Module 2:Extracting , Cleaning and Preprocessing Data

Case Study

Case Study

1.Write a program to enter a string from user and perform following tasks

- Write a python function named “Tokenize” which returns the tokenized string
- Print tokens along with the frequency of each token using the above function
- Print the 5 least occurring tokens

2. Write a program to enter a string from user and perform following tasks.

- Write a python function named “RemoveStopWords” which returns the string after removing stop words
- Count frequency of each stop word present in a string using the above function
- Plot a bar graph depicting stop words and their frequencies

3. Write a program to enter a string from user and perform following tasks

- Write a python function named “Lemmatize” which returns a string after lemmatizing the string.
- Write a python function named “Stemmed” which returns a string after stemming the string.
  (Use any stemmer of your preference)
- Print all the words along with their lemmatized and stemmed form using the above functions
- Save these results in a csv file having 3 columns:

Original Word	Lemmatized Form	Stemmed Form

4. Create a python file named “PreProcess” and perform the following tasks.

- Copy the function “Tokenize” in this file from question 1
- Copy the function “RemoveStopWords” in this file from question 2
- Copy the function “Lemmatize” in this file from question 3

Create a function named “Refine” which accepts a string and call the above 3 functions in the same order i.e. first Tokenize then RemoveStopWords then Lemmatize.

Remember:

>Inputted string will be input to Tokenize Function
> Tokenized String will be input to RemoveStopWords function
> StopWordsRemoved string will be input to Lemmatize function

Save this python file as PreProcess and you can use this for upcoming assignments.