TECHNICAL DOCUMENTATION

FIRST FIGURE: Web Scraping and counting

I have performed the web scraping on Reddit which gives you completely access to internal subreddits and comments.
To do that I have used a python library called **praw**, here's a brief description taken from the documentation that you can find here
[https://praw.readthedocs.io/en/stable/getting_started/quick_start.html]

PRAW (Python Reddit API Wrapper) is a Python library that provides a simple, organized way to access and interact with Reddit's API. Designed for ease of use, PRAW abstracts the complexities of making API requests, enabling users to authenticate, retrieve, and post Reddit content with minimal code. It supports both read-only and authorized modes, allowing users to read public Reddit data, submit new posts, comment, and perform other actions as permitted by the API. PRAW is commonly used for building Reddit bots, data collection scripts, and automation workflows within Python projects

In this part I have created a dataset based on 5 different drugs, from the most relevant subreddits based on number of participants and comments.

Output:
A dataset with +400k samples, each sample composed of:
- text
- date
- tag
- number of comments below

You can find the notebook for this part in this directory:
github.com/emaalberti/UISProject_NLPAnalysis/Notebooks/RedditScraper.ipynb

The dataset is uploaded on our personal google drive.

SECOND FIGURE: Semantic Analysis with Bertopic

In this part we focused on understanding the most discussed topics for drugs. Our goal is to figure out if there are some differences or patterns between them to point out.

For this part I implemented an already existing model called **Bertopic,** here's a brief description:

BERTopic is a **topic modeling framework** for Python that utilizes **transformer-based embeddings** and class-based TF-IDF (c-TF-IDF) to generate interpretable topic clusters from text data.
BERTopic offers **modularity** in its **workflow**, allowing users to customize each step:
- embedding,
- dimensionality reduction,
- clustering,
- topic representation
to suit diverse analytical requirements.

I have selected each step after tuning and validating, using appropriate quantitative metrics relevant to its stage. For **embedding** and **dimensionality reduction**, intrinsic metrics such as clustering quality (e.g., **silhouette score**).
For clustering, both **intrinsic measures (Adjusted Rand Index (ARI) and silhouette score)** and **topic assignment** stability.
For the vectorizer, I reviewed the official documentation and selected the implementation that best aligned with our project's objectives. After validating its effectiveness with our data and evaluation metrics, I confirmed that it provided high-quality topic representations, so we proceeded to adopt it for the final pipeline.

- **Embedding:** model "all-mpnet-v2" from hugging face
- **Dimensionality Reduction:** PCA
- **Clustering:** HDBSCAN
- **Vectorizer** for topic representation