Semester : VI Subject : DAV Academic Year: 2023- 2024

REPRESENTING TEXT :-

* It includes phase 3 and phase 4 of Data Analytics Life cycle.

* Tokenization is the task of separating (also called tokenizing) words from the body of text.

* Raw text is converted into collections of tokens after the tokenization, where each token is generally a word.

A common approach is tokenizing on spaces. For example, with the tweet shown previously:

I ~~only~~ once had a gf back in the day. Then the bPhone came out lol.

tokenization based on spaces would output a list of tokens.

{I, once, had, a, gf, back, in, the, day.,
Then, the, bPhone, came, out, lol}.

Note that token "day." contains a period. This is the result of only using space as the separator. Therefore, tokens "day." and "day" would be considered different terms in the downstream analysis unless an additional lookup table is provided.

The problems faced:

Can't → Can't should be writen as can and t

We'll → We'll should be writen we and ll

Wi-Fi → Wi-Fi should be written wi and Fi.

Text Normalization Technique - Case Folding:

It reduces all letters to lowercase (or the opposite if applicable). For the previous tweet, after case folding the text would become this:

Semester : VI Subject : DAV Academic Year: 2023 2024

I once had a gf back in the day, then the bphone came out lol.

Problems faced:

For example,
General motor → The gGeneral motor ~~can~~ be ~~was~~ will be interpreted as general and motors.
WHO → WHO stands for World Health Organization but here WHO changes to who.

Text Normalization Technique - Stopwords.

- Stop words are a set of commonly used words in a language.
- Examples of stop words in English are "a", "the", "is", "are" etc
- It is used to eliminate words that are so widely used that they carry very little useful information.

Example:

```
import nltk
from nltk.corpus import stopwords.
nltk.download('stopwords')
print (stopwords.words('english'))
```

Output:

```
['i', 'me', 'myself', 'we', 'our', 'ours', 'ourselves', .... ].
```

Example: Stopwords

Sample text with Stop Words.	Without Stop Words
GeeksforGeeks - A Computer science Portal for Geeks. Can listening be exhausting? I like reading, so I read	GeeksforGeeks, Computer Science, Portal, Geeks. Listening, Exhausting. Like, Reading, read.

Semester : VI Subject : DAV Academic Year: 2023 - 2024

Bag of Words (BOW) model in NLP :

Bag of Words model is used to preprocess the text by converting it into a bag of words, which keeps a count of a total occurences of most frequently used words.

This model can be visualized using a table, which contains the count of words corresponding to the word itself.

Example :

Let us consider the three sentences:

(1) Hello and Welcome everyone.

(2) John loves teaching.

(3) John loves to play cricket.

Lets do Bag of words for the following sentences: We have to generate vectors for these sentences. We will consider the unique words in the corpus given.

	Hello	and	welcome	everyone	John	loves	teaching	to	play	cricki
(1)	1	1	1	1	0	0	0	0	0	0
(2)	0	0	0	0	1	1	1	0	0	0
(3)	0	0	0	0	1	1	0	1	1	1

remove stopword remove stopword

For every sentences, we subititute with 1 and 0 according to the occurance of the word:

To reduce the size, we remove the stopwords from the given table. In the given example "and and to" is stopword. Let us remove it and we get 8 words. This is how bag of words model work.