

# Big Data, Machine Learning e Text Mining em Economia: Estudos Recentes e Análise de Sentimento do BACEN

Hudson Chaves Costa <sup>1</sup>   Sabino Porto Júnior <sup>2</sup>   Fernando A. B. S. da Silva <sup>3</sup>

<sup>1</sup> IBMEC - MG

<sup>2</sup>Faculdade de Ciências Econômicas - UFRGS

<sup>3</sup>Instituto de Matemática e Estatística - UFRGS

# Sumário

- Introdução/Motivação
- Metodologia
- Resultados
- Conclusão

# Introdução/Motivação

- A Ciência Econômica tem evoluído ao longo de várias décadas em direção a uma maior ênfase em trabalhos empíricos. Hamermesh (2013) revisou publicações das principais revistas para o período de 1963 a 2011:
  - Até meados da década de 1980, a maioria dos artigos eram teóricos e após isso a participação de artigos empíricos subiu para mais de 70%
  - A maioria usou dados montados ou obtidos pelos autores ou gerados através de experimentos controlados.
- Atualmente há maior disponibilidade de dados o que pode afetar a pesquisa econômica
  - Mensuração da inflação e do mercado de trabalho
  - Comportamento do consumidor, produtividade e *job search*
- *Big Data* torna acessível abordagens estatísticas (*Classification Trees*, *regression Trees*, *Random Forest*) e computacionais (*Machine Learning* e *Text Mining*) já comumente utilizadas em outros campos de pesquisa;

# Introdução/Motivação

- Objetivos deste artigo:
  - Apresentar pesquisas que utilizaram *Big Data*, *Machine Learning* e *Text Mining* em macroeconomia;
  - Discutir principais técnicas e tecnologias;
  - Analisar o sentimento do Banco Central do Brasil (BCB) sobre a economia usando *Web Scraping* e *Text Mining*
- Alguns resultados e conclusões:
  - Criamos um algoritmo que acessa as atas divulgadas em inglês pelo Copom no site do BCB e retira do PDF as palavras usadas na escrita das atas;
  - Construímos um índice de sentimento para a autoridade monetária tendo como base o dicionário de sentimentos Inquirer (Harvard)
  - Nossos resultados confirmam que tal abordagem pode contribuir para a avaliação econômica dado que a série temporal do índice parece estar relacionada com importantes variáveis macroeconômicas

# Metodologia - Web Scraping

- *Web Scraping*:

- As informações disponíveis na internet (*Web*) raramente estão no formato adequado para uso;
- *Web Scraping* facilita o processo de coleta de tais dados;
- Trata-se de escrever algoritmos que executam automaticamente o que fazemos manualmente:
  - As páginas são construídas usando uma linguagem de estruturação (HTML)
  - Este código tem *tags*, tais como `<title>` e `<p>`;
  - Estas *tags* tendem a permanecer constantes ao longo do tempo enquanto a informação dentro delas (preço de um produto ou a ata do Copom) são dinâmicas;
  - O algoritmo é ensinado a utilizar tais *tags* para localizar as informações e guardá-las em um banco de dados

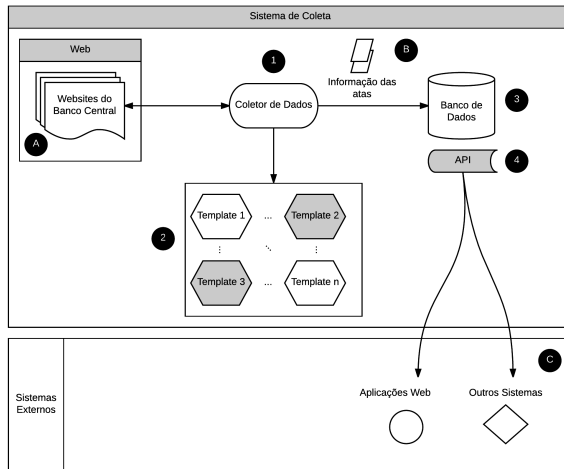
# Metodologia - Web Scraping

## Exemplo do algoritmo de coleta das atas

```
> main.page = read_html(x = "http://www.bcb.gov.br/?MINUTES") > urls =  
main.pagehtml_nodes("cronoAnoa")html_attr("href") > ano =  
main.pagehtml_nodes("cronoAnoa")html_text() >> 2016http://www.bcb.gov.br/?id =  
MINUTESano = 2016
```

- Desenvolvemos um coletor capaz de arquitetar e executar de forma lógica e escalável todo esse processo.
- Ele itera com as páginas da Web, extrai a informação e armazena os dados
- Exemplos de coletores: Google e Buscapé

