CSCI - 4146 - The Process of Data Science - Summer 2022

Assignment 3

The submission must be done through Brightspace.

Due date and time as shown on Brightspace under Assignments.

- To prepare your assignment solution use the assignment template notebook available on Brightspace.
- The detailed requirements for your writing and code can be found in the evaluation rubric document on Brightspace.
- Questions will be marked individually with a letter grade. Their weights are shown in parentheses after the question.
- Assignments can be done by a pair of students, or individually. If the submission is by a pair of students, only one of the students should submit the assignment on Brightspace.
- We will use plagiarism tools to detect any type of cheating and copying (your code and PDF).
- Your submission is a single Jupyter notebook and a PDF (With the compiled results generated by your Jupyter notebook). File names should be:
    - **A3-<your_name1>-<your_name2>.ipynb**
    - **A3-<your_name1>-<your_name2>.pdf**
- **Forgetting to submit both files results in 0 markings for both students.**

In this assignment, you will build models for text classification on the corpus. We will use the Amazon Product Reviews dataset:
http://deepyeti.ucsd.edu/jianmo/amazon/categoryFilesSmall/Books_5.json.gz

This dataset might be too large to fit into your computer's memory. Therefore, use 1 million entries during the data understanding and exploration phase, and 100 thousand entries while building a model. The data is quite large in size and might not fit in your memory. Thus, you have two options:

- Use our already provided subsample of 1 million entries. https://www.kaggle.com/datasets/parvezmrobin/amazon-book-review-1m-sample
- Create your own subsample from the original dataset. Remember that you cannot just use the first 1 million entries. If you create your own subsample, **you will have 5% bonus marks**. Note that, this bonus marks will saturate on 100% which means anything you get above 100% will have no effect.

1. **Data understanding (0.15).** The majority of the features in the dataset are textual data, for which a general data quality report doesn't provide a lot of insights. Therefore, for the purpose of building a data quality report, we will substitute the actual text items with their properties such as:
    - Text length (i.e., the number of characters).
    - The number of words.
    - Presence of non-alphanumeric characters.
    - Any additional properties that you find useful in understanding text.

    1. Build the data quality report.
    2. Identify data quality issues and build the data quality plan.
    3. Preprocess your data according to the data quality plan.
    4. Answer the following questions:
        1. What is the distribution of the top 50 most frequent words (excluding the stop words) for each of the textual features?
        2. What is the proportion of each format in the dataset?
        3. What is the most/least common format of the books?
        4. What patterns can you find in your data? E.g., if you look at the counts for each overall score, people tend to give more positive reviews than negatives. (you are encouraged to find different patterns to the one proposed here as an example)

2. **Text normalization and feature engineering (0.2)**
    1. Create a new column merging review summary and text.
    2. Remove stop words.
    3. Remove numbers and other non-letter characters.
    4. Perform either lemmatization or stemming. Motivate your choice.
    5. Convert the corpus into a bag-of-words TF-IDF weighted vector representation.

3. **Build a model to predict overall score (0.3)**
    1. Use score as the target variable. Explain what is the task you're solving (e.g., supervised x unsupervised, classification x regression x clustering or similarity matching x etc).
    2. Use a feature selection method to select the features to build a model.
    3. Select the evaluation metric/metrics. Justify your choice.
    4. Perform hyperparameter tuning if applicable.
    5. Train and evaluate your model.

6. How do you make sure not to overfit?
7. Plot a visualization of the learning process or the learned information of the model.
8. Analyze the results.

4. **Perform part-of-speech tagging (0.35)**
    1. Perform part-of-speech tagging on the raw data (i.e. prior to Q2), and after processing (after Q2), and extract the nouns only to obtain a bag-of-words tf-idf weighted vector. representation using only the nouns.
    2. Repeat question Q3.
    3. Compare the performance with what you received in Q3 and Q4 with a statistical significance test. Discuss your findings.