安倍首相のツイートと日経平均株価の関係

- テーマ背景
  　某日夜中に何となくTwitterAPIの申請をしたところ通ってしまい、折角なのでこちらを利用して課題をやってみようと思いました。

- 補足
  日経平均株価とは … **日経平均株価**は、日本の株式市場の代表的な株価指標の一つ。単に**日経平均**や**日経225**とも呼ばれる。東証第一部上場銘柄のうち取引が活発で流動性の高い225銘柄を選定し算出する。(by Wikipedia)

- 用いたツールなど
  - Python
  - TwitterAPI
  - 日経平均株価の時系列データ
  - 安倍首相のTwitter(https://twitter.com/AbeShinzo)

- データの詳細

category	ツイートの分類 プログラムで大体の単語をツイート内容を分類した。 0 … その他、1 … RT、2 … 災害関係、3 … 政治関係 (RTはデータを切る時になくなってしまいました)
fav	ツイートがお気に入りされた数
retweet	リツイート数
0m	その時点での日経平均株価。値自体ははそこまで重要ではない。
3m	ツイートされてから３分後の日経平均株価の差分。１分間の高値と安値を平均した値を利用した。 e.g. (3分後の日経平均株価) - (ツイート時点での日経平均株価)
5m	3分後の説明を5分後に変える

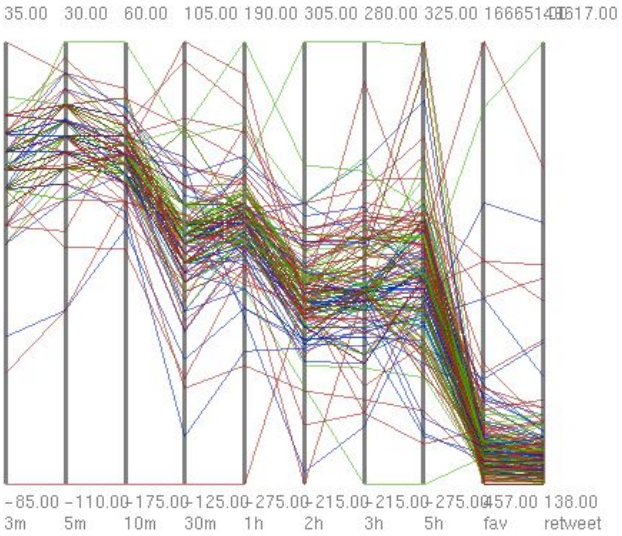
10m	3分後 -> 10分後
30m	3分後 -> 30分後
1h	3分後 -> 1時間後
2h	3分後 -> 2時間後
3h	3分後 -> 3時間後
5h	3分後 -> 5時間後

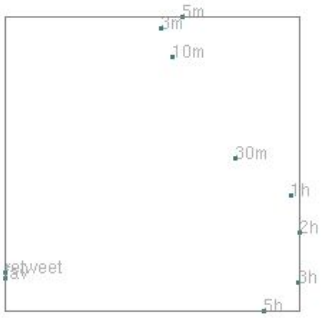
■ データ収集手法

Pythonを使って集めました。ソースコードは割愛しますが手法としては,

- TwitterAPIを用いて安倍首相のツイートを400件ほど取得します
    - 取引外の時間や時間的に近いデータを切るため欲しいデータ量よりも多く用意しました。
- そのままではデータとして使えないため、ツイート内容を整理して使える形に加工し、csvに保存します。
    - TwitterAPIにはリクエストの回数制限があり、Twitter社に迷惑をかけないようにする意味も込めて一度csvにします。
- 安倍首相のツイートが入ったcsvを読み出し、ツイート時刻から日経平均株価の時系列データを公開しているサイトのダウンロードできるリンクをプログラムでひらすら叩きます。
    - データの公開先のサイトは「地獄へ行く方法」とか公開していてちょっと怪しいです。
- ダウンロードしたcsvのデータから安倍首相のツイート時刻を見つけて、その10分後、20分後....のデータを集めていきます
    - ここでX時間後が存在しないデータについては捨てていきました。
    - 例えば首相がツイートした時刻が土曜、日曜だった場合はデータが存在しないのでそのツイートに関しては捨てます。
- 集めたデータと安倍首相のcsvから欲しいデータを抜き出してcsvを作成し、完成です。
    - ここまでのプログラム書くのに丸々２日くらいかかりました(涙)

