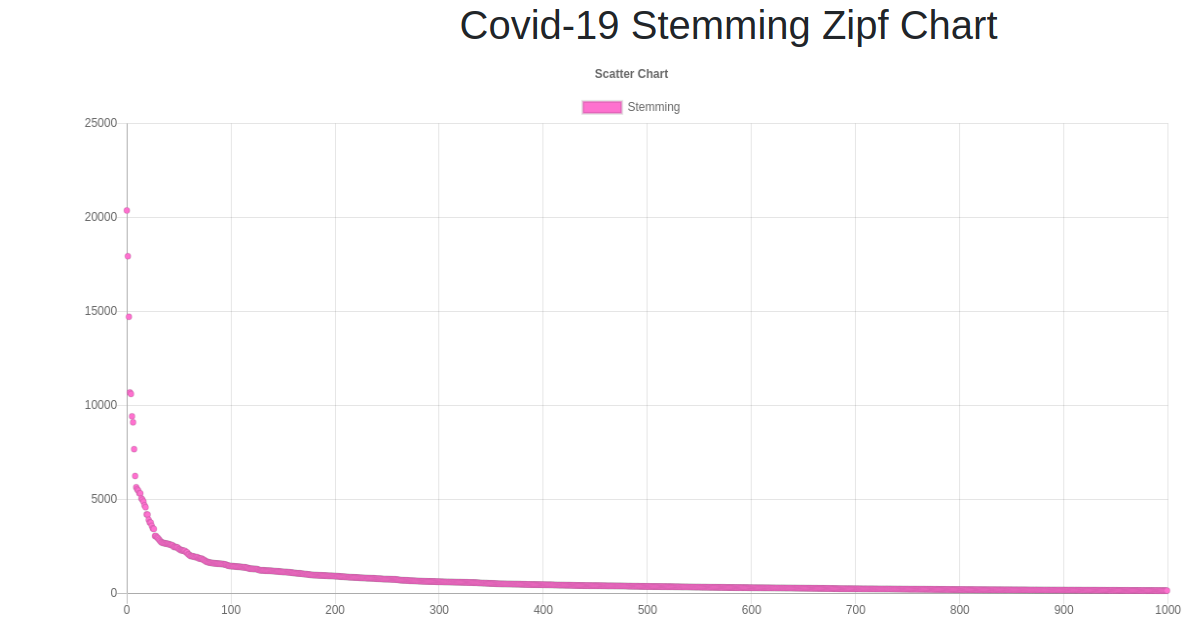
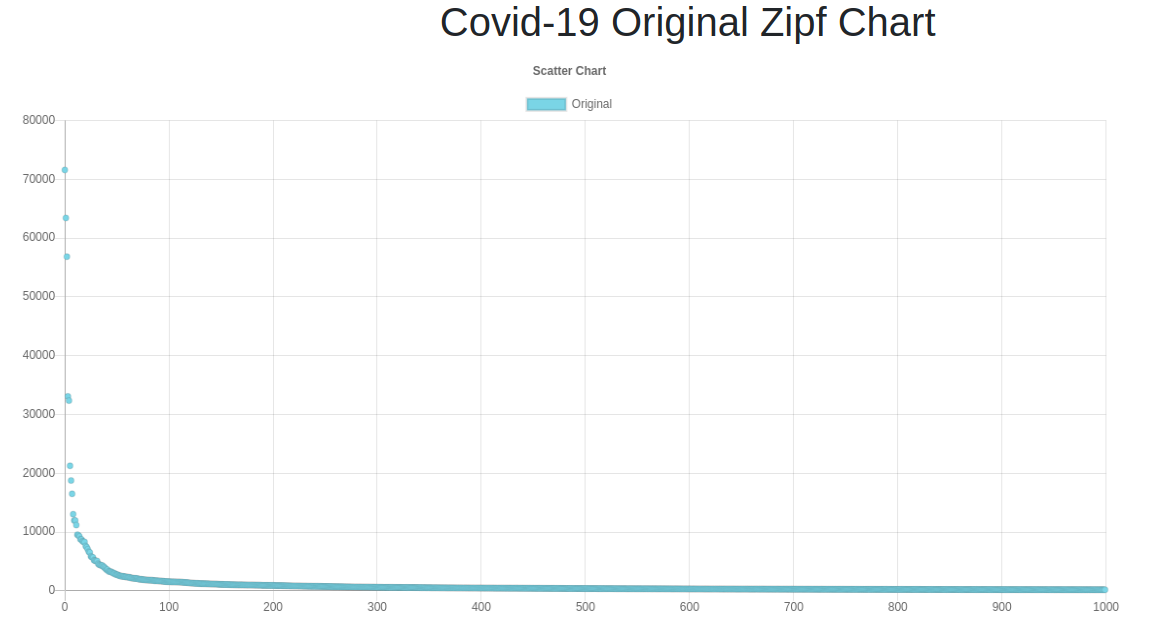
# IIR Homework 2 方郁文

