Assignment No. 4

* Problem Statement :
Consider a suitable text dataset. Remove Stop words, apply Stemming & feature Selectio techniques to represent documents as vecto classify documents & evaluate Precision, recall.

* Objective : Learn how to tokenize & filter a document into its diff. words and then do word count for each word in doc. Apply stemming, feature Selection on doc (txt)

* outcome : demonstrated text Processing using nltk, understood vectorizing & removing Stop words.

* Theory :

⊙ NLP is a Subfield of linguistic computer Science & AI concerned with interaction b/w computer & Human language, in Particular how to Program computer to Process & Analyse large amt of Natural language dat

⊙ In computing stop words are words that are filtered out before or after the natural language data (text) are

Processed. while 'stop words' typically re to the most common word, there is no universal list of stop words.

⊙ STEMMING ⇒ for grammitical reasons, do are going to use diff. forms of a word (e.g. organized, organizing, etc). Additionall there are families of derivitavelly relati words with similar meaning like democra democresy & democratization.

→ The goal of stemming (and lemmatization) is to reduce inflectional forms & sometim derivationally linked forms of a word to its common base form

am, are, is → be

car, cars, cars', car's → car

→ when applied to a doc, result will be somewhat like this

the boy's cars are diff. color

the boy car be differ color

→ Stemming is a more crudde process, a heuristic that chops off the end of a word in the hope of achieving the goal correctly more of time, and often includes removal of derivational affixes

→ Feature selection is a process of selecting a subset of the terms occurring in the remaining set and using only this Subset as feature in text classification.

→ It serves 2 purpuses. first, it makes training & applying a classifior more efficient by decreasing the size of vocab. this is of perticulor importance to classifion that are expensive to train.

→ Second, feature Selection often increases classification accuracy by eleminating noise featu these features when added to docs representati increases the classification error on new data.

→ Vectorization is a Process of converting text into machine redable form. words are represented as 'Vectors'(numerically)

→ Count vectorizer(one-not encoding) involves counting the number of occurence for each word occuring in the doc.

→ Idea behind it is Simple. vector is created having as many dimentions as there are distinct words in text/doc/collection of docs being used. each unique word has a unique dimention & is represented by a 1 in the dimention with 0s everywhere else. This results in huge & sparse vectors that capture no relational data.

→ Tf IDF vectors are related to one ho encoded vectors, but instead of just featur a count, they feature numerical repres-ation, where ~~they~~ words aren't just prese or not Present instead, they are represe by their term freq. multiplied by inverse doc. freq.

→ In simpler terms, words that occur everywhere should be given very littl weight or significance, coz they dont provide a large amt. of value, Howev if a word appears very little or ~~freq~~ frequently but only in specific Place, then they are Possible of higher significanc

→ the downside is there in no capture of Semantic relatedness, this is solved with the co-occurence matr or a ~~[illegible]~~ neutral ~~predictive~~ Probabillistic model.

* Conclusion :

Text docs were tokenized, filtered, vectorized, & thus successfully Processed basics of NLP were understood.