Lab Assignment 10

UCS420 Cognitive Computing

Assignment Title: NLP using Python-II
(Feature extraction from text, sentiment analysis and text generation)

Q1. Write a unique paragraph (5-6 sentences) about your favorite topic (e.g., sports, technology, food, books, etc.).

1. Convert text to lowercase and remove punctuation using re.
2. Tokenize the text into words and sentences.
3. Split using split() and word_tokenize() and compare how Python split and NLTK's word_tokenize() differ.
4. Remove stopwords (using NLTK's stopwords list).
5. Display word frequency distribution (excluding stopwords).

Q2. Using the same paragraph from Q1:

1. Extract all words with only alphabets using re.findall()
2. Remove stop words using NLTK's stopword list
3. Perform stemming with PorterStemmer
4. Perform lemmatization with WordNetLemmatizer
5. Compare the stemmed and lemmatized outputs and explain when you'd prefer one over the other.

Q3. Choose 3 short texts of your own (e.g., different news headlines, product reviews).

1. Use CountVectorizer to generate the Bag of Words representation.
2. Use TfidfVectorizer to compute TF-IDF scores.
3. Print and interpret the top 3 keywords from each text using TF-IDF.

Q4. Write 2 short texts (4–6 lines each) describing two different technologies (e.g., AI vs Blockchain).

1. Preprocess and tokenize both texts.
2. Calculate:
   a. Jaccard Similarity using sets
   b. Cosine Similarity using TfidfVectorizer + cosine_similarity()

c. Analyze which similarity metric gives better insights in your case.

Q5. Write a short review for a product or service.

1. Use TextBlob or VADER to find polarity & subjectivity for each review.
2. Classify reviews into Positive / Negative / Neutral.
3. Create a word cloud using the wordcloud library for all positive reviews.

Q6. Choose your own paragraph (~100 words) as training data.

1. Tokenize text using Tokenizer() from keras.preprocessing.text
2. Create input sequences and build a simple LSTM or Dense model
3. Train the model and generate 2–3 new lines of text starting from any seed word you provide.