# IIR Homework 3

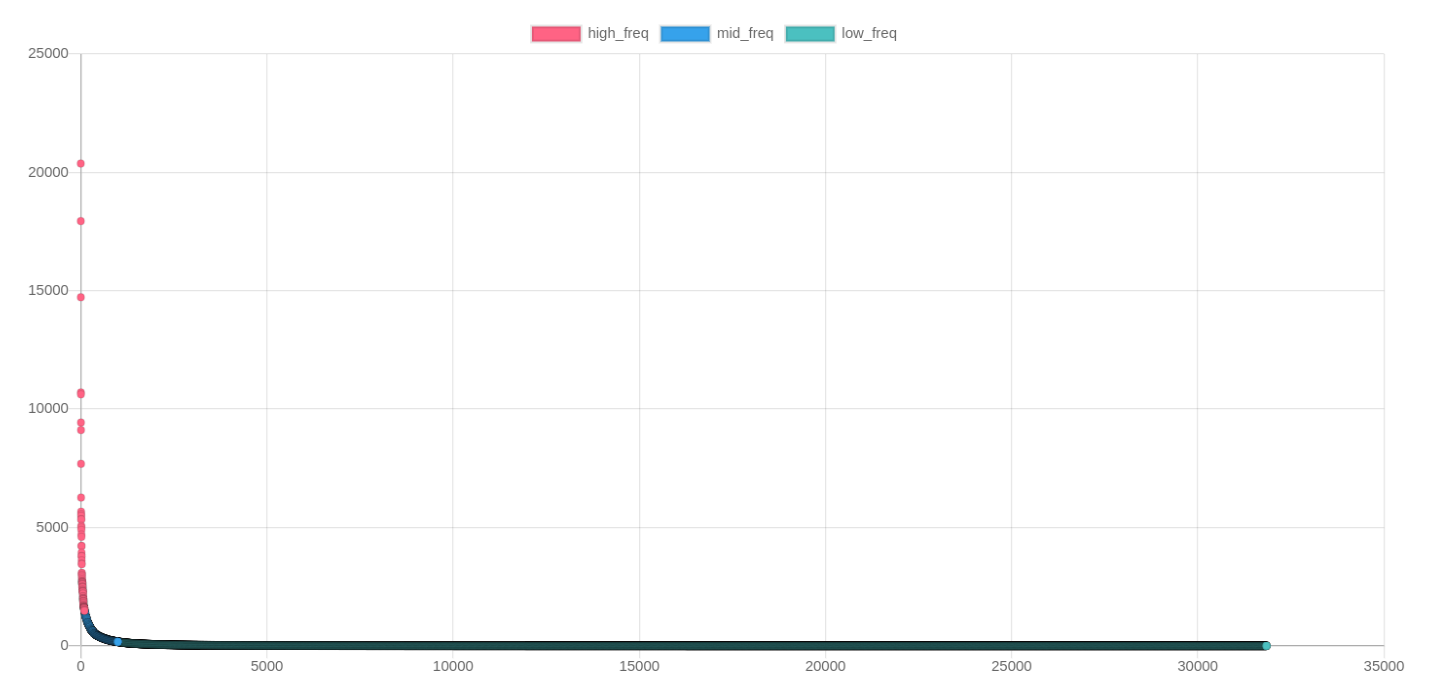
## 系統環境

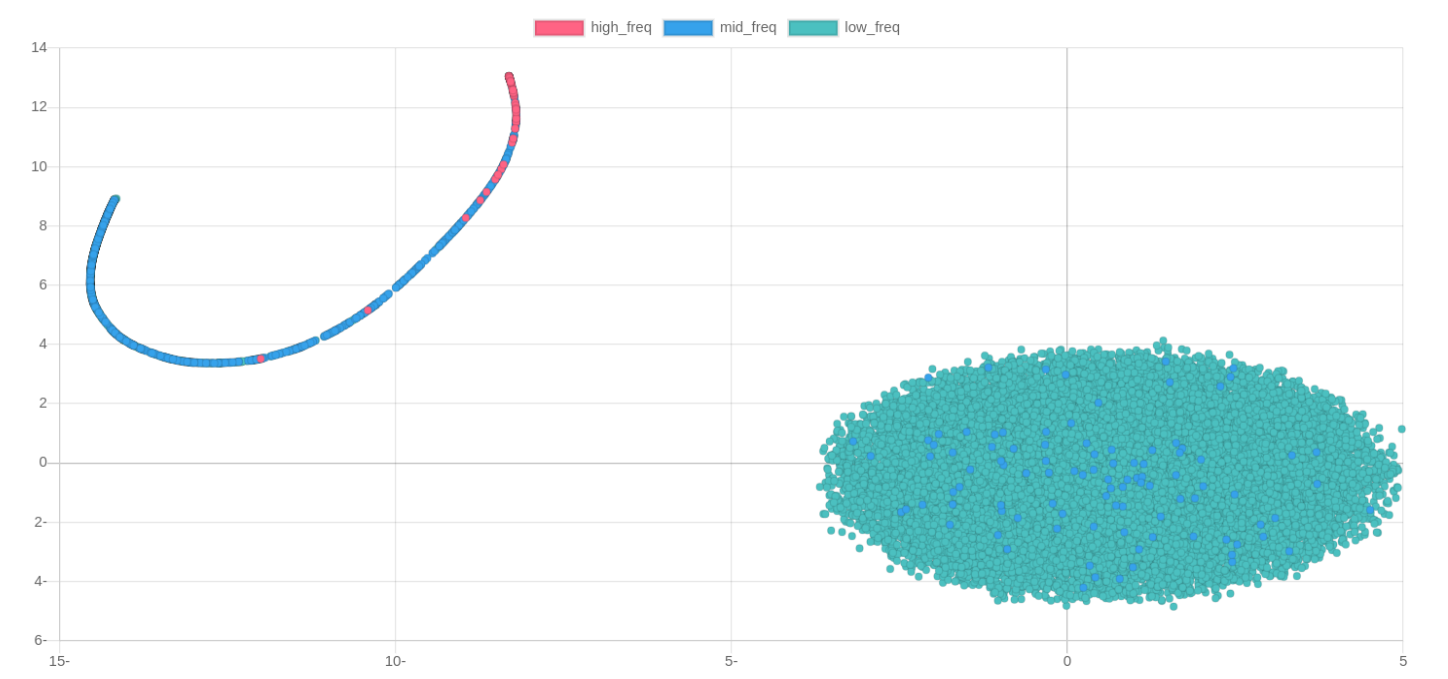
* 作業系統 ： Ubuntu 20.04 LTS
* 程式語言 ： Python 3.7.9
* 網站架構 ： Django 3.1.1
* Porter Stemmer : NLTK 3.5
* 讀取csv檔 ： Pandas 1.1.3
* Word2Vec : gensim 3.8.0
* 資料降維 ： tsnecuda 2.1.0

