Coleta automática de dados para índices de preço e ajustamento de qualidade utilizando web scraping

Tiago Mendes Dantas (DPE/COMEQ)

Lincoln Teixeira da Silva (DPE/COINP)

Como coletar dados da web ?

- **Web scraping**

- **APIs**

O que é web scraping ?

Web scraping é o ato de capturar a informação disponível na internet e estruturá-la em um conjunto de dados de forma automática

Website (www.globo.com)

globo.com g1 ge gshow tech vídeos encontre na globo.com

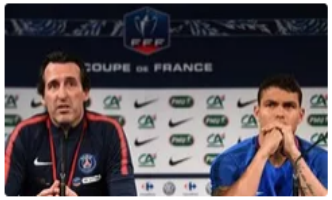
Presos da Lava Jato no RJ são transferidos após supostas regalias

- Fachin prorroga inquérito contra Jucá após delação da Odebrecht

Arcada dentária é de homem que caiu com prédio

Okamotto depõe e relata visitas a sítio em Atibaia

Tiro atinge poltrona na torre do Rio Sul, no RJ



Pergunta sobre Ney irrita técnico do PSG: 'Isso é importante?'

- Técnico: 'Não está adaptado'
- Thiago Silva minimiza polêmica

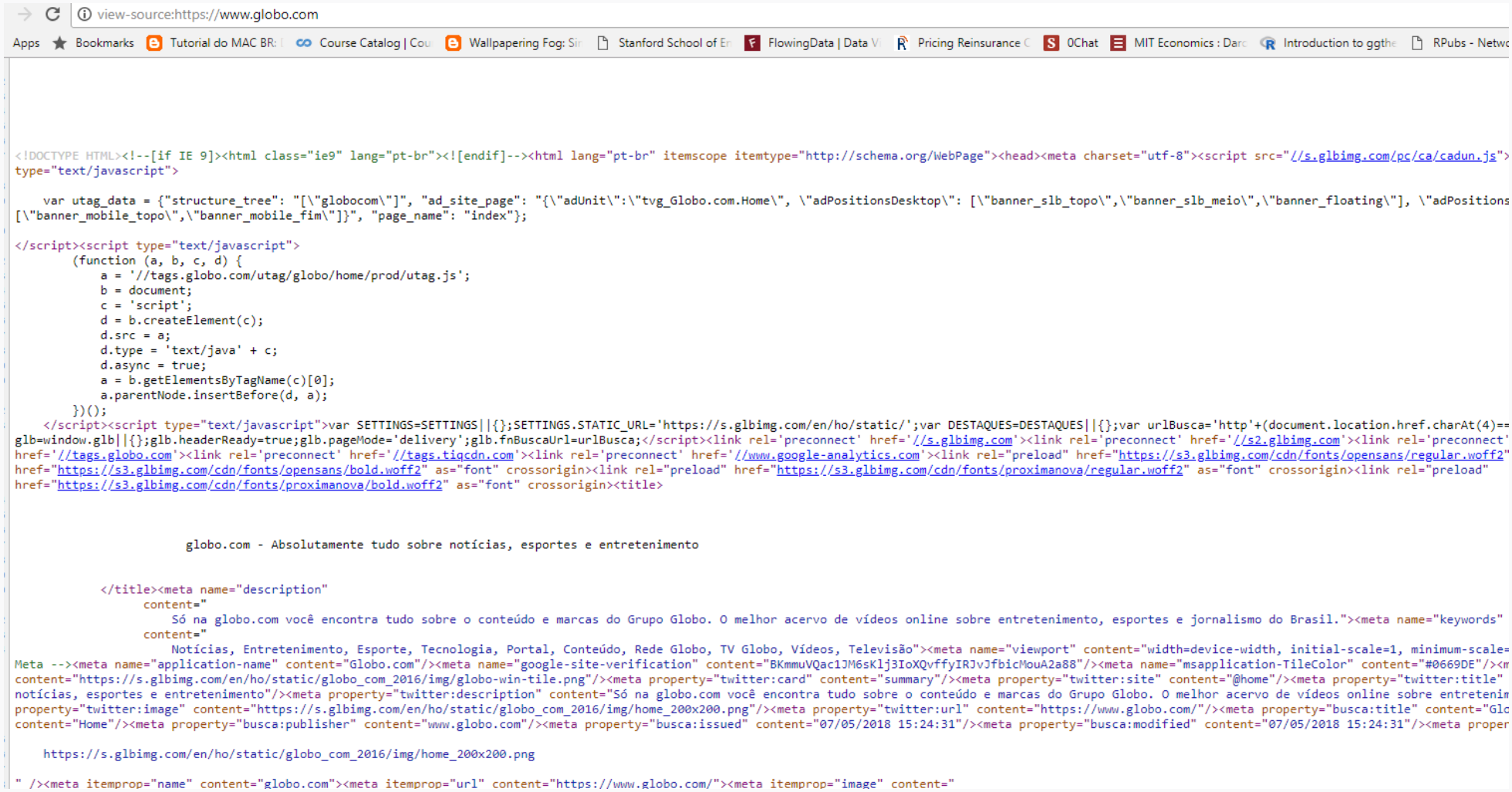
Ex-vices do Vasco denunciam sumiço da pasta sobre Paulinho

