ASSIGNMENT

Natural Language Processing with NLTK, SpaCy, Word2Vec, and TF-IDF

Objective: In this assignment, you will learn how to perform Natural Language Processing tasks using Python libraries such as NLTK, SpaCy, Word2Vec, and TF-IDF. You will practice techniques such as tokenization, stemming, lemmatization, and document similarity calculation using these libraries. You will use the BBC News dataset available on Kaggle.

Tasks:

1. Import the necessary libraries: Start by importing the required libraries, including NLTK, SpaCy, gensim, and scikit-learn.
2. Load the dataset: Load the BBC News dataset(BBC_DATA.csv) into a pandas DataFrame using the read_csv() function. The dataset contains 2,225 rows and 2 columns, with the first column containing the text of the news articles.
3. Tokenization with NLTK: Implement tokenization using NLTK's word_tokenize() and sent_tokenize() functions. Apply these functions to a sample news article from the dataset.
4. Stemming and Lemmatization with NLTK: Implement stemming and lemmatization using NLTK's PorterStemmer and WordNetLemmatizer functions. Apply these functions to a sample news article from the dataset.
5. Named Entity Recognition with SpaCy: Use SpaCy's pre-trained model to perform named entity recognition on a sample news article from the dataset. Visualize the named entities using displaCy.
6. Word2Vec with gensim: Implement Word2Vec using gensim's Word2Vec function on the entire dataset. Train the model and get the vector representation of a sample word.
7. TF-IDF with scikit-learn: Implement TF-IDF using scikit-learn's TfidfVectorizer function on the entire dataset. Transform the dataset using the fitted vectorizer and calculate the cosine similarity between two news articles.
8. Bonus: Choose a different dataset and perform NLP tasks using different techniques. Be creative!

Submission:

Submit a Jupyter Notebook file (.ipynb) containing your code, a brief explanation of your thought process for each task, and any necessary comments for clarity. Make sure to test your code and provide examples of the output for each task

Data set Link-

https://drive.google.com/file/d/1TbTtdkJ1T-vnk3UdTEJcM9fCsIxdNouo/view?usp=drive_link