How Big Data, Machine Learning and Text Mining can help predicting economic activity?

Hudson C. Costa, Sabino P. da Silva Júnior, Fernando A. B. Sabino da Silva

Appus/HR Analytics, PPGE/UFRGS, Department of Statistics/UFRGS

Summary

1. Introduction and Motivation

2. Methodology

3. Results

4. Conclusions

Introduction and Motivation

- Economic Science has evolved over several decades toward a greater emphasis on empirical work. Hamermesh (2013) reviewed publications of the leading journals for the period 1963 to 2011:
  - Until the mid-1980s, most articles were theoretical. After that the share of empirical articles rose more than 70%
  - Most data are assembled or obtained by the authors or generated through controlled experiments
- Currently, there are a greater availability of data which may affect economic research
  - Measurement of inflation and the labor market
  - Consumer behavior, productivity and *job search*
- *Big Data* makes accessible statistical approaches (*Classification Trees*, *Regression Trees*, *Random Forest*) and computational (*Machine Learning* and *Text Mining*) already commonly used in other fields of research;

Introduction and Motivation

- Main Goals:
  - Discuss computational natural language processing (NLP) techniques
  - Investigate whether the sentiment conveyed by Central Bank of Brazil (BCB) about the economy in their statements using *Web Scraping* (the process of extracting text from a web page) and *Text Mining* affects the short-term inflation
- Some Results and Conclusions:
  - We create an algorithm that extracts all text from minutes of Copom (Monetary Policy Committee of the Central Bank of Brazil) from BCB website and transform the pdf format into textual format to create a corpus;
  - We constructed a sentiment index for the monetary authority based on a dictionary of feelings (Inquider) maintained by Harvard University
  - Our results confirm that such approach can contribute to the economic evaluation given that the time series of the index seems to be related to important macroeconomic variables

Methodology - Web Scraping

- *Web Scraping*:
  - The information available on the Internet (*Web*) is rearely in a suitable format to be extracted in an automatic and efficient way
  - *Web Scraping* makes easier the process of collecting such data;
  - We design an algorithm to extract information from unstructured text:
    - The web pages are constructed using HTML language (HTML)
    - The codes have *tags*, such as <title> and <p>;
    - These *tags* to remain constant over time while the information inside them (minutes of Copom, for example) is dynamic;
    - The algorithm is taught to use such *tags* to locate information and store it in a database

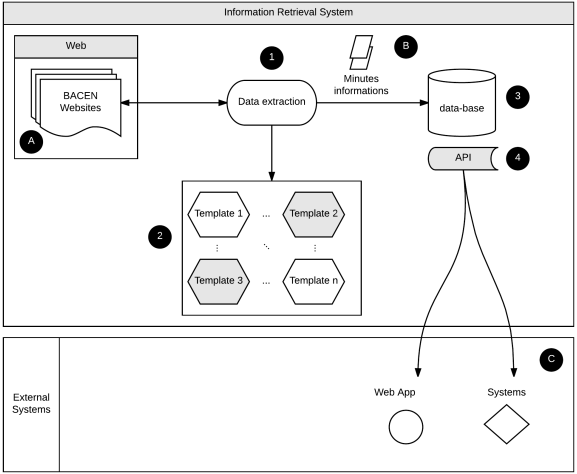
Methodology - Web Scraping

Example of the minutes collection algorithm

```
> main.page = read_html(x = "http://www.bcb.gov.br/?MINUTES")
> urls = main.page %>%
+   html_nodes("#cronoAno a") %>%
+   html_attr("href")
> ano = main.page %>%
+   html_nodes("#cronoAno a") %>%
+   html_text()
>
> # 2016 http://www.bcb.gov.br/?id=MINUTES&ano=2016
```

- We have developed a system that carries out massive collection and then uploads the information to a database, based on a logical and scalable architecture solution
- It interacts with the web pages, extracts the information and stores the data
- Data Collectors: Google, Facebook.

Methodology - Web Scraping



Methodology - Sentiment Analysis

- It is the computational study of opinions, attitudes and emotions behind a series of words expressed within an online mention;
- The applications of sentiment analysis are broad and powerful. It allows us to gain an overview of the wider public opinion behind certain topics.;
- Two main approaches: (1) based on lexical resources and neutral language processing (it uses semantic dictionaries) and (2) employing machine learning algorithms;
- We follow the simpler approach of Hu and Liu (2004) and Kim and Hovy (2004), i.e., we focus in a dictionary-based approach for polarity identifications;

Methodology - Sentiment Analysis

- The strategy is to first collect a small set of opinion words manually with known orientations, and then to grow this set by searching in an online dictionary for their synonyms and antonyms.
- The newly found words are added to the seed list. Then, the next iteration starts. It stops when no more words are found.
- We use the Inquirer dictionary provided by Harvard University (Stone, Dumphy and Smith (1966)) which classifies 11,788 words into semantic groups (*positive*, *negative*, *strong*, *weak*, among others).

