

DS4CS Final Project Proposal

資管三甲 曹瀨之 108306095

Topic: Chinese Spam Message Classification

Introduction

自從疫情開始後，許多人經常收到垃圾簡訊，之前的作業中做了 SMS spam classification，因此這次想延伸此課題實作中文的垃圾簡訊分類模型。

Methodology

• Dataset

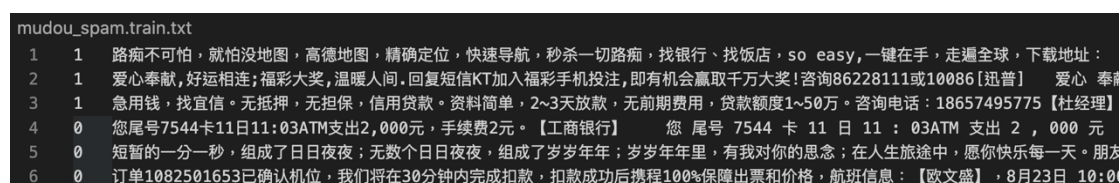
資料集使用一個叫做 mudou_spam 的中國垃圾簡訊資料集 (<https://github.com/shaonianruntu/Spam-Message-Classification/>)，訓練集共有 10872 筆資料，測試集共有 1382 筆資料，總計 12254 筆資料。

每一筆資料共有三個欄位：

垃圾簡訊標籤 (0 或 1)	簡訊內容	預先斷詞結果
----------------	------	--------

其中非垃圾簡訊 (標記為 0) 的資料約有 4630 筆，垃圾簡訊與非垃圾簡訊的比例約為 2:1。

圖 1: 資料集部分截圖



mudou_spam.train.txt		
1	1	路痴不可怕，就怕没地图，高德地图，精确定位，快速导航，秒杀一切路痴，找银行、找饭店，so easy，一键在手，走遍全球，下载地址：
2	1	爱心奉献，好运相连；福彩大奖，温暖人间。回复短信KT加入福彩手机投注，即有机会赢取千万大奖！咨询86228111或10086【迅普】 爱心 奉献
3	1	急用钱，找宜信。无抵押，无担保，信用贷款。资料简单，2~3天放款，无前期费用，贷款额度1~50万。咨询电话：18657495775【杜经理】
4	0	您尾号7544卡11日11:03ATM支出2,000元，手续费2元。【工商银行】 您 尾号 7544 卡 11 日 11 : 03ATM 支出 2 , 000 元
5	0	短暂的一分一秒，组成了日日夜夜；无数个日日夜夜，组成了岁岁年年；岁岁年年里，有我对你的思念；在人生旅途中，愿你快乐每一天。朋友
6	0	订单1082501653已确认机位，我们将在30分钟内完成扣款，扣款成功后携程100%保障出票和价格，航班信息：【欧文盛】，8月23日 10:00

• Preprocessing

1. 斷詞時除了本身資料集提供的斷詞，也有考慮嘗試中研院 CKIP，因此可能要將資料集轉為繁體字才能使用。
2. 因為資料集有些許不平衡，可以嘗試作 under sampling 排除部分垃圾簡訊，達成資料集的平衡。
3. 可以嘗試排除 stop words 看看訓練效果 (<https://github.com/goto456/stopwords>)

• Algorithms

因為問題本身屬於二元分類問題，預計使用五種分類問題常用的演算法進行比較：

1. Random Forest: 決策樹可以比較清楚看到節點跟判斷依據
2. Naive Bayes Classifier: 比較經典的分類問題演算法
3. RNN (neural network, supervised): 在文本分析上效果不錯
4. CNN (neural network, supervised): 參考此篇論文做法(A Study of the Chinese spam Classification with Doc2vec and CNN)，想進行嘗試
5. BERT (self-supervised): 較新的技術，比較自監督式學習與監督式學習模型的差異

Expected Result

預計使用 cross entropy loss 與 test accuracy 評估模型。BERT 可能會是效果最好的，其次可能是 RNN 與 Random Forest。但因為模型參數需要大量調整，也有可能看不出模型間的明顯差異，需要多多嘗試不同組合測試。