Generating Word Clouds in Python

A word cloud is an image filled with texts in different sizes which represent the frequency or the importance of each word.

The objectives of this assignment are as follow:

- Use web crawling to extract text from Wikipedia article
- Apply text preprocessing pipeline to clean and extract the most frequent words in the text
- Visualize the most frequent words using word clouds

Assignment Tasks:

Task 1 – Data Acquisition

- Use BeautifulSoup to extract texts from Wikipedia Article
- Hint: use the below link so we should get similar outputs
- https://en.wikipedia.org/wiki/Large_language_model

Deliverable: A function that return the extracted text

Task 2 – Text Preprocessing Using spaCy

- Clean the text (extracted in task 1):
  - Converting to lowercase
  - Removing punctuation, numbers, and special characters
  - Removing stop words
- Extract all tokens after lemmatization
- Extract just noun tokens after lemmatization

Deliverable: A function that return list of tokens and a list of just noun tokens

Task 3 – Word Frequency Analysis

- Count occurrences and plot the most 20 frequent ones

Deliverable: Two bar charts for the most 20 frequent tokens and noun tokens

Hint: use Counter from collection to count occurrences and get the most frequent ones and matplotlib to plot the bar charts

Task 4 – Word Cloud Visualization

- Generate and display a word cloud showing the most frequent words

Deliverable: Word cloud visualization for the most frequent tokens and another word cloud visualization for the most frequent noun tokens

Hint: use wordcloud and matplotlib libraries

Submission Guidelines

- Submit a Jupyter notebook named: **WordCloud_Assignment_<YourName>.ipynb**
- Ensure each task is performed in a different cell
- Ensure plots and images are show in your notebook when submitting the file