Homework Assignment 1: Python Basics & Text Handling in NLP

Natural Language Processing

School of Engineering and Technology,
K.R. Mangalam University

August 6, 2025

Instructions

- Submit your solutions as a Jupyter Notebook (`.ipynb`) file.
- Include comments in your code explaining each step.
- Use meaningful variable names and proper formatting.
- Submit all generated output files (`.txt` and `.csv`) along with your notebook.
- Deadline: ____________________

Homework Tasks

Task 1: Reading and Writing Files

- Download a public-domain English book or article (e.g., from Project Gutenberg: `https://www.gutenberg.org/`).
- Save it as `input_text.txt`.
- Write a Python program to:
  a) Read the file and print the first 20 lines.
  b) Convert the text to lowercase.
  c) Save the lowercase text to `lowercase_output.txt`.

Task 2: Text Cleaning

- Remove all punctuation, digits, and extra spaces from the lowercase text.

- Save the cleaned text to `cleaned_output.txt`.

Task 3: Tokenization and Analysis

- Tokenize the cleaned text into words using NLTK.
- Remove stopwords using the NLTK stopwords list.
- Count and display:
  a) Total number of tokens before and after stopword removal.
  b) Vocabulary size (number of unique words).

Task 4: Word Frequency Analysis

- Count the frequency of each word (after stopword removal).
- Display the top 20 most common words with their counts.
- Save the frequency table to `word_frequency.csv`.

Task 5: Bonus (Optional)

- Create a bar plot showing the frequencies of the top 10 words using `matplotlib`.
- Save the plot as `top_words.png`.

Deliverables

- `.ipynb` file containing all code and output.
- Generated text files:
  - `lowercase_output.txt`
  - `cleaned_output.txt`
- `word_frequency.csv` file.
- `top_words.png` (if Bonus task attempted).