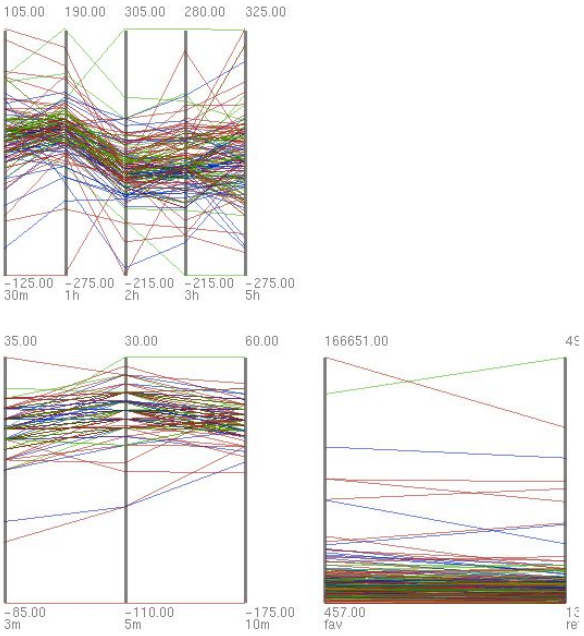
■ 可視化結果

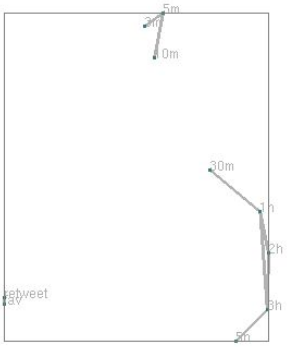




こちらはツイートのカテゴリー別に分類したPCP図です。
ざっと見る感じだと、10分と30分, 30分と1時間で相関関係が少し変わることがわかります。

ある程度の相関係数で繋げてみるとこのようになりました。

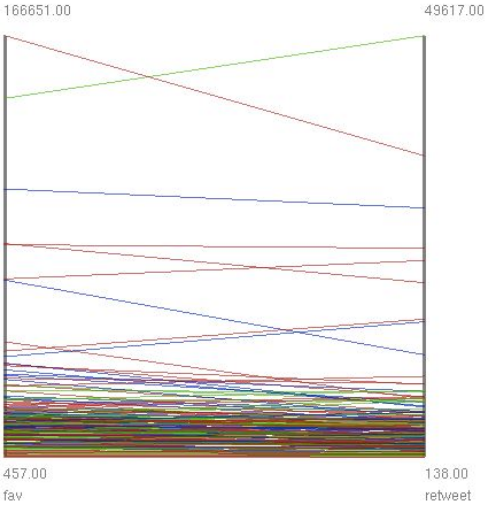




次のページから詳しく見ていきたいと思います。

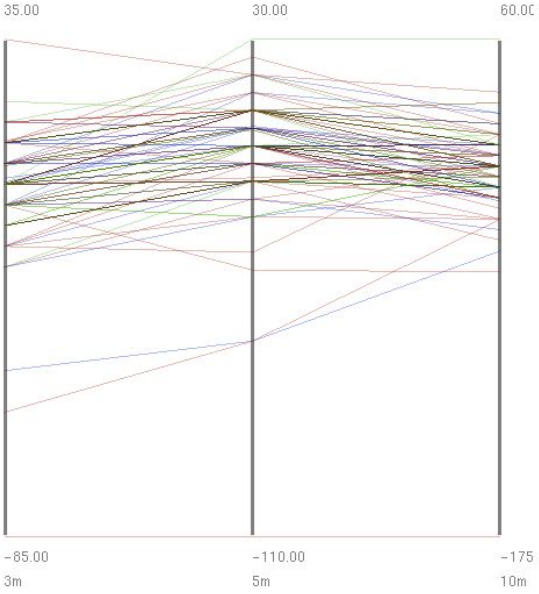
■ 各項目について

各項目について詳しく見ていきたいと思います。
まずはリツイートといいねの関係です。

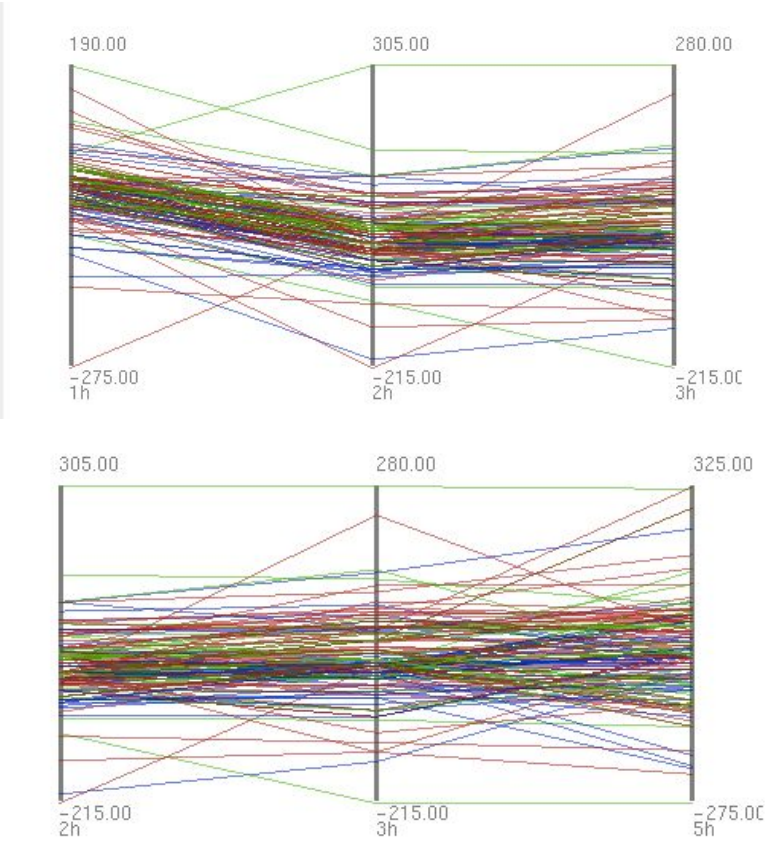


リツイートといいねは綺麗な正の相関をしています。
リツイートが多いほど他人からツイートを見られる確率が上がるため、いいねされる回数が増加する傾向にあります。これは予想と一致していました。

次にツイートして3分後～10分後を見てみます。
先ほどの図ほど綺麗な相関関係はないですが、正の相関関係があることがわかります。
外れ値によって図が歪んでしまってることが原因であまり綺麗に見えないのだと思います。実際に相関係数を測ったところ3分~5分は0.86、5分~10分は0.80ありました。



1時間-5時間部分です。
ツールの関係で図を分離しています。



1時間～2時間は正の相関が少し薄いですが、2時間～3時間は相関がありそうです。3時間～5時間は結構バラバラな感じです。
相関係数を測ったところ1時間の部分から順に0.56, 0.75, 0.60で確かに図の見た目と一致していました。

正直このあたりの相関がどうして変わるのかはわかりません。
ツイートしてから2時間、3時間後は落ち着くのですかね。

