SMA: Capstone Project

1. Overview

Carry out a social media analytics project on a topic and dataset of your choice. The project must use a selection of the techniques covered in the lecture: e.g. web scraping, sentiment analysis, text mining, text classification, LLM applications, retrieval augmented generation, network analysis, or related techniques. You should certainly NOT cover all of these techniques, but focus on whatever is useful for the problem that you are addressing.

I want to inspire your creativity and give room for different kinds of projects (exploratory, machine learning, LLMs, etc.). If you are not sure whether your project idea is suitable, then ask me!

2. Choosing a data source and topic

- Your project must be based on a real-world data set
- The data context can be a social media platform, but also other forms/sources of text data (e.g. product reviews, news articles, company website, etc.) or network data are fine.
- You can either use (1) an existing data set or (2) engineer your own data set via APIs, web scraping, combining data from multiple sources, or annotation.
- If you project has a significant data engineering part this may involve considerable upfront efforts. Such **data engineering efforts will be acknowledged in the grading** (see below). Conversely, if you use an already fully cleaned data set, I expect a larger scope, complexity or innovativeness in terms of the actual visualizations and the story.
- If you intend to train a supervised ML model, then keep in mind that you need a data set that comes with labels, which is often not automatically the case with online text data. A possible solution is to annotate some of the data yourself, either manually or via a semi-automatic process. Gaining experience with this process is a valuable skill and can be a part of your project.

In the beginning of the semester, choosing the "right" data source and topic can be difficult. To give you some inspiration, here is a **selection of topics from past semesters** (data source in brackets):

- Analysis of Lyrics of Eminem Songs (Genius.com)
- Analysis of Patterns in Youtube Comments related to US Elections (Youtube)

- Analysis and Development of a Recommender System based on Coffee Reviews
- Analysis and Prediction of User Satisfaction for Hotels (Booking.com)
- Predicting sentiment and emotions in mental health forums (Reddit)
- Predicting Stock Movements Based on Press Releases (Nvidia; Yahoo Finance)
- Development of a Market Screener for Consumer Products (Ebay and Reddit)
- Development of a Semantic Search Engine for Supermarket Products (Lidl)
- Development of a RAG-based Personal Podcast Recommender
- Development of a RAG Chatbot Providing Medical Information (Wikipedia)
- Development of a RAG System on a Network Dataset of Lecture Notes

3. Deliverables

Submit a zip file containing the following:

PDF project report

The project report is the main output of your Capstone Project. It guides the reader through your Capstone project: What is the motivation and goal? How do you collect, process, and analyze/model the data? What are your results? What are the main insights and learnings? The report should be concise (about 5 pages long), well-structured, and explain the project in a clear language. Results should be condensed into tables or figures that make it easy to take away the most important findings.

Jupyter notebook(s)

- The Jupyter Notebook includes code, and additional/technical explanations not covered in the project report.
- If executed from top to bottom, the notebook should reproduce all results and visualizations that are presented in the project report. However, it may also contain additional analyses that provide further background or insights.
- Here you can provide relevant background information about e.g. the data, technical details, or assumptions using Markdown. Code-related explanations can be made in the form of code comments.
- The notebook should be well-structured and cleaned up.
- You may split your code into 2 notebooks, e.g. if you have a data engineering part and an analytical part.

Further resources

- Data: provide the data set(s) used in your project. Also provide the data obtained from web scraping, APIs, etc. I need to be able to reproduce your results without having to scrape the data myself.
- Helper scripts
- etc.

4. Allowed resources and languages

- Cite all relevant resources on which your project is based or from which you draw inspiration.
- You may use all the code from the lectures. Copying and adapting from other sources is allowed in small quantities.
- 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.
- Allowed languages: English and German.

5. Grading

Your project is graded based on a holistic evaluation of the following aspects:

- Data engineering efforts
- Complexity and innovativeness
- Correctness of approach
- Thorough evaluation and correct interpretation
- Convincing Storytelling
- Well-structured, concise and clean submission
- "ChatGPT buzzword bingo" will be considered a significant malus. Write in your own words!