Big Data Project

Big Data Project - Part 1.2

Authors:

Meritxell Jordana Gavieiro
Marc Piñol Pueyo
Carolina Romeu Farre

April 24, 2017

Universitat de Lleida

Contents

1 Introduction

In this project we attempt to correlate bitcoin prices with several external factors and make an informed prediction based on the collected data.
In addition to the own bitcoin price trends, the different data we aim to use are:

1. Twitter activity
    - Prominent political figures (Trump, Merkel, Àngel Ros...)
    - Laymen and other unimportant people (like us)
    - Corporations, foundations and enterprises, both global and local (FSF, Apple, Samsung...)
    - General famous people (Mark Hamill, Mark Zuckerberg, Evan Spiegel, Richard Stallman...)
2. Related reddit subs (r/bitcoin, r/economics, r/girlsgonebitcoin...)
3. Gold historical prices
4. Stock market fluctuations

2 Obtaining the data

2.1 Bitcoin prices

In order to extract the data from the bitcoin prices we use the API from coindesk and to call the API and collect the data we create the code called bitcoin_data.py that saves the data from 2010-07-17 to today in a file called bitcoin_historic.json. The call to the API used is the following: `http://api.coindesk.com/v1/bpi/historical/close.json?start=2010-07-17&end="+str(date.today())` when you only have to put the date of today and starting date is fixed to the first day the API have data.

2.2 Twitter activity

After researching the Twitter Official API, we saw that we could not use it because it had limitations on extracting the data before a week. There

are some tools which provide access to older tweets but most of them are commercial products. Topsy was an alternative, but they closed after Apple bought it. For this reason we opted to use Twitter's search filter. To collect the data found an interesting project that provided us with one of the main advantages of Twitter Search on browsers: we could search the oldest tweets. When entering on a Twitter page a scroll loader starts and, if you scroll down, it keeps returning tweets, all through calls to a JSON provider. It exports directly to a CSV file structured with the most relevant fields.
The script has the following arguments:

- `username` (str): An optional specific username from a twitter account. Without "@".
- `since` (yyyy-mm-dd): A lower bound date to restrict search
- `until` (yyyy-mm-dd): An upper bound date to restrict search
- `querysearch` (str): A query text to be matched
- `toptweets` (bool): If True only the Top Tweets will be retrieved
- `maxtweets` (int): The maximum number of tweets to be retrieved. If this number is not set or lower than 1 all possible tweets will be retrieved

Example:

```
python Exporter.py --querysearch "trump" --username "bitcoin" --since
2010-07-17
```

2.3 Reddit subs

We wrote a Python script to consume the already available Reddit API. As a helper we used the PRAW wrapper to make the calls easier and manage certain Reddit quirks (like request frequency). A small caveat of the API is that the different calls return a maximum of 1000 posts, so a workaroud had to be found: we used the search methods to get the different posts between certain dates. This way, by going from our desired start date to the current day's epoch, we were able to find all the posts and circumvent this limit.
The script works straightforwardly and creates a CSV file ready for processing for both posts and comments. It also retries any failed request to make sure all the info is obtained.

The required information to use the script is a Reddit user/pass, client ID and client secret.

2.4 Gold prices data

To extract the gold price more or less is the same operation as we do to extract the bitcoin data. We extract data from one of the multiple datasets offered by quandl and we have a python script called gold_historic.py. In this case the JSON is not as simply as before, because have elements as:

`["2016-10-24",1267.0,1265.55,1034.886,1035.6,1163.613,1163.04]`

because for each date they offer us the gold price in multiple currencies. We decided not to delete any of the currencies because in a future can be interesting to work not only in euros or dollars.

It was not planned at the beginning to look for the gold price but we need some different variables to study the dependency of the bitcoin price and at the beginning we thought in the weather but we saw that there were no API's that offered us all the info necessary for a period time, in this case from 2010-07-17 to today that is the range of dates for the bitcoin prices.

Also, (is not asked for but) we create a plot in order to be able to see graphically the rises and slopes of both prices and be able to see if there were any dependency. The call to the API used is the following:

```
https://www.quandl.com/api/v3/datasets/LBMA/GOLD.json?api_key="+
key+"&start_date=2010-07-17
```

where obtaining a key for the API you're able only fixing the starting date to obtain the data from this date until today.

2.5 Stock market data

Next data obtained was the exchange historical data from the 12 most important stock markets of the world (From Spain Ibex35, from Germany dax, from Japan nikkei, etc). We found this interesting because we need to know

how is the world going economically to be able to see if the bitcoin price follows the states of this stock markets values. To obtain this data we had to download manually from the `http://finance.yahoo.com/`.

3 Data transformation

3.1 Bitcoin prices

This file did not need any filtering or cleaning because it was a simple JSON where the only data contained is the date and the price of the the bitcoin in that date, so it was just a file with two columns (both needed).

3.2 Twitter activity

The format was already in CSV, but some changes had had to be made. Fields were transformed into numeric, date, etc, some columns were erased, and empty values were changed into "N/A".

3.3 Reddit subs

To prepare the data we got from Reddit we sanitized it (it was not really necessary) and converted and formatted the different dates from epoch to a human-readable format. We also extracted the domain of the different links. After this was done we wrote a small script to count the amount of posts, up/downs, score and number of comments each day, by each author and referring to each domain so we can have statistics to work with in the future and correlate the Reddit activity with bitcoin prices.

3.4 Gold prices data

In this case also we didn't need to clean anything because data es complete and without apparent errors like bad outliners. What we did with the data was compare throught a plot, the gold price and the bitcoin price in order to have visual references of the data we collect. This maybe is not for this part but it useful to starting to see if there are any relation between both.

3.5 Stock market data

This files contain the data, the opening price, the closing price and some variations during the dayand the volume, but what we found interesting was to add a new column that indicates us if each day the value increases or decreases. To do this we work with OpenRefine, adding a new column(Gain) that can have two values: Loss (if this day the close price was lower than the open price) or Gain (if this day the close price was higher than the open price).

All the cells contains a value so we don't need to complete information and the cvs files are correct so we don't need to check outliners, blank spaces or some rare characters that would be ad for our future study.

What we have to do is to modify the date column in order to add a new date facet because it was an string instead of a date format, all this done also with OpenRefine.

Also as we commented before, we keep all the columns of the original files because in a future would be interesting for us to use them for some calculation.