Python for Computer Science and Data Science 2 (CSE 3652)

**Minor Assignment-3: NATURAL LANGUAGE PROCESSING**

1. Define Natural Language Processing (NLP). Provide three real-world applications of NLP and explain how they impact society.

**Ans:-** Natural Language Processing (NLP) is a field of artificial intelligence (AI) that enables computers to understand, interpret, and respond to human language in a way that is both meaningful and useful. It combines computational linguistics, machine learning, and deep learning to process and analyze large amounts of natural language data.

Three Real-World Applications of NLP:

1. Chat-bots and Virtual Assistants (e.g., Siri, Alexa, Google Assistant):

Impact: These tools enhance convenience by handling tasks like setting reminders, answering questions, and controlling smart home devices. They make technology accessible to non-technical users and support people with disabilities.

1. Sentiment Analysis (e.g., Monitoring Social Media):

Impact: Businesses use sentiment analysis to understand public opinion, gauge customer satisfaction, and manage brand reputation. It helps in decision-making by providing insights into market trends and consumer behavior.

1. Language Translation (e.g., Google Translate):

Impact: Real-time translation breaks language barriers, facilitating communication in international business, travel, and education. It promotes cultural exchange and helps people understand content in different languages.

NLP transforms how we interact with technology, making machines more intuitive and responsive to human needs.

1. The Explain the following terms and their significance in NLP:

• Tokenization

• Stemming

• Lemmatization

**Ans:-** a) Tokenization:

Definition: Tokenization is the process of breaking down text into smaller units called tokens. These tokens can be words, phrases, or even characters. For example, the sentence "NLP is fascinating" becomes ["NLP", "is", "fascinating"] after tokenization.

Significance: It's a fundamental preprocessing step in NLP. Tokenization helps in analyzing text by isolating meaningful components, enabling tasks like text analysis, search, and machine translation.

b) Stemming:

Definition: Stemming reduces words to their root or base form by removing suffixes. It uses simple rules without considering the word's meaning. For instance, "running," "runner," and "runs" may all be reduced to "run."

Significance: Stemming speeds up NLP processes by normalizing words, but it may produce non-real words (e.g., "studies" to "studi"). It's useful for applications where accuracy is less critical, like search engines.

1. Lemmatization:

Definition: Lemmatization reduces words to their base or dictionary form (lemma) while considering context and grammar. For example, "running" becomes "run," and "better" becomes "good."

Significance: It's more accurate than stemming, preserving the meaning of words. Lemmatization is preferred in NLP tasks requiring semantic understanding, such as machine translation and sentiment analysis.

1. What is Part-of-Speech (POS) tagging? Discuss its importance with an example.

**Ans:-** Part-of-Speech (POS) Tagging is a fundamental step in Natural Language Processing (NLP) where each word in a sentence is labeled with its grammatical role, such as noun, verb, adjective, adverb, etc. The process helps machines understand the syntactic structure of sentences.

Importance of POS Tagging:

1. Text Analysis and Understanding: POS tagging enables deeper linguistic analysis, helping machines grasp context and meaning in text.
2. Named Entity Recognition (NER): Identifying proper nouns (e.g., names, places) becomes easier with accurate POS tagging.
3. Syntactic Parsing: Helps in identifying grammatical relationships and sentence structure.
4. Machine Translation and Chat-bots: Ensures accurate translation and response generation by recognizing the role of each word.

Example: Consider the sentence: “She enjoys a long walk in the morning.”

After POS tagging, it would look like:

She (PRON) - Pronoun

enjoys (VERB) - Verb

a (DET) - Determiner

long (ADJ) - Adjective

walk (NOUN) - Noun

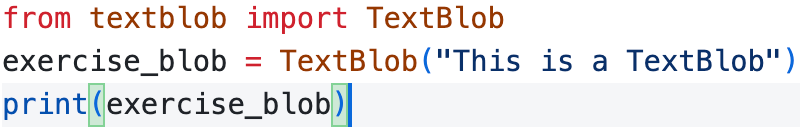
in (PREP) - Preposition

the (DET) - Determiner

morning (NOUN) - Noun

1. Create a TextBlob named exercise blob containing “This is a TextBlob”.

**Ans:-**



Output:- Screenshot 2025-04-02 at 8.02.33 PM

1. Write a Python script to perform the following tasks on the given text:

• Tokenize the text into words and sentences.

