Extraction of Reddit Data

1. **Without creating a Reddit account/credentials (did not use the API either)**
   Made use of the python requests library to directly pass a URL (e.g.: `https://www.reddit.com/r/all?q=gentrification+of&type=link.json`) to get results. This only gave results where the terms “gentrification of” were part of a subreddit title. It did not search in any individual posts. The output did not show any interesting results.

2. **Using the PRAW library**
   Usage of the Python Reddit API Wrapper (PRAW) library required creation of an account and requesting a token through which an API request could be made. This gave a wider spectrum for search since it looked for the terms “gentrification of” in individual posts rather than subtitles. A total of over 200 results was obtained. The years over which this data was obtained range from 2014 to 2023. Manual filtering was done to get the most interesting results. Here is a step by step process along with a guide to the output files and what each of them contain:
   a. Reddit account creation and requesting an access token.
   b. Writing code to make API calls with search keywords “gentrification of”.
   c. Making API calls for multiple features (relevance, hot, top, new, best). This gave all results including the usage of the word gentrification in its literal sense (e.g.: gentrification of Bangkok, gentrification of neighborhood, etc.). These results are stored in files all_{filter}.csv where “filter” stands for hot, top, new, best, and relevance.
      Feature descriptions:
      i. Relevance: Most relevant to the search keywords
      ii. Top: Posts shown according to highest upvote-downvote ratio
      iii. Hot: Posts that have been getting upvotes recently
      iv. New: Posts sorted by time of submission
      v. Best: Special algorithm which shows different order of posts everytime feed is refreshed
   d. For each filter, NER (Named Entity Recognition) was run. It removed all proper nouns following the words “gentrification of”, thus helping to eliminate posts that contained gentrification of a particular town/city/country. Manual filtering was then done to keep the most interesting contexts in which gentrification has been used (e.g.: gentrification of democracy). These files are stored as unique_{filter}.csv where “filter” stands for the 5 filters mentioned above.
   e. These filters have overlapping unique results which are all combined and stored in the ‘unique_relevance (master).csv’ file.

3. **Search for comments**
   The PRAW library was also used to try and extract comments in which the terms

“gentrification of” has been used. Although the UI itself gives results for this search condition, the Python API does not have a working feature to search and extract comments for a particular keyword. Thus, comments could not be fetched.