Foundations of Data Science
Mini-project 2
Due date: 10 January 2022

Tasks:

Using the Drug Review dataset described here
https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29#

- Predict effectiveness rating based on the benefits review.
- Predict side effects rating based on the side effects review.
- Classify the review texts as belonging to the benefit, side effect, or general category.
- For each task, try feature selection (e.g. n-gram size, how many / which features to use, based on IDF / Information Gain, …) and grid-search to find the best model and parameters.
- Evaluate the final models on the test data.
- Apply your classification models to this unclassified dataset and analyse your results: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29
- Apply a sentiment analysis tool to identify the drugs that receive most negative reviews. You may use spaCyTextBlob[1], VADER[2], or other tools that you identify.

Submission:

Submit a Jupyter notebook with your project or share your (private) git repository.
The notebook should be well organised and documented, describing all data preparation and analysis steps.

(1)
https://textblob.readthedocs.io/en/dev/api_reference.html#textblob.blob.TextBlob.sentiment
https://textblob.readthedocs.io/en/dev/_modules/textblob/en/sentiments.html
https://github.com/clips/pattern

(2)
https://github.com/cjhutto/vaderSentiment