you should continue working on the project throughout the semester.
- As part of your problemsets you will be asked to (1) investigate data access options, and (2) write a project proposal, present it in person to the lecturer and collect feedback from him.

Grading

Due to the different nature of the projects, the grading will be based on a holistic evaluation of the project. The following aspects will be considered:

- Creativity, complexity and innovativeness of the project
- Correctness of the approach
- Data engineering efforts
- Thorough evaluation and correct interpretation of the results
- Convincing Storytelling: purpose of the project, insights, reflection on learnings and limitations
- Well-structured, concise and clean submission
- "ChatGPT buzzword bingo" will be considered as a malus. Write in your own words!

Selected Projects from past semesters

- Analyzing trends and patterns in Youtube Comments on US Elections
- Implement a Retrieval Augmented Generation (RAG) system on a network dataset
- RAG Chatbot providing medical information to patients based on data scraped from Wikipedia
- Predicting relationship between Reddit Sentiment and Share Prices
- Developing a search engine prototype focussing on newly published papers
- Predict sentiment, emotions, and psychological content in mental health forums on Reddit
- Eminem song lyrics analyis
- Predicting NVIDIA stock movements based on press releases
- Analysis and Recommender System based on Coffee Reviews