Quarterly results of Brazilian public companies (Bovespa)

By Javier Alperte Pérez-Rejón

Dataset title

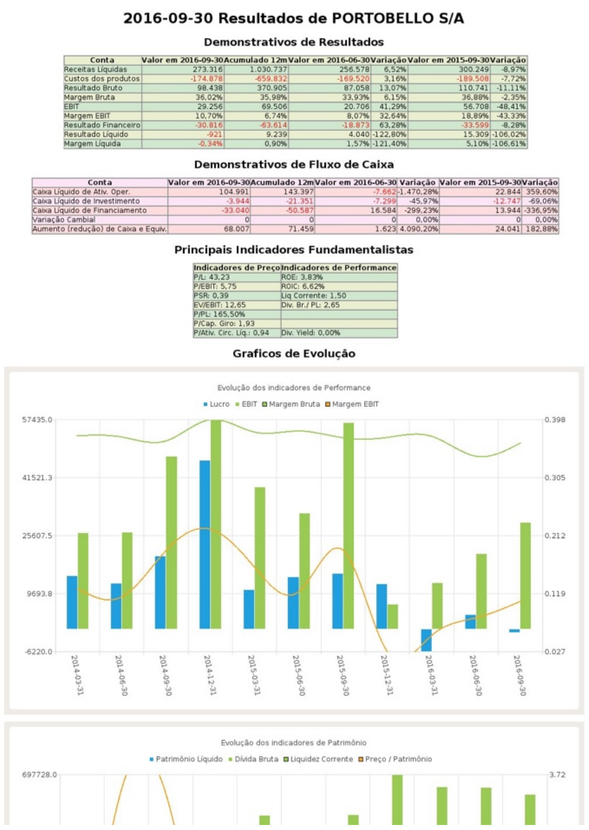
Quarterly results of Brazilian public companies (Bovespa)

Dataset subtitle

This dataset contains the quarterly economic results delivered by companies listed on the Brazilian stock exchange.

These results are presented in the form of a Balance Sheet, Profit and Loss, and Cash Flow.

Imatge



2016-09-30 Resultados de PORTOBELLO S/A

Demonstrativos de Resultados

Conta	Valor em 2016-09-30	Acumulado 12m	Valor em 2016-06-30	Variação	Valor em 2015-09-30	Variação
Receitas Líquidas	273.316	1.030.737	256.578	6,52%	300.249	-8,97%
Custos dos produtos	-174.878	-659.832	-169.520	3,16%	-189.508	-7,72%
Resultado Bruto	98.438	370.905	87.058	13,07%	110.741	-11,11%
Margem Bruta	36,02%	35,98%	33,93%	6,15%	36,88%	-2,35%
EBIT	29.256	69.506	20.706	41,29%	56.708	-48,41%
Margem EBIT	10,70%	6,74%	8,07%	32,64%	18,89%	-43,33%
Resultado Financeiro	-30.816	-63.614	-18.873	63,28%	-33.599	-8,28%
Resultado Líquido	-921	9.239	4.040	-122,80%	15.309	-106,02%
Margem Líquida	-0,34%	0,90%	1,57%	-121,40%	5,10%	-106,61%

Demonstrativos de Fluxo de Caixa

Conta	Valor em 2016-09-30	Acumulado 12m	Valor em 2016-06-30	Variação	Valor em 2015-09-30	Variação
Caixa Líquido de Ativ. Oper.	104.991	143.397	-7.662	-1.470,28%	22.844	359,60%
Caixa Líquido de Investimento	-3.944	-21.351	-7.299	-45,97%	-12.747	-69,06%
Caixa Líquido de Financiamento	-33.040	-50.587	16.584	-299,23%	13.944	-336,95%
Variação Cambial	0	0	0	0,00%	0	0,00%
Aumento (redução) de Caixa e Equiv.	68.007	71.459	1.623	4.090,20%	24.041	182,88%

Principais Indicadores Fundamentalistas

Indicadores de Preço	Indicadores de Performance
P/L: 43,23	ROE: 3,83%
P/EBIT: 5,75	ROIC: 6,62%
PSR: 0,39	Liq Corrente: 1,50
EV/EBIT: 12,65	Div. Br./ PL: 2,65
P/PL: 165,50%	
P/Cap. Giro: 1,93	
P/Ativ. Circ. Líq.: 0,94	Div. Yield: 0,00%

Graficos de Evolução

Evolução dos indicadores de Performance
■ Lucro ■ EBIT ■ Margem Bruta ■ Margem EBIT
57435.0
41521.3
25607.5
9693.8
-6220.0
0.398
0.305
0.212
0.119
0.027
2014-03-31
2014-06-30
2014-09-30
2014-12-31
2015-03-31
2015-06-30
2015-09-30
2015-12-31
2016-03-31
2016-06-30
2016-09-30
Evolução dos indicadores de Patrimônio
■ Patrimônio Líquido ■ Dívida Bruta ■ Liquidez Corrente ■ Preço / Patrimônio
697728.0
3.72

Context

There are different approaches to investment analysis. One of these approaches is the Value Investing method.

This method tries to discover important gaps between the market price of the shares of the companies (official price on the stock exchange), and the real price of the company, known as the [Intrinsic Value] (https: / /www.investopedia.com/terms/i/intrinsicvalue.asp).

The ultimate goal is to invest in those companies with a very low or much higher intrinsic value than the public one (stock exchange).

In order to be able to calculate the intrinsic value of Brazilian companies, we need to have access to the quarterly economic reports that companies deliver every three months.

And that is the reason for this utility, to extract the economic reports presented by the Brazilian companies on the stock market of São Paulo.

