Homework 1

The goal of this assignment is to give you a sense of the challenges an NLP practitioner often faces. Additionally, this is an opportunity to experiment with Python. No knowledge of NLP is required to complete this assignment. However at least basic knowledge of Python is needed. Many details of this assignment are left intentionally vague.

You will be working with movie review data, which you can download here:

https://github.com/dennybritz/cnn-text-classification-tf/tree/master/data/rt-polaritydata

You will find two files there (one with positive and one with negative reviews). Your task is to build a binary classifier in Python that will perform movie review classification automatically. Most likely, it will be a rule-based classifier implementing such rules as:

- If I see the word 'amazing', the review is positive
- If I see the word 'awful', the review is negative

Once you finalize your classifier, evaluate by calculating how accurate it is (i.e. how many samples it classifies correctly out of the total number of samples).

I am intentionally leaving the details of this assignment vague, to mimic the situation you may encounter in your future jobs. You will often be given a dataset and asked to model it without much additional information. I will be happy, though, to answer any questions you might have on Slack, in class, or during my office hours.

I do not expect you to solve this problem. However, you must give it a try even if it is just to understand how difficult this problem is.

The ability to summarize your findings is extremely important when doing research or working as a data scientist. Please summarize your findings in a (up to) 1-page paper. Include the results of your evaluation. This assignments will be evaluated primarily based on your writeup.

Please submit the assignment using Sakai. The information about what to submit is included in the syllabus.

Note: for this assignment, you are not allowed to use machine learning libraries (e.g. scikit-learn)!

Good luck!