## 實作方法

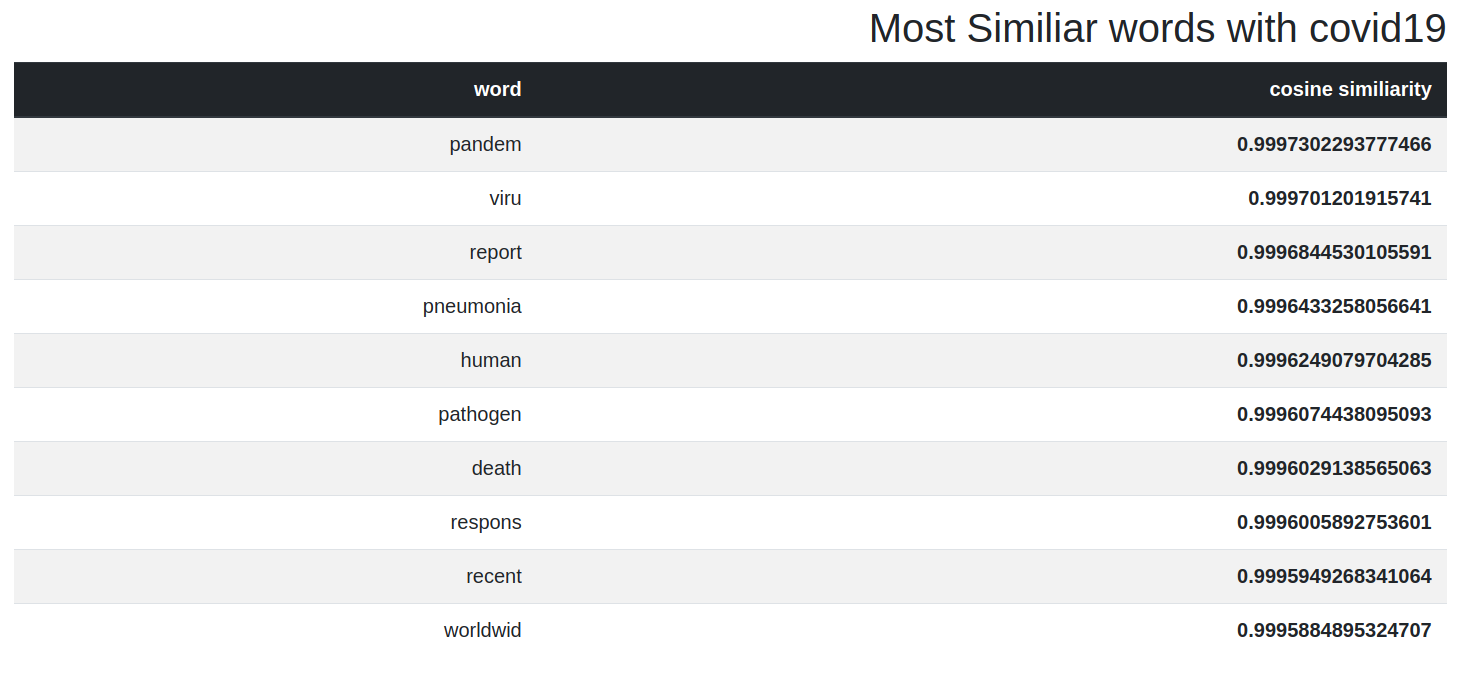
* 利用python的gensim套件建立word2vec，採用skip-gram方式。
* 先試著將IR hw2蒐集到的所有文件作前處理（刪除stop-wrods, stemming)，再將處理完的文字丟入word2vec做training。
* word2vec model的Hyperparmeter設定如下：
* 使用skip-gram
* Dimension為1000
* Training的window size設定為3
* 頻率大於1的字都加入training (及所有文字都會training)
* 除了使用全部文檔之外，參考ir hw2，將所有文檔分subset，利用這些subset再train其特有的word2vec Model

## 實驗

* 利用t-SNE將word vector做降維 



上圖中 紅、藍、綠 分別對應下面zipf圖中的高頻字、中頻字和低頻字，可看出這3種字可被大致分成3種cluster

* 計算cosine similarity

上圖為比較covid-19以及和其相似度高(cosine similarity)的字