COMP41680 Assignment 2: Text Classification

Deadline: Friday 26th April 2019

Overview:

The objective of this assignment is to scrape consumer reviews from a set of web pages and evaluate the performance of text classification on the data. The reviews have been divided into five categories here:

http://mlg.ucd.ie/modules/yalp

Each review has a star rating. For this assignment, we will assume that 1-star to 3-star reviews are "negative", and 4-star to 5-star reviews as "positive".

The assignment should be implemented as a single Jupyter Notebook (not a script). Your notebook should be clearly documented, using comments and Markdown cells to explain the code and results. The assignment can be completed either individually or in pairs.

Tasks:

In this assignment you should complete all of the following tasks:

1. Select two review categories of your choice. Scrape all reviews for each category and store them as two separate datasets. For each review, you should store the review text and a class label (i.e. whether the review is "positive" or "negative").
2. For both category datasets:
    a. From the reviews in this category, apply appropriate preprocessing steps to create a numeric representation of the data, suitable for classification.
    b. Build a classification model using a classifier of your choice, to distinguish between "positive" and "negative" reviews.
    c. Test the predictions of the classification model using an appropriate evaluation strategy. Report and discuss the evaluation results in your notebook.
3. Evaluate how well your two classification models transfer between category. That is, run experiments to:
    a. Train a classification model on the data from "Category A", and evaluate its performance on the data from "Category B".
    b. Train a classification model on the data from "Category B", and evaluate its performance on the data from "Category A".

Guidelines:

- The assignment can be completed either individually or in pairs. Any evidence of plagiarism will result in a 0 grade.
- For the assignment, only these third-party packages can be used: NumPy, Pandas, Scikit-learn, NLTK, Gensim, SciPy, Requests, BeautifulSoup, Matplotlib, Seaborn.
- Submit your assignment via the COMP41680 Moodle page. Your submission should be in the form of a single ZIP file containing the notebook and your data. In the

notebook please clearly state your full name(s) and student number(s). Students working in pairs only need to make one submission from either student.

- Hard deadline: Submit by the end of Friday 26th April 2019
  - 1-5 days late: 10% deduction from overall mark
  - 6-10 days late: 20% deduction from overall mark
  - No assignments accepted after 10 days without extenuating circumstances approval and/or medical certificate.