B. K. Birla College(Empowered Autonomous Status)
Department of Computer Science
MSc – AI SEMESTER II / IV (PRACTICAL LIST): 2024 – 25
SUBJECT – NLP

NLP_LAB_1

Q.1 Demonstrate different CORPORA available in NLTK and how to access and use the contents of these CORPORA.

NLP_LAB_2

Q.1 Demonstrate the use of the following concepts in NLP

1. Concept 1. Phonology (Study of Sound Patterns - Concepts Covered: Phonemes, rhyming words.)
2. Concept 2. Morphology (Study of Word Formation - Tokenization, Lemmatization, POS tagging, Morphological analysis.)
3. Print a structured table containing Tokenization, Lemmatization, Part of Speech (POS) tagging, and Morphological features of each word in a given sentence.
4. Concept 3. Syntax (Sentence Structure & Grammar – Sentence structure, dependency parsing.)
5. Concept 4. Semantics (Meaning of Words & Sentences – Word embeddings, semantic similarity.)
6. Concept 5. Pragmatics (Context & Implicature – Context-based analysis, sentiment analysis.)

NLP_LAB_3

Q. 1 Demonstrate how to perform word and sentence tokenization in NLP by

1. Taking
    a. A sentence
    b. A paragraph
    c. Text document from any Corpus of your choice.
2. Accessing the texts from the Books available in NLTK

Q. 2 Demonstrate the following tasks in NLP

1. Task 1: A concordance view shows us every occurrence of a given word, together with some context. To find other words appearing in a similar range of contexts, we append the term 'similar' to the name of the text in question.
2. Task 2: Common_contexts allows us to examine just the contexts that are shared by two or more words.
3. Task 3: Draw Lexical dispersion plot for words is used to investigate changes in language use over time.
4. Task 4: Count the vocabulary and display the sorted list of vocabulary
5. Task 5: Calculate a measure of the lexical richness of the text by performing
    a. How many times each word in text repeats.
    b. Count how often a word occurs in a text, and compute what percentage of the text is taken up by a specific word
6. Task 6: Find the Lexical diversity of various genres(tokens, Types, counts, lexical diversity, percentage) in the Brown Corpus which
    a. Loads the Brown Corpus from NLTK.
    b. Extracts words from each genre.
    c. Calculates Tokens: Total words in the genre, Types: Unique words, Type-Token Ratio (TTR): Unique words / Total words, Lexical Diversity Percentage:(Types / Tokens) * 100 and Stores results in a Pandas DataFrame for better readability

NLP_LAB_4

Q. 1 Perform the following Language Computing tasks

1. Task 1: Creation of N - gram Model by taking a sample text
    a. Tokenize the text using NLTK.
    b. Generate N-grams (unigrams, bigrams, trigrams) using nltk.util.ngrams().
    c. Use CountVectorizer from scikit-learn to extract n-grams automatically.
    d. Display the generated unigrams, bigrams and tri-grams.

2. Task 2: Create a N-gram Model as above for the Text paragraph taken as a list.

NLP_LAB_5

Q.1 Demonstrate how to perform the following Operations performed on Text considered as a list

- a. Concatenation
- b. Appending the text
- c. Indexing
- d. Sorting
- e. Finding lexical diversity
- f. Finding Collocations and Bigrams

Q.2 Take any sentence and a paragraph and demonstrate the computational statistical elements associated with the paragraph including

- a. Statistical Summary
- b. Frequency Distribution
- c. Freq Plots

NLP_LAB_6_1

Q.1 Take a string and a paragraph and perform the following text processing tasks

1. Drawing raw text
2. Take a sentence and perform following basic operations on text:
    - a. Getting multiline text
    - b. Concatenating the multiple strings
    - c. To find the length of the string
    - d. To check for the substring in a string, which returns true or false
    - e. To demonstrate string indexing and slicing
    - f. Getting details of the string / text
    - g. case conversion: Capitalize, upper and lower case, title case
    - h. string replacement
    - i. Checking for the strings/word- numeric, alpha, alphanumeric
    - j. to split, join, and strip the given strings
    - k. Joining the words by a blank space instead a ','
    - l. Removing the blank space using 'strip'
    - m. Printing every character on the new line
    - n. Formatting the string with new formatting method
    - o. s - using string format {}
3. Take a paragraph with multiple sentences and display the sentences separately when they are separated by '.' and join them back and display them on the two new lines.
4. Take two or more sentence and demonstrate working(creating and using) with regular expressions(regexes) using the module 're' for
    - a. Pattern matching
    - b. Text substitution are useful to find and replace specific text tokens in strings.
5. Print the Emojis using Unicode and CLDR short names in Python

NLP_LAB_6_2

Q. 1 Take any text file from the ‘gutenberg’ corpus for demonstration of text processing tasks and read the text using readLines() and perform the following:

1. Task 1: Basic preprocessing - remove all the empty newlines in the corpus and strip any newline characters from other lines using strip()
2. Task 2: Basic frequency analysis of the corpus including

   a. Task 2.1: Computing the length of each sentence and then visualize this using a histogram.
   b. Task 2.2: Visualize the overall distribution of typical sentence or line lengths across the selected file say blake/hamlet/bible.
   c. Task 2.3: Tokenize each sentence by splitting it into words and compute the length of each sentence to get the total words per sentence.