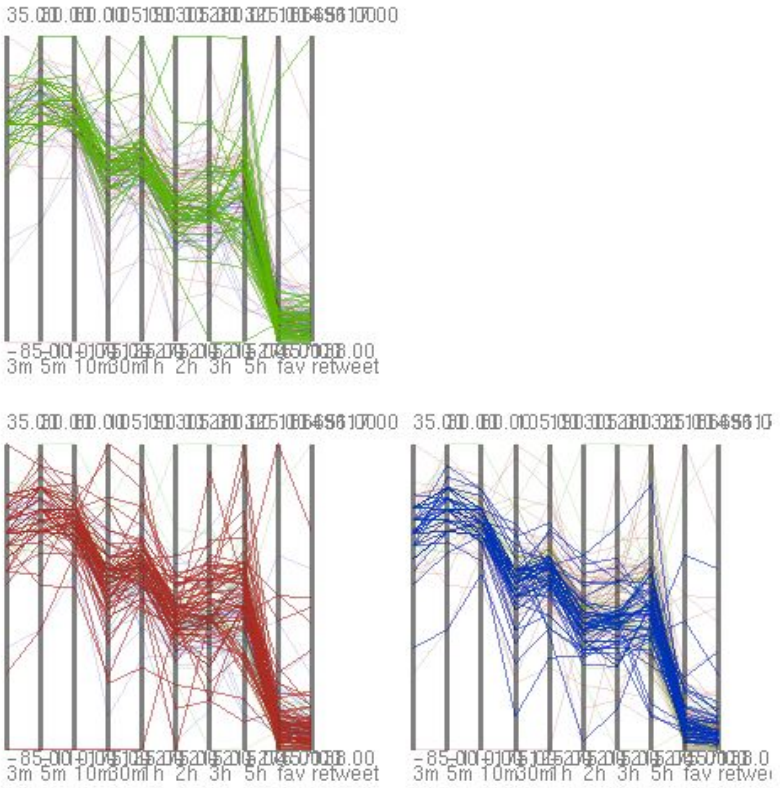
全体の相関関係はPythonで調べてみるとこのようになりました。

```
# 全ての要素に対しての相関関係
[[ 1.        0.85636582 0.66891765 0.3652304  0.42563506 0.10650342, -0.1291547  -0.03228344] # 3
 [ 0.85636582 1.        0.8072089  0.44784758 0.45463658 0.22447879, -0.08450737 -0.01801085] # 5
 [ 0.66891765 0.8072089  1.        0.58779264 0.58661144 0.24064274, -0.07843891  0.00127132] # 10
 [ 0.3652304  0.44784758 0.58779264 1.        0.72992073 0.36186143, 0.20736579  0.22719209] # 30
 [ 0.42563506 0.45463658 0.58661144 0.72992073 1.        0.55767801, 0.38648586  0.25083564] # 1
 [ 0.10650342 0.22447879 0.24064274 0.36186143 0.55767801 1., 0.75162151  0.41955799] # 2
 [-0.1291547  -0.08450737 -0.07843891 0.20736579 0.38648586 0.75162151, 1.        0.59501865] # 3
 [-0.03228344 -0.01801085  0.00127132 0.22719209 0.25083564 0.41955799, 0.59501865 1.        ]] # 5
```

