**Subtitle Processor**

This program can be used by language students who watch movies as a way of learning English. The program takes the name of a movie and outputs the useful words and collocations as well as their meaning from Wiktionary. Specifically, the following information is provided by the program:

1. Word clouds of most frequent and most difficult words
2. Most difficult words:
3. A complete list of phrasal verbs (e.g. wipe out, come to terms with, used to
4. A list of 2-gram collocations (e.g. human race, traffic jam, whole thing, split second)
5. A list of 3-gram collocations (e.g. behind the wheel, hell of a, no matter what)
6. A list of 4-gram collocations (e.g. for the first time, with the exception of, over and over again)
7. A list of 5-gram collocations (e.g. beat the crap out of)

This is basically a summary of all the important information a language student can learn from a movie. The program provide the definitions for the above mentioned words, phrasal verbs, and collocations from Wiktionary.org website as well as the line in the movie where these are used.

The data workflow starts from scraping the subtitle srt file from <http://www.moviesubtitles.org/> website that contains subtitle files in many languages for around 22,000 movies. This is done using *get\_subtitle.py* script. Natural language processing libraries such as *nltk* to pase the srt files, provide word frequencies, and determine part of speech information. A mathematical formula was designed to map the frequency of a given word to its difficulty on the scale of 0 to 10. The word frequencies of all English words were determined by *nltk* using all words in around 22,000 English movie subtitles. n-gram collocations were determined using *nltk.* The preprocessing scripts can be found in Scripts/Preprocessing\_scripts directory. Flask library was used to build a simple website.

**How to run:**

1. Install the following dependencies:

* nltk
* numpy
* pandas
* matplotlib
* flask
* requests
* BeautifulSoup
* shutil, json, zipfile, wordcloud, and csv

1. Modify line 15 of Subtitle.py and add the correct directory
2. Run Subtitle.py
3. Insert the name of the movie
4. Hit Submit

**File Directory:**

The code and data of this program can be found at https://github.com/mostafa-razavi/Subtitle

Data/

|\_\_\_\_\_ mso\_index\_name\_year\_subindex.csv name and index values of all movies

|\_\_\_\_\_ word\_freq.json frequency of all English words

|\_\_\_\_\_ wiktionary.json Wiktionary data (definitions, examples, etc)

|\_\_\_\_\_ wiktionary\_ngrams.json a dataset of n-grams in English language

Scripts/

|\_\_\_\_\_ Subtitle.py runs the website locally

|\_\_\_\_\_ dictionary.py finds the meaning from Wiktionary

|\_\_\_\_\_ get\_subtitle.py scrapes the movie subtitles from moviessubtitles.org

|\_\_\_\_\_ pretty\_table.py converts a Python dictionary into a nice html table

|\_\_\_\_\_ srt\_to\_wordfreq.py finds the word frequencies

|\_\_\_\_\_ find\_ngrams\_in\_subtitle.py finds all English n-grams from among Wiktionary titles