1. 目前資源

Paper

Site

1. 研究方向
   1. 找出夠具有代表性的新聞  
      學校專題只有用這個部分，下面將會解說已經操作的步驟
   2. 新聞分群:主題分類
   3. 影響漲跌的字詞
   4. 影響漲跌的文章
2. 找出夠具有代表性的新聞-研究步驟
   1. 廣集資料  
      心得:本階段需要資料集夠廣，夠多，不能只有只單一或一兩個消息來源……
      1. Oilprice
      2. Google
      3. Cnbc
      4. Economic
   2. 初步探看  
      心得:本階段千萬不能省略，免得後期猶如大海撈針，專題這方面幾乎因為時間略過很不好
      1. 操作論文所述方法
      2. 繪製圖表
      3. 自動加入更新文章
      4. 觀察價格走勢，均線圖等等
   3. 標註(tagging)  
      心得:本來使用絕對波動，即新聞出來之後，rolling std大於一定值就標註為1，但看論文都會跟大盤做比較，石油期貨沒有大盤，我們就使用與過去五天的波動性比較的想法
      1. 新聞推出後絕對波動性(與所有資料的波動性std比較)
      2. 新聞推出後相對波動性(與過去五天的波動性std比較)
      3. 單方向的價格變化(漲或是跌percentage change)
      4. 加入sentiment的標註(效果不好)
   4. 建立字典  
      心得:目前是第四種方法成效最好，但有可能是因為資料集太小，導致overfit，但也可能是因為主題(石油期貨)相關的詞本來就少，這個方法正好適用。
      1. 基本詞頻
      2. Unigram、bigram
      3. 金融相關字典(網路資源)
      4. 預訓練字典(Glove)
      5. 訓練集中目標文章(被標註為1者)與其他文章(被標註為0者)字典的差集
   5. 向量空間(embedding)

心得:本來期待3.4.3 Predict base會有最好的結果，畢竟可以處裡字詞間順序性，但結果沒有比較好

* + 1. One-hot(Tf or TfiDf)
    2. BOW(Tf or TfiDf)
    3. Predict base產生向量空間(Word2Vec、Doc2Vec)，using keras or gensim
  1. 模型

心得:NB意外的高分，反而SVM沒有那麼強大，恐怕是因為字詞的向量空間本來維度就相當可觀，能夠投射到高維度的能力在這裏反而沒有那麼強大。之後希望可以用LSTM等RNN模型嘗試

* + 1. SVM
    2. NB
    3. Randomforest

1. 預處理操作事項與方法
   1. Stemming  
      nltk
   2. Summarize  
      gensim
   3. Stopword  
      nltk
   4. Sentiment

gensim

1. 結果與未來方向

如今可以建構肉眼看起來挺有意義的 bigram 字典，但預測出來的結果分布不是很好，選出來的文章，很難打中test data，但可以看出都集中在test data標註的日期附近，如果用交易日當作判斷得分基準的話，可以得到比較好的結果，但還不到可靠的等級，可以理解為一天當中真正影響價格的文章並不多，但也許只要一兩篇就會影響很大，而且我們的資料集不夠大，學習的對象有限，以下幾點構思:

* 1. 更多元的資料集，不能只使用單一來源，但難處在，越大的資料集裡面垃圾也越多，需要”乾淨的”資料集難度很大
  2. 標註的方法不能過於粗糙，但人工標記也不是可行解…
  3. 可能可以縮小範圍，限制與某主題相關(新聞進來之前先用分群)，例如中美貿易戰，的新聞，再給予預測。
  4. 逐漸建立原油相關字典，慢慢完善