Content

The dataset is composed by more than 400 fields:

- the first three fields of the dataset identify the company code in Bovespa (CCVM), the fiscal period (period), and the version of the deliver documents.
- the rest of the fields are a reference to each account of the Brazilian official accounting plan.

You can see an example inside the **data** folder. We can see an example of the dataset in the file dataset.csv, and its data dictionary in the file dictionary.csv.

Acknowledgments

The data comes from the CVM (Comissao de Valores Mobiliarios), the Securities and Exchange Commission of Brazil. Its objective is oversee, regulate, discipline and develop the securities market in Brazil.

Resources:

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.

- Post Automatização de balanços explaining the documents available from Bovespa, and how to access to them.

- http://www.potuz.net/denet/. An implementation in C++ of a system to do fundamental analysis using the information from Bovespa.

Inspiration

I have read some books on Value Investing and I found them very interesting.

After those readings, I decided to focus my project of the end of this master in develop a set of models to help me identify investing patterns.

License

This software has been licensed under CC BY-SA 4.0.

With this license you are free to:

- **Share** — copy and redistribute the material in any medium or format
- **Adapt** — remix, transform, and build upon the material for any purpose, even commercially.

With the following conditions:

- **Attribution**: you must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

- **ShareAlike**: if you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

Code

crawl.py: is the main function of the crawler. The crawler is divided intro three phases:

1. **crawling_parts/listed_companies.py**: crawl all the listed companies in Bovespa (ancient and new)

2. **crawling_parts/company_files.py**: crawl the list of delivered files by the companies to Bovespa.

3. **crawling_parts/download_file.py**: download the delivered files, extract the financial information from them, and generate the `data/dataset.csv` and `data/dictionary.csv` files

The **throttle.py** contains an implementation of a decorator to be used in the functions we want to throttle. The functions that we are actually throttling are:

- `update_listed_companies`(crawling_parts/listed_companies.py): 20 requests by minute.
- `obtain_company_files` (crawling_parts/company_files.py): 20 requests by minute.
- `download_file` (crawling_parts/download_file.py): 20 requests by minute.

The **utils.py** contains some utility methods to manage the control files and date time objects.

The **ctl** folder will be used by the crawler to generate checkpoint files for each of the previous stages. This control files will help us to continue the process to continue the process from where it was left in case the process exited unexpectedly.

The **crawler_cache** folder will be used by the script as the default cache folder to leave all the downloaded files, and to unzip the files for futher processing.

The **data** folder will contains the csv files generated by the crawler. We can now see an example of:

- `companies.csv`: the listed companies in Bovespa. Generated by the first stage of the crawling process.
- `dataset.csv`: the dataset with all the financial information for all the listed companies. Generated by the last stage of the crawling process.
- `dictionary.csv`: the data dictionary associated with the `dataset.csv` file. Generated by the last stage of the crawling process.

Some details about the implementation:

- We use multiprocessing Pool to crawl faster.
- We use of a custom Throttler to make sure we call the source using a request/second policy (20 request per minute).
- We use checkpoint files to take control of the already crawled parts and start the proces from the previos checkpoint if something wrong happen during the process.

Installation

To run the crawler we will need to install the PhantomJS library.

- In MacOS we can use brew to install it: *`brew install phantomjs`*

- In this page we can also found the installers for all the platforms.

We will need to remember the installation folder because a reference to the `..../bin/phantomjs` folder will be needed.

The crawler is implemented for Python 3.7. Then we will need a 3.7 environment ready to install the dependencies.

We can use Anaconda to manage the environment. Once Anaconda is present in our system, we can do the following steps:

```
$ conda create -n bovespa_crawler python=3.7

$ conda activate bovespa_crawler

$ pip install -r requirements.txt
```

Now the crawler should be ready to run.

Run the crawler

To run the crawler we only need to run th `crawl.py` file. The crawler can be invoked using the following config parameters:

Arguments:

- `--phantomjs-path`: The path where we can found the PanthomJS library installed. Default: `None`. Ex: "/phantomjs-2.1.1-macosx/bin/phantomjs".
- `--from-date`: Extract only the data after an specific date. Default: `None`. Ex: "2018-01-01".
- `--cache-folder`: The folder we want to use to save the downloaded files. Default: `./crawler_cache`. Ex: /data/crawlers/bovespa.
- `--workers-num`: The number of parallel threads crawling. Default: `10`. Ex: 20.
- `--force-crawl-listed-companies`: If we want to bypass the checkpoint control and crawl the basic data of all the open companies in bovespa. Ex: --force-crawl-listed-companies.
- `--force-crawl-company-files`: If we want to bypass the checkpoint control and crawl the company files since the begining. Ex: --force-crawl-company-files.
- `--include-companies`: If we want to focus only on a specific companies. Ex: 35 94 1384.

Examples:

- Crawl all the companies using 5 parallel processes

```
python crawl.py \
    --phantomjs-path "/phantomjs-2.1.1-macosx/bin/phantomjs" \
    --workers-num 5
```

- Crawl all the companies using 5 parallel processes and only files published after 2018-06-01

```
python crawl.py \
    --phantomjs-path "/phantomjs-2.1.1-macosx/bin/phantomjs" \
    --workers-num 5 \
    --from-date "2018-06-01"
```

- Crawl companies with the following CCVM codes only:
    - 86 - ACOS DO BRASIL SA IND E COM
    - 13773 - PORTOBELLO S/A
    - 9512 - PETRÓLEO BRASILEIRO S.A. - PETROBRAS

```
python crawl.py \
    --phantomjs-path "/phantomjs-2.1.1-macosx/bin/phantomjs" \
    --workers-num 5 \
    --include-companies 86 13773 9512
```

Dataset

The dataset has been generated using 4200 companies and from 12296 financial reports.

Dataset: dataset.csv

Data Dictionary: dictionary.csv