Assignment 5: Natural Language Processing

In this assignment we will work on analyzing text data. A tremendous amount of data is available as text data, from digitized books, documents, articles, to customer reviews and social media exchanges. However, analyzing this data requires different processing steps, as compared to quantitative data, that we will explore here.

Tasks / Learning Goals

- To work with analyzing text data, exploring natural language processing
- Explore different methods to encode text data
    - In particular, 'Bag of Words' and TF/IDF approaches
- Practice classification models, in particular, using SVMs

Due Date

11:59 pm on Sunday, March 11th, submitted on TritonED.

Submitting Assignments

You must submit the provided Jupyter notebook file (.ipynb) to TritonED. Make sure that the file you submit has the following filename (filled in with your course ID number - first letter of your last name, followed by the last 4 numbers of your student ID number):

'A5_$####.ipynb'

Grading Rubric

This assignment is worth 12% of your grade (12 points).
There are 3 parts to this assignment, with the following point values:

Part 1	Bag of Words	5.5 points
Part 2	TF/IDF	2.5 points
Part 3	Customer Reviews	4.0 points

Associated Data Files

- data/rt-polarity.tsv
- data/custrev_train.tsv
- data/custrev_test.tsv