Methodology - Sentiment Analysis

- Using two databases (matrix of words of Copom minutes and semantic dictionary), we look for the words that are present in both bases for each of the minutes
- Thus, we know how many **negative** and **positive** words are present in each record of the minutes.

Sentiment Index

$$I_t = \frac{NP_t - NN_t}{N}$$

where I_t is the sentiment index for each minute disclosed in t, NP_t is the number of **positive** words present in the minutes reported in t while NN_t is the number of **negative** words and N is the total number of words in the minutes. A higher value of the index I_t, signifies that the monetary authority has better expectations about the economyr.

Methodology - Data

- We use Copom's minutes to evaluate the BCB's sentiment about the economy;
- We use the English version available on the Internet and in PDF format since the 42nd Meeting;
- Total minutes available: 159 (until July 2016 meeting);
- We remove the minutes of meetings 42 and 43 from the sample due to differences in the layout of the files compared to the other minutes
- In addition, using Quandl R package we collect the following time series to compare how well the sentiment index performs:
    - Annual IPCA (National Consumer Price Index);
    - Nominal interest rate (Selic);
    - Annual inflation target

Results

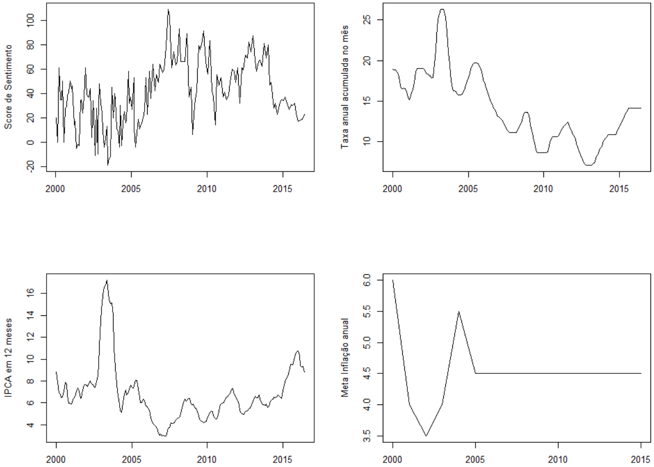
- Once all minutes are stored, it is necessary to transform them so that each word can be extracted from its contents.
- Next, we go through the following steps:
  1. First we read and store PDF files in the format of a **Corpus** (a collection of textual documents);
  2. Here, our **Corpus** contains 157 textual documents, each document being a textual representation of the corresponding minute;
  3. Next, we format each text document to get rid from probable ”impurities”:
     - Remove numbers, punctuation characters, meaningless words (the, you, we, for example), blanks;
     - Perform a word *stemming* process, i.e., remove suffixes from words to get the common origin and transform all characters to lowercase.
  4. Create a matrix where we have in each column the distinct words of all documents that belong to the **Corpus**. Each row represents a different document.

Results

Tabela: Document-term matrix (Documents x Words)

Minute	absorption	acceleration	accommodative	accordance	according	account
199th	2.00	1.00	1.00	1.00	20.00	1.00
198th	2.00	1.00	1.00	1.00	19.00	1.00
197th	2.00	1.00	1.00	1.00	18.00	1.00
196th	2.00	1.00	1.00	1.00	18.00	1.00
195th	2.00	2.00	1.00	1.00	18.00	1.00
194th	2.00	2.00	1.00	1.00	18.00	1.00
193rd	2.00	2.00	1.00	1.00	23.00	1.00
192nd	2.00	2.00	1.00	1.00	19.00	1.00
191st	2.00	1.00	1.00	0.00	20.00	1.00
190th	2.00	1.00	1.00	0.00	17.00	1.00

Results


Score de Sentimento
-20
0
20
40
60
80
100
2000
2005
2010
2015
Taxa anual acumulada no mês
10
15
20
25
2000
2005
2010
2015
IPCA em 12 meses
4
6
8
10
12
14
16
2000
2005
2010
2015
Meta Inflação anual
3.5
4.0
4.5
5.0
5.5
6.0
2000
2005
2010
2015

Conclusions

- We report an application of *Text Mining* in the Copom minutes that are published on the BCB website;
- Using *Web Scraping* and *Text Mining* techniques we show that the sentiment index may help predicting economic activity. Thus, the arrival of economic information such as press-releases may be used to improve estimates of financial risk, for example;
- Future research will include econometric specifications augmented with news and sentiment indexes. We also will consider *Machine Learning* techniques.