Report: Spurious relationships in social media data

Abstract

... Claudio ...

1 Introduction (Anna)

... Proposal ...

2 Data (Claudio)

2.1 Keywords

... List keywords ...

2.2 Dataset

... Twitter dataset ...

2.3 Data analysis

... Describe how we reduce Twitter data with pyspark ...

3 Time series analysis (Andrina)

... Correlation measures, k-means clustering, binning/smoothing ...

The main results of the data reduction explained in the previous section, are time-ordered datasets of the occurence of each of the 100 keywords in the queried twitter dataset. In

order to identify relationships between different keywords, we compare their occurences using three different methods based on the Pearson correlation coefficient and k means clustering.

Given two time series X_1 and X_2 we can determine the amount of linear dependence with the time dependent correlation coefficient $\rho(\Delta t)$, which is defined as

$$\rho(\Delta t) = \langle X_{1,t} X_{2,t+\delta t} \rangle, \tag{1}$$

where Δt denotes the time lag between the two time series. For $\Delta t = 0$ we recover the usual correlation coefficient, whereas for $\Delta t \neq 0$, $\rho(\Delta t)$ quantifies the amount of linear dependence between shifted time series. In order for 1 to be a valid correlation measure, the values of both time series at all time t need to identically distributed (IID) random variables. This means that both time series X_1, X_2 need to be stationary which means that their mean, variance and covariance do not depend specifically on time

$$\mu = \langle X_{i,t} \rangle, \tag{2}$$

$$\sigma^2 = \langle (X_{i,t} - \bar{X})^2 \rangle \tag{3}$$

$$\rho(\Delta t) = \langle (X_{i,t} - \bar{X})(X_{i,t+\Delta t} - \bar{X}) \rangle. \tag{4}$$

We do not expect the occurences of keywords in twitter data to be stationary time series, since the mean occurence of any keyword will for example depend on world events, its popularity and also on the number of people using twitter. We can assess if any time series is stationary by evaluating its autocorrelation function $\rho(\Delta t)$. For an IID time series, $\rho(\Delta t)$ will be close to zero whereas non stationary time series are characterised by slowly decaying autocorrelation functions. In order to be able to quantify the correlation between non stationary time series we need to make the series stationary prior to evaluating 1. This process is called pre-whitening. The most basic form of prewhitening of the differencing method, which allows to transform time series with a slow trend to stationary series. A time series with a trend can be written as

$$X_{1,t} = T_{i,t} + R_{i,t}, \tag{5}$$

where $T_{i,t}$ denotes the time series' trend and $R_{i,t}$ are random fluctuations around this trend. If the time series varies slowly, we can remove the trend by considering the differences time series instead of the initial one. The differenced time series is defined as

$$D_{1,t} = X_{1,t+1} - X_{1,t}. \tag{6}$$

If $T_{i,t}$ is slowly varying, this transformed time series will only capture the random variations around the global trend. The cross correlation between two differenced time series

thus encodes the amount of linear dependence between changes in one time series from its global trend and changes in the second.

In order to find relationships between the keywords we will proceed in two steps: for illustration purposes we will compute the cross correlation between all the time series regardless of stationarity. We will then compute the autocorrelation function of all the series and recompute the correlation coefficient using the differenced time series for non stationary processes and compare theses two measures.

4 Results (Anna, Andrina)

...

5 Conclusion (All)

...

6 Acknowledgements

...