一番上の行をみてみると、安倍首相のツイートをみてから10分後くらいまでは割と相関がありそうです。30分後に相関が下がって1時間後に何故かまた相関が上がります。30分後と1時間後の相関が薄いのはここから少々納得できるような気もします。

■ カテゴリー別に分類

先ほどは全ての要素に対して相関を求めていたので、次はカテゴリー別に分類してみました。(赤... その他のツイート、緑...災害関係のツイート、青...政治関係のツイート)



正直、形が全部同じように見えるため全く相関関係がわかりません。
一つ一つ比較してみていくのも大変そうです。

そこで先ほど相関係数を出したように、Pythonのnumpy.corrcoef関数で相関係数を出してみます。

【各カテゴリー別の相関係数】

```
# カテゴリーがその他の相関係数
[[ 1.   0.90150832  0.77458329  0.43638665  0.58862995  0.09049361, -0.30024194 -0.09218312]
 [ 0.90150832  1.    0.84384225  0.46658968  0.5642579   0.23910035, -0.28822755 -0.06424171]
 [ 0.77458329  0.84384225  1.    0.61291829  0.68309191  0.20393882, -0.32053089 -0.06584698]
 [ 0.43638665  0.46658968  0.61291829  1.     0.81741674  0.37668339, 0.11119981  0.15783391]
 [ 0.58862995  0.5642579   0.68309191  0.81741674  1.     0.41154522, 0.15252136  0.23871872]
 [ 0.09049361  0.23910035  0.20393882  0.37668339  0.41154522  1., 0.55973975   0.31083886]
 [-0.30024194 -0.28822755 -0.32053089  0.11119981  0.15252136  0.55973975,  1.   0.60106657]
 [-0.09218312 -0.06424171 -0.06584698  0.15783391  0.23871872  0.31083886,  0.60106657  1.       ]]

# 災害関係
[[ 1.      0.67779394  0.2883002   0.19383088  0.00192     0.03783418, -0.01107374 -0.06301552]
 [ 0.67779394  1.    0.65492596  0.35836007  0.17603122  0.22778456, 0.18520882 -0.09122846]
 [ 0.2883002   0.65492596  1.      0.43280186  0.43654379  0.3806069, 0.35704424 -0.03684294]
 [ 0.19383088  0.35836007  0.43280186  1.    0.66882698  0.48126189, 0.38820794  0.18626841]
 [ 0.00192     0.17603122  0.43654379  0.66882698  1.      0.68903636, 0.63545758  0.24945167]
 [ 0.03783418  0.22778456  0.3806069   0.48126189  0.68903636  1., 0.89616006  0.48981867]
 [-0.01107374  0.18520882  0.35704424  0.38820794  0.63545758  0.89616006, 1.    0.60799757]
 [-0.06301552 -0.09122846 -0.03684294  0.18626841  0.24945167  0.48981867, 0.60799757  1.       ]]

# 政治関係
[[1.        0.81212303 0.5117608  0.25567963 0.27290183 0.22786498, 0.21786213 0.1714625 ]
 [0.81212303 1.        0.78806195 0.46532308 0.37447037 0.24274767, 0.20763994 0.21855952]
 [0.5117608  0.78806195 1.        0.67499777 0.44353321 0.26381073,  0.2354776  0.34597727]
 [0.25567963 0.46532308 0.67499777 1.        0.51673279 0.26157377, 0.24154204 0.43126881]
 [0.27290183 0.37447037 0.44353321 0.51673279 1.        0.72949116, 0.62367445 0.23341112]
 [0.22786498 0.24274767 0.26381073 0.26157377 0.72949116 1., 0.89636737 0.48231384]
 [0.21786213 0.20763994 0.2354776  0.24154204 0.62367445 0.89636737, 1.        0.50641811]
 [0.1714625  0.21855952 0.34597727 0.43126881 0.23341112 0.48231384,  0.50641811 1.       ]]
```

