Implementation of a method to identify the language a document is written in.

April 23, 2019

In [ ]:
```
"""
title: Implementation of a method to identify the language a document is written in.
author: Shakeel Ahmad Sheikh
date: 23-April, 2019
"""
```

In [14]:
```
import sys
from nltk import wordpunct_tokenize as tok
from nltk.corpus import stopwords as sw
#Function to detect Language
def language_detection(input):
    """
    Probabilty of input text in several languages is being
    calculated and the languge with highest probabilty score is returned
    Using Stop words technique to detect Language
    input: Input Text for which we want to detect language """
    #Computing Language Probability
    languages_with_ratios = {}
    #Tokenization
    tokens = tok(input)
    words = []
    for word in tokens:
        words.append(word.lower())
    for language in sw.fileids():
        stopwords_set = set(sw.words(language))
        #print(stopwords_set)
        #print("\n")
        words_set = set(words)
        #print(words_set)
        #print("\n")
        common_elements = words_set.intersection(stopwords_set)
        #print(common_elements)
        #print("\n")
        #print(len(common_elements))
        #print("\n")
        languages_with_ratios[language] = len(common_elements)
```

```
            #print(languages_with_ratios)
            #print("\t")
        language_with_high_prob = max(languages_with_ratios, key=languages_with_ratios.get)

        return language_with_high_prob
```

In [15]:
```
###Main Function
if __name__=='__main__':

    #Reading Text Files
    file_en = open("english.txt","r+")
    text_en = file_en.read()
    file_en.close()
    file_gr = open("german.txt","r+")
    text_gr = file_gr.read()
    file_gr.close()
    file_gk = open("greek.txt","r+")
    text_gk = file_gk.read()
    file_gk.close()
    file_sp = open("spanish.txt","r+")
    text_sp = file_sp.read()
    file_sp.close()

    ##Sample Input Text
#     input = '''
#     There's a passage I got memorized. Ezekiel 25:17. "The path of the righteous man
#     by the inequities of the selfish and the tyranny of evil men. Blessed is he who,
#     '''

    language1 = language_detection(text_en)
    language2 = language_detection(text_gr)
    language3 = language_detection(text_gk)
    language4 = language_detection(text_sp)

    print("Language Detected in document  english.txt is " + language1)
    print("Language Detected in document german.txt is " + language2)
    print("Language Detected in document greek.txt is " + language3)
    print("Language Detected in document spanish.txt is " + language4)
```

```
Language Detected in document  english.txt is english
Language Detected in document german.txt is german
Language Detected in document greek.txt is greek
Language Detected in document spanish.txt is spanish
```

In [ ]: