# CE706 - Information Retrieval

# Assignment 1

# Report

# Name: Yaowei Lyu

# Tutor: Professor Udo Kruschwitz

# Student Number: 1802697

## Software Requirement

In this case i suppose to use the Pycharm as my compiler. This programme also need to use the nltk, BeautifulSoup, StanfordNERtagger, TfidfVectorizer, RegexpParser and ssl to complete the function.

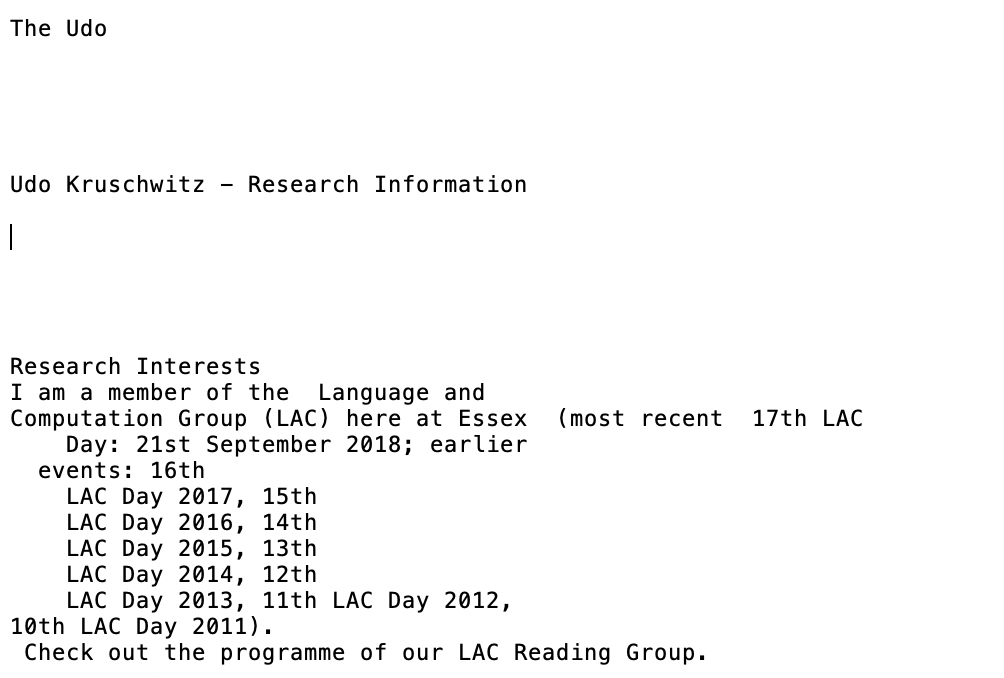
## **Function Explanation**

* + **Get\_text()**

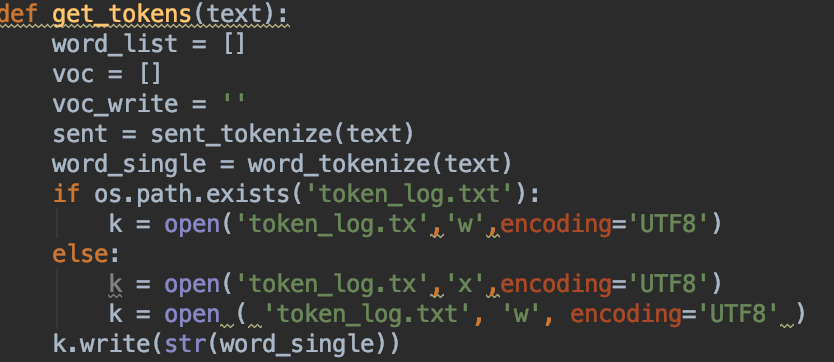


This function is going to abstract the information from URL, in order to complete this function, I use BeautifulSoup to implement this. Beautifulsoup can get the text from the URL by its own get\_text