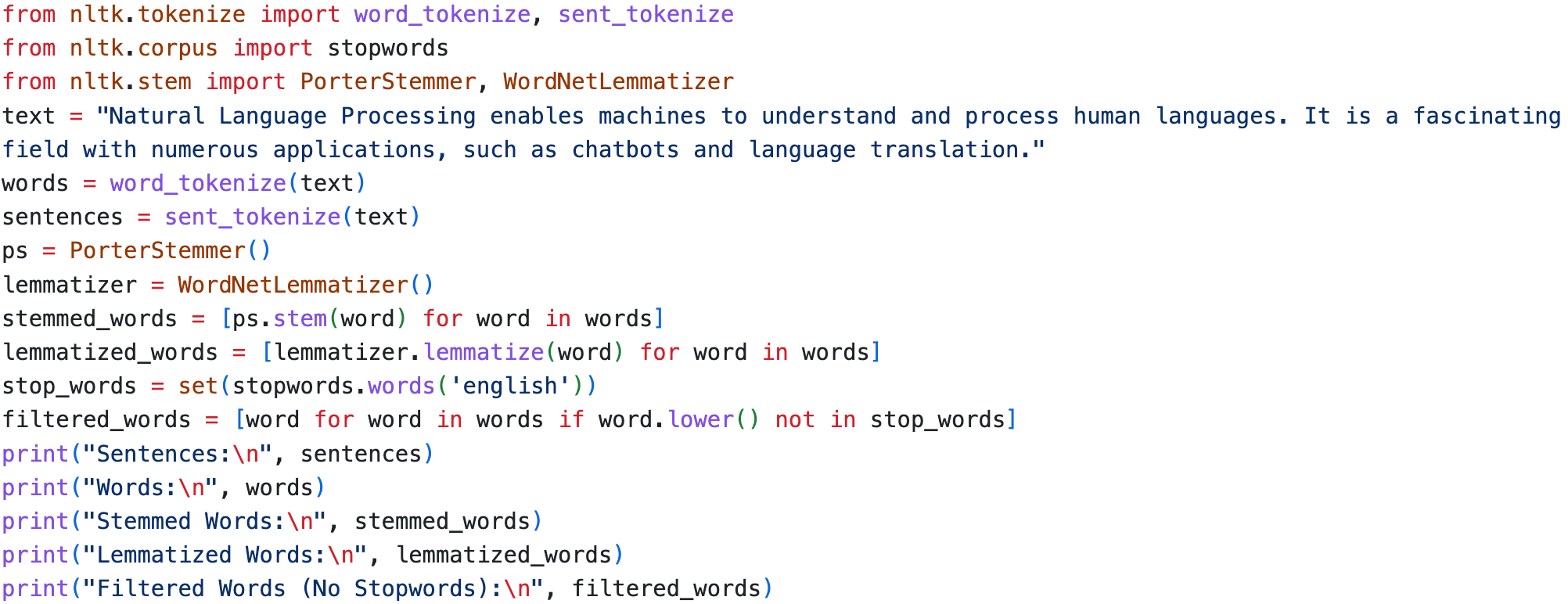
• Perform stemming and lemmatization using NLTK or SpaCy.

• Remove stop words from the text.

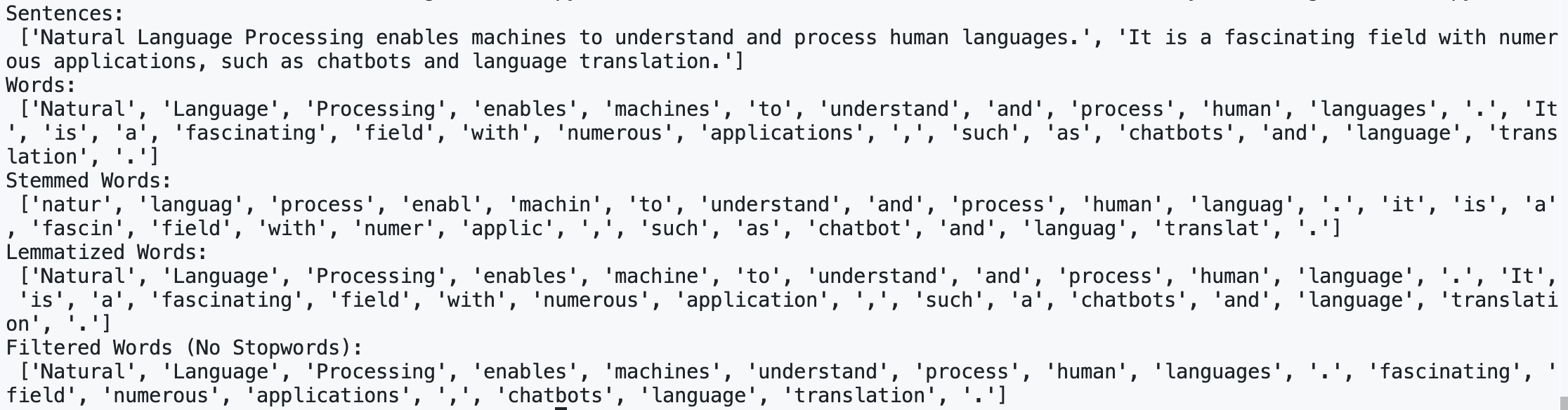
• Sample Text:

