Problemset 1

The goal is to - collaboratively - explore how structured and unstructured data can be accessed from social media platforms, online portals, and forums:

- Each student chooses a specific source (social network, online portal, forum, . . .). We distribute the sources among all students during the lecture. Please indicate your choice in the following table: https://bit.ly/sma_source_allocation
- A maximum of 2 students are allowed per source. You may team up with another student, or work individually on the task. If you do it as a team, state this in the beginning of your submission.
- Each student investigates available options for data collection, and provides a concise summary of the insights gained.
- The submissions will be shared with the entire class. Thus, make your submission as informative and useful as possible for your fellow students. But keep it short, so that the key takeaways become clear.

1. Overview

1. Briefly describe your source: What are the main contents / what is the main purpose of your source? What kind of data is available? (Keep it short, focus on the core aspects!)
2. Investigate whether it is possible to access data from your source, via (1) APIs, (2) Python Packages, or (3) web scraping. Provide a structured overview (e.g. a table) of available options, e.g. with links to relevant information, and a comment on potential pecularities/problems of these options.
3. Based on your initial investigation, what is a typical or recommended way of accessing data from this source?

Important note: For this exercise you do not need to test the options yourself. It is sufficient to summarise your insights from your initial investigation.

2. Mini Tutorial

Choose a suitable data access option from above, and provide a mini tutorial for your fellow stud ents:

- What are the main steps involved? (login, API key, Python package installation, etc.)
- Provide demo code that illustrates how data from this source can be accessed. Do not scrape large amounts of data. Only provide a small code example. Return the first few rows of the retrieved data.
- Briefly discuss: Can this method be extended to further type of information, or easily scaled to access more data of the same type? Did you experience any limitations? Are there ethical concerns to be considered?

3. Use case

Describe **properties of the data** that are particularly interesting in the context of our course, and describe **one use case** for how these data could be used. What would be the goal or benefit of this use case?

Consider for example:

1. Are there **labels** (ratings, likes, tags, etc.) that could be used for **supervised ML**?
2. Are there **network structures** (e.g., replies, mentions, shares) that could be analyzed?
3. Could a **retrieval-augmented generation (RAG)** system be built based on your data?