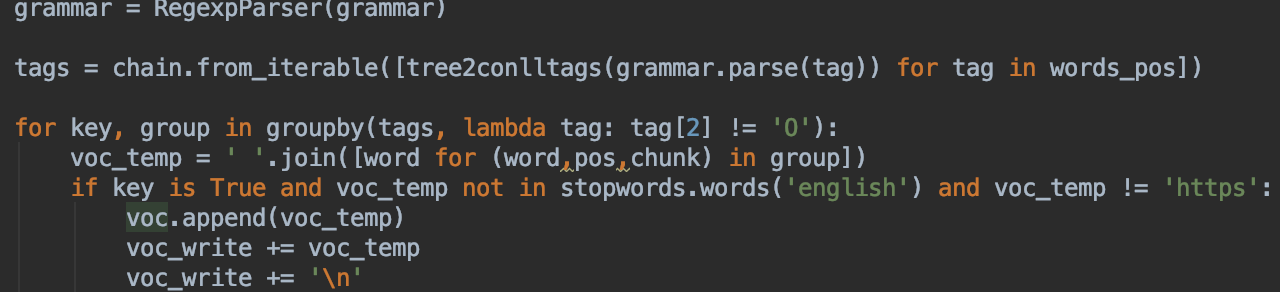
* + **Result**



* + **Get\_tokens()**

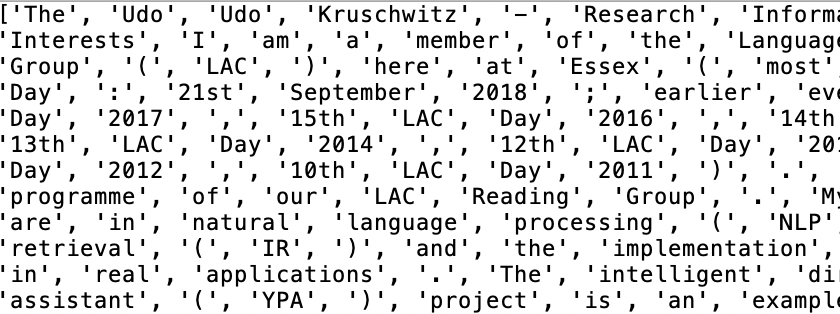


In this part I use the word\_tokenize to get the tokens of the sentence and the sent\_tokenize to get the splitting sentence.

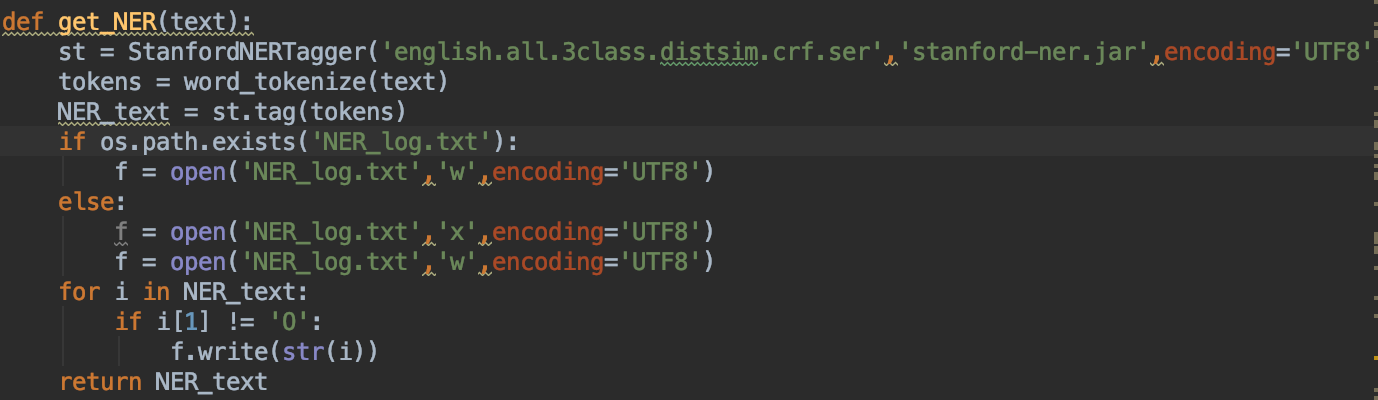


I suppos to create a grammar rule and get the phrase from the tokens. So I decide to use the tree2conlltags to complete this requirement. If the word is stopwords like: a, the, is etc, it will be removed。

