To complete this lab, you might use the following tools:

* File operations
* String operations
* Text processing: NLTK
* Machine learning: scikit-learn
* Numpy, Pandas, regular expression

Large Movie Review Dataset

This dataset contains movie reviews along with their associated binary sentiment polarity labels. It is intended to serve as a benchmark for sentiment classification. This document outlines how the dataset was gathered, and how to use the files provided.

There are two directories [pos/, neg/] for the reviews with binary labels positive and negative. Within these directories, reviews are stored in text files named following the convention [[id]\_[rating].txt] where [id] is a unique id and [rating] is the star rating for that review on a 1-10 scale. For example, the file [pos/200\_8.txt] is the text for a positive-labeled example with unique id 200 and star rating 8/10 from IMDb.

*References*

Potts, Christopher. 2011. On the negativity of negation. In Nan Li and David Lutz, eds., Proceedings of Semantics and Linguistic Theory 20, 636-659.

You need to build a machine-learning model (binary classifier) to predict sentiment for a given unlabeled text. This task includes the following major components:

1. Preprocess the data: remove stopwords, non-word characters, all lowercase, etc.

2. Convert text into feature representations and build the training / testing set: uni-gram, bi-gram, or tri-gram? Count or TFIDF? Some other features (be creative)?

3. Select a good model and test the performance: which model? How to evaluate?

4. Predict sentiment for a new (unlabeled) document / sentence: make prediction

You can also build a multi-class classifier (each rating is a label) rather than the simple binary situation.