Shopback Coding assignment
Search Backend Engineer

1. [Word Tokenization] Tokenization is the process to split a sentence into smaller parts called tokens. Please choose one of your favours open source tokenizer project and write a simple program to tokenize input file .
    a. Download Google news RSS feed(https://news.google.com/rss?hl=zh-TW&gl=TW&ceid=TW:zh-Hant) as input, you need to tokenize the description field for each news article. (please ignore other fields like title, link, guid, pubDate, ...)
    b. The description field might include Chinese characters, English word, number and even html tag. please make sure your tokenizer can handle them properly
    c. Please include the following 5 files in your question1 directory:
        - **readme.txt** : 1) Introduction to the open source projects you used(ie, url, github, document page). 2) How to install these libraries in your development environment
        - **news.rss** : The rss feed you downloaded from google.
        - **description.txt** : The description field for each news article, one line per news article.
        - **source code of your program.** (you can use any filename)
        - **output.txt**: The tokenized version of description, one line per article.
    d. Estimate efforts: less than 2 hours.

2. [TF-IDF] TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.(from wikipedia : https://en.wikipedia.org/wiki/Tf%E2%80%93idf). Please write a simple program to generate TF-IDF vector from **output.txt** from Question #1 above.
    a. Please choose any one of your favour open source libraries to calculate TF-IDF vector.(Don't implement a TF-IDF algorithm by yourself) for **output.txt** from question #1.
    b. You need to understand how to calculate the TF-IDF vector and get familiar with any option or parameters supported by your library.
    c. Please include the following files in your question2 directory:
        - **readme.txt** : same as question #1.

- **Source code of your program**. (you can use any filename)
- **output.txt** : Print the original input sentence(tokenized description) first then output the TF-IDF vector for each news article.
  The TF-IDF vector format is doesn’t matter, use the most convenient method provided by your library.

d. Estimate efforts: less than 2 hours.

Note:

1. You are encouraged to reference/copy the answer from internet(github, stack overflow, open source projects and etc). However, please make sure you understand how and why they work. We will have further questions to evaluate your understanding.
2. You need to understand the API you called and the purpose of every parameter they supported.
3. You are feel free to use any programming language for this assignment.