Homework 4

You can submit in groups of 2.
Due 4/17, 1pm.

All assignments need to be submitted via github classroom:
https://classroom.github.com/g/xng4nbv
and via gradescope.

In this homework, we try to solve the problem of moderating the science subreddit r/science.
You can find the data here:
https://www.kaggle.com/areeves87/rscience-popular-comment-removal

While you can find several kernels on kaggle already, I highly recommend you start your own solution from scratch. For this homework, only use the reddit_200k training and test set, and only use the "body" and "removed" columns.
Be careful about the encoding when loading the data.
Pick an appropriate evaluation metric for imbalanced binary classification.

Task 1 Bag of Words and simple Features

1.1 Create a baseline 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 Explore other features you can derive from the text, such as html, length, punctuation, capitalization or other features you deem important from exploring the dataset

Task 2 Word Vectors

Use a pretrained word-embedding (word2vec, glove or fasttext) instead of the bag-of-words model. Does this improve classification?