Assignment 3: sentiment analysis

Text mining course

This is an **individual hand-in assignment**. Your report should not be longer than 3 pages

Goals of this assignment

- You can perform a sentiment classification task with doc2vec feature representation
- You can make sensible adaptations to the pre-processing pipeline, feature representations, and classifier learning steps for text classification tasks
- You can evaluate sentiment classification with the suitable evaluation metrics

Preliminaries

- You have followed the tutorial on sentiment analysis with doc2vec: https://github.com/linanqiu/word2vec-sentiments
- You have all the required Python packages installed

Tasks

1. Download the *Large Movie Review Dataset v1.0* (aclImdb_v1.tar.gz) from http://ai.stanford.edu/~amaas/data/sentiment/ and unzip it
2. Read the documentation (README) to make sure you understand the data
3. Add a function to the script provided in the tutorial that reads the positive and negative examples and stores them as training data and test data
4. Pre-process the data by lowercasing and removing punctuation
5. Run a baseline run using doc2vec with the code from the tutorial (note that it will cost time to get your data exactly in the form that the script requires)
6. Experiment by making adaptations to your code. You might consider changing the pre-processing, changing the word embeddings model, changing the feature representation, adding features, changing the classifier, or tuning the classifier.

Write a two-page report (3 pages is the hard maximum) in which you:

- describe the data (give a few statistics)
- describe the method adaptations that you made
- show a results table with both the baseline results and results for at least three experimental settings. Describe what changes to made to obtain the presented results.
- write your conclusions.