# Machine Learning Final Project Proposal

## Selected Topic

News Retrieval – AI CUP 2019 Competition

## Team Information

NTU_B04 戰隊

B04901147 黃健祐 B04901166 陳培鳴

B04104040 解正安 R07943176 高禎謙

## Problem Study

■ TF-IDF 及 BM25

TF-IDF 是以 Term Frequency (TF) 及 Inverse Document Frequency (IDF) 評估 similarity 的方法，公式如下：

$$similarity = log\left(\frac{numDocs}{docFreq+1}\right) * \sqrt{tf} * \left(\frac{1}{\sqrt{length}}\right)$$

BM25 則是基於 TF-IDF 改良而來的演算法。傳統的 TF Score 理論上是可以無限大的，而在 BM25 中，TF Score 的計算公式引入常數$k$來限制其增長極限：

$$TF\ \ Score = \frac{(k+1)*tf}{k+tf}$$

BM25 的 TF Score 會被限制在$0$~$k+1$，這樣的作法相當合理：某一個關鍵字的影響強度不能是無限。

此外，BM25 還引入了平均文件長度的概念，單一文件長度對相關性的影響力與它和平均長度的比值有關。考慮這項因素後，TF Score 的公式可以被寫為：

$$TF\ \ Score = \frac{(k+1)*tf}{k*(1-b+b*L)+tf}$$

其中$L$是長度比值，$b$是常數，用來調節$L$的影響力。

綜合以上，完整的 BM25 計算 similarity 的公式如下：

$$similarity = IDF * \frac{(k+1)*tf}{k*\left(1-b+b*\left(\frac{|D|}{D_{avg}}\right)\right)+tf}$$

## Proposed Method

我們使用上文中提及的 BM25 來進行 relevance ranking，參數尚未進行完善的 fine-tuned，目前在 public leaderboard 上的成績為 0.2214147（通過

simple baseline )。之後預計將嘗試[4]中所提及的 deep learning 方法，使用 context-sensitive term encoding 以及 multiple views of terms 來提升準確率。

**Reference**


[1] Joachims, Thorsten. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. No. CMU-CS-96-118. Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[2] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." Proceedings of the first instructional conference on machine learning. Vol. 242. 2003.

[3] Robertson, Stephen, and Hugo Zaragoza. "The probabilistic relevance framework: BM25 and beyond." Foundations and Trends® in Information Retrieval 3.4 (2009): 333-389.

[4] McDonald, Ryan, Georgios-Ioannis Brokos, and Ion Androutsopoulos. "Deep relevance ranking using enhanced document-query interactions." arXiv preprint arXiv:1809.01682 (2018).