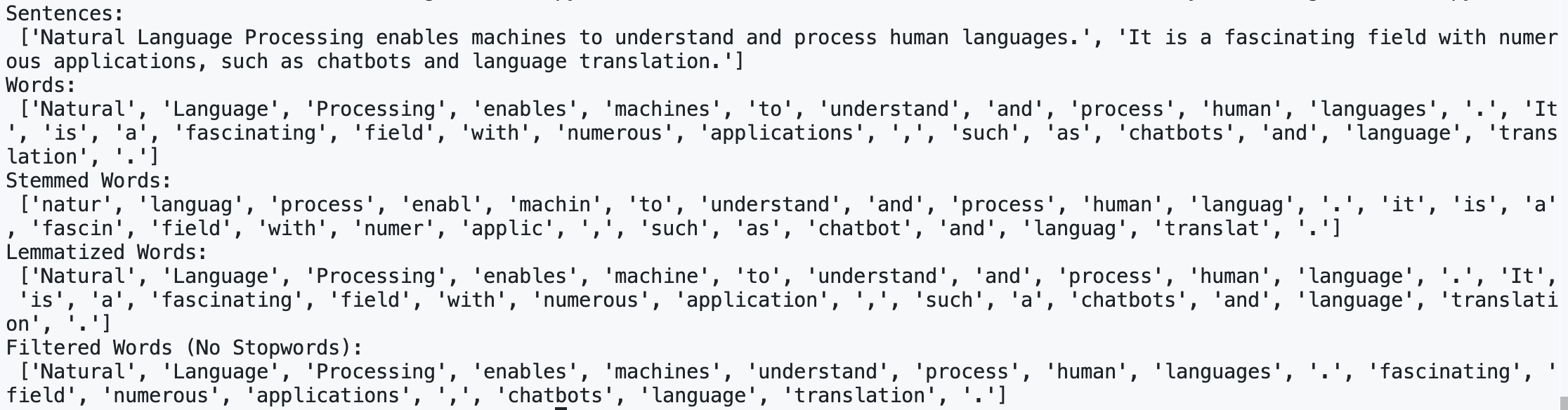
“Natural Language Processing enables machines to understand and process human languages. It is a fascinating field with numerous applications, such as chatbots and language translation.”

**Ans:-**



Output:-





1. Web Scraping with the Requests and Beautiful Soup Libraries:

• Use the requests library to download the www.python.org home page’s content.

• Use the Beautiful Soup library to extract only the text from the page.

• Eliminate the stop words in the resulting text, then use the wordcloud module to create a word cloud based on the text.

**Ans: -**



Output -



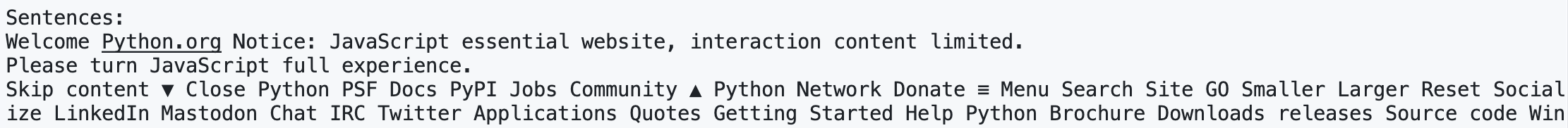
1. (Tokenizing Text and Noun Phrases) Using the text from above problem, create a TextBlob, then tokenize it into Sentences and Words, and extract its noun phrases.