'Lado': Caetana é ameaçada com faca

- Renato tem planos para Tomaz
- Ex-BBB Gleici grava cenas hoje

'Fortes': Aurora vê

Código Fonte



view-source:https://www.globo.com

Apps Bookmarks Tutorial do MAC BR: Course Catalog | Cou Wallpapering Fog: Sir Stanford School of Er FlowingData | Data V Pricing Reinsurance C 0Chat MIT Economics : Darc Introduction to ggthe RPubs - Netwo

```
<!DOCTYPE HTML><!--[if IE 9]><html class="ie9" lang="pt-br"><![endif]--><html lang="pt-br" itemscope itemtype="http://schema.org/WebPage"><head><meta charset="utf-8"><script src="//s.glbimg.com/pc/ca/cadun.js">
type="text/javascript">

    var utag_data = {"structure_tree": "[\"globocom\"]", "ad_site_page": "{\"adUnit\":\"tvg_Globo.com.Home\", \"adPositionsDesktop\": [\"banner_slb_topo\",\"banner_slb_meio\",\"banner_floating\"], \"adPositions
[\"banner_mobile_topo\",\"banner_mobile_fim\"]}", "page_name": "index"};

</script><script type="text/javascript">
        (function (a, b, c, d) {
            a = '//tags.globo.com/utag/globo/home/prod/utag.js';
            b = document;
            c = 'script';
            d = b.createElement(c);
            d.src = a;
            d.type = 'text/java' + c;
            d.async = true;
            a = b.getElementsByTagName(c)[0];
            a.parentNode.insertBefore(d, a);
        })();
    </script><script type="text/javascript">var SETTINGS=SETTINGS||{};SETTINGS.STATIC_URL='https://s.glbimg.com/en/ho/static/';var DESTAQUES=DESTAQUES||{};var urlBusca='http'+(document.location.href.charAt(4)==
glb=window.glb||{};glb.headerReady=true;glb.pageMode='delivery';glb.fnBuscaUrl=urlBusca;</script><link rel='preconnect' href='//s.glbimg.com'><link rel='preconnect' href='//s2.glbimg.com'><link rel='preconnect'
href='//tags.globo.com'><link rel='preconnect' href='//tags.tiqcdn.com'><link rel='preconnect' href='//www.google-analytics.com'><link rel="preload" href="https://s3.glbimg.com/cdn/fonts/opensans/regular.woff2"
href="https://s3.glbimg.com/cdn/fonts/opensans/bold.woff2" as="font" crossorigin><link rel="preload" href="https://s3.glbimg.com/cdn/fonts/proximanova/regular.woff2" as="font" crossorigin><link rel="preload"
href="https://s3.glbimg.com/cdn/fonts/proximanova/bold.woff2" as="font" crossorigin><title>


                    globo.com - Absolutamente tudo sobre notícias, esportes e entretenimento


            </title><meta name="description"
                  content="
                      Só na globo.com você encontra tudo sobre o conteúdo e marcas do Grupo Globo. O melhor acervo de vídeos online sobre entretenimento, esportes e jornalismo do Brasil."><meta name="keywords"
                  content="
                      Notícias, Entretenimento, Esporte, Tecnologia, Portal, Conteúdo, Rede Globo, TV Globo, Vídeos, Televisão"><meta name="viewport" content="width=device-width, initial-scale=1, minimum-scale=
Meta --><meta name="application-name" content="Globo.com"/><meta name="google-site-verification" content="BKmmuVQac1JM6sKlj3IoXQvffyIRJvJfbicMouA2a88"/><meta name="msapplication-TileColor" content="#0669DE"/><m
content="https://s.glbimg.com/en/ho/static/globo_com_2016/img/globo-win-tile.png"/><meta property="twitter:card" content="summary"/><meta property="twitter:site" content="@home"/><meta property="twitter:title"
notícias, esportes e entretenimento"/><meta property="twitter:description" content="Só na globo.com você encontra tudo sobre o conteúdo e marcas do Grupo Globo. O melhor acervo de vídeos online sobre entretenim
property="twitter:image" content="https://s.glbimg.com/en/ho/static/globo_com_2016/img/home_200x200.png"/><meta property="twitter:url" content="https://www.globo.com/"/><meta property="busca:title" content="Glo
content="Home"/><meta property="busca:publisher" content="www.globo.com"/><meta property="busca:issued" content="07/05/2018 15:24:31"/><meta property="busca:modified" content="07/05/2018 15:24:31"/><meta proper

    https://s.glbimg.com/en/ho/static/globo_com_2016/img/home_200x200.png

" /><meta itemprop="name" content="globo.com"><meta itemprop="url" content="https://www.globo.com/"><meta itemprop="image" content="
```