## 系統環境

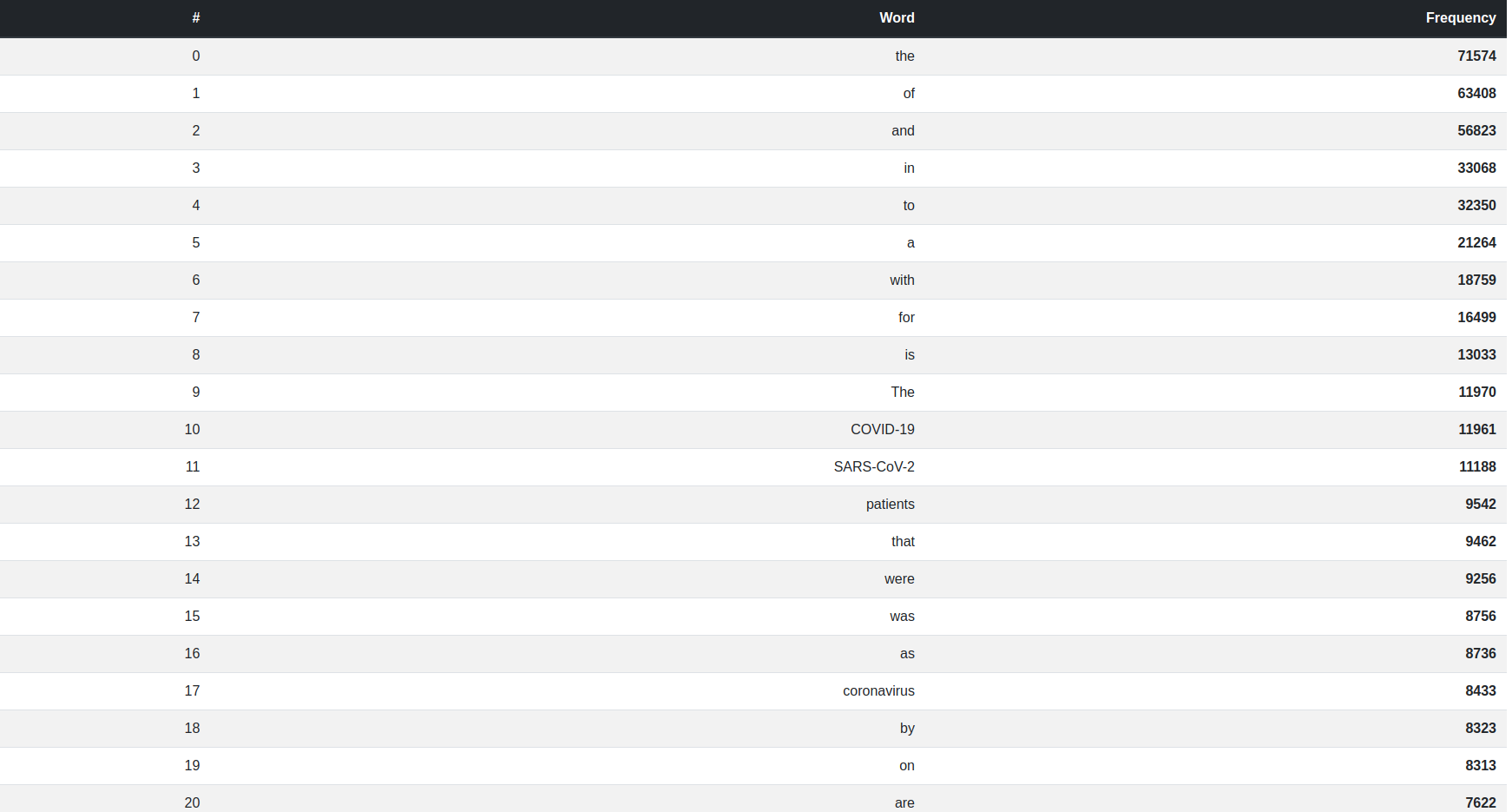
* 作業系統 ： Ubuntu 20.04 LTS
* 程式語言 ： Python 3.7.9
* 網站架構 ： Django 3.1.1
* Porter Stemmer : NLTK 3.5
* Edit distance : PySpellChecker 0.5.5
* 讀取csv檔 ： Pandas 1.1.3
* Pubmed爬蟲 ： Requests 2.24.0 and Beautifulsoup 4.9.1

## 實作方法

* 首先利用Requests及Beautifulsoup4對Pubmed針對 “Covid-19” 這個關鍵字進行爬蟲，擷取了9850篇文章的Title和Abstract。
* 對於爬下來的文章進行處理，得到有1416則文章沒有abstract，對剩下的8436篇文章進行分析。
* 延續Homework1的作法，將8436篇文章的Abstract以 原文 以及 處理後 的方式存入資料庫。處理方式如Homework1，將標點、空白、以及Stopwords刪除後，對剩餘的字套用Porter Stemming演算法，存入資料庫。
* 延續Homework1，將字建立inverted-index，將每個字從『哪篇文章』中的『哪個位置』出現建立，以供搜尋時候使用
* 為繪製Zipf圖表，統計每個字在『Stemming前以及Stemming後』出現的頻率，即可針對『 未處理的字 』和『 處理後 』的字繪製Zipf圖



x軸為出現的字，y軸為此字出現的頻率

* 將每個字出現的頻率輸出成table
* 能夠以User搜尋找到的文章為subset，繪製其專屬的Zipf 
* 分析Subset中的高頻字在全部的文章中出現的頻率高低，繪製成Table