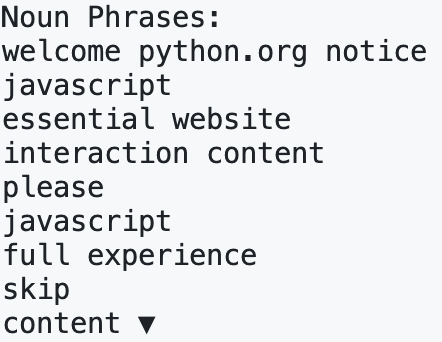
**Ans:-**





Output:-





1. (Sentiment of a News Article) Using the techniques in problem no. 5, download a web page for a current news article and create a TextBlob. Display the sentiment for the entire TextBlob and for each Sentence .

**Ans:-**



Output:-



1. (Sentiment of a News Article with the NaiveBayesAnalyzer) Repeat the previous exercise but use the NaiveBayesAnalyzer for sentiment analysis.

**Ans:-**

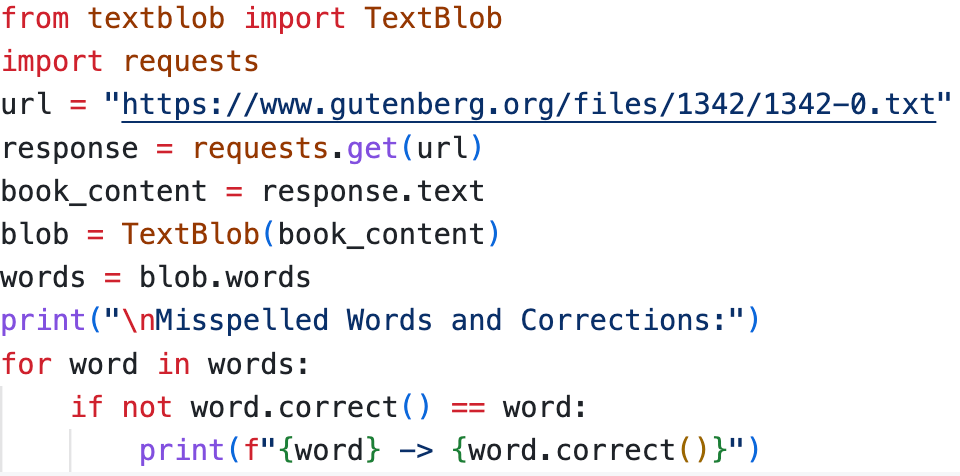


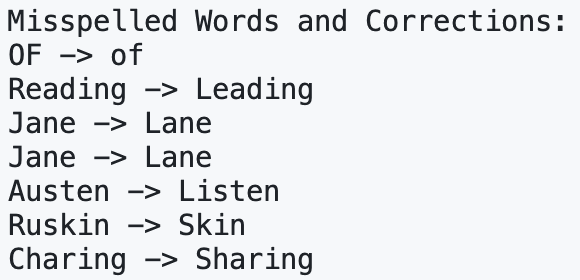
Output:-



1. Use (Spell Check a Project Gutenberg Book) Download a Project Gutenberg book and create a TextBlob. Tokenize the TextBlob into Words and determine whether any are misspelled. If so, display the possible corrections.

**Ans:-**



Output:- 

1. • Write a Python program that takes user input in English and translates it to French, Spanish, and German using TextBlob.

• Create a program that takes multiple user-inputted sentences, analyzes polarity and subjectivity, and categorizes them as objective/subjective and positive/negative/neutral.

• Develop a function that takes a paragraph, splits it into sentences, and calculates the sentiment score for each sentence individually.

• Write a program that takes a sentence as input and prints each word along with its POS tag using TextBlob.

• Create a function that takes a user-inputted word, checks its spelling using TextBlob, and suggests top 3 closest words if a mistake is found.

