Lab Assignment 9

UCS420 Cognitive Computing

Assignment Title: NLP using Python

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.
2. Tokenize the text into words and sentences.
3. Remove stopwords (using NLTK's stopwords list).
4. Display word frequency distribution (excluding stopwords).

Q2: Stemming and Lemmatization

1. Take the tokenized words from Question 1 (after stopword removal).
2. Apply stemming using NLTK's PorterStemmer and LancasterStemmer.
3. Apply lemmatization using NLTK's WordNetLemmatizer.
4. Compare and display results of both techniques.

Q3. Regular Expressions and Text Splitting

1. Take their original text from Question 1.
2. Use regular expressions to:
    a. Extract all words with more than 5 letters.
    b. Extract all numbers (if any exist in their text).
    c. Extract all capitalized words.
3. Use text splitting techniques to:
    a. Split the text into words containing only alphabets (removing digits and special characters).
    b. Extract words starting with a vowel.

Q4. Custom Tokenization & Regex-based Text Cleaning

1. Take original text from Question 1.
2. Write a custom tokenization function that:
    a. Removes punctuation and special symbols, but keeps contractions (e.g., "isn't" should not be split into "is" and "n't").
    b. Handles hyphenated words as a single token (e.g., "state-of-the-art" remains a single token).
    c. Tokenizes numbers separately but keeps decimal numbers intact (e.g., "3.14" should remain as is).

3. Use Regex Substitutions (re.sub) to:
    a. Replace email addresses with '<EMAIL>' placeholder.
    b. Replace URLs with '<URL>' placeholder.
    c. Replace phone numbers (formats: 123-456-7890 or +91 9876543210) with '<PHONE>' placeholder.