Roll no:- 41310

Assignment:- 4

Page No.:
Date:

Title:- Consider suitable text data set, Remove stop words, apply stemming and feature selection techniques to represent document as vectors. classify documents & evaluate precision, recall.

* objective:-
* Implementation of the problem statement using python.
* Remove stop words, apply stemming and feature selection.

* Theory:-

In Computing stop words are the words which are filtered out before and after processing of natural language data. Though stop words refer to most common words in a language, there is no single universal list of stop words used by all natural language processing tools and indeed not all tools even use such a list. Any group of word can be chosen as the stop words for a given purpose.

Code to remove stop words.

```
from nltk.tokenize import sent_tokenize,
word_tokenize.

from nltk.corpus import stopwords.

data="data mining is a subject."
```

```
stopwords = set(stopwords.words('english'))
words = word_tokenize(data)
wordsFiltered = []
for w in words:
    if w not in stopwords:
        wordsFiltered.append(w)
print(wordsFiltered)
```

* **stemming:-**

stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form - generally a written word form. The stem need not be identical to the morphological root of the word it is usually sufficient that related words map to same stem, even if this stem is not in itself a valid root.

Code for stemming:-

```
from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize

ps = PorterStemmer()
ex_words = ["python", "data", "mining"]

for w in ex_words:
    print(ps.stem(w))
```

Page No.:
Date:

* Feature Extraction:-

In machine learning and statistics feature selection, also known as variable selection, attribute selection or variable subset selection is the process of selecting a subset of relevant feature for use in model construction.

(1) Feature selection techniques are used for four reasons.

(2) simplification of models to take them easier to interpret by researchers/users.

(3) shorter training times.

(4) To avoid the curse of dimensionality.

(5) Enhanced generalization by reducing over fitting.

Conclusion:-

We succesfully removed stop words, apply stemming and feature extraction techniques to represent documents as vectors.