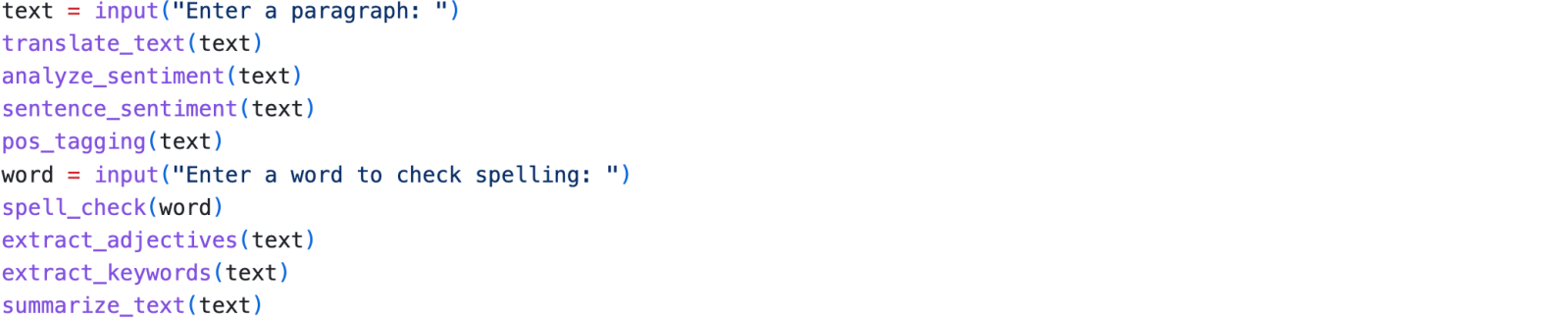
• Build a Python script that extracts all adjectives from a given paragraph and prints them in order of occurrence.

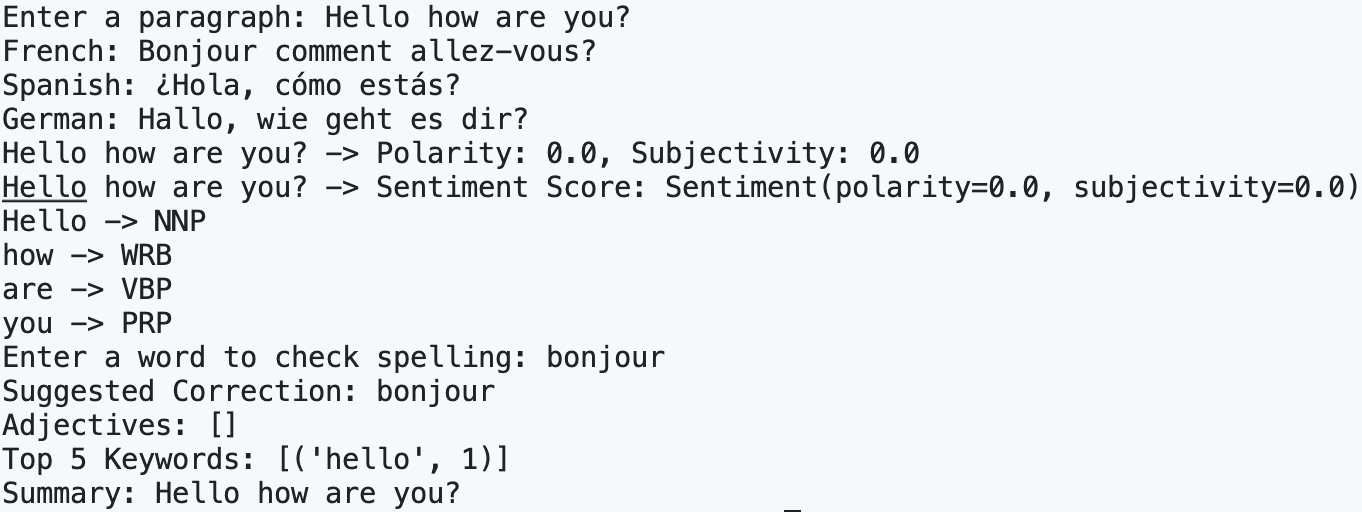
• Write a program that takes a news article as input and extracts the top 5 most common noun phrases as keywords.

• Write a program that takes a news article as input and extracts the top 5 most common noun phrases as keywords.

• Write a program that summarizes a given paragraph by keeping only the most informative sentences, based on noun phrase frequency.

