# Programming Assignment 1

COMP 550, Fall 2021

Due: **Friday, October 1$^{st}$**, 2021, 9:00pm.

You must do this assignment individually. You may consult with other students orally, but may not take notes or share code, and you must complete the final submission on your own.

## Sentiment Analysis

You will train models that classify a sentence into either a positive or negative sentiment. These sentences come from a movie review dataset constructed by the authors of this paper:

Bo Pang and Lillian Lee, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, *Proceedings of ACL 2005*.

The goal of this assignment is to give you experience in using existing tools for machine learning and natural language processing to solve a classification task. Before you attempt this assignment, you will need to install Python 3 on the machine you plan to work on, as well as the following Python packages and their dependencies:

- NLTK: `http://www.nltk.org/`
- NumPy: `http://www.numpy.org/`
- scikit-learn: `http://scikit-learn.org/stable/`

Download the corpus of text available in the attached file. This corpus is a collection of movie review sentences that are separated into positive and negative polarity. Your task is to train classifiers to distinguish them.

### Data storage and format

The raw text files are stored in *rt-polarity.neg* for the negative cases, and *rt-polarity.pos* for the positive cases.

### Research question

The main research question being asked by your experiments will be: what preprocessing decisions work well for sentence-level sentiment classification?

### Preprocessing and feature extraction

Your responsibility is to design and run the correct experiments in order to answer the research question above. You must explore at least 3 preprocessing decisions that we have discussed in class. You may use scikit-learn's feature extraction module to help you, as well as any other tool from NLTK or NumPy. Reading scikit-learn's documentation will be of great help in your experimentation.

### Setting up the experiments

Design and implement experiments to draw reasonable conclusions about the research question above. This will require creating subsets of the dataset as we discussed in class. There are multiple correct ways to set up your experiments (as well as many incorrect ways). Stick to the logistic regression model for this assignment.

### Report

Write a *short* report on your method and results, carefully document i) the problem setup, ii) your experimental procedure, iii) the range of parameter settings that you tried, iv) the results and conclusions, and v) the limitations of your study. It should be no more than 1.5 pages long. Report on the performance in terms of accuracy, and speculate on the successes and failures of the models.

Your assignment will be marked on i) how well it satisfies the requirements stated in this handout, ii) whether your experiments adequately and correctly address the research question, iii) how well written your report is. It will NOT be marked based on the performance that you achieve with your models on this dataset.

### Submitting code

Submit your code in a file named “a1.py”.

## What To Submit

Submit your report as a single pdf on myCourses called “a1-answers.pdf”. In addition, you should submit one plaintext file with your source code called “a1.py”. All work should be submitted to myCourses under the Assignment 1 folder.