Name of Group: JSEN
Group Members: Sarah Zhang, Evan Xiang, Jingwen Zhang, Nancy Li

Description of Project: Scraping college subreddits on Reddit and creating lists of most popular keywords from different colleges. Users enter the college, time frame, and task they want accomplished. The scope of the project will be limited to the ten top colleges according to USNews.

Goals of Project:

- The goal is to be able to scrape college subreddits
- It will return most popular keywords/phrases from the subreddits for a given time period
    - This will give us an idea of the interests of the student body at a particular college during a particular time period.
- It will also compare percentage similarities across colleges
    - This same functionality can also compare percentage similarities of the same college across time periods
- It will also analyze word prevalence over time
    - It can create a graph that will show the word prevalence as a percentage of total words over time
- It will also analyze upvotes

Data Sources We Plan to Use:
HTML and API from subreddit links

Helpful Resources:
https://nycdatascience.com/blog/student-works/web-scraping-reddit-analyzing-user-behavior-and-top-content-from-a-marketing-perspective/?fbclid=IwAR0_Tj_xhSY6nXfLOMoTeewchjOmp-Y_sCWC_i4fq-aq8NRg5kdR2ZjQL9c

https://www.datacamp.com/community/tutorials/wordcloud-python?fbclid=IwAR2uJhVRDPQ1aH0FeS_9FiZfDsGL7VlTQd_FVeKba064QCI8DTQwTu7RvEU

Lists of Tasks to Complete and Timeline for Completing them:

Each of us is responsible for one of the above tasks.
Nancy: Analyzing word prevalence over time
Evan/Sarah: Percentage Similarities/ Top Keywords
Jingwen: Analyzing upvotes

Week 4
- Presentation

Week 5
- Collect data (all ten schools)
- Research characteristics of database that we will need to accomplish each of our tasks
    - Have meeting about this, and discuss how best to proceed in creating database (what characteristics will the tables have? How will we be representing the data?)
- Begin building database

Week 6
- Create algorithm to clean data (for one school)
- Create corpus of insignificant/repetitive words (from one school, limited data)
- Set up database for one school
- Get database working for one school, so that the method can be applied to the rest of the data
- Clean remaining data and add to database
- THE DATABASE SHOULD BE FINISHED AT THE END OF WEEK 6

Week 7-8
- Each team member works on their designated task separately
- Then, we will integrate all the programs so that it runs cohesively