Started at : June 25, 2019

目標

1. 構思 & 規劃 2019/06/23 - 2019/06/30
2. 技術開發 2019/06 - 08
3. 文章撰寫 2019/07 - 09
4. 文章修改 2019/08 - 09
5. 比賽 2019/09/02 - 2019/10/01

→ 同步放至 medium

Topic: PTT 文章分類

<前期可能要做的事>

1. ~~下載資料 code~~
    - ~~Company list~~
    - ~~Value~~
    - ~~PTT articles~~
2. ~~良好訓練集 (正確的資料)~~
    - ~~人工標記 GUI tools~~
3. Dataset
    - 將中文編碼 (FastText) ?
4. Wiki 各分類法評估

---

(1) = Introduction = x 3

1.1 說明專案來源 (為何想做?)

1.2 股票基本名詞

1.3 說明好處 → 統計結果/大家的說法, 證明初步理念以經過驗證

~~**Day 01: 工人智慧**~~

~~**Day 02: 入手前，你需要先了解**~~

~~**Day 03: 高手真的在名間 →** 說明 PTT 各 id 發表的狀況, 讓人相信文章內有寶藏~~

(2) = develop = x 1

2.1 開發環境 etc. 說明

- 程式語言

- 開發環境 (Eclipse, Mac)
- 預計開發哪些項目？→ 列點

~~**Day 04: Let's Go →** 使用開發環境 & 程式語言~~

(3) = Data Source =

A. 下載PPT文章 (自動 vs. 半自動) x 2
B. 下載公司名稱 代號 x 1
    a. Statistical x 2
C. 擷取股價數值紀錄 (特定時間區間) x 1

~~**Day 05 : 先來當個爬蟲 ; 文章資料來源 1 →** ptt 文章擷取 (怎麼下載)~~
~~**Day 06 : 爬蟲回家後 ; 當文章資料來源 2 →** ptt 文章處理 (抓什麼資料)~~
~~**Day 07 : 數據資料來源大部分解 (上集) →** 人 : 人名 & 推文數~~
~~**Day 08 : 數據資料來源大部分解 (下集) →** 文章 : 發文量 & 被推文數~~
~~**Day 09 : 今天漲停還是跌停 →** 大盤指數收集~~

(4) = Tagging = x 5

A. 開發標記工具 x 2
    a. 介面說明
    b. 存檔規則
B. 人工紀錄規則說明 x 2
    a. 文章標記 (大方向)
        i. 遇到哪些文章需要標記？<文章類型>
        ii. 哪些不需要標記? (哪些需要濾掉)
            1. 奇怪的標題
    b. 文章標記 (內容問題：公司名 & 代號 tagging)
        i. 遇到重複公司資料怎麼辦？
        ii. 沒有資料怎麼辦？
            1. 沒有公司資料 (名稱\代號)
            2. 沒有內文 or 內文很弱
C. 好壞判斷規則, (搭配 股價數值紀錄)
    說明: 標記後的資料不是只為了本次的目的, 做一個公開的資料集讓大家測試
    同: http://thuctc.thunlp.org/

~~**Day 10: 真正的工人智慧上線 .. 前** (資料需要標記) → 說明資料為何要標記~~
~~**Day 11: 打造自己的小天地 →** tagging 畫面 ; 怎麼開發？Swing, 元件~~
~~**Day 12: 小天地裡的遊戲規則** (A.a)(A.b) ; Swing, 畫面呈現 (元件位置)~~
~~**Day 13: 小天地裡的遊戲 (上集)** (B.a)~~
~~**Day 14: 小世界裡的遊戲 (下集)** (B.b)~~
~~**Day 15: 小世界裡的好人與壞人** (4.C)~~

(5) = DataSet =

A. 建構完整的 Dataset
    a. arff 格式說明 (說名採用 weka 的 arff 格式, 來源, 額外解釋多少人使用相關資料)
B. 中文字 轉 數值
    a. Fasttext (解釋 fasttext 優缺點)
    b. Article to arff by fasttext
        i. Text to vector by Fasttext (非本標記文)
        ii. Text to vector by Fasttext (本標記文)

~~**Day 16: 電腦也要懂妳** (5.A.a)~~

為何用 wordembedding?

1) 如果使用 One-Hot 的方式 (一個字詞一個維度), 維度會隨著字詞量跟著增大, 導致訓練困難, 且無法表達字詞間的關聯
2) 克服稀疏矩陣的問題
3) arff 誰提出? 多少人使用? 公用資料及有哪些?

~~**Day 17: 文字與數字之間的戀愛** (5.B.a) → fasttext~~

說明 fasttext 使用方式

- 下載現有model
- 建構
- 用一個測試 當範例 (範例 code)

~~**Day 18: 要餵對食物** (5.B.b.i)~~

~~1) 先解釋用了哪些 stop word, 或是過濾掉哪些詞 (網址, 數字, 英文?)~~
~~2) 之後有哪些字詞不需要 (須先濾掉) ?~~
~~3) 要先進行繁體轉簡體, 才能使用 stanford 的 Chinese Segmentation <給範例>~~
4) 如何斷詞? 用stanford or CKIP ? <快速評估 or ?>

~~**Day 19: 文字 vs. 數字** (5.B.b.ii)~~

1) Code 說明, 如何操作 fasttext ?
2) 各別單詞, vector
3) 經由 fasttext 會產生出什麼樣子的東西 ? <以 PTT文章為例>
4) 累計 & 平均

(6) =Classification =

A. 各分類器說明 & 評估 (Wiki)
    a. Wika 簡單說明
    b. 評估方式
B. 評估
    a. 標準版
    b. 人工預設 (rules?)
C. 效能改進
    a. 提出改進的方式
        i. Rules 設定 (資料標記)
    b. 比較改進前與改進後
        i. 不同 articles 數量
        ii. 不同 algorithms
        iii. 不同的時間範圍標記

