Prasad Jawale DIGAD 20

NLP Experiment 2

Aim: Apply various text processing techniques to any given text : tokenisation & script validation

Objective :

The objective of this practical is to explore and apply various text processing techniques to a given text.

Outcome :

Thus we applied various text processing techniques and script validation on a given text and also generated tokens out of given text

Problem statement:

we are going to perform various text processing tasks on a given text. The task involves tokenisation, regional language filtration, stop word filtration, punctuation filtration, email validation, name validation & phone number validation

Theory

① Tokenisation : It is a fundamental text processing technique which involves breaking down a piece of text into a smaller unit called tokens. It allows us to easier to do actions such as counting of words, analyze sentence structure, perform sentiment analysis etc.

Teacher's Sign.: _____

② Text Processing : Text processing techniques encompass a wide range of methods and operations applied to text data to extract useful information clean & preprocess the text.

ⓐ Stop word removal : Removing common words like "the", "is" and "in" that don't carry significant meaning in analysis.

ⓑ Punctuation removal : Eliminating punctu-ation marks such as comma, period & Question marks to focus on content.

ⓒ Regional language filtration : It ensures that only content in the desired language is retained for analysis.

③ Script validation : Script validation refers to process of ensuring that text or content adhere to a specific script or writing system. In a multicultural and multilingual words, it is crucial to validate and ensure that correct script is used in correct context.

Results & Discussion :

① Result : After applying these text processing techniques to the given text, we obtained the following results

ⓐ Tokenisation : the text was succesfully tokenised into individual words or tokens making it ready for future analysis

Teacher's Sign.: _____

(b) Regional language filtration: Non relevant language content was removed from the text to ensure consistency.

(c) stop word filtration: Common stop words were filtered to reduce noise in text.

(d) Punctuation filtration: Punctuation marks were removed to focus on content words.

(e) Email validation: It determined wether the email provided is valid or not

(f) Name validation: It validated if the given name is valid or not

(2) Discussion: Text processing is an esseptial technique in various NLP and data processing tasks, enabling more accurate & meaningful analysis of textual data

Mamta

Teacher's Sign.: _____

Prasad Jawale D16AD 20

NLP Experiment 3

Aim:

A) Implement Porter stemmer algorithm using Python
B) Use NLTK stemming function.

Objective:

The objective of this practical is to implement Porter's stemmeing algorithm manually using regular expression in python and compare its request with NLTK stem function.

Outcome:

Thus we implemented :

a) Porter stemmer's algorithm using regular expression in python
b) NLTK stemming function to perform stemming on a set of words

Problem statement:

a) Write a program to implement Porter's stemmers algorithm using regular expression in Python
b) Write a program to use NLTK stemming function

Theory

1) Stemming: It is a process in NLP that aims to reduce words to their base or root form. It involves removing suffixes from word to obtain common or simplified form.

Teacher's Sign.: _____

Example : For the words "running" & "runner", stemming would reduce both of them to common root 'run'.

Stemming is commonly used to do NLP tasks such as text search, information retrieval and text analysis. It helps in finding relevant document or information.

② Porter stemmer algorithm : This algorithm was developed for the purpose of stemming. The Porter Stemmer was developed by Martin Porter in 1980 and is one of the most widely used stemming algorithm. It consists of several stages each involving a series of regular expression and rules to transform words

Results & Discussion:

Both implementations produce very similar results for each given words, however there are some minor differences

happily → happli	'happy'
owned → own	'own'

Porter stemmer algorithm is used to perform stemming operation and it produces some non standard stems for certain words

Mantra

Teacher's Sign.: _____

Prasad Jawale DIGAD 20

NLP Experiment 4

Aim: Perform morphological analysis and word generat-ion for any given text

Objective: The objective of this experiment is to identify the root word for a given word, Identify its tense, POS etc also generate singular, plural, comparitive and superlative versions for the given word.

Outcome: Thus we were able to get the root word for any given word and were also to identify its tense, Part of speech etc. we were also able to generative singular, plural versions.

Problem Statement:

A) write a program to perform morphological analysis and identify the tense, POS, root word etc for a given word.

B) write a program to generate superlative, comperative, singular → plural versions of a given word.

Theory

① Morphological analysis: It involves breaking down words into their smaller meaningful units called morph-emes. Morphemes are smallest unit of language that carry meaning

Eg: Unhappiness

suffix → ness prefix → un

happy → root word

Teacher's Sign.: _____

Morphological analysis helps computer understand how words are formed and their grammatical structure

(2) word generation : word generation in NLP involves creating new words or word forms based on rules & patterns. One common use of word generation is to create different forms of word such as singular, plural etc

For example: word → Fast

Singular → Fast Plural → -

Comparative → Faster Superlative → Fastest

This is useful in NLP for task such as text generation, language translation & language understanding

Results & Discussion: Morphological analysis helps us in understanding how words are constructed for similar units (morphemes) and word generation is the process of creating new words from based on rules, which is especially useful for generating different variations of words in NLP tasks

Manika

Teacher's Sign.: _____

Prasad Jawale D16AD 20

NLP Experiment 5

Aim: Ngram models

A) display the list of unigram, bigram and trigram probabilities

B) display bigram probabilities & count table.

