Neha Mathew
Netid: nehaam2

CS 410: Text Information Systems
Project Documentation

1) An overview of the function of the code (i.e., what it does and what it can be used for).
This project provides a brief overview of the opinions of students at UIUC about the courses available. This is done by scraping the UIUC subreddit and presenting the results as a sentiment analysis text classification task. This can be used to learn about past students' experiences at a glance, compare courses (2 at a time) and make a decision on whether to take a course or not.

2) Documentation of how the software is implemented with sufficient detail so that others can have a basic understanding of your code for future extension or any further improvement.

1. getPostsForCourse()
   This function takes in a course number. It checks if the existing master dictionary already contains the course number as a key and if not, it builds a search query with the course number as the parameter. It takes all the posts returned by the query and adds it to the master dictionary. If the master dictionary already contains the course number, it means the data has already been scraped.
   The query can be changed to get more recent posts or limit the number of posts.

2. getCommentsForPosts()
   This function takes a list of posts and uses the Reddit API to get all the top level comments for each post. Top level comments are the first comments in a thread. Replies to comments are not included.
   It returns the comments as a list.

3. cleanText()
   This function cleans the data according to observed patterns and model requirements as the pre-trained sentiment analysis model already takes care of some of the data cleaning. Notable functionality includes removing links, user mentions, special characters and truncating long comments.

4. getSentiment()

Neha Mathew
Netid: nehaam2

This function runs the pretrained sentiment analysis model 'finiteautomata/bertweet-base-sentiment-analysis' (originally trained on tweets) and returns a sorted and zipped list of the comments and their respective sentiments and scores.

5. OneCourseView()
   This function prints a color coded list of the top five comments with the highest sentiment score.

6. TwoCourseView()
   This function compares the reddit sentiment of two courses by drawing a bar graph of the number of positive, negative and neutral comments for each course.

The functionality only works for courses that have been talked about on reddit and can be picked up by the search query used in the API. Here are some examples of such courses to test functionality:
'kin 104', 'ling 100', 'fshn 120', 'mcb 150', 'math 221', 'cs 410', 'cs 225'.

3) Documentation of the usage of the software including either documentation of usages of APIs or detailed instructions on how to install and run a software, whichever is applicable.

- To run the project, users must set up a reddit app and obtain a client id and secret. This can be done by following the instructions here.
- Once the client id and secret have been obtained, download the notebook titled 'CS 410 Project Reddit Course Sentiment Finder'. It can be opened on Google Colab and does not require any additional files.
- Input the client id and secret in the space provided in cell #3.
- Hit Run All Cells
- The last cell will prompt the user for input. Choose between learning about the sentiment on the UIUC subreddit for a particular course or comparing two courses.
- Enter the course number(s) as 'subject tag <space> number' . Eg: 'kin 104'.

4) Brief description of contribution of each team member.
This is a solo project.