Homework 4

You can submit in groups of 2.
Due 4/22, 1pm.

All assignments need to be submitted via github classroom:
https://classroom.github.com/g/692pxJCi
and via gradescope.

Change in procedure: this is a "programming assignment" on gradescope.
Please upload your jupyter notebooks directly, not PDFs.
You can upload them directly from your github. Also, please separate the three tasks into three different notebooks, called task1.ipynb, task2.ipynb and task3.ipynb (optional).
The total points for the homework are capped at 100 as before. Task3 provides the option to make up points you didn't get in Task 1 and Task 2. It's likely to be more work but also will be a good learning experience.

In this homework, we try to solve the problem of predicting wine quality from review texts and other properties of the wine. You can find the dataset here:
https://www.kaggle.com/zynicide/wine-reviews

While you can find several kernels on kaggle already, I highly recommend you start your own solution from scratch. For this homework, only use wine from the United States.
Feel free to subsample the data for building your model.

Task 1 Bag of Words and simple Features [50pts]

1.1 Create a baseline model for predicting wine quality using only non-text features.
1.2 Create a simple text-based model using a bag-of-words approach and a linear model.
1.2 Try using n-grams, characters, tf-idf rescaling and possibly other ways to tune the BoW model. Be aware that you might need to adjust the (regularization of the) linear model for different feature sets.
1.3 Combine the non-text features and the text features. How does adding those features improve upon just using bag-of-words?

Task 2 Word Vectors [50pts]

Use a pretrained word-embedding (word2vec, glove or fasttext) for featurization instead of the bag-of-words model. Does this improve classification? How about combining the embedded words with the BoW model?

Task 3 Transformers (bonus / optional) [50pts]

Fine-tune a BERT model on the text data alone using the transformers library.
How does this model compare to a BoW model, and how does it compare to a model using all features?