COMP41680 Assignment 2

Deadline: Sunday 24th April 2022

Overview:

The objective of this assignment is to scrape a collection of product reviews from a set of web pages, preprocess the data, and evaluate the performance of different classifiers in the context of two related text classification tasks: (i) predicting review sentiment; (ii) predicting review helpfulness.

The assignment should be implemented as two Jupyter Notebooks (not script files). Your notebooks should be clearly documented, using comments and Markdown cells to explain the code and interpret the results of your analysis.

Complete the three tasks below in two separate notebooks. Note that Task 1 should be completed in the 1st notebook, while Tasks 2 and 3 should be in the 2nd notebook.

Task 1. Data Collection

1. Scrape the complete set of web pages from your personal website address:

   http://mlg.ucd.ie/modules/python/assign2/<STUDENT_NUMBER>/

   For example, if your student number is 195023491, your web pages are at:

   *http://mlg.ucd.ie/modules/python/assign2/*195023491*/*
2. From the web pages above, parse every review across all years 2016-2021. For each product review, extract the following information:
   i) The star rating of the review
   ii) The title text of the review
   iii) The main body text of the review
   iv) Review helpfulness information
3. Store the parsed review data in an appropriate format.

Task 2. Review Sentiment Classification

1. Load the data from Task 1 and create a set of documents, one per review. Each document should consist of the concatenation of the review's title and body text.
2. Assign a class label ("positive" or "negative") to each review. We will assume that 1-star to 3-star reviews are "negative", and 4-star to 5-star reviews are "positive".
3. Apply appropriate preprocessing steps to create a numeric representation of the documents, suitable for classification.
4. Build two different binary classification models using two classifiers of your choice, to distinguish between "positive" and "negative" reviews.
5. Compare the performance of the classification models using an appropriate evaluation strategy. Report and discuss the evaluation results.

Task 3. Review Helpfulness Classification

1. Assign a class label (“helpful” or “unhelpful”) to each review in your dataset from Task 2, based on its associated helpfulness information.
2. Build two different binary classification models using two classifiers of your choice, to distinguish between “helpful” and “unhelpful” reviews.
3. Compare the performance of the classification models using an appropriate evaluation strategy. Report and discuss the evaluation results.
4. Based on the evaluation results from both Tasks 2 and 3, compare and discuss the differences in performance for the two classification tasks (i.e. sentiment and helpfulness classification).

Guidelines:

- The assignment should be completed individually. Any evidence of plagiarism will result in a 0 grade.
- The grade awarded will depend on the complexity of the analysis and level of detail, i.e., data preprocessing, classifier evaluation and comparison etc.
- Submit your assignment via the COMP41680 Brightspace page. Your submission should be in the form of a single ZIP file containing the two notebooks (i.e. IPYNB files) and your data as stored in Task 1.
- Hard deadline: Submit by end of **Sunday 24th April 2022**
    - 1-5 days late: 1 grade point deduction, e.g. B to B-
    - 6-10 days late: 2 grade point deduction, e.g. B to C+
    - Assignments will not be accepted after 10 working days without Extenuating Circumstances formally approved by UCD.