Page No.
Date

Experiment No. : Group 1 - ③

- **Title :**

Perform text cleansining, perform lemmatization (any method), remove stop words (any method) label encoding.

- **Problem statement :**

Perform text cleaning, perform lemmatization (any method), remove stop words (any method), label encoding. create representations using TF-IDF. save outputs.

- **Practical objective :**

To understand Text cleaning, lemmatization and label encoding using TF-IDF.

- **Practical outcome :**

CO2 : Use tools and techniques in the area of software development to build ~~mini~~ practical.

- **software and Hardware requirements :**

Sr. no.	softwares / Hardwares	specifications.
1.	Laptop / Desktop	64-bits, 8 GB RAM.
2.	Jupyter Notebook	version 7.3.3.

Page No.
Date

• Theory :

1. In any machine learning task or data analysis task the first most step is to clean and process the data.

2. cleaning is important for model building. well, cleaning of data depends on the type of data. if the data is text uni then it is one more vital to clean the data.

3. well, the area various type of the text processing technique that we can apply to the text data, but we need to be careful while applying and chousing the processing steps.

4. forexample:
in sentiment analysis, we don't need to ~~remove~~ have a strong ided about what they want their end result to be the and even review the data to see what is exactly can achieve.

Text cleaning :

1. Text cleaning is task-specific and ~~one~~ ones need to have strong idea about what they want their and end result to be and even review the data to see ~~ex~~ what is exactly they can achieve.

Page No.
Date

- Having too many types of spelling mistakes in the text.
- Having too many number and punctuations (E.g. Love!!!)
- Text is full of emoji and motion and the user name and like too
- some of the text parts are not in the english language. Data is having a mixture of more than one language.
- some of the words are combined with the hyper phenor data having contractions words. (E.g. text-processing).
- Repetitions of words (E.g. Data)

Clean the textual data for the following methods.

- Lower caseing the data.
- Removing punctuations.
- Removing Numbers.
- Removing extra spaces
- Replacing the repetitions of punctuations.
- Removing Emojis
- Removing emotions.
- Removing contractions Importing the library.

Page No.
Date

Lower casing the Data:

from the first glance we just lower-case the data, the idea is to convert the input text into the same casing format so that it converts 'DATA', 'data', 'DaTa', into 'data'. In some casing, like the tokenizer and vectorization processes, the lower casing it done before hand. But choose the lower casing is done before hand.

Removing Punctuations :

The second most common text processing techniques. is removing punctuations from the texual data. The punctuation removal process will help to treat each text equally.

Removing Numbers :

some times numbers doen't hold any vital information in the text depending upon the use cases. so it is better to remove them than to keep them. for example, when we are doing sentiment analysis.

Removing Extra spaces :

Well, removing the extra spaces is good as it doesn't store extra memory and even we can see the data clearly.

Page No.
Date

Replacing the repetitions of punctuation:

Having knowledge of regular expression will help to code faster and easier. to remove the repetition of punctuation is very helpful because. it doesn't hold any vital information if we keep more than one punctuation in word.

Removing Emoticons:

while doing the text analysis of twitter and the Instagrams data we often find this emoticons and nowadays, there is hardly any text which doesn't contain any emoticons in them.

~~ID~~
TF - IDF

- Term frequency - inverse document frequency is a text vectorize that transforms the text into a usable vector.
- It combines 2 concepts, Term frequency (TF) and Document frequency (DF).
- The term frequency is the number of occurrences of a specific term in a document.

Page No.
Date

- **Conclusion :**

In this practical, I have performed the Text cleaning technique, and Lemmatization and the label encoding using TF- IDF.