Amany Mohammed Al Luhaybi

amluhaybi@uqu.edu.sa

Project Proposal

1. Methodology:

This project will be based on binary text classification problem to predict the ratings of Coursera reviews from their text content. Multiple classification algorithms will be used as noted in tools section. Each reviewer can rate from 1 to 5, these rates will be separated into two classes:

- **High rating (class 1):** if rating > = 3
- **Low rating (class 0):** if rating < 3

2. Question/need:

Q1: Which institution had the most courses?

Q2: Which institution had the highest rates by reviewers?

Q3: Which institution had the lowest high rates by reviewers?

Q4: Which institution had the highest number of reviews?

Q5: Which course had the highest rates by reviewers?

Q6: Which course had the lowest rates by reviewers?

Q7: Which course had the highest number of reviews?

Q8: How many reviews per month from February 2020 to the end of the year for all courses on Coursera basically during coronavirus official announcement as pandemic?

- Text classification will be performed on reviews feature to predict the ratings of Coursera reviews from their text. This project is beneficial as it could be used as part of bigger applications. It can be included in a personal website for university's professors as external resources for their courses and it will only have the courses from Coursera with high rates.

3. Data Description:

a) Dataset source:

https://www.kaggle.com/imuhammad/course-reviews-on-coursera?select=Coursera_reviews.csv

b) Description of the Dataset:

Two datasets will be used one for reviews and the other one for the courses after merging these two datasets we will have 8 features and about 1454711 instances

- The features are: Reviews, Reviewers, date_reviews, rating, course_id, name, institution, course_url
- I'm expecting to be working on the following features: Reviews, Reviewers, date_reviews, rating, course_id, name, institution, new column for class.
- I will not be working this feature: course_url

4. Tools:

Sklearn for classification with SVM and logistic regression, as this library will be used for building a confusion matrix to evaluate the models. Seaborn, matplotlib and pandas. Keras for neural network with applying transfer of learning technique to do text classifications using the reviews feature, NLTK and TextBlob

5. MVP Goal:

- Basic data cleaning:

a) Removing punctuation, numbers, extra whitespace, stop-words and duplicates

b) Converting text to lower case (no capital letters)

c) Handling missing data

- Basic Data Exploration such as:

a) Merge two datasets

b) Create new columns such as the label column.

c) The shape, head, info, describe, and Summary of the dataset

d) Convert some data types if needed

e) Statistical Insight

f) Plotting the distribution of the data to handle imbalanced distribution

- Visualizations to the previous questions in Question/need section.