Código Fonte


```
▼<a href="https://g1.globo.com/rj/rio-de-janeiro/noticia/presos-da-lava-jato-do-rio-vao-mudar-de-presidio.ghtml" title="Presos da Lava Jato no RJ são transferidos após supostas regalias" class="hui-premium__link
            hui-highlight__link

    ">
      <p class="hui-premium__title">Presos da Lava Jato no RJ são transferidos após supostas regalias</p> == $0
```

Código Fonte

globo.com g1 ge gshow tech vídeos encontre na globo.com

IBGE contrata 2 mil especialistas em web scraping

- Fachin prorroga inquérito contra Jucá após delação da Odebrecht

PGR questiona se caso de Dudu da Fonte fica no STF

Temer tem 'lista de pedidos' para liberar votações

MEC recebe 190 mil inscrições no Enem em 2 horas

Pergunta sobre Ney irrita técnico do PSG: 'Isso é importante?'

- Técnico: 'Não está adaptado'
- Thiago Silva minimiza polêmica

Ex-vices do Vasco denunciam sumiço da pasta sobre Paulinho

'Lado': Caetana é ameaçada com faca

- Renato tem planos para Tomaz
- Ex-BBB Gleici grava cenas hoje

'Fortes': Aurora vê

Big Data – Coordenação de Metodologia e Qualidade

COMEQ criou um grupo para estudar o assunto em Abril de 2017:

- Trabalho inicial de entender o que os países já estavam fazendo
- Maior parte dos trabalhos, de fato, em execução estavam relacionados a coleta de dados da web
- Proposta de tema (maio de 2017)

Coleta de preços utilizando web scraping (2017)

Objetivos Gerais:

Propor um método alternativo de coleta de informações para o SNIPC, no qual o uso de fontes alternativas (por exemplo, sites de companhias aéreas e de supermercados) é feito através de técnicas de web scraping

Coleta de preços utilizando web scraping (2017)

Objetivos específicos:

- Reduzir custos e tempo no processo de coleta de preços para o Índice de Preços ao Consumidor Amplo (IPCA).
- Mostrar a relevância da aplicação de fontes alternativas de informação para produção estatística.
- Desenvolver algoritmos para automatizar o processo de extração de dados sites selecionados.

Projeto-Piloto 1: Coleta de preço de passagens aéreas

Justificativa

- Hoje o processo feito de forma manual (servidores fazem a busca em diferentes regiões do país)
- Processo custoso e existe o risco de não coletar todas as possíveis passagens.

Vantagens

- Praticamente elimina o processo manual de coleta
- Permite ampliar a abrangência geográfica (dados de várias cidades)
- Permite criar séries de variação com novos recortes temporais

Projeto-Piloto 1: Coleta de preço de passagens aéreas

Produtos

- Função genérica em linguagem open source que coleta os preços de passagem aérea das mesmas empresas de aviação utilizadas no SNIPC
- Aplicação Web que permite calcular variações de preços por diferentes recortes (temporais e geográficos)

Projeto-Piloto 1: Coleta de preço de passagens aéreas

Produtos

- Função genérica em R que coleta os preços de passagem aérea das mesmas empresas de aviação utilizadas no SNIPC – Finalizado e expandido para novas empresas aéreas não coletadas pelo SNIPC. Projeto ampliado ao Programa de Comparação Internacional.

