Perfil de usuários e viagens do aplicativo BlaBlaCar na região de Lavras, MG: um estudo utilizando a técnica *web scraping*

Lucas B. Patrício[1], Julia R. de Oliveira[2], Paula R. Santos[3], Rodrigo Marçal Gandia[4], Fábio Antonialli[5], Izabela R. Cardoso de Oliveira[6]

1. Introdução

Nos últimos anos, os estatísticos têm enfrentado novos desafios no que diz respeito à gestão, análise de dados e geração de informação. A era de *data science* exige desse e de outros profissionais um perfil diferenciado, que inclui, por exemplo, a capacidade de lidar com bases de dados gigantescas. Segundo estudo recente da *Business Software Alliance (BSA)*, realizada em 2015, as pessoas criam 2,5 quintilhões de bytes de dados, ou quase 2,3 milhões de terabytes de dados todos os dias, tanto que 90% dos dados atualmente no mundo foram criados apenas nos últimos dois anos. Além disso, muito desse fluxo consiste em dados sobre informações semelhantes, gerando enormes conjuntos de dados com bilhões de observações. O termo *big data* refere-se não apenas ao dilúvio de dados gerados, mas também ao exorbitante tamanho dos conjuntos de dados que vêm de todos os lugares: sensores físicos, sensores humanos, como as redes sociais, GPS e aplicativos de telefone celular, entre outros.

A infinidade de dados gerados juntamente com o tamanho dos conjuntos de dados criam desafios e oportunidades para os cientistas de dados. Eles são uma nova geração de especialistas responsáveis pelo estudo disciplinado dos dados, e informações inerentes ao negócio e todas as visões relacionadas a um determinado assunto. É uma ciência que estuda as informações, seu processo de captura, transformação, geração e, posteriormente, análise de dados. O processo envolve, dentre outras etapas, a transformação de dados em informações, formulação de problemas e escolha de métodos estatísticos apropriados.

Para realizar estas tarefas e conseguir gerar insights interessantes, o cientista de dados precisa reunir habilidades como *data storytelling*, aprendizado de máquina (*Machine learning)* (James et al., 2013), mineração de dados (*data mining)* (Aggarwal, 2016) e *web scraping*. Esse último, que é o foco desse trabalho, consiste na raspagem de dados de sites da internet para, posteriormente, transformá-los em um formato mais simples e analisá-los. Essa técnica tem sido usada em diversas áreas do conhecimento como, por exemplo, na extração de dados do site Tripadvisor como suporte na elaboração de indicadores do turismo de Minas Gerais (Oliveira e Porto, 2016).

Aliado ao grande volume de dados, a ascensão dos smartphones trouxe também novos formatos de negócios baseados em aplicativos. A ideia de compartilhamento passa a combinar com

[1]Graduando em ABI/Engenharia Química e estudante de Iniciação Científica no Departamento de Estatística. Universidade Federal de Lavras. lucas.patricio@estudante.ufla.br
[2]Graduanda em ABI/Engenharia Química e estudante de Iniciação Científica no Departamento de Estatística. Universidade Federal de Lavras. juliiaro@hotmail.com
[3]Mestranda em Estatística e Experimentação Agropecuária. Universidade Federal de Lavras. paullasant_s@hotmail.com
[4]Doutorando em Administração. Universidade Federal de Lavras/CentraleSupélec. romgandia@gmail.com
[5]Pós-doutorando em Engenharia Industrial. CentraleSupélec. fantonialli@gmail.com
[6]Professora Adjunta. Departamento de Estatística. Universidade Federal de Lavras. Izabela.oliveira@ufla