

# Python Assignment

## Introduction to Software Systems

April 10, 2023

### 1 Task 1

This task will teach you how to use Python to extract data from websites and create visualizations by plotting the data on graphs.

1. Write a Python script to scrape the top 1000 highest-grossing movies of all time from the IMDB website. After scraping the data, save it to a file named 'q1text.txt' for further analysis. You must also use the data obtained to plot the following graphs:
  - (a) A bar graph of the **genre** to frequency of movies belonging to that genre.
  - (b) A line graph of the top 100 movies and their **gross value**.
2. Create a python program that first prints 100 movies in the descending order of their **Imdb rating** (if two movies have the same Imdb rating, they must be sorted based on increasing **runtime**). Additionally, this program should enable the user to filter the list of top 1000 movies based on the following criteria.
  - (a) **Duration**
  - (b) Imdb Rating
  - (c) **Year of Release**
  - (d) **Genre**

Take an input from the user on which of the criteria they want to filter based on. For the first 3, the user must be prompted a range(i.e. 2 values), and the entries that come within the range must be displayed. For the genre, the user must be prompted to enter a genre and all the movies belonging to that genre must be printed.



Write a Python script to perform the following tasks.

1. First, you must read the two files 'sh.txt' and 'stopwords.txt'. The 'sh.txt' file contains the text that will be used to generate the word cloud. The 'stopwords.txt' file contains a list of English stopwords, one per line. Remove the stop words from the text in 'sh.txt' and create a word cloud with the leftover words.
2. Next, print the most common word in sh.txt (after removing the stop words).
3. Finally, print average word length of the words in sh.txt .

Note: For this task, you must manually remove the stopwords from the given story. You are not allowed to use any libraries or pre-built functions for this task, except for the generation of the word cloud. This will help you understand the process of stopword removal and develop your skills in working with text data in Python. Once you have removed the stopwords, you can use a library of your choice to generate the word cloud.

### 3 Submission Format

Your submission should consist of the following files.

1. q1.py : Task 1.1, i.e. scraping the website, and saving the data, and printing the graphs.
2. q2.py : Task 1.2, i.e. sorted list of movies, and filter system.
3. q3.py : Task2, i.e. the word cloud, most common word, and average word length.
4. q1text.txt
5. graph1.png
6. graph2.png
7. wordcloud.png

The submission will be automatically downloaded directly from GitHub classroom, so push your progress to your repository. **Note: It is imperative that you adhere to the following submission format for your assignment. Failure to do so may result in a loss of points.**