- Aplicação Web que permite calcular variações de preços por diferentes recortes (temporais e geográficos) – Em andamento

Estrutura de captura de informação:

- Dados não estruturados e carregados via JavaScript
- Necessário utilizar um automatizador do navegador (Selenium)
- R conectado ao Selenium faz o trabalho
- Funções também escritas em Python: Python conectado ao Selenium
- Produção em C# conectado ao Selenium

Exemplo R:

```r
# Função ------------------------------------------------------------------
# Argumentos:
# origem - string com o código do aeroporto de partida. Ex.: aero_origem = "GIG"
# destino - string com o código do aeroporto de chegada. Ex.: aero_destino = "CGH"
# dataida - string com a data de ida na forma aaaammdd. Ex.: data_ida = "20171017"
# datavolta - string com a data de volta na forma aaaammdd. Ex: data_volta = "20171023"
# espera - tempo de espera para a página carregar. Ex: espera = 20
coleta     <- function(origem,
                       destino,
                       dataida,
                       datavolta,
                       espera = 20){


  dataida_certo <- dataida
  datavolta_certo <- datavolta

  # Janela do chrome
  sessao$open()


  # Acesso ao site de compra de passagens da
  sessao$navigate("https://              com.br/")
  # sessao$maxWindowSize()
  Sys.sleep(espera) # pausa para dar tempo do site carregar completamente


  # # Troca de FOR por Fortaleza e SSA por Salvador
  origem_aux <- ifelse(origem == "SSA", "Salvador", origem)
  destino_aux <- ifelse(destino == "SSA", "Salvador", destino)
  # origem_aux = origem
  # destino_aux = destino


  # Preenchimento dos campos para compra de passagem

      # Origem
      aux_origem <- sessao$findElement(using = 'id', value = 'ticket-origin1')
      aux_origem$clickElement()
      # aux_origem$sendKeysToElement(list(origem_aux, key = "tab"))
      aux_origem$sendKeysToActiveElement(list(origem_aux, key = "tab"))
      Sys.sleep(5)

      # Destino
      aux_destino <- sessao$findElement(using = 'id', value= 'ticket-destination1')
      aux_destino$clickElement()
      # aux_destino$sendKeysToElement(list(destino_aux, key = "tab"))
      aux_destino$sendKeysToActiveElement(list(destino_aux, key = "tab"))
      Sys.sleep(5)

      # Data de ida
      aux_dataida <- sessao$findElement(using = 'id', value= 'ticket-departure1')
      aux_dataida$clickElement()
      # aux_dataida$sendKeysToElement(list(format(as.Date(dataida, "%Y%m%d"), "%d%m%Y"), key = "ta
      aux_dataida$sendKeysToActiveElement(list(format(as.Date(dataida, "%Y%m%d"), "%d%m%Y"), key =
      Sys.sleep(5)
```

Exemplo Python:

```
# FUNCAO     ------------------------------------------------------------
# Argumentos:
# origem - string com o código do aeroporto de partida. Ex.: aero_origem = "GIG"
# destino - string com o código do aeroporto de chegada. Ex.: aero_destino = "CGH"
# dataida - string com a data de ida na forma aaaammdd. Ex.: data_ida = "20171017"
# datavolta - string com a data de volta na forma aaaammdd. Ex: data_volta = "20171023"
# espera - tempo de espera para a página carregar. Ex: espera = 20

def coleta     (origem, destino, dataida, datavolta, espera = 20):

    # Importando funções necessárias
    import os
    from time import sleep
    from time import strftime, localtime  # manipulação de datas
    #from selenium.webdriver.common.keys import Keys # ex.: tab, space, page up, etc
    #import math
    import pandas as pd # trabalhando com data frames
    import numpy as np
    from datetime import datetime, timedelta


    #
    dataida_certo = dataida
    datavolta_certo = datavolta


    #
    cidades = {'VIX' : "Vit%C3%B3ria", 'AJU' : "Aracaju", 'BEL' : "Bel%C3%A9m", 'BSB' : "Bras%C3%ADlia",
               'CGR' : "Campo%20Grande", 'CNF' : "Belo%20Horizonte", 'CWB' : "Curitiba", 'FOR' : "Fortaleza",
               'GIG' : "Rio%20de%20Janeiro", 'GRU' : "S%C3%A3o%20Paulo", 'GYN' : "Goi%C3%A2nia",
               'POA' : "Porto%20Alegre", 'RBR' : "Rio%20Branco", 'REC' : "Recife",
               'SLZ' : "S%C3%A3o%20Lu%C3%ADs", 'SSA' : "Salvador", 'FLN' : "Florian%C3%B3polis"}


    # Inicializando webdriver e chrome
    from selenium import webdriver
    sessao = webdriver.Chrome("chromedriver.exe")


    # Entrando no site da Latam
    endereco = '                                    booking?fecha1_dia=" + dataida[6:8] + "&fecha1_anomes="
    endereco = endereco + dataida[0:4] + "-" + dataida[4:6] + "&fecha2_dia=" + datavolta[6:8] + "&fecha2_anomes="
    endereco = endereco + datavolta[0:4] + "-" + datavolta[4:6] + "&from_city2=" + destino + "&to_city2=" + origem
    endereco = endereco + "&auAvailability=1&ida_vuelta=ida_vuelta&vuelos_origen=" + cidades[origem] + "&from_city1=" + origem
    endereco = endereco + "&vuelos_destino=" + cidades[destino] + "&to_city1=" + destino + "&flex=1&vuelos_fecha_salida_ddmmaaaa="
    endereco = endereco + datetime.strptime(dataida, '%Y%m%d').strftime("%d/%m/%Y") + "&vuelos_fecha_regreso_ddmmaaaa="
    endereco = endereco + datetime.strptime(datavolta, '%Y%m%d').strftime("%d/%m/%Y") + "&cabina=Y&nadults=1&nchildren=0&ninfants=0"

    sessao.get(endereco)
    sessao.maximize_window() # maximizando a janela
    sleep(espera) # pausa para dar tempo do site carregar completamente
```

