Web Retrieval and Mining 2020 - Final Project Proposal

Team Members

- 隊名： `賈裴根陳尹症候群`
- 隊員組成

學號	姓名
B06902001	陳義榮
B06902029	裴梧鈞
B06902039	賈本晧
B06902103	尹聖翔

Introduction to the Problem

我們想要探討 Deep Learning 的技術如何套用在 Information Retrieval 的領域上面，探討其表現，並討論其與傳統的 Vector Space Model 與 Probabilistic Model 等方法的差異、比較。

Methodology

傳統 IR 方法

- Vector Space Model：套用 Programming Assignment 1 的方法，嘗試一些 relevance feedback 的方式。
- Language Model：使用 unigram, bigram 等 language model 做 query-generation model。

Deep Learning 方法

- Doc2Vec：仍須 survey。
- BERT：嘗試 state-of-the-art 的 deep model。

Experiments

Dataset

使用 TREC 的 Deep Learning Track 使用的 MS-Macro dataset。我們將實驗在 **Document Full Ranking** 這個 task 上面。

Reference

1. https://trec.nist.gov/data/deep2019.html
2. https://microsoft.github.io/TREC-2019-Deep-Learning/

Evaluation

除了這個 track 使用的 MRR (Mean Reciprocal Rank) 之外，我們也將使用其他 evaluation metrics，如 MAP、nDCG。此外，我們也將標註 training / testing 時間，並分析其成效。

Schedule

週次	日期	事件
0	06/01 ~ 06/07	1. 討論、完成 Proposal。 2. 與教授討論題目。
1	06/08 ~ 06/14	1. 取得、分析並整理 Dataset。 2. Survey DL 方法並確定細節。
2	06/15 ~ 06/21	1. 嘗試傳統 IR 方法。 2. 訂定 baseline。
3	06/22 ~ 06/28	1. 嘗試 Deep Learning 方法。
4	06/29 ~ 06/30	1. 完成 Report。