Spurious relationships in Twitter data

Data Science in Techno-Socio-Economic Systems

Spring Semester 2016

Claudio Bruderer
cbruderer@phys.ethz.ch

Andrina Nicola
nicolaa@phys.ethz.ch

Anna Weigel
anweigel@phys.ethz.ch

Abstract

... Claudio ...
include link to website

1 Introduction (Anna)

We have entered the era of Big Data and the wealth of information available to us entails its own new challenges. Relationships between quantities, whether real or without any obvious causal link, will naturally arise if the data sets are large enough (for examples see [1]). Even though we might be inclined to interpret apparent links between measurements, a correlation does not imply causation. One needs to find a way to distinguish spurious relationships between measurements from those with a causal link, and quantify the frequency of chance correlations occurring.

Based on a list of pre-defined keywords and Twitter data, we aim to show that relationships between quantities without any obvious causal link can be identified in large data sets. We use a Twitter dataset which contains unfiltered tweets between **time range**. For each day, we determine the number of tweets containing one of our pre-defined keywords. We then measure and quantify the time-correlation functions between the popularity of various keywords. To test our method we recover and quantify expected correlations between synonyms or related keywords. We additionally investigate and discuss the performance of different correlation measures.

2 Data (Claudio)

2.1 Keywords

... List keywords ...
maybe also mention why keywords are not suited, e.g. who or right

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 series of the occurrence of each of the 100 keywords in the queried twitter dataset. In order to identify relationships between different keywords, we compare their occurrence frequencies using three different methods based on the Pearson correlation coefficient and k means clustering.

We can determine the amount of linear dependence between two time series X_i and X_j using the correlation coefficient ρ. The correlation $\rho(\Delta t)$ between two time series is a function of the time lag Δt between the two and is defined as [2]

$$\rho_{ij}(\Delta t) = \langle (X_{i,t} - \bar{X}_i)(X_{j,t+\Delta t} - \bar{X}_j) \rangle, \tag{1}$$

where $\langle ... \rangle$ denotes the ensemble average. For $\Delta t = 0$ we recover the usual correlation coefficient, whereas for $\Delta t \neq 0$, $\rho_{ij}(\Delta t)$ quantifies the amount of linear dependence between shifted time series. If the two time series are equal i.e. $i = j$, then we recover the autocorrelation function of the time series, which quantifies the amount of linear dependence in the time series itself. In order to compare the correlation between different time series, it is customary to normalise Eq. 1 by the variance of the two time series i.e. [2]

$$\rho_{ij}(\Delta t) = \frac{\langle (X_{i,t} - \bar{X}_i)(X_{j,t+\Delta t} - \bar{X}_j) \rangle}{\sqrt{\langle (X_{i,t} - \bar{X}_i)(X_{i,t} - \bar{X}_i) \rangle \langle (X_{j,t} - \bar{X}_j)(X_{j,t} - \bar{X}_j) \rangle}}. \tag{2}$$

When performing time series analyses we generally do not have access to the ensemble averages of the series, but we need to estimate the auto and cross correlations from the samples we have at hand. In order to be able to estimate the auto- or cross correlation using Eq. 1, the ensemble average needs to equal the sample mean i.e. the values of both time series at all times t need to identically distributed random variables. This is equivalent to requiring both time series X_i, X_j to be stationary, which means that their mean μ, variance σ^2 and autocorrelation function $\rho_{ii}(\Delta t)$ do not depend explicitly on time t

$$\mu = \langle X_{i,t} \rangle, \tag{3}$$

$$\sigma^2 = \langle (X_{i,t} - \bar{X})^2 \rangle \tag{4}$$

$$\rho_{ii}(\Delta t) = \frac{\langle (X_{i,t} - \bar{X}_i)(X_{i,t+\Delta t} - \bar{X}_i) \rangle}{\sqrt{\langle (X_{i,t} - \bar{X}_i)(X_{i,t} - \bar{X}_i) \rangle \langle (X_{i,t+\Delta t} - \bar{X}_i)(X_{i,t+\Delta t} - \bar{X}_i) \rangle}}. \tag{5}$$

