Assignment No. 7

Aim :- Test analytics:

1) Extract sample document & apply following document preprocessing methods:
   Tokenization, pos tagging, stop words removal, stemming & lemmatization.

2) Create representation of document by calculating Term frequency & inverse document frequency.

Theory :-

TF - IDF ~~from~~ scratch in python on a real-world ~~dataset~~.

Table of contents:-

- what is TF - IDF?
- preprocessing data.
- weights to title & body
- Document retrieval using TF-~~I~~DF matching score
- Document retrieval using TF-IDF cosine similarity.

Introduction:- TF - IDF.

TF - IDF stands for "Term frequency - Inverse Document frequency". This is a technique to quantify words. we generally compute a score.

TF - IDF = Term ~~F~~frequency (TF) * Inverse Document frequency (IDF)

Terminology:-

- t - term
- d - document
- N - count of corpus
- corpus - total document set.

Term frequency:-

This measures frequency of word in a document. Highly depends on length of document.

$$TF(t,d) = \frac{\text{count of } t \text{ in } d}{\text{no. of words in } d}$$

Document frequency :-

This measures the importance of documents in a whole set of corpus. This is very similar to TF but difference is that DF is count of occurrences of term t in document set N.

$df(t) =$ occurrences of t in N documents.

Inverse document frequency :-

IDF is inverse document frequency which measures the informativeness of term t.

$$idt(t) = \frac{N}{df}$$

if few words of vocab might be absent,

$$idf(t) = \log(N/(df+1))$$

by taking multiplicative value,

$$tf\text{-}idf(t,d) = tf(t,d) * \log(N/(df+1))$$

Implementing on a real world dataset:-

1) step1:- Analysing dataset.

The first step in any machine learning task is to analyse the data.

2) step 2:- Extracting title & body

This totally depends on problem statement at hand and on the analysis, we do on the dataset.

3) step3: Preprocessing

Preprocessing is one of the major steps when we are dealing with any kind of text model. During this stage we have to look at the distributn of our data, what techniques are needed & how deep we should clean.

- stop words:-

stop words are most commonly occuring words that don't give any additional value to the document vector.

- punctuation:-

punctuation is the set of unnecessary symbols that are in our corpus documents.

- Apostrophe:-

Note that there is no 'apostrophe' in the punctuation symbol.

Because when we remove punctuation first it will convert don't to dont. & it stop word that wont be removed.

- Single characters :-

single characters are not much useful in knowing the importance of the document & few final singal characters might be irrelevant symbol.

- stemming :-

This is final & most important part of preprocessing stemming converts words to their stem.

- Lemmatisation :-

Lemmatisation is a way to reduce the word the root synonym of a word.

= step 3: calculating TF-IDF.

document = body + title

TF-IDF = TF-IDF (title) * alpha + TF-IDF (body) + (1-alpha)

- calculating DF :-

DF will have word as key & list of doc id's as value.

⇒ conclusion:- Hence in this manner we test the analytics.