Projeto-Piloto 1: Coleta de preço de passagens aéreas

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Empresa	Data da consulta	Hora da co	Usuario	Data de id	Data de re	Origem	Destino	Horário de	Horário de	Duração	Conexão	Promo				ID1	ID2
2		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	06:10	09:25	3h 15min	1 parada		435,27	465,27	550,27	1	1
3		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	10:30	14:00	3h 30min	1 parada		698,27	733,27	843,27	1	1
4		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	11:00	15:15	4h 15min	1 parada		254,27	279,27	506,27	1	1
5		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	06:10	11:30	5h 20min	1 parada		435,27	465,27	550,27	1	1
6		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	19:05	01:10	6h 5min	1 parada		254,27	279,27	506,27	1	1
7		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	11:00	18:00	7h	1 parada		285,21	310,21	537,21	1	1
8		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	10:40	18:00	7h 20min	2 paradas		391,27	421,27	506,27	1	1
9		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	10:30	18:00	7h 30min	1 parada		698,27	733,27	843,27	1	1
10		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	06:10	14:00	7h 50min	1 parada		466,21	496,21	581,21	1	1
11		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	06:10	14:00	7h 50min	2 paradas		1792,27	1752,27		1	1
12		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	06:10	14:00	7h 50min	2 paradas		1792,27	1752,27		1	1
13		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	11:00	18:55	7h 55min	1 parada		254,27	279,27	506,27	1	1
14		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	06:10	14:15	8h 5min	2 paradas		1792,27	1752,27		1	1
15		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	10:40	18:55	8h 15min	2 paradas		285,54	310,54	537,54	1	1
16		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	10:30	18:55	8h 25min	1 parada		729,54	764,54	874,54	1	1
17		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	11:00	19:50	8h 50min	1 parada		285,21	310,21	537,21	1	1
18		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	06:10	15:15	9h 5min	1 parada		435,27	465,27	550,27	1	1
19		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	06:10	15:20	9h 10min	2 paradas		1792,27	1752,27		1	1
20		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	10:40	19:50	9h 10min	2 paradas		391,27	421,27	506,27	1	1
21		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	10:30	19:50	9h 20min	1 parada		698,27	733,27	843,27	1	1
22		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	14:25	01:10	10h 45min	2 paradas		254,27	279,27	506,27	1	1
23		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	14:25	01:10	10h 45min	2 paradas		267,27	292,27	506,27	1	1
24		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	14:25	01:10	10h 45min	2 paradas		285,21	310,21	537,21	1	1
25		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	10:40	22:10	11h 30min	2 paradas		285,21	310,21	537,21	1	1
26		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	10:40	22:10	11h 30min	2 paradas		316,48	341,48	568,48	1	1
27		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	10:40	22:10	11h 30min	2 paradas		391,27	421,27	506,27	1	1
28		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	10:40	01:10	14h 30min	2 paradas		316,48	341,48	568,48	1	1
29		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	VIX	CWB	06:10	22:30	16h 20min	2 paradas		1823,21	1783,21		1	1
30		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	CWB	VIX	10:40	14:10	3h 30min	1 parada		391,27	421,27	506,27	2	1
31		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	CWB	VIX	08:00	11:35	3h 35min	1 parada		254,27	279,27	506,27	2	1
32		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	CWB	VIX	05:25	09:55	4h 30min	1 parada		254,27	279,27	470,27	2	1
33		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	CWB	VIX	11:20	15:50	4h 30min	1 parada		267,27	292,27	506,27	2	1
34		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	CWB	VIX	17:00	21:50	4h 50min	1 parada		435,27	465,27	550,27	2	1
35		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	CWB	VIX	10:15	15:50	5h 35min	1 parada		267,27	292,27	506,27	2	1
36		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	CWB	VIX	14:40	20:30	5h 50min	2 paradas		637,27	672,27	772,27	2	1
37		24/09/2018	14:15:35	ingrid.oliveira	24/11/201	02/12/201	CWB	VIX	15:45	21:50	6h 5min	1 parada		435,27	465,27	550,27	2	1

