Assignment No.7

Name - kudharan Sumit Dattatraya
Class - TE
Div - 4
Subject - DSBDAL

Problem Statement -

1. Extract Sample document and apply following document Preprocessing methods: Tokenization, POS Tagging, Stop words removal, Stemming & Lemmatization.
2. Create representation of document by calculating Term Frequency and Inverse Document Frequency.

Theory -

1) Explain text analysis with different steps

→ Text Analysis is about parsing texts in order to extract machine readable facts from them. The purpose of text analysis is to create structured data out of free text content.

Steps -

① Reading the text file -

```
filename = "C:\Users\Sumit\Desktop\example.txt"
text = open(filename, "r").read()
```

② Printing the text

```
Print(text)
```

③ Installing the library for NLP.

We will use the Spacy library which is an open-source software library for advanced NLP written in the programming languages python & Cython. Spacy also supports deep learning workflows that allow connecting statistical models trained by machine learning libraries like Tensorflow.

Page No.
Date

```
pip install -U pip setuptools wheel
pip install -U spacy
```

Since we are dealing with the English language so we need to install the en_core_web_sm package for it.

```
python -m spacy download en_core_web_sm
```

④ Tokenization -

Tokenization is the process of converting the entire text into an array of words known as tokens.

```
text_doc = nlp(text)
print([token.text for token in text_doc])
```

⑤ Sentence Identification -

Identifying the sentences from the text is useful when we want to configure meaningful parts of text that occur together

```
about_doc = nlp(about_text)
sentences = list(about_doc.sents)
```

⑥ Stopwords Removal -

Stopword are defined as words that appear frequently in language. So stopword must be removed from text to get a clearer picture of the text.

⑦ Punctuation Removal -

There are punctuation marks that are of no use to use.

⑧ Lemmatization -

Lemmatization is the process of reducing a word to its original form.

⑨ Word Frequency Count -

It is the process of finding top ten words according to their frequencies in the text.

⑩ Sentiment Analysis -

Sentiment Analysis is the process of analyzing the ~~sentimate~~ sentiment of text.

2) Explain the different document pre-processing methods.

→ ① Rescale Data - When our data consists of attributes with different scales mainly ML algorithm can be benifited from rescaling attributes. It means that all attributes of dataset have some scale so that measuring parameter of dataset maintains uniformity.

② Binarize data - Binarization is process that is used to transform data features of any entity into binary numbers.

③ Data Augmentation -

Data Augmention is strategy that allows practitioners or scientists to increase diversity of available data for training models, even without collecting or gathering new data.

There are various types of data augmentation given below:

① Flip ③ Crop

② Scale ④ Translation

3) Explain term frequency & inverse document frequency.

→ Term Frequency (TF) -

Suppose we have a set of English text documents & wish to rank which document is most relevant to the query, "Data science is awesome!" A simple way to start out is by eliminating documents that do not contain all three words "Data", "is", "Science", and "awesome" but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document the number of times a term occurs in a document is called its term frequency.

Formula - tf (t, d) = count of t in d / number of words in d

Document Frequency (DF) -

This measures the importance of document in whole set of

of corpus this is very similar to TF. DF is the number of documents in which the coord is present.

df(t) = occurrance of t in documents.

Inverse Document Frequency (IDF) -

IDF is the inverse of the document frequency which measure the informativeness of term t. When we calculate IDF, it will be very low for the most occuring coords such as stop coords.

Formula - idf(t) = ~~N+df~~ log (N/df+1))

4) How do you visualize text data in python?

→ ① Scatter Text -

Scatter Text is a powerful pytho based tool for extracting terms in a body of text & visualizing them in a interactive HTML display. ~~The official Github repo~~

② Word cloud -

A coord cloud is a text visualization technique that focuses on the frequency of coords and correlates the sized opacity of a coord to its frequency within a body of text. The output is usually an image that depicts different words in different sizes & opacities relative to the coord frequency.

5) What are the applications of text analysis?

→ ① Fraud detection

② social Media Analysis

③ Customer care Service

④ knowledge Management

⑤ Risk Management.

Conclusion - ① Extracted sample document & applied preprocessing methods. ② Created representation of document by calculating term frequency & Inverse document frequency.