# Metodologia - Web Scraping



# Metodologia - Análise de Sentimento

- É o estudo computacional das opiniões, atitudes e emoções em relação a uma entidade (indivíduos, empresas, etc);
- O objetivo é encontrar opiniões ou identificar os sentimentos expressos em um texto;
- Existem muitas aplicações de algoritmos de análise de sentimento. Em resumo, podemos classificar em abordagens baseadas em *Machine Learning* e orientadas por Lexicon.
- Aplicamos neste estudo, a segunda abordagem que faz uso de dicionários semânticos;
  - Apresentada por Kim e Hovy (2004) e Hu e Liu (2004)
  - Uma pequena quantidade de palavras de opinião são coletadas manualmente e a partir da manutenção de pesquisadores aumenta-se a coleção de palavras do dicionário
  - Usamos o dicionário Inquirer disponibilizado pela Universidade de Harvard (Stone, Dumphy e Smith (1966)) que classifica 11.788 palavras em grupos semânticos (*positive*, *negative*, *strong*, *weak*, entre outros)



# Metodologia - Análise de Sentimento

- Usando duas bases de dados (matriz de palavras das atas e dicionário semântico), buscamos quais palavras estão presentes nas duas bases para cada uma das atas;
- Assim, sabemos quantas palavras **negativas** e **positivas** estão presentes na escrita de cada ata.

## Índice de Sentimento

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

onde  $I_t$  é o índice de sentimento para cada ata divulgada em  $t$ ,  $NP_t$  é a quantidade de palavras **positivas** presentes na ata divulgada em  $t$  enquanto  $NN_t$  é a quantidade de palavras **negativas** e  $N$  a quantidade de palavras na ata. Quanto maior o valor  $I_t$ , mais **positiva** é a ata e, conseqüentemente, a expectativa para a economia pela autoridade monetária.

# Metodologia - Dados

- Usamos as atas do Copom para avaliar o sentimento do BCB sobre a economia;
- Utilizamos a versão em inglês disponível na internet e em formato PDF desde a 42ª Reunião;
- Total de atas disponíveis: 159 (até a reunião de julho de 2016);
- Eliminamos da amostra as atas das reuniões 42 e 43 em função de diferença no layout dos arquivos em comparação com as demais atas (**amostra com 157 atas**)
- Além disso, para efeito de comparação da série temporal do Índice de Sentimento:
  - IPCA anual e acumulado no mês;
  - Taxa de juros nominal (Selic);
  - Meta anual do IPCA

# Resultados

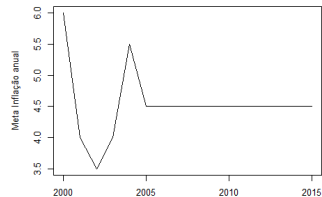
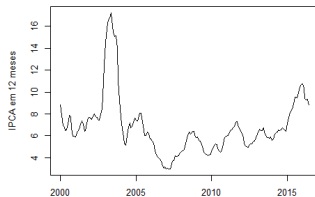
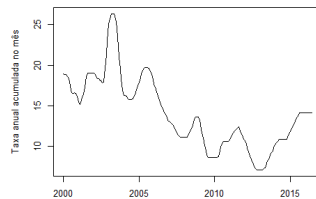
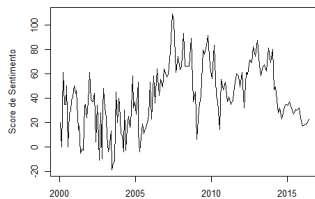
- Uma vez que todas as atas estão armazenadas, faz-se necessário transformá-las de forma que seja possível extrair cada uma das palavras do seu conteúdo.
- Em linhas gerais, temos o seguinte processo:
  - 1 Leitura e armazenamento dos arquivos PDF no formato de um **Corpus** que é uma coleção de documentos textuais;
  - 2 O **Corpus** contém 157 documentos textuais sendo que cada documento é uma representação textual da ata;
  - 3 Após isso, o texto de cada documento é reformatado para eliminar prováveis impurezas:
    - Remoção de números, caracteres de pontuação, palavras sem sentido (the, you, we, por exemplo), espaços em branco;
    - Aplicação do procedimento de *stemming* (reduzir palavras relacionadas) a uma forma mínima comum e transformação de todos os caracteres para minúsculo
  - 4 Criar uma matriz onde temos em cada coluna as distintas palavras de todos os documentos do **Corpus** e nas linhas cada um dos documentos

# Resultados

Tabela: Matriz de dados (Documentos x Palavras)

Ata	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

# Resultados



# Conclusões

- Expomos a aplicação de *Text Mining* nas atas das reuniões do Copom que são divulgadas no site do BCB;
- Utilizando técnicas de *Web Scraping* e *Text Mining* mostramos o índice de sentimento está relacionado com séries temporais importantes para a autoridade monetária (IPCA, Selic);
- Tal resultado fortalece a importância do uso das técnicas apresentadas em pesquisa ecocômicas aplicadas;
- No que tange aos resultados empíricos, testes comumente utilizados em séries temporais como Causalidade de Granger podem contribuir para a robustez dos resultados assim como o uso da série temporal do índice de sentimento em modelos VAR, SVAR, VEC ou SVEC;
- Além disso, trabalhos futuros podem usar a mesma base de dados em busca de prever as decisões do BCB por meio de técnicas de *Machine Learning*;
- Por fim, a mesma metodologia de acesso aos dados pode ser empregada em outros documentos divulgados pela autoridade monetária (relatórios de inflação, comunicados,