Projeto-Piloto 1: Aplicação Web

Passagens Aéreas | Consulta | Dados

Origem

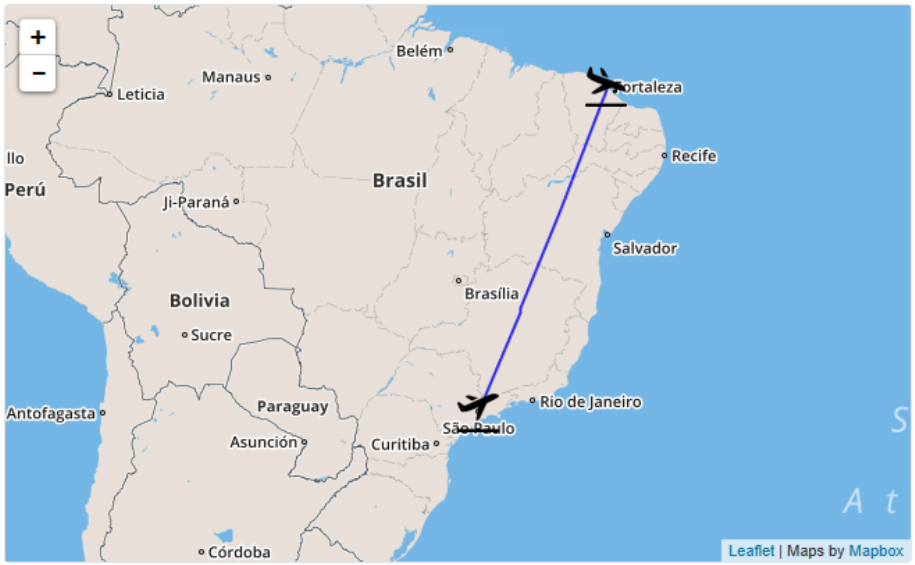
Sao Paulo

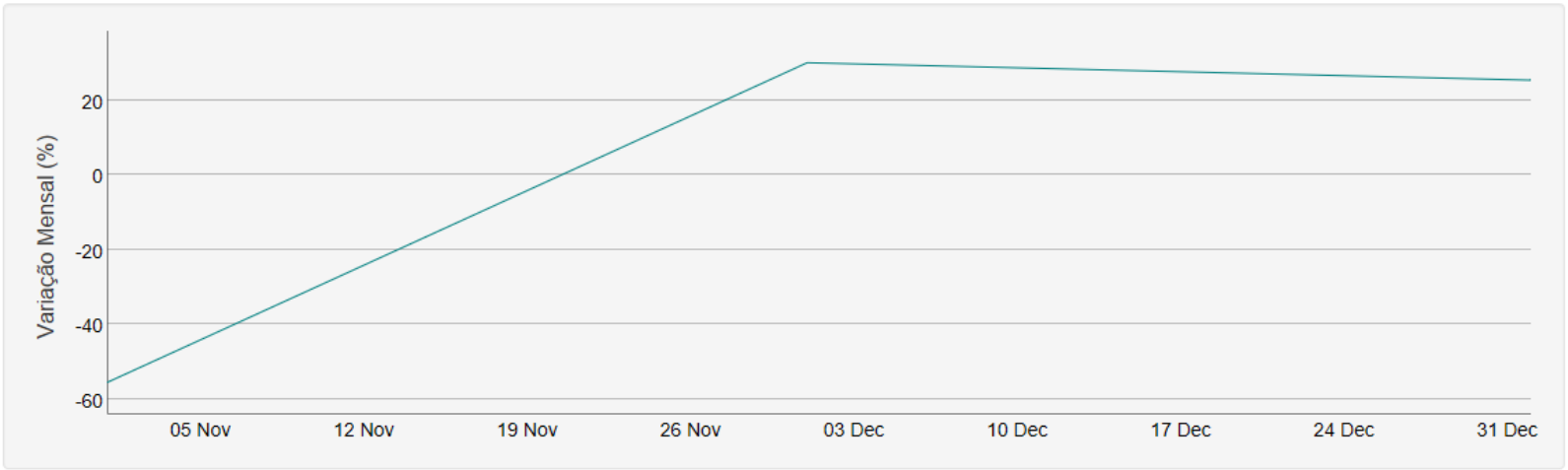
Destino

Fortaleza

Informações Gerais

Origem: Sao Paulo
Destino: Fortaleza
Valor médio (R$): 2040.87
Variação Mensal (%): 25.37 (Jan/18)