Only in this case, can we replace the ensemble averages in Equations 1 and 2 with the sample means. This leads to the following estimators for the auto and cross correlation functions [2]:

$$\rho_{ii}(\Delta t) = \frac{\sum_{t=1}^{n-\Delta t} (X_{i,t} - \bar{X}_i)(X_{i,t+\Delta t} - \bar{X}_i)}{\sum_{t=1}^{n} (X_{i,t} - \bar{X}_i)^2} \tag{6}$$

$$\rho_{ij}(\Delta t) = \frac{\sum_{t=1}^{n-\Delta t} (X_{i,t} - \bar{X}_i)(X_{j,t+\Delta t} - \bar{X}_j)}{\sqrt{\sum_{t=1}^{n} (X_{i,t} - \bar{X}_i)^2 \sum_{t=1}^{n} (X_{j,t} - \bar{X}_j)^2}}. \tag{7}$$

We do not expect the occurrences of keywords in twitter data to be stationary time series, since the mean occurrence of any keyword will for example depend on world events, its popularity and also on the number of people using twitter. We can roughly assess if any time series is stationary by visually inspecting the time series plot and also by computing its autocorrelation function $\rho(\Delta t)$. Non stationary time series are generally characterised by clear patterns in the autocorrelation functions because consecutive values are strongly correlated through the common trend.

In order to be able to quantify the correlation between non stationary time series we need to make the series stationary prior to evaluating Equations 1 and 2. The most basic form is called differencing, which allows us to transform time series with a slow trend to stationary series. A time series with a trend can be written as [2]

$$X_{i,t} = T_{i,t} + R_{i,t}, \tag{8}$$

where $T_{i,t}$ denotes the time series' trend and $R_{i,t}$ are random fluctuations around this trend. These fluctuations $R_{i,t}$ are generally correlated random variables with mean zero. If the time series varies slowly, we can remove the trend by considering the differenced time series instead of the initial one. The differenced time series is defined as [2]

$$D_{i,t} = X_{i,t+1} - X_{i,t}. \tag{9}$$

If $T_{i,t}$ is slowly varying, this transformed time series will only capture the random variations around the global trend and will thus approximately be stationary.

For non stationary time series, we can thus first perform differencing and then compute the auto and cross correlations between the differenced time series. This will encode the amount of linear dependence between changes in one time series from its global trend and changes in the second.

In order to quantify the significance of both the auto- and cross correlations, we need to compare the measured values to the expected values for uncorrelated time series. For uncorrelated time series, which are a sequence of IID variables we expect both for the auto and cross correlations $\rho(\Delta t) = 0$ with a variance of $\sigma^2(\rho(\Delta t)) = \frac{1}{n}$, where n denotes the number of samples used to estimate the correlation [2]. For simplicity, we will call a cross-correlation between two time series significant if its value exceeds $1.96\sigma(\rho(\Delta t))$ (i.e. the 95% confidence limits), even though we do not expect the differenced time series to be perfect IID processes.

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

This project was part of the lecture *Data Science in Techno-Socio-Economic Systems* given by Dr. Evangelos Pournaras, Prof. Dr. Dirk Helbing and Dr. Izabela Moise at ETH Zurich during the spring semester 2016. We would like to thank Dr. Izabela Moise and Dr. Rok Roskar for the valuable guidance and helpful discussions. We used Apache Spark [3] to query the Twitter data, and Python [4] and bokeh [5] for the data analysis and visualization. The website was built on a html5up [6] template.

References

[1] `http://tylervigen.com/spurious-correlations`

[2] Marcel Dettling, Applied Time Series Analysis, SS 2014, Zurich University of Applied Sciences, `https://stat.ethz.ch/education/semesters/ss2014/atsa/Scriptum_v140523.pdf`

[3] `http://spark.apache.org`

[4] `https://www.python.org`

[5] `http://bokeh.pydata.org/en/latest/`

[6] `http://html5up.net`