**Ans:-**



Output: - 

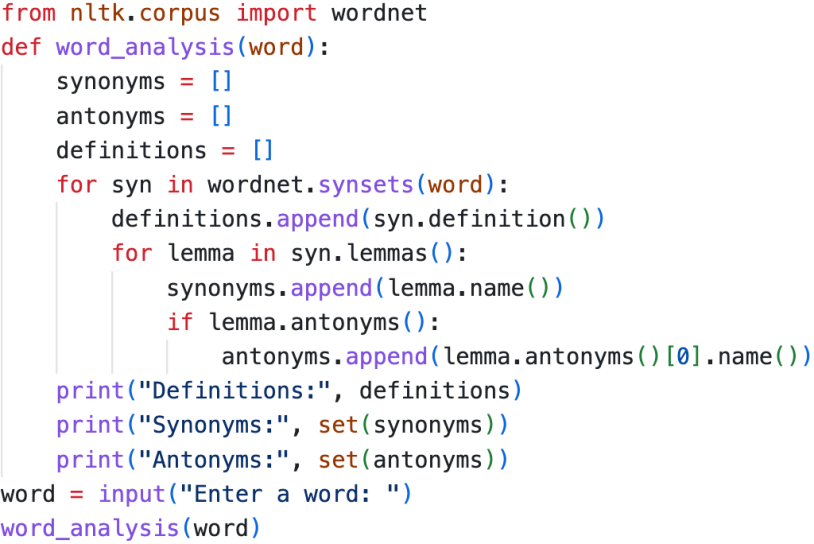
1. Write a Python program that takes a word as input and returns:

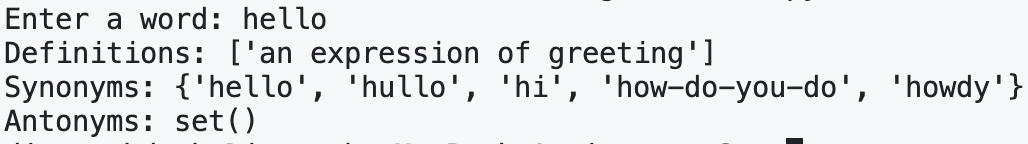
• Its definition

• Its synonyms

• Its antonyms(if available)

**Ans: -**



Output - 

1. • Write a Python program that reads a .txt file, processes the text, and generates a word cloud visualization.

• Create a word cloud in the shape of an object (e.g., a heart, star) using WordCloud and a mask image.

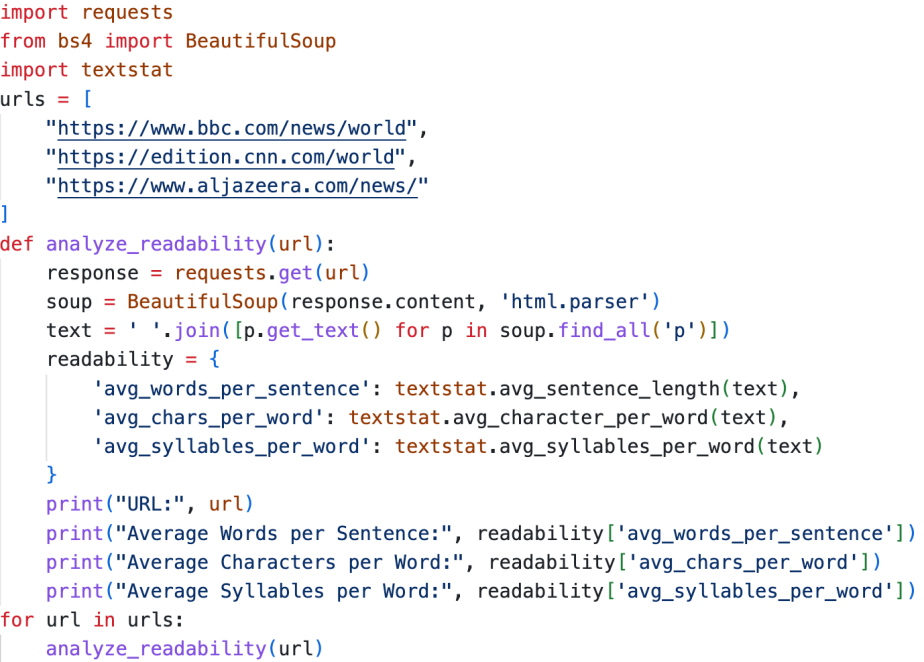
**Ans: -**

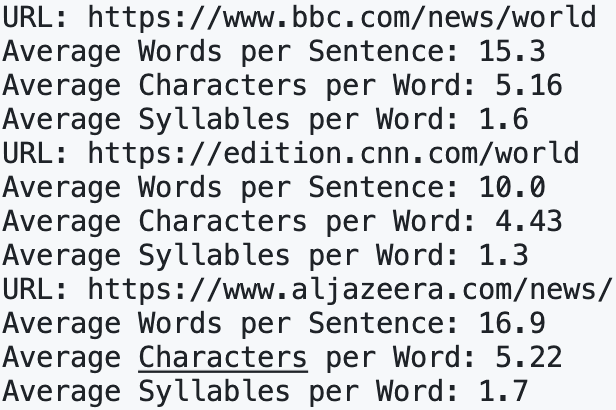


Output - 

1. (Textatistic: Readability of News Articles) Using the above techniques, download from several news sites current news articles on the same topic. Perform readability assessments on them to determine which sites are the most readable. For each article, calculate the average number of words per sentence, the average number of characters per word and the average number of syllables per word.