Projeto-Piloto 1: Aplicação Web

Passagens Aéreas | Consulta | Dados

Origem: Sao Paulo

Destino: Fortaleza

Periodicidade: Semanal

- Mensal
- Semanal

Show 25 entries

Search:

Data Ida	Data Volta	Valor em reais
07/10/2017	15/10/2017	2973.07
04/11/2017	12/11/2017	1320.51
02/12/2017	10/12/2017	1717.14
06/01/2018	14/01/2018	2152.75
Data Ida	Data Volta	Valor em reais

Showing 1 to 4 of 4 entries

Previous 1 Next

Conclusões preliminares

- Redução drástica no tempo de coleta
- Aumento no número de informações obtidas
- Poucas diferenças nos valores coletados manualmente
- Controle e registro completo sobre o processo de coleta

Cesta fixa

Índices de preços ao consumidor são baseados em cesta fixa. Portanto, os mesmo produtos devem ser comparados período a período.

Substituições de produtos

Um mesmo produto numa mesma loja deve ser coletado em períodos subsequentes. Método modelo pareado

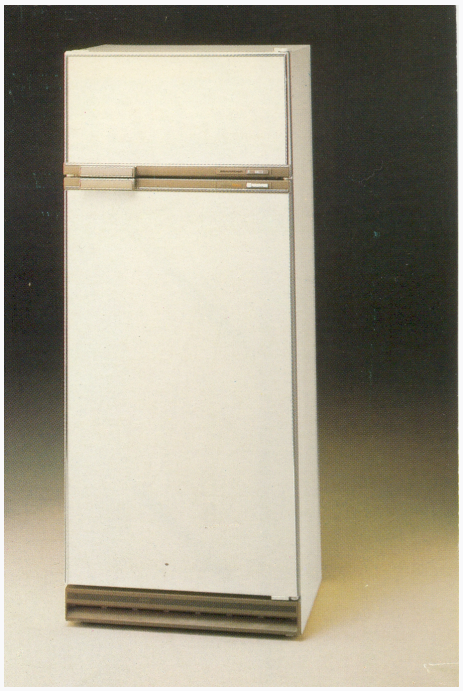
Mês: t – 1 t

O mercado é dinâmico e produtos podem entrar e sair de circulação. Consequentemente, substituições são necessárias.

A substituição pode significar a inclusão de um produto com características e qualidades diferentes. Mudança na utilidade.

Comparabilidade entre os produtos

A mudança de qualidade dos eletrodomésticos, por exemplo, impactam na qualidade de vida dos consumidores.



Não podem/devem ter seus preços comparados de forma direta. Método modelo pareado.

Comparabilidade entre os produtos

Suponhamos que a geladeira m não é mais encontrada no mercado e a geladeira n é a substituta.

Item/period	t	$t+1$	$t+2$	$t+3$	$t+4$
l	p_l^t	p_l^{t+1}	p_l^{t+2}	p_l^{t+3}	p_l^{t+4}
m	p_m^t	p_m^{t+1}	p_m^{t+2}		
n				p_n^{t+3}	p_n^{t+4}

$$R_n^{t+3,t+2} = p_n^{t+3}/p_m^{t+2}$$

Viés! As geladeiras não são comparáveis porque possuem atributos diferentes. Não estaríamos medindo variação pura de preços.

Atribuir valor aos atributos

Como medir a mudança na qualidade entre os produtos m e n já que quase nunca dispomos do preço de cada atributo isoladamente para torná-los comparáveis?

O mercado, em geral, não informa o valor de cada atributo.

Para isso, o manual internacional de índices de preços ao consumidor recomendam os modelos hedônicos.

Modelos hedônicos:

1. Patching quando as substituições não são rotineiras.
2. Índices hedônicos quando as substituições são rotineiras. (ex: carros usados)

Patching

Regressão múltipla entre os preços e as características (z) dos itens.

$$Price = \beta_0 \beta_1^{z_1} \beta_2^{z_2} \beta_3^{z_3} \ldots . \beta_n^{z_n} \varepsilon$$

$$\ln \text{Price} = \ln \beta_0 + z_1 \ln \beta_1 + z_2 \ln \beta_2 + z_3 \ln \beta_3 + \ldots . . z_n \ln \beta_n + \ln \varepsilon$$

Através da regressão, é possível atribuir valor/variação de preço para cada característica significativa z.

