CSCI - 4146/6409 - The Process of Data Science - Fall 2022
Assignment 4

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:
    - **A4-<your_name1>-<your_name2>.ipynb**
    - **A4-<your_name1>-<your_name2>.pdf**
- **Forgetting to submit both files results in 0 mark for both students.**

In this assignment, you will build models for text classification on the corpus. We will use the Wikipedia movie plots data set: https://www.kaggle.com/jrobischon/wikipedia-movie-plots.

1. Data understanding (0.2). 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 purposes of building a data quality report, we will substitute the actual text items with their properties such as:

- Text length
- The number of words
- Presence of non-alphanumeric characters
- Any additional properties that you find useful in understanding text

a. Build the data quality report

b. Identify data quality issues and build the data quality plan

c. Preprocess your data according to the data quality plan

d. Answer the following questions:

i. What is the distribution of the top 50 most frequent words (excluding the stop words) for each of the textual features?

ii. What is the proportion of each genre in the dataset?

iii. What is the most/least common origin of the movies?

iv. What trends can you find in your data?

2. Genres selection and understanding. (0.2)

a. Select the following five movie genres (classes): Drama, comedy, adventure, romance, western.

Some of the instances in the dataset belong to multiple genres. Be sure to motivate your selection strategy for those cases.

b. What is the most frequent word in the title of the movies for each of the genres?

c. What is the distribution of words of the plot description between the genres?

d. Calculate vocabulary alignments between the genres. Vocabularies alignment is defined as the percentage of the top 1000 most frequent words from one corpus present in the top 1000 most frequent words of another corpus.

3. Text normalization and feature engineering (0.1)

a. Remove stop words

b. Remove numbers and other non-letter characters

c. Perform either lemmatization or stemming. Motivate your choice.

d. Convert the corpus into a bag-of-words tf-idf weighted vector representation.

4. Build a model to predict movie genres (0.2)

a. Explain what is the task you're solving (e.g., supervised or unsupervised, classification or regression or clustering or similarity matching etc)

b. Use a feature selection method to select the features to build a model.

c. Select the evaluation metric. Justify your choice.

d. Perform hyperparameter tuning if applicable.

e. Train and evaluate your model. Report the confusion matrix.

f. How do you make sure not to overfit?

g. Plot learning curve

h. Analyze the results

5. Apply part-of-speech tagging (0.1)

a. Filter the text to only retain the nouns and repeat Q3 and Q4; see **nltk.tag** for description on part-of-speech tagging in python.

b. Discuss the advantages (or, disadvantages thereof) of using the nouns only in representing the text by comparing and contrasting the results you obtained in Q4 and Q5(a).

6. Apply K-means Clustering algorithm (0.2)

a.

b. Apply K-means clustering on the same dataset using the better of the two representations (from Q4 and Q5) with k = number of genres. Experiment with different initializations, and evaluate the quality of the resulting clusterings against the ground truth (the known genre of each movie).

c. Discuss how well clustering can reproduce the classes of the dataset.