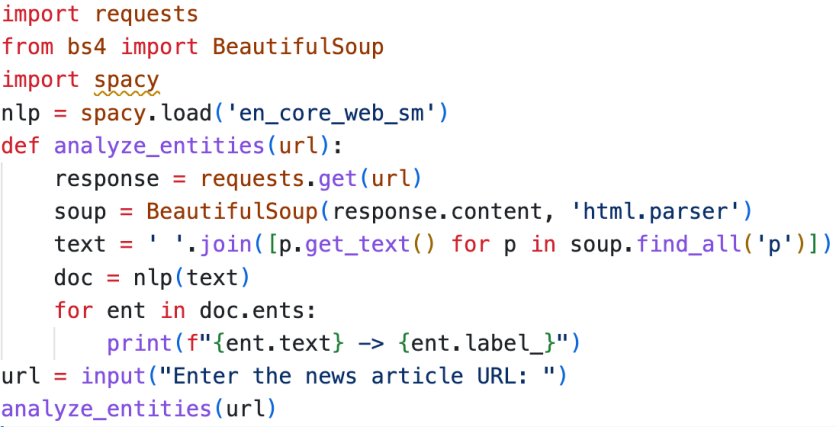
**Ans: -**



Output - 

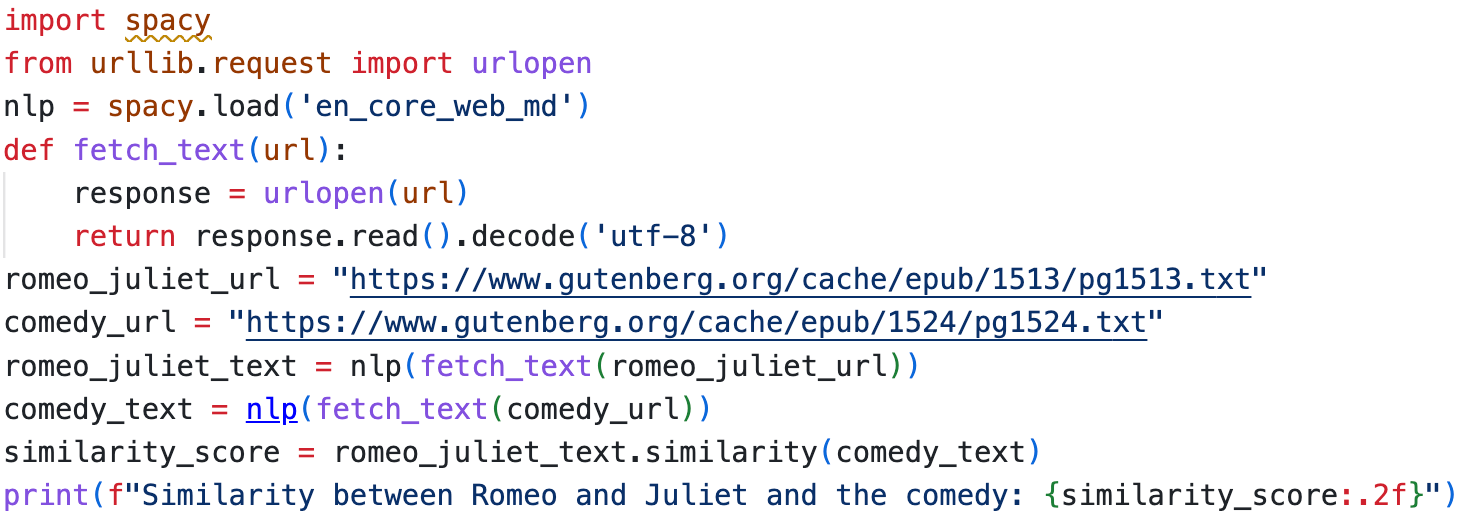
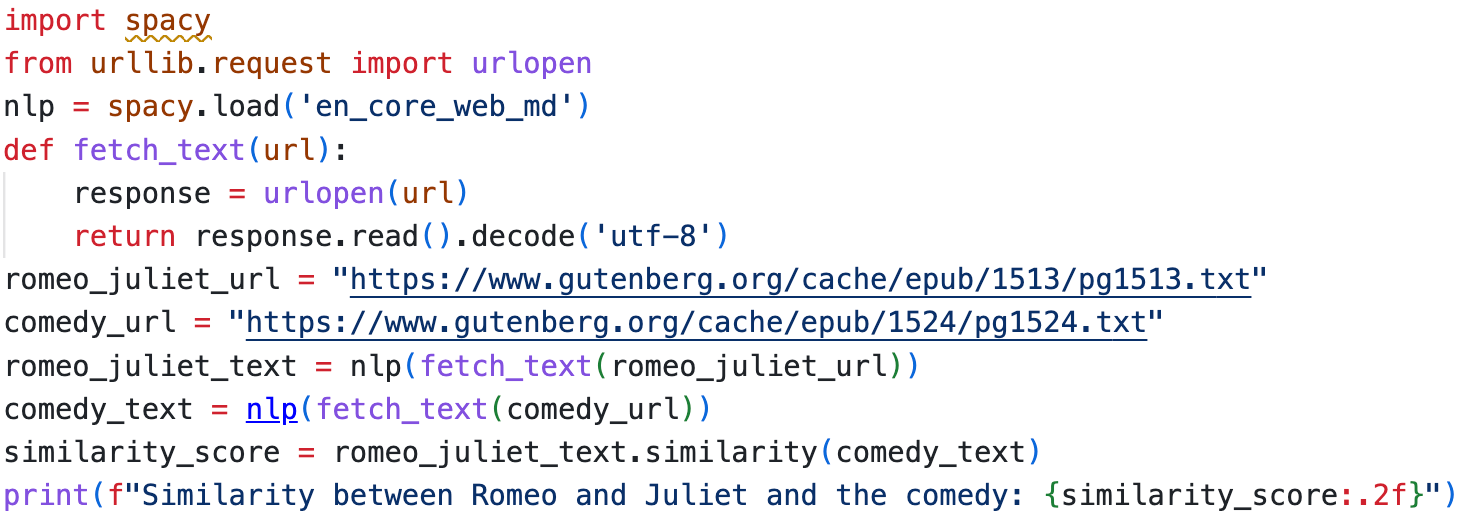
1. (spaCy: Named Entity Recognition) Using the above techniques, download a current news article, then use the spaCy library’s named entity recognition capabilities to display the named entities (people, places, organizations, etc.) in the article.