Objective: The objective of this experiment is to analyze the linguistic structure of a chosen corpus by generating list of unigram, bigram & trigram.

Outcome: Thus

A) we displayed the list of unigram, bigram and trigram probabilities.

B) we displayed bigram probabilities table and bigram count table.

Problem Statement:

In this experiment, we aim to seek and explore the patterns & relationships within a selected text corpus by generating a list of unigram, bigram & trigram.

Theory: N-gram is a term used in NLP to describe a sequence of n items typically words or characters that appear together in a piece of text. These items are usually adjacent to each other.

① unigram (1-gram): These are individual words in a text for example "I love cats", unigrams are 'I', 'love', 'cats'.

Teacher's Sign.: _____

② Bigram (2 gram) : Bigram consist of pair of adjacent words in a text. For example, "I love cats", Bigrams are "I love", "love cats".

③ Trigram (3 gram) : Trigram are sequence of 3 consecutive words in a word. For example "I love cats & dogs", trigrams are "I love cats", "cats and dogs".

④ N-gram (N > 3) : N gram can have more than 3 items and are esentially a sequence of n consecutive words

N-grams are used in NLP for reasons such as

→ Language modelling
→ Text analysis
→ Information retrieval

Results & Discussion : N grams are a way to break down sequence of words or characters helping computers analyze language patterns and improve various NLP tasks such as prediction, analysis & information retrieval

Mamta

Teacher's Sign.: _____

Prasad Jawale DIGAD 20

NLP Experiment 6

Aim: Perform POS (Part-of-Speech) tagging for given sentence

Objective: The objective of this practical is to perform POS tagging for a given sentence, enabling partici-pants to understand and how POS tagging assigns grammatical categories to word in a text,

Outcome: Thus we have identified the POS tag for each word in given sentence

Problem Statement: Write a program to identify POS tagging for each word in given sentence

Theory: Part of Speech (POS) tagging is a fundamen-tal task in natural language processing that involves categorizing each word in a text into specific grammatical category or part of speech. These categories include things like noun, verb, adjective, adverb etc.

POS tagging is used in NLP for several reasons such as
- → Language understandings
- → Text analysis
- → Parsing
- → Information retrieval

Teacher's Sign.: _____

Date
Page

Results & Discussion:

- **Results**

Word	POS Tag
The	DD
Quick	JJ
Brown	NN
Fox	NN
Jumps	VBZ
Over	IN
The	DT
Lazy	JJ
Dog	NN

- **Discussion**

We implemented a python program that uses NLTK library to perform POS Tagging on a sample sentence.

Manya

Teacher's Sign.: _____

Prasad Jawale D16AD 20

NLP Experiment 7

Aim: Exploratory data analysis of given text (word cloud) using python

Theory:

1) Exploratory data analysis (EDA) is a crucial phase in the analysis of textual data. It involves the initial examination and visualisation of data to uncover patterns, ~~the~~ trends & meaningful insights.
2) Word cloud is a graphical representation of text data where words are displayed in varying size based on their frequency within the text

Steps

1) Text Selection : Choose the target text or a document that you want to analyze.
2) Text Preprocessing : Preprocess the text to remove any irrelevant or noisy information
3) word frequency count : count the frequency of each word in preprocessed text
4) word cloud generation : Create the word cloud visualisation using a suitable software library or tool
5) Visualisation : Display the word cloud to visualise the distribution of words within text

Teacher's Sign.: _____

3) Applications of wordcloud
   → Identify prominent themes
   → Discover themes
   → Highlight keywords
   → Visualise data

Conclusion:
Wordcloud for a given dataset has been generated using python

Teacher's Sign.: _____

Prasad Jawale D16AD 20

NLP Experiment 8

Aim: Implement named entity recogniser for the given text input

Theory:

1) Named entity recognition (NER) is a crucial NLP task that aims to identify and categorise named entities, such as names of people, organ-isations, locations, dates and more within a given text input.

2) The key components of NER are

→ Tokenisation: The word is split into individual words or tokens.

→ Feature Engineering: It relies on a set of features to distinguish named entities from another words

→ Named entity categories: Named entities can belong to various categories, such as person, organisation, location, date or more

→ Machine learning model: models such as Conditional Random fields (CRF), Hidden markov models (HMM), LSTM's and transformers (BERT) are used to classify tokens into named entity categories

Eg: I hear Berlin is wonderful in the winter

Place → Berlin

Time → Winter

Teacher's Sign.: _____

3) Named entity recognition can be used in fields such as human resources, customer support, recommendation engines, content classification, healthcare etc

Conclusion:

Named entity recognition has been implemented in python

Teacher's Sign.: _____