注目したいのは一番上の3分後の値とX分後の相関関係なのですが、その他や災害関係のツイートはX分後に対してあまり相関がみられないのに対して、政治関係のツイートはそれなりの相関が長時間に渡って出ていることがわかりました。また、政治関係や災害関係は10分後には相関関係が薄くなっているのに対し、その他のツイートでは1時間くらいまでは0.50以上を維持するなど、しばらく相関関係が高いことがわかります。

ただ、データ数が少なく偶然このような結果が出た可能性もあることを考慮して、ここからはデータ数を増やしてみることにしました。

ツイートを2000件取ってみたのですが日経株価平均の公開データセットが2018年は以前は使っているPCでサポートしていないデータフォーマットであったため、実際に取れた件数は315件(うち、その他:166件、災害:57件、政治:90件)になりました。(先ほどのデータは161件であるため倍くらいです。)

【データ数を倍にした場合の相関係数】

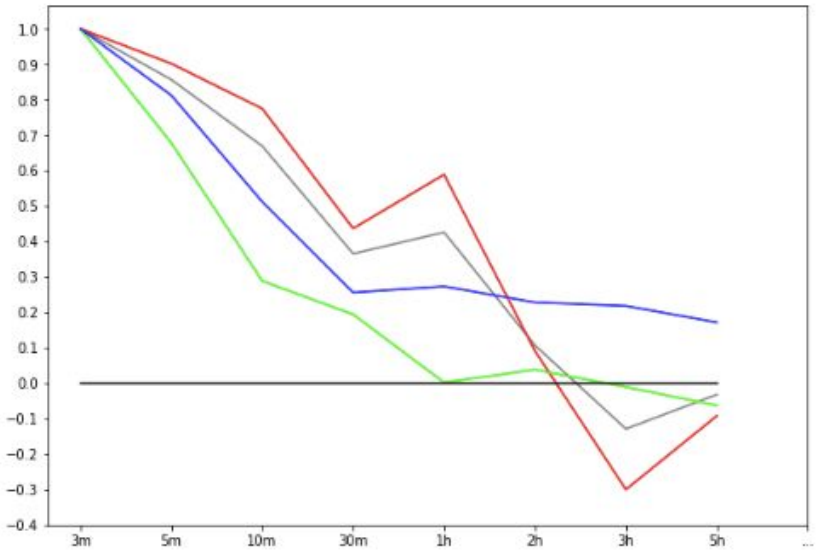
```
# カテゴリーがその他の相関係数
# データ数:166
[[ 1.         0.84179723  0.72152449  0.37243329  0.45867633  0.19295808,  -0.01734786  0.12352826]
 [ 0.84179723  1.         0.7833596   0.31519469  0.41058124  0.17868414,  -0.11116063  0.09602633]
 [ 0.72152449  0.7833596   1.         0.5125565   0.55172761  0.25633048, 0.03825426  0.09280736]
 [ 0.37243329  0.31519469  0.5125565   1.         0.76478411  0.43825686, 0.33439179  0.00457118]
 [ 0.45867633  0.41058124  0.55172761  0.76478411  1.         0.53482301,  0.37232023 -0.04031374]
 [ 0.19295808  0.17868414  0.25633048  0.43825686  0.53482301  1., 0.71513087 -0.10246357]
 [-0.01734786 -0.11116063  0.03825426  0.33439179  0.37232023  0.71513087, 1.         -0.1025474 ]
 [ 0.12352826  0.09602633  0.09280736  0.00457118 -0.04031374 -0.10246357,  -0.1025474   1.        ]]

# 災害関係
# データ数:57
[[ 1.         0.65833158  0.33748495  0.25762128 -0.01854626 -0.00683466,   -0.0706354  -0.03665038]
 [ 0.65833158  1.         0.67355632  0.26337132  0.06424402  0.07884045,  0.07187594 -0.06476173]
 [ 0.33748495  0.67355632  1.         0.43334098  0.31467844  0.32105503, 0.22691852 -0.01677659]
 [ 0.25762128  0.26337132  0.43334098  1.         0.69097784  0.52374225,  0.43517816  0.14752457]
 [-0.01854626  0.06424402  0.31467844  0.69097784  1.         0.71166287,  0.69000278  0.21955093]
 [-0.00683466  0.07884045  0.32105503  0.52374225  0.71166287  1.,  0.8589381   0.45817601]
 [-0.0706354   0.07187594  0.22691852  0.43517816  0.69000278  0.8589381,    1.         0.56383441]
 [-0.03665038 -0.06476173 -0.01677659  0.14752457  0.21955093  0.45817601,  0.56383441  1.        ]]

# 政治関係
# データ数:90
[[1.         0.83685555 0.59604192 0.44623809 0.38055563 0.37351109,   0.13315192 0.27525254]
 [0.83685555 1.         0.75240571 0.49037482 0.46705107 0.33518322, 0.18364528 0.25399241]
 [0.59604192 0.75240571 1.         0.70813877 0.61554394 0.41030328,  0.2208116  0.32757366]
 [0.44623809 0.49037482 0.70813877 1.         0.63047178 0.45315542,  0.36772033 0.43054591]
 [0.38055563 0.46705107 0.61554394 0.63047178 1.         0.77510961,  0.6345815  0.40310035]
 [0.37351109 0.33518322 0.41030328 0.45315542 0.77510961 1. ,  0.75030557 0.55828849]
 [0.13315192 0.18364528 0.2208116  0.36772033 0.6345815  0.75030557,  1.         0.52383684]
 [0.27525254 0.25399241 0.32757366 0.43054591 0.40310035 0.55828849, 0.52383684 1.        ]]
```

