Machine Learning Final Project

AI CUP 2019 - 新聞立場檢索技術獎金賽

TA:解正平

AI CUP 2019 Competition

- Link : https://goo.gl/xPEmRZ
- 註冊 -> 登入 -> 參賽 -> 組隊
  - 6/30 sign up deadline
- 2 Stage
  - 3/25 - 7/8 (testing 1)
  - 7/8 (testing 2)

名次	獎金
第一名	10 萬元
第二名	6 萬元
第三名	4 萬元
佳作 10 名	各 1 萬元

Task Introduction

參與本競賽之隊伍需開發一搜尋引擎，找出「與爭議性議題相關」且「符合特定立場」的新聞。應用「資訊檢索」及「機器學習」技術於檢索模型的訓練，期望所開發之搜尋引擎能有效找出相關新聞，並依照相關程度由高至低排列。

Competition Data - NC-1 (部分新聞語料庫) - Label

- 100,000 News_URL and download news by ourselves
- Only based on **News Title** and **News Text Article**

News_Index	News_URL
news_000001	http://www.chinatimes.com/newspapers/20150108001507-260107
news_000002	http://tw.sports.appledaily.com/daily/20110623/33479530/
...	...
news_100000	http://tw.news.appledaily.com/headline/daily/20160311/37103743/

Competition Data - QS-1 (測試查詢題目) - TestData

- 20 query topics
- Need to find the **Top 300** most relative News

Query_Index	Query
q_01	通姦在刑法上應該除罪化
q_02	應該取消機車強制待轉或二段式左轉
...	...
q_20	反對旺旺中時併購中嘉

Competition Data - TD (訓練標記語料) - TrainData

- Query / News / Relevance (0-3)
- Similar topic but different point of view => **Relevance = 0**

Query	News_Index	Relevance
贊成流浪動物零撲殺	news_000109	3
核四應該啟用	news_000156	1
...	...	...
遠雄大巨蛋工程應停工或拆除	news_000684	0
拒絕公投通過門檻下修	news_000091	2

Evaluation

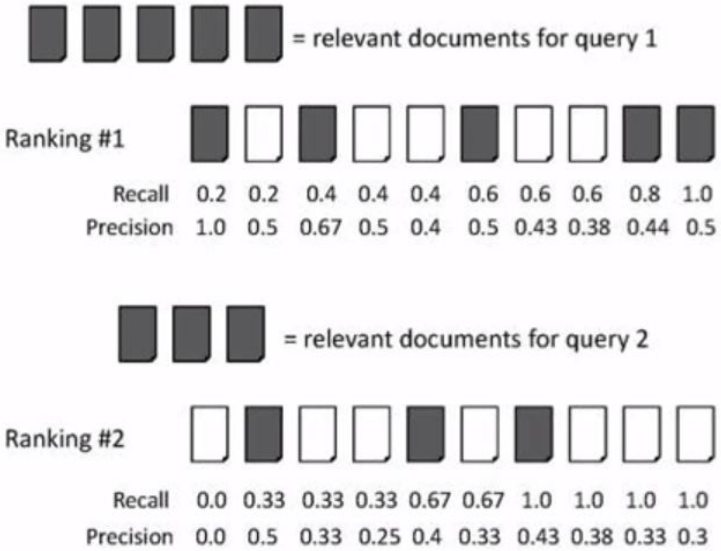
本競賽採用 $MAP@300$（Mean Average Precision at 300）指標來評估參賽隊伍之系統效能，並以此成績高低作為評估最後獎金賽名次之依據。$MAP@300$ 的值介於 0 到 1 之間，值愈高表示搜尋結果愈好，詳細計算方式定義如下：

$$MAP@300 = \frac{1}{|Q|}\sum_{q \in Q} AveP(q)@300$$

其中 Q 代表測試查詢題目的集合，$|Q|$ 是測試查詢題目的個數，而 q 表示某一個測試查詢題目；$AveP(q)$ 的計算則定義為：

$$AveP(q)@300 = \frac{1}{min(|R(q)|, 300)}\sum_{k=1}^{300}(P(k) \times rel(k))$$

Mean Average Precision: example



$$\textit{average precision query 1} = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$\textit{average precision query 2} = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$\textit{mean average precision} = (0.62 + 0.44)/2 = 0.53$$

Simple Baseline

Simple Baseline：0.1568653

9	成大westbrook	0.1716376	2019/04/23 22:57:15	8
10	ASTLM	0.1682468	2019/04/22 17:46:22	8
11	[×_×]	0.1662414	2019/04/30 19:29:48	16
12	StandorFall	0.1662414	2019/05/02 03:05:40	32
13	shan840930	0.1557588	2019/04/15 15:01:44	1
14	nphard001	0.1556762	2019/03/30 20:53:08	2
15	A1_105502502	0.1556762	2019/04/16 01:35:19	1

Hints - News Data

- News Data (only for education)
- Dictionary => {News URL : News}
- Some News may not exist
- Use multiprocessing to accelerate prediction

Hints - Sentence Representation

- Neural Network (RNN/BERT) (maybe need more training data)
- TF-IDF Bag of words
- Average word embedding (BERT/FastText/Word2vec/Glove)
- Average keyword-word embedding
- Weighted TF-IDF word embedding
- Google universal sentence encoder
- Gensim Doc2vec

Hints - Sentence Similarity

- Neural Network (Linear Layer)
- Sklearn classifier (XGBoost/SVM)
- Cosine Similarity
- Word mover’s distance (Gensim WmdSimilarity)
- BERT Next Sentence prediction