ML final project proposal

1. 隊名: NTU_b06901087_87
   隊員: B06901087 翁瑋襄
   B06901063 黃士豪
   B06901020 張恆瑞
   B07901069 劉奇聖

2. 題目: Intent Retrieval from Online News

3. problem study:
   TF-IDF原理:
   (1)TF----計算特定的詞語在文章中出現的頻率，公式如下:

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

   (2)IDF----計算特定的詞語出現在所有文章中所佔的篇數，其中出現的機率越低表示該詞語對文章語意有重大影響，公式如下:

$$\mathrm{idf_i} = \lg \frac{|D|}{|\{j : t_i \in d_j\}|}$$

   最後TF-IDF由上面兩項指標相乘，代表該詞語在文章中的重要性。公式如下:

$$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \times \mathrm{idf_i}$$

   scikit-learn的TF-IDF套件:
   https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

4. proposed method:
   目前的方法:
   將新聞內容讀進來後，用scikit-learn TF-IDF分析文章的關鍵字出現頻率，然後將要比較關聯性的標題作詞語分解，並分別搜尋每一個詞在文章裡面出現的頻率以判斷兩者關聯性。
   上傳結果: 0.1817609
   未來可能的改進方式:
   (1) 使用word2vec分析同義詞跟反義詞，能更準確抓住文章關鍵字並分析立場
   (2) 先將文章透過RNN+attention抓出文章重點再進行比較，避免垃圾資訊的干擾。
   (3) 使用word mover's distance來比較兩個句子的相關程度。