3. Task 3: To determine the most common words in the blake corpus. We already have our sentences tokenized into words (lists of words).
   a. Task 3.1: The first step involves flattening this big list of lists (each list is a tokenized sentence of words) into one big list of words.
   b. Task 3.2: Find the most frequent words invoke the counter from the collections - module
4. Task 4: To remove unwanted symbols and special characters in some ofthe words, use "re.sub"
5. Task 5: Text Wrangling
   a. Task 5.1: Web Scrapping
   b. Task 5.2: Tokenization : demonstration of word and sentence tokenization for (i) inbuilt text (ii) sample text (iii) text from some other language
   c. Task 5.3 Word Tokenization verification for both default and pre-trained WordTokenizer(WT)
6. Task 6: Text Normalization
   a. Task 6.1: Removing special characters for any given sentence before and after tokenization
   b. Task 6.2: Replacing contractions with expanded form
   c. Task 6.3: Matching the contraction in the sentence and replacing it with its expanded form
   d. Task 6.4: Removing Stopwords
   e. Task 6.5: Correcting words

NLP_LAB_6_3

Q.1 Download any html file and perform the following Text Normalization tasks

1. Removing HTML Tags
2. Removing special characters
3. Replacing contractions with expanded form
4. Removing Special Characters
5. Performing Stemming
6. Performing lemmatization
7. Removing stop words

Q.2 Generate a paragraph of your own choice and write a **text normalization routine** to perform all the seven text normalization tasks :

1. Removing HTML Tags
2. Removing special characters
3. Replacing contractions with expanded form
4. Removing Special Characters
5. Performing Stemming
6. Performing lemmatization
7. Removing stop words

NLP_LAB_7_1

Q.1 Demonstrate the working of Phrase Structure Grammar (PSG) / Constituency Grammar and display the phase structure with rules and productions generated using NLTK for the sentence with

1. Only NP(Noun Phrase)
2. Only VP(Verb Phrase)

3. Noun Phrase (Noun + Adjective)
4. Verb Phrase (Verb + Adverb)
5. Determiner, NP and VP
6. Two NPs, one VP and Determiner
7. Two NP and one VP
8. One NP, one VP and Adverb
9. Auxillary, Continuous present VP and a Noun
10. Auxillary, Continuous present VP and a Pronoun
11. a NP & adjective, VP & Adverb
12. a NP, two or more adjectives, a determiner and a conjunction
13. Two NP, a conjunction, auxillary
14. NP, VP, Det, Adjective, Auxillary, Pronoun, and a Conjunction with a single sentence with main and subordinate clause.
15. NP, VP, Det, Adjective, Auxillary, Pronoun, and a Conjunction using composition of two sentences

NLP_LAB_7_2

Q. 1 Demonstrate the working of Dependency Grammar for the sentences of your choice with

1. Only NP(Noun Phrase)
2. Only VP(Verb Phrase)
3. Noun Phrase (Noun + Adjective)
4. Verb Phrase (Verb + Adverb)
5. Determiner, NP and VP
6. Two NPs, one VP and Determiner
7. Two NP and one VP
8. One NP, one VP and Adverb
9. Auxillary, Continuous present VP and a Noun
10. Auxillary, Continuous present VP and a Pronoun
11. a NP & adjective, VP & Adverb
12. a NP, two or more adjectives, a determiner and a conjunction
13. Two NP, a conjunction, auxillary
14. NP, VP, Det, Adjective, Auxillary, Pronoun, and a Conjunction with a single sentence with main and subordinate clause.
15. NP, VP, Det, Adjective, Auxillary, Pronoun, and a Conjunction using composition of two sentences

and

a. print the parsing dependency
b. draw the dependency graph

NLP_LAB_7_3

Q.1 Demonstrate the working of Context Free Grammar and generate a structure using NLTK for the sentences with

1. Only NP(Noun Phrase)
2. Only VP(Verb Phrase)
3. Noun Phrase (Noun + Adjective)
4. Verb Phrase (Verb + Adverb)
5. Determiner, NP and VP
6. Two NPs, one VP and Determiner
7. Two NP and one VP
8. One NP, one VP and Adverb
9. Auxillary, Continuous present VP and a Noun

10. Auxillary, Continuous present VP and a Pronoun
11. a NP & adjective, VP & Adverb
12. a NP, two or more adjectives, a determiner and a conjunction
13. Two NP, a conjunction, auxillary
14. NP, VP, Det, Adjective, Auxillary, Pronoun, and a Conjunction with a single sentence with main and subordinate clause.
15. NP, VP, Det, Adjective, Auxillary, Pronoun, and a Conjunction using composition of two sentences