* Result

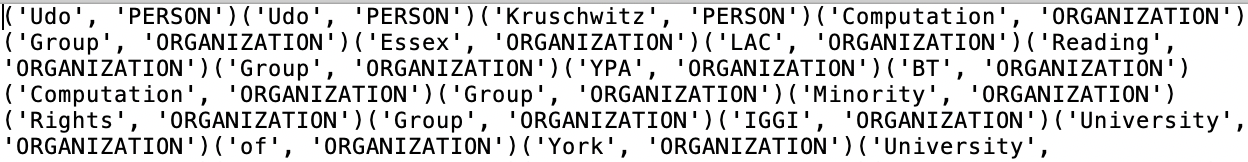


* + **Get\_NER()**

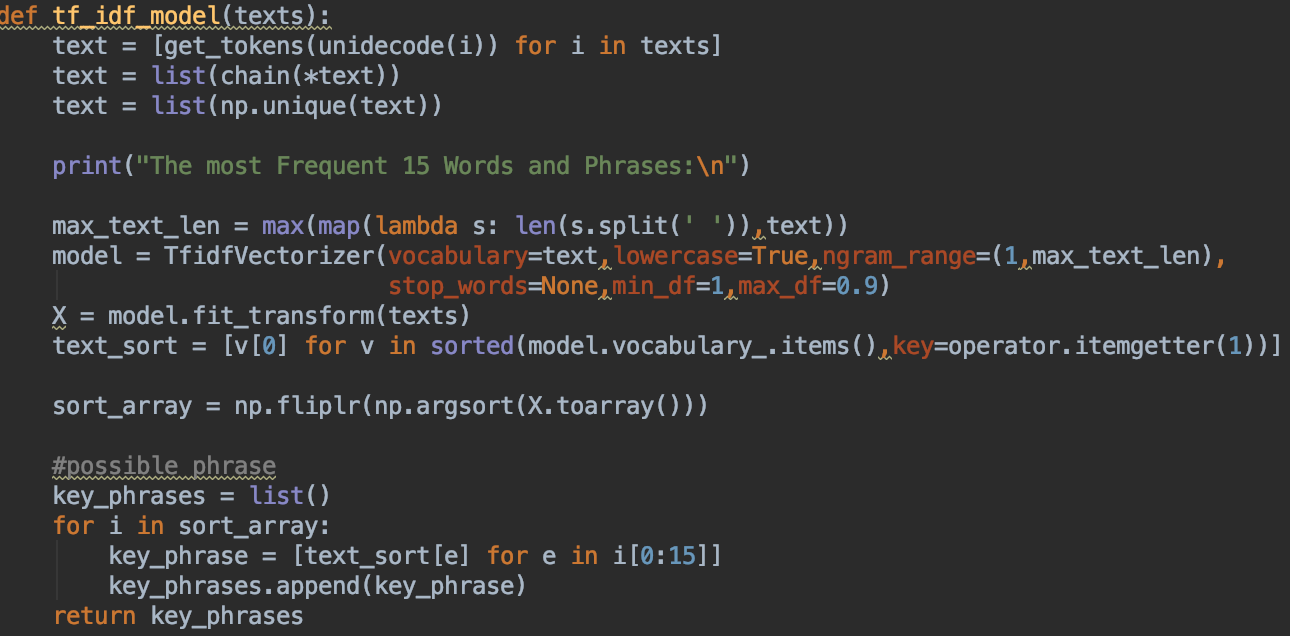


In this part, I am going to use the Stanford to get the name entity recognition. First get the tokens of the text and then put in the model. After that it can get the NER.

* Results



* + **Tf\_idf model**



This part is going to build the model of TF-IDF. We need to get the text of the URL Then input the text as the Input of the model. After that sort the result and get the phrases.

* Results

