CSE 2050 – Programming in a Second Language
Fall 2023
Homework 3: Python Programming with Functions

Total Points: 30
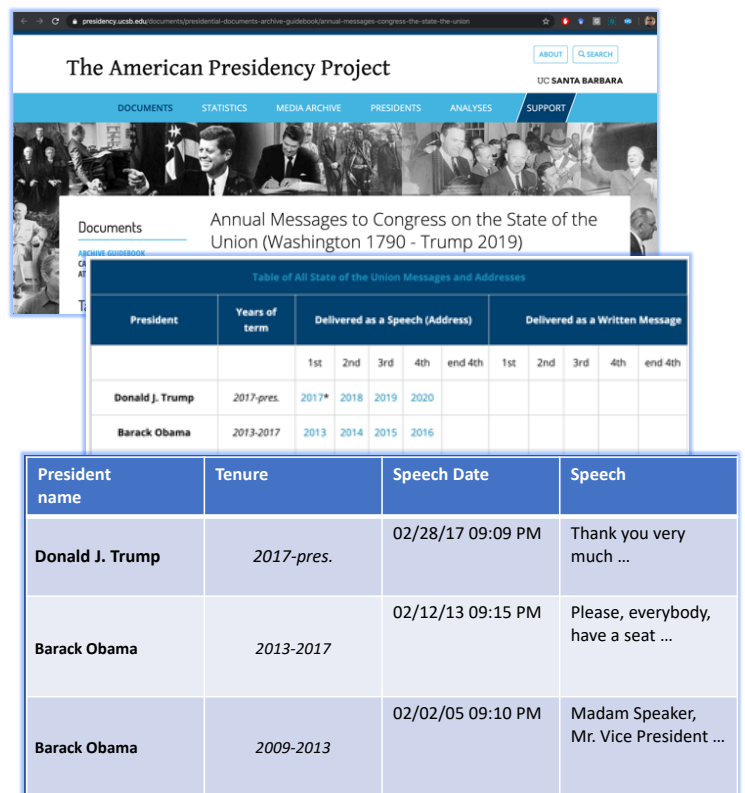
Date Assigned: Friday, Sept 22, 2023   **Due Date:** Sunday, Oct 1, 2023

Instructions: Please use functions as necessary throughout the assignment. Submit your work on Canvas as a Jupyter Notebook `ipynb` file named `cse2050_your_group_number_hw2.ipynb`. Remember to only use concepts covered up to Week 5 of classes to solve the problems. For example, do not use regular expressions, packages outside of the Python standard library, or other concepts not covered in class. See the *Tips for Success* and *Course Requirements* for more information.

Key Programming Concepts Demonstrated
- Web Scraping
  - Scraping textual data from a webpage (Q1)
- Programming experience
  - Structuring code and maintaining program flow (Q1)
  - Designing functions that work together to solve a problem (Q2)
- Working with data
  - Manipulating tabular data with different fields (Q1-Q3)
  - Formatting 2-dimensional data (Q1-Q3)
  - Classifying article readability and grade level based on Flesch Index and Flesch-Kincaid Grade Level Formula (Q2)
- Bonus (Visualization)
  - Creating a bar chart from data in a table (Q3)

President name	Tenure	Speech Date	Speech
Donald J. Trump	*2017-pres.*	02/28/17 09:09 PM	Thank you very much ...
Barack Obama	*2013-2017*	02/12/13 09:15 PM	Please, everybody, have a seat ...
Barack Obama	*2009-2013*	02/02/05 09:10 PM	Madam Speaker, Mr. Vice President ...

Recall the web scraping activities demonstrated in class during the lectures from week 4-5. This assignment will build on some of that code. Particularly, we will scrape a collection of U.S. presidential State of the Union addresses from the web and use the Flesch Index to classify their readability according to grade levels.

What is Flesch Index?