**Ans: -**



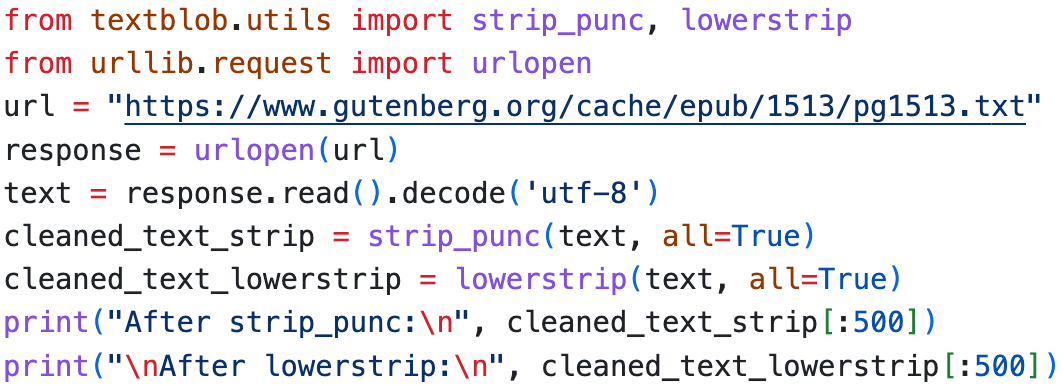
1. (spaCy: Shakespeare Similarity Detection) Using the spaCy techniques, download a Shakespeare comedy from Project Gutenberg and compare it for similarity with Romeo and Juliet.

**Ans: -**



1. (textblob.utils Utility Functions) Use strip punc and lowerstrip functions of TextBlob’stextblob.utils module with all=True keyword argument to remove punctuation and to get a string in all lowercase letters with whitespace and punctuation removed. Experiment with each function on Romeo and Juliet.

**Ans: -**



Output: - 