注意(可寫入): 在THUCTC中选取二字串bigram作为特征单元，特征降维方法为Chi-square，权重计算方法为tfidf
http://thuctc.thunlp.org/

~~**Day 20：分類器 與 分類氣** (6.A)~~: 目前常用的分類器
解釋目前監督式學習常使用的分類器，並比較效能 (reference 別人 paper)

~~**Day 21：分類氣 (一)**~~ (6.A.a) : weka 上的分類器
將 Day 20 的分類器對應到 Weka 上的使用

~~**Day 22：分類氣 (二)**~~ (6.A.b) : 評估方式
- 各種 evaluation 方式
- 切割資料方式

參考
https://pdfs.semanticscholar.org/6174/3124c2a4b4e550731ac39508c7d18e520979.pdf

~~**Day 23：你好還是我好？**~~ (6.B.a)(6.B.b)
- 1) Taaging 記錄轉 數值 (arff)
- 2) 跑 自己的資料 分類測試
- ~~我們自己如何評估？(股市版)~~

~~**Day 24：我錯了, 怎麼改？**~~ (6.C) → 說明
- 簡單說明 Day 25, Day 26, Day 27, Day 28 會採用的方式
    - a.) Day 25: 抓取資料問題；
    - b.) Day 26: 分類演算法

- c.) Day 27: 資料面 (word to value)
- d.) Day 28: 資料降維 (feature selection)

-

~~Day 25：第一招 改進資訊辨識成果~~

- <列出遇到的問題點>
- A. Data
    1) 擷取資料 (regular expression) ; 辨識資料
        a) 聯發, 聯發科 ; 南亞, 南亞科
        b) 已 title 為主, 但有些人不寫清楚 (只有代號)
        c) 不同寫法 : 不寫全名, 暱稱, and 簡稱
    2) 公司多維(多個公司)問題
    3) 過濾資料 (依據 title 做 filter)
    4) 股價變動過高, 忽高忽低
- B. manual tagging
    1) 依據 一、二、三 個月平均值做好壞判斷
- C. context 不平均,
    1) 有些太少, 有些太多　<沒寫入, 懶惰了>
    2) 只貼圖 (只有圖)　<沒寫入, 懶惰了>
    3) PTT 擷取資料時本身問題 (空白)

-

→ 以上問題都是邊標記編修正
＊＊ 列出改善後的結果

~~Day 26：第二招 分類器最佳化~~, classification algorithms

- 用不同 classification algorithms <NB, KNN etc.>
- 調整分類器參數
  https://blog.csdn.net/qiao1245/article/details/51005797

＊＊ 列出改善後的結果

~~Day 27：第三招 資料面改善~~, data

- 改變 wordembedding model (自己 train), 因為原本用 facebook fasttext
    - https://ithelp.ithome.com.tw/articles/10201537
    - https://github.com/zake7749/word2vec-tutorial/blob/master/README.md
- 改變 wordembedding model, (中變英? or others)
- wordembedding 改成 Glove ?
    - 因為是 Python 版, 這以後再執行
    - 偏向 deep learning 方式, 這邊先不考量?

- TF-IDF ?
- articles 只用單一類別的資料
- Training & Testing 數量 ?　　<不寫, 怕就沒資料了>
    - 資料量太少 ?
- N fold-validation ?
- 正規化 ? (Normalization) ?

\+ 資料正規化後加上 Parameter optimization

＊＊ 列出改善後的結果

~~**Day 28：第四招 資料降維, feature selection ?**~~

- 加入 feature selection 嘗試改進
    - Information Gain
    - Optimization (algorithms)

Fasttext
Word2vec

參考這篇 paper (放入本文)
https://arxiv.org/ftp/arxiv/papers/1612/1612.08669.pdf

＊＊ 列出改善後的結果

= Conclusion =　　x 1

A. 總結
    a. 下一步：使用新的 algorithm (Seq2seq)
    b. 三個分類演算法沒有做 參數最佳化 (之後 Auto-Weka 才有)

B. 說明此 project 目的
    a. 收集資料, 標記資料 (所以花了五天在講怎麼標記資料)

C. Reference
    https://www.itread01.com/content/1546158995.html

~~**Day 29：發大財了沒 ?**~~
<股股雞 發大財 sticker>
總結 Da5+y 25 - Day 28 ; 四種改善的方, 比較結果

一定要提到這篇

convolutional neural network for sentence classification
https://www.aclweb.org/anthology/D14-1181
為何不用 CNN 做分類？
https://github.com/gaussic/text-classification-cnn-rnn
之後會繼續進行

還可以做：參數最佳化 (但資料會一直進來, 先訓練好不一定是好事)

未來：

1) 釋放出標記的訓練資料集
2) 接下來用 python 撰寫
3)

> 是否不用用數學來解釋?
人工就好 (跟著操作)：

- 抄底王
- 經濟日報 比賽

Day 30：後會有期

<個人心得>

- 準備的時間, 從何時到何時
- 照片. git 紀錄 etc.
- 將 Google doc. 的紀錄輸出成 pdf, 放到本文
-

寫文章前先看:

該寫入資訊
https://www.twse.com.tw/zh/page/products/stock-code2.html

Text classification:
https://monkeylearn.com/text-classification/

上市公司名單 https://isin.twse.com.tw/isin/C_public.jsp?strMode=2
上櫃公司名單

Develop
= FastText =
安裝與使用
https://blog.csdn.net/HappyRocking/article/details/80668012

同比賽類似的文章
https://ithelp.ithome.com.tw/articles/10204200

其他人做過類似的事情
http://stock.kuankuan.nctu.me