In 1949, Dr. Rudolf Flesch published *The Art of Readable Writing*, in which he proposed a measure of text readability known as the Flesch Index. This index is based on the average number of syllables per word and the average number of words per sentence in a piece of text. Index scores usually range from 0 to 100, and indicate readable prose for the following grade levels:

Score	School level	Notes
100.00–90.00	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–0.0	College graduate	Very difficult to read. Best understood by university graduates.

Table 1: Article Readability based on Flesch Index

Flesch's formula to calculate the index F is the following:
$F = 206.835 - 1.015 \times (words.count/sentences.count) - 84.6 \times (syllables.count/words.count)$

The Flesch-Kincaid Grade Level Formula is used to compute the Equivalent Grade Level G:
$G = 0.39 \times (words.count/sentences.count) + 11.8 \times (syllables.count/words.count) - 15.59$

Definitions	
Word:	Any sequence of non-whitespace characters.
Sentence:	Any sequence of words ending in a period, question mark, exclamation point, colon, or semicolon.
Syllable:	Any word of three characters or less; or any vowel (a, e, i, o, u) or pair of consecutive vowels, except for a final -es, -ed, or -e that is not -le

Table 2: Definition of terms related to the Flesch Index

First, access the code located at https://repl.it/@fnembhard/souspeechscraper. Copy the code to your Jupyter notebook. This code scrapes data from the webpage of The American Presidency Project at UC Santa Barbara and extracts links to State of the Union (SOU) addresses. The code only selects the first SOU delivered by each president during each term if available. Further, the code prints out a table containing the following fields for each president: `[president_name, tenure, speech_link, speech_date, speech]`. (`speech_date` and `speech` are left for you to update using Python code).

1. **Get the Data** (10 points)
   Update the code by adding a function named `get_speech_and_speech_date(url)` to extract each speech and speech date from the page referenced by the link in the table and update your table (2-D list). Remember you can return multiple values from a function as a tuple. This can be done by accessing the object containing the desired text by *class*. Figure 1 shows the layout of the page for each speech. For example, to extract the element containing the speech date, use the following code:

```
speech_date_elmnt = html_elem.cssselect('[class="date-display-single"]')
```

Figure 1: The American Presidency Project SOU page format

Use the *content* attribute to extract the actual speech date (full date with time), which is the actual date the speech was delivered. Next, extract the speech transcript and store it in your table. Print out a table of your results. Only print the first 50 characters of each speech you extract when displaying the output. Please pay attention to the reformatted and more readable date in the table.

Expected Output

```
President's Name        Term            Date of Speech          Excerpt
Joseph R. Biden         2021-present    04/28/21 09:06 PM       The President. Thank you. Thank you. Thank you. Go
Donald J. Trump         2017-pres.      02/28/17 09:09 PM       Thank you very much. Mr. Speaker, Mr. Vice Preside
Barack Obama            2013-2017       02/12/13 09:15 PM       Please, everybody, have a seat. Mr. Speaker, Mr. V
George W. Bush          2005-2009       02/02/05 09:10 PM       Mr. Speaker, Vice President Cheney, Members of Con
William J. Clinton      1997-2001       02/04/97 09:15 PM       Mr. Speaker, Mr. Vice President, Members of the 10
George Bush             1989-1993       02/09/89 12:00 AM       Mr. Speaker, Mr. President, and distinguished Memb
Ronald Reagan           1985-1989       02/06/85 09:05 PM       Mr. Speaker, Mr. President, distinguished Members
Lyndon B. Johnson       1965-1969       01/04/65 09:04 PM       Mr. Speaker, Mr. President, Members of the Congres
John F. Kennedy         1961-1963       01/30/61 12:30 PM       [As delivered in person before a joint session]Mr.
Dwight D. Eisenhower    1957-1961       01/10/57 12:00 AM       To the Congress of the United States:I appear befo
Harry S Truman          1949-1953       01/05/49 01:00 PM       Mr. President, Mr. Speaker, Members of the Congres
Franklin D. Roosevelt   1945            01/06/45 12:00 AM       Today, in pursuance of my Constitutional duty, I s
Warren G. Harding       1921-1923       12/06/21 12:00 AM       MR. SPEAKER AND MEMBERS OF THE CONGRESS:It is a ve
Woodrow Wilson          1917-1921       12/04/17 12:00 AM       Gentlemen of the Congress:Eight months have elapse
John Adams              1797-1801       11/22/97 12:00 AM       Gentlemen of the Senate and Gentlemen of the House
George Washington       1793-1797       12/03/93 12:00 AM       Fellow-Citizens of the Senate and House of Represe
```

