**HW2**

**310706043肇綺筠**

1. **文字探勘前處理方法**

用迴圈遍歷Text欄位裡面每個row裡的詞

* 1. decontracted函數清除停頓詞(re.sub)
  2. 移除含有html的tag (有用到beautiful soup)
  3. 移除標點符號
  4. 英文字母全轉lowercase + 用snowballstemmer做文字stemming

一張含有 文字 的圖片

自動產生的描述

1. **TF-IDF / Word2Vec比較**
   1. **TF-IDF參數設定與運行結果：**

一張含有 文字 的圖片

自動產生的描述

* 1. **Word2Vec參數設定與運行結果**

一張含有 文字 的圖片

自動產生的描述

一張含有 文字 的圖片

自動產生的描述

* 1. **ACC結果比較：TF-IDF效果 > Word2Vec**

在kaggle上分數表現是一樣的

一張含有 文字 的圖片

自動產生的描述

一張含有 文字 的圖片

自動產生的描述

1. **額外實驗（kaggle達到0.72UP的分數）**
   1. 把helpfulness%考慮進去，移掉一些幫助不大的評論，train的資料量是“全部”而非只有一萬筆

一張含有 文字 的圖片

自動產生的描述



* 1. 使用Logistic Regression + TF-IDF + K-fold

一張含有 文字 的圖片

自動產生的描述