やはり先程の結果と少し変わりましたが、災害関係は相変わらず低い値を取っていたり、政治関係がしばらくずっと相関関係が高いことは変わらないようです。

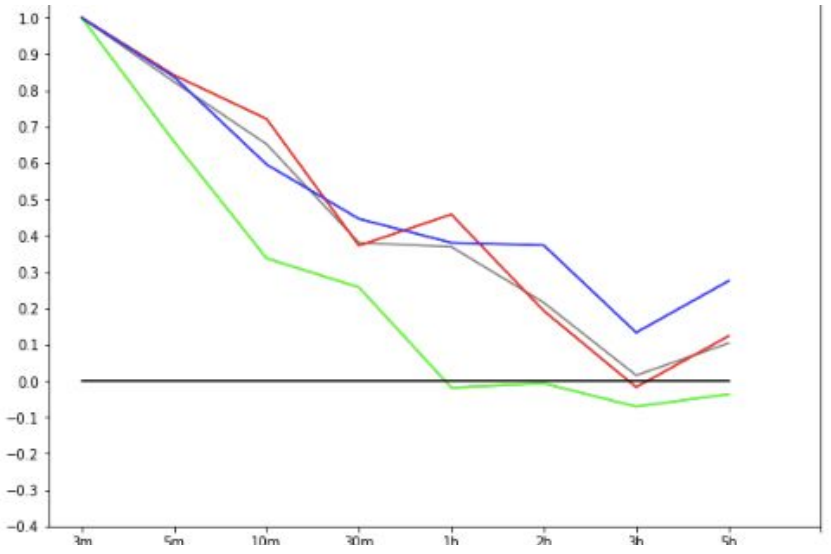
つまりは、このデータから言えるのは安倍首相が政治関係のツイートして日経平均株価の3分後の動きがプラスだった場合、1時間後、2時間後もプラスであることが多いということです。(相関が正の相関であるため)

数値だけだとわかりにくいため相関関係も可視化してみました。

【３分後の日経平均株価の変動とX分後の関係】(データ数:161)
灰色 … 全体データ、赤...その他、緑...災害関係、青...政治関係
黒...相関0ライン



【３分後の日経平均株価の変動とX分後の関係】(データ数:315)



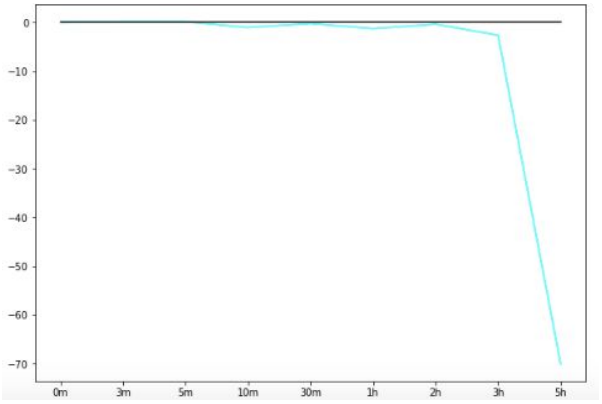
可視化して初めて気づきましたが、災害関係のツイートだけは他と比べてすぐに相関がなくなってしまうことがわかります。災害時の日経平均株価の変動は政治以外の要因が大きいのかもしれません。

