# 数据下载

从tweet上搜索包含“trade war”或#tradewar的推文，时间范围20180320-20191015。得到100W条。从中过滤含有china的，还有597709 （all\_tweets\_china.jl）条。

# 数据清洗

都小写；

去掉http,pic等超链接；

去掉hashtag的#和@someone的@

去掉所有cashtag, eg: $WTC

使用nltk.stem 对单词做词形还原lemmatizer from nltk.stem import WordNetLemmatizer

过滤掉为空的，还剩597333 （all\_tweets\_cleaned.jl）

去重，去掉不足7个字的，还剩396560.

# LDA

使用gensim包的LDA模型主体提取，模型训练时去掉了 trade, war, tradewar, china, us这几个几乎都会出现的词，以及常见停用词。模型输出结果如下

[(0,

'0.017\*"economy" + 0.015\*"say" + 0.013\*"state" + 0.013\*"economic" + '

'0.013\*"world" + 0.012\*"trump" + 0.012\*"global" + 0.009\*"president" + '

'0.009\*"united" + 0.008\*"two" + 0.008\*"chinese" + 0.008\*"election" + '

'0.008\*"talk" + 0.007\*"agreement" + 0.007\*"recession"'),

(1,

'0.036\*"market" + 0.017\*"stock" + 0.012\*"amid" + 0.012\*"year" + '

'0.011\*"global" + 0.010\*"kong" + 0.010\*"hong" + 0.008\*"price" + '

'0.008\*"growth" + 0.008\*"point" + 0.007\*"week" + 0.007\*"investor" + '

'0.006\*"rate" + 0.006\*"gold" + 0.006\*"news"'),

(2,

'0.055\*"tariff" + 0.038\*"trump" + 0.017\*"chinese" + 0.016\*"billion" + '

'0.014\*"new" + 0.013\*"$" + 0.011\*"good" + 0.010\*"export" + 0.008\*"company" + '

'0.008\*"import" + 0.007\*"business" + 0.007\*"news" + 0.007\*"president" + '

'0.006\*"ivanka" + 0.006\*"economy"'),

(3,

'0.033\*"trump" + 0.017\*"farmer" + 0.007\*"like" + 0.006\*"one" + 0.006\*"get" + '

'0.006\*"win" + 0.006\*"going" + 0.006\*"country" + 0.006\*"people" + '

'0.006\*"don't" + 0.005\*"think" + 0.005\*"started" + 0.005\*"job" + '

'0.005\*"economy" + 0.005\*"know"'),

(4,

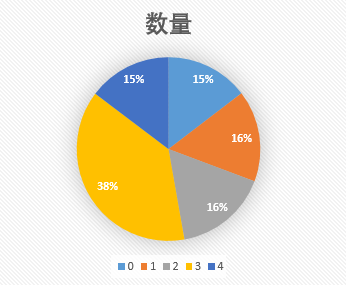
'0.030\*"deal" + 0.026\*"trump" + 0.007\*"oil" + 0.006\*"talk" + 0.006\*"end" + '

'0.006\*"want" + 0.005\*"iran" + 0.005\*"losing" + 0.005\*"ongoing" + '

'0.005\*"get" + 0.005\*"2020" + 0.004\*"usa" + 0.004\*"investigate" + '

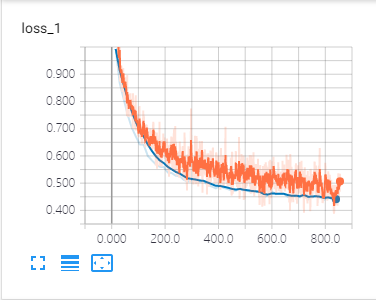
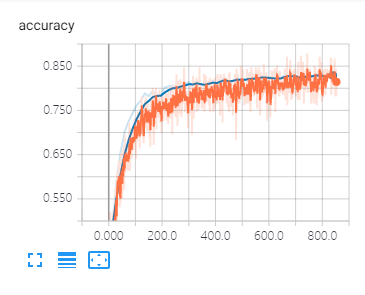
'0.004\*"paying" + 0.004\*"done"')]

|  |  |  |
| --- | --- | --- |
| 类别 | 解释 | 关键词 |
| 0 | 世界经济/中美对世界经济的影响 | economy, economic,recession, trump |
| 1 | 金融以及股票市场 | market,stock,hong kong, growth, investor,gold |
| 2 | 关税以及进出口 | tariff,billion, new(s), good, export, import,business |
| 3 | 农业，工作，美国内部 | trump,farmer,win,people,job |
| 4 | 关心，希望和解 | deal,oil, talk,end,losing |

主题对应tweet数量统计

|  |  |
| --- | --- |
| 类别 | 数量 |
| 0 | 57912 |
| 1 | 63875 |
| 2 | 65492 |
| 3 | 150970 |
| 4 | 58311 |

然后我们标注了4W多条推文，之后使用近几年提出的一个文本分类模型TextCNN做情感分类。下两个图是运行过程，大概准确率83%，还可以。



得到模型后，把模型在在标注的数据上预测，得到所有数据的情感。

分类统计如下

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 类别 | 总量 | CENTRAL | NEGATIVE | POSITIVE | 中性占比 | 负向占比 | 正向占比 |
| 0 | 57912 | 33961 | 17846 | 6105 | 58.64% | 30.82% | 10.54% |
| 1 | 63875 | 37068 | 21358 | 5449 | 58.03% | 33.44% | 8.53% |
| 2 | 65492 | 44084 | 14687 | 6721 | 67.31% | 22.43% | 10.26% |
| 3 | 150970 | 75751 | 51594 | 23625 | 50.18% | 34.18% | 15.65% |
| 4 | 58311 | 33263 | 17238 | 7810 | 57.04% | 29.56% | 13.39% |

大概可以看出，大部分推文都是陈述性的中性。具有情感态度的，几个类别都是负向多于正向。但是正向占比可以反应出信心的多少，类别1股市最低，类别3农业或者内生工作最高。

然后就是挑几条典型的推文，说一下。