隊名NTU_b06705057_
成員 b06705001楊力行 b06705009任恬儀 b06705057 黃資翔 b06705058劉品枘
2. 選擇的題目是新聞的分析

3.
參考TF-IDF算法提取關鍵字
TF-IDF 演算法包含了兩個部分：詞頻跟逆向文件頻率。詞頻指的是某一個給定的詞語在該文件中出現的頻率，而逆向文件頻率則是用來處理常用字的問題。而一個字對於一篇文件重要性的分數就可以透過TF與IDF兩個指標計算出來，當詞彙 t 很常出現在文件 d 時，他的 $tf_{t,d}$ 就會比較大，而如果詞彙t 也很少出現在其他篇文章，則 idf_t 也會比較大，使 $w_{t,d}$整體來說比較大，也就是說詞彙t 對於文件 d 來說是很重要的。如此一來，我們就可以計算出 TF-IDF 矩陣
https://blog.csdn.net/Nonoroya_Zoro/article/details/80342532
用jieba 進行中文詞彙斷詞
https://github.com/fxsjy/jieba
word2vec進行詞向量的轉換https://radimrehurek.com/gensim/models/word2vec.html
運用RNN架構進行詞向量的訓練
https://arxiv.org/abs/1808.03314

4. 目前就只是將所有文本作word2vec取出詞向量取平均再對計算query的距離，之後可能會使用rnn對她所給的訓練資料進行訓練，以及多模型ensemble。