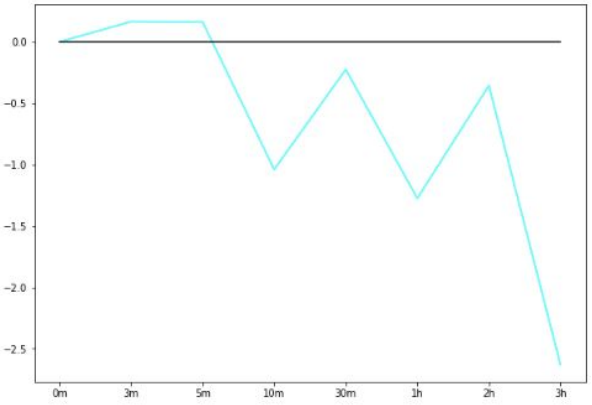
■　　実際の数値を軽くみてみる

PCPによる可視化は各X軸の要素に関する相関しか分からないため、実際の数値がどうのような値を取っているのかが個人的に少し気になります。

そこでPythonのmatplotlibを使って「各時刻に対する日経平均株価の増減分の平均」を見てみました。データ数は315件です。

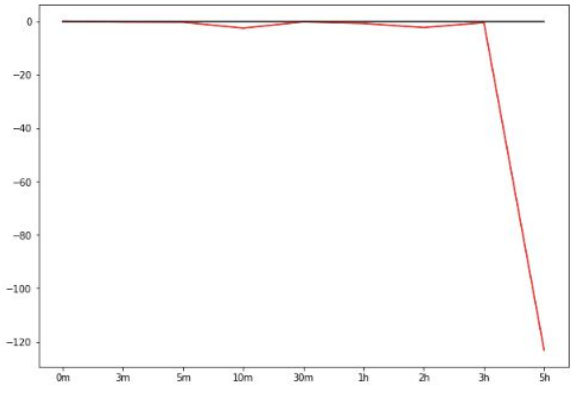
【全体：各時刻に対する日経平均株価の増減分の平均】

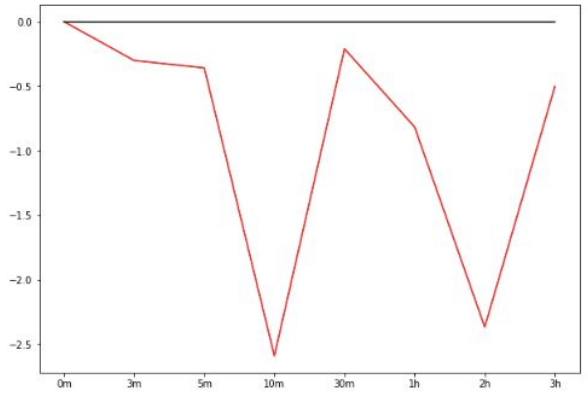




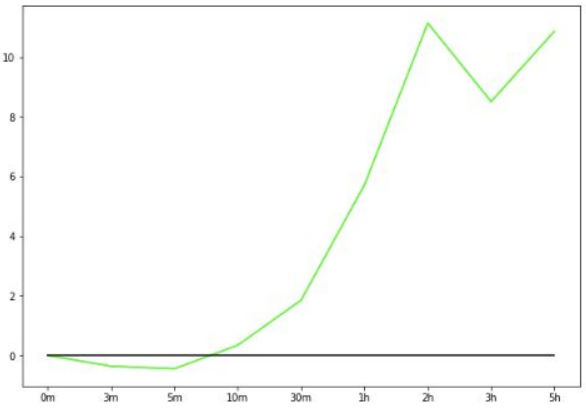
左が全体の分布です。右のグラフは一部の外れ値でグラフの詳細がわかりにくいため、5hを消去しました。
と言っても右のグラフのy軸をみるとデータの範囲が-2.5~0.5であるため、ほぼ0に近いと言っても良さそうです。

【カテゴリー別:各時刻に対する日経平均株価の増減分の平均】
赤...その他、緑...災害関係、青...政治関係、黒...0ライン

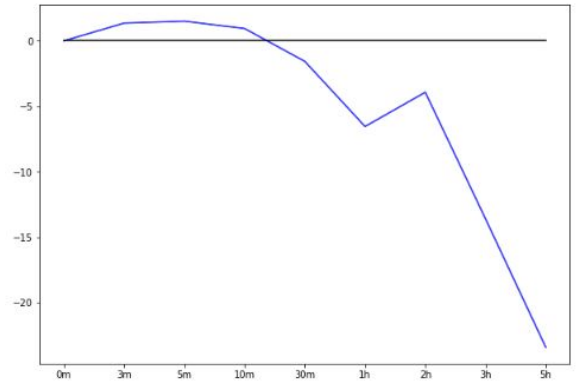




こちらも5hで平均-120あたりを取っていて見にくいため5hを消去したのが右のグラフです。y軸をみるとデータの範囲が-2.5~0.5であるためかなり小さいです。
全体の形とほぼ同じ形をしていることがわかります。



