Module 2:Extracting , Cleaning and Preprocessing Data

Case Study

Case Study

You are provided with a file named "Brexit.docx". This file contains the introductory paragraph from Wikipedia on Brexit. (https://en.wikipedia.org/wiki/Brexit)

You have to perform some analysis on this paragraph by performing the following tasks:

1. Read the file "Brexit.docx" and write a function in Python named "GetNGrams" which takes a string and a number 'n' as input and returns n-grams from the string.

Example:

String: "John met with an accident"

Output:

When n=2 => [(John, met), (met, with), (with, an), (an, accident)]

When n=3 => [(John, met, with), (met, with, an), (with, an, accident)]

When n=4 => [(John, met, with, an), (met, with, an, accident)]

2. Read the file "Brexit.docx" and write python functions which take a string as an input and returns:

- Number of Nouns (all forms of noun). Take function name as "NounsCount"
- Number of Pronouns (all forms). Take function name as "PronounsCount"
- Number of Adjectives (all forms). Take function name as "AdjectivesCount"
- Number of Verbs (all forms). Take function name as "VerbsCount"
- Number of Adverbs (all forms). Take function name as "AdverbsCount"

Plot a pie chart showing the distribution of nouns, pronouns, verbs, adverbs and adjectives.

3. Read the file "Brexit.docx" and write python functions which take a string as an input and returns:

- Number of geo-Political entities present in the file. Take function name as "GeoPoliticalCount"
- Number of Persons present in the file. Take function name as "PersonsCount"
- Numbers of Organizations mentioned in the file. Take function name as "OrganizationsCount"

4. Answer the following questions:

- Most frequent bi-gram from the data
- Most frequent Noun
- Most frequent GeoPolitical Entity
- Most frequent person