A partir do modelo, podemos imputar o preço estimado para o novo produto n no mês anterior t + 2.

Item/period	t	t+1	t+2	t+3	t+4
l	p_l^t	p_l^{t+1}	p_l^{t+2}	p_l^{t+3}	p_l^{t+4}
m	p_m^t	p_m^{t+1}	p_m^{t+2}		
n			$\hat{p}_n^{t+2}$	p_n^{t+3}	p_n^{t+4}

$$R_n^{t+3,t+2} = p_n^{t+3} / \hat{p}_n^{t+2}$$

Patching

Dificuldades da aplicação desse método:

- Coletar informações sobre características dos produtos é custoso porque exige mais do entrevistador e incomoda o respondente.

- Controle na garantia das informações de atributos.

Webscraping

Webscraping para coleta de características:

- Barato
- Controlável
- Eficiente.

Não constitui acesso massivo às páginas e, assim, evita possíveis bloqueios.

Webscraping

Webscraping utilizando o R para coleta de preços e características para geladeiras.

Exemplo de atributos para geladeira.

Marca		Brastemp
Capacidade Total	? O que é isso?	443 Litros
Capacidade do Refrigerador		318 Litros
Capacidade do Congelador/Freezer		125 Litros
Tipo de Porta	? O que é isso?	Inverse
Dispenser Externo		Não possui
Tipo de Controle	? O que é isso?	Painel Eletrônico
Acabamento Externo da Porta		Inox

Webscraping

A coleta dessas informações por webscraping utilizando o software R dura cerca de 1 minuto.

São coletados aproximadamente:

- 1900 preços
- 160 produtos
- 14 atributos

Regressão log-lin

Mínimos quadrados ordinários.

Patching - Resultados

$$ln(Preco) = \beta_0^* + \beta_1^* \cdot \text{AEP} + \beta_2^* \cdot \text{CapT} + \beta_3^* \cdot \text{DExt} + \beta_4^* \cdot \text{Mar} + \beta_5^* \cdot \text{Deg} + \beta_6^* \cdot \text{TiPor}$$

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    6.5861484  0.0790365  83.330  < 2e-16 ***
Acabamento.Externo.da.PortaInox 0.1224168  0.0235635   5.195 7.92e-07 ***
Capacidade.Total               0.0021376  0.0001743  12.266  < 2e-16 ***
Dispenser.ExternoSim           0.2936710  0.0626698   4.686 7.06e-06 ***
MarcaConsul                   -0.1291417  0.0360732  -3.580 0.000488 ***
MarcaElectrolux               -0.0278870  0.0300748  -0.927 0.355553
MarcaPanasonic                -0.0554155  0.0391107  -1.417 0.158964
MarcaSamsung                   0.2369567  0.0439374   5.393 3.26e-07 ***
Tipo.de.DegeloSim              0.1426981  0.0393934   3.622 0.000421 ***
Tipo.de.PortaDuplex            0.2314976  0.0513358   4.509 1.46e-05 ***
Tipo.de.PortaFrench Door Inverse 0.7145420  0.0820120   8.713 1.36e-14 ***
Tipo.de.PortaInverse           0.4331857  0.0646102   6.705 5.95e-10 ***
Tipo.de.PortaSide by Side      0.7460366  0.0974326   7.657 4.22e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1259 on 127 degrees of freedom
Multiple R-squared:  0.9347,    Adjusted R-squared:  0.9285
F-statistic: 151.4 on 12 and 127 DF,  p-value: < 2.2e-16
```

Patching - Testes

```
        Shapiro-Wilk normality test

data:  modelofinal$residuals
W = 0.98536, p-value = 0.1424


        Breusch-Godfrey test for serial correlation of order up to 1

data:  modelofinal
LM test = 2.8367, df = 1, p-value = 0.09213


        studentized Breusch-Pagan test

data:  modelofinal
BP = 21.02, df = 12, p-value = 0.05009
```

	GVIF	Df	GVIF^(1/(2*Df))
Acabamento.Externo.da.Porta	1.220338	1	1.104689
Capacidade.Total	2.037771	1	1.427505
Dispenser.Externo	2.301904	1	1.517203
Marca	1.680467	4	1.067035
Tipo.de.Degelo	1.311825	1	1.145349
Tipo.de.Porta	4.095833	4	1.192732

Considerações finais e próximos passos

A adoção do patching utilizando webscraping é promissor.

Índices hedônicos ainda requerem mais estudos e aprofundamentos.

A técnica de webscraping também permite mapear produtos que estejam perdendo/ganhando representatividade no mercado.

Obrigado

tiago.dantas@ibge.gov.br

lincoln.silva@ibge.gov.br