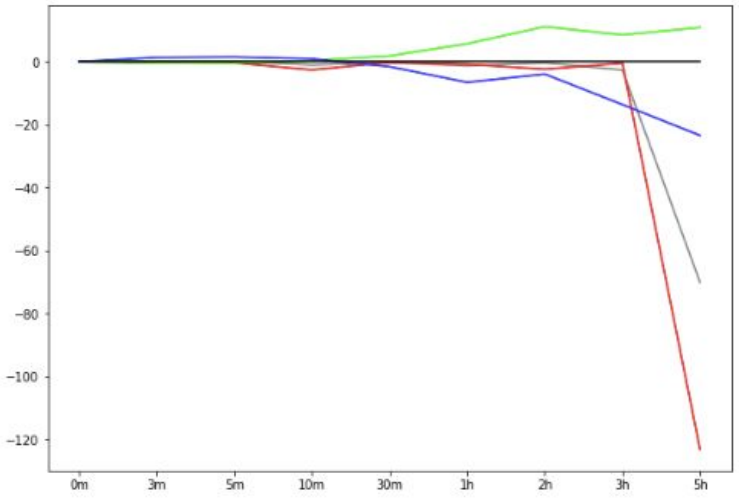
こちらは災害関係のツイートをした時の動きです。全体平均やその他ツイートの平均とは全く違う形をしています。確かに、災害ツイート時だけ相関係数の取り方が他とは違ったためこのような結果になったのかもしれません。



最後は政治系のツイートをした時の動きです。これもまた少し違って、全体やその他ツイートの図では0~10mがマイナスであったのに対し、こちらはプラスです。しかし5時間経つと急激にガクンと下がるのは全体やその他ツイートの傾向と同じです。

最後に全部の図をまとめてみました。

【各時刻に対する日経平均株価の増減分の平均】



全体の数値です。(一行に表示したい関係で桁数を一部減らしました)

```
# 全体平均
[0.0, 0.1645, 0.1613, -1.0399, -0.2252, -1.274, -0.3562, -2.6294, -70.0767]

# その他ツイート
[0.0, -0.3012, -0.3584, -2.5903, -0.2108, -0.8162, -2.3674, -0.5030,   -123.1656]

# 災害ツイート
[0.0, -0.3596, -0.4473, 0.34210, 1.85087, 5.7105, 11.1403, 8.5087, 10.8684]

# 政治ツイート
[0.0, 1.3556, 1.5056, 0.9444, -1.5667, -6.5444, -3.9276, -13.6056, -23.4222]
```

平均はあくまで平均であり、大きい数値などに持ってかれて数値が変わってしまうのは仕方がないと個人的には思います。

■ 発展として

時間の都合上行いませんでしたが他に調べたかったこととしては、

- RT・いいね数と日経平均株価の関係
- 相関関係をみたときに10分と30分で相関係数が急に下がったが、15分または20分ではどうなるか
- トランプ大統領とアメリカ市場の関係

■ 結論、感想など

相関関係や平均のグラフから、ツイートしてから10分以内は動きが小さいことがわかりました。ツイートしてから3時間後、5時間後は平均株価の動きが大きく変わりそうなので大損したくなければ10分以内に売れば良さそうです。3時間後、5時間後でも相関は薄いとは言いつつも平均値は大きくマイナスになることから安倍首相がツイートしたのを見て売りで動いたら勝てる可能性が高いかもしれないですね。また、今回の分析で災害時の平均株価は特殊な動きをすることが判明したため、その時期に着目して研究することも面白いなと思いました。

■ 終わりに

PCPで可視化は多次元のデータをぱっと見るのには向いていますが、もう少し詳しく見たい場合には別の手法を使うのが最適だなとデータを加工していて思い、欲しいデータの種類によって適切な手法を使うことが分析において重要だと思いました。
(ただこの課題はHiddenがメインであるのに、サブの図となってしまった感が否めなく申し訳なく思っています。)

それと、このデータは私個人で作成したものでありデータの信憑性については保証しませんし、このデータを元に日経平均株価の取引を行って何かあっても一切の責任をとりません。

■ 参考文献

安倍首相のTwitter(https://twitter.com/AbeShinzo)
日経平均株価のcsvサイト(http://www.mujinzou.jp/)
TwitterAPI実装で参考にしたサイト(http://ailaby.com/twitter_api/)