2. **Writing Functions** (20 points)

   Write code to create the following functions to find the number of sentences, words, and syllables in each speech, as well as the speech's Flesch Index and Grade Level Equivalent. Refer to the formulas previously defined and the definitions in Table 2. Update your table with the Flesch Index and grade level as shown in the expected output below.

   count_sentences(speech): Counts the number of sentences in text.

   count_words(speech): Counts the number of words in text.

   count_syllables(speech): Counts the number of syllables in text.

   compute_flesch_index(syllable_count, word_count, sentence_count): Computes the Flesch Index for the given numbers of sentences, words, and syllables.

   compute_grade_level(syllable_count, word_count, sentence_count): Computes the Grade Level Equivalent for the given numbers of sentences, words, and syllables.

   classify_article_readability(flesch_index): Returns the school/grade level in worded format based on Table 1.

Expected Output

President's Name	Term	Date of Speech	Excerpt	Flesch	Grade Level #	Grade Level
Joseph R. Biden	2021-present	04/28/21 09:06 PM	The President. Thank you. Than	57.64	9	10th to 12th grade
Donald J. Trump	2017-pres.	02/28/17 09:09 PM	Thank you very much. Mr. Speak	53.97	10	10th to 12th grade
Barack Obama	2013-2017	02/12/13 09:15 PM	Please, everybody, have a seat	54.46	10	10th to 12th grade
George W. Bush	2005-2009	02/02/05 09:10 PM	Mr. Speaker, Vice President Ch	45.84	12	College
William J. Clinton	1997-2001	02/04/97 09:15 PM	Mr. Speaker, Mr. Vice Presiden	56.20	10	10th to 12th grade
George Bush	1989-1993	02/09/89 12:00 AM	Mr. Speaker, Mr. President, an	58.46	9	10th to 12th grade
Ronald Reagan	1985-1989	02/06/85 09:05 PM	Mr. Speaker, Mr. President, di	49.72	10	College
Lyndon B. Johnson	1965-1969	01/04/65 09:04 PM	Mr. Speaker, Mr. President, Me	53.14	10	10th to 12th grade
John F. Kennedy	1961-1963	01/30/61 12:30 PM	[As delivered in person before	36.56	14	College
Dwight D. Eisenhower	1957-1961	01/10/57 12:00 AM	To the Congress of the United	34.90	13	College
Harry S Truman	1949-1953	01/05/49 01:00 PM	Mr. President, Mr. Speaker, Me	42.21	12	College
Franklin D. Roosevelt	1945	01/06/45 12:00 AM	Today, in pursuance of my Cons	46.84	12	College
Warren G. Harding	1921-1923	12/06/21 12:00 AM	MR. SPEAKER AND MEMBERS OF THE	34.33	15	College
Woodrow Wilson	1917-1921	12/04/17 12:00 AM	Gentlemen of the Congress:Eigh	52.71	12	10th to 12th grade
John Adams	1797-1801	11/22/97 12:00 AM	Gentlemen of the Senate and Ge	30.95	17	College
George Washington	1793-1797	12/03/93 12:00 AM	Fellow-Citizens of the Senate	39.37	15	College

Bonus (7 points)
Research and use matplotlib to display a plot of the Flesch Indexes for all the presidents in our sample dataset as shown in the figure below. Do not hard-code any data. Extract data from your data-table and use it to create your plot.

Figure 2: Plot of Flesch indexes for a selection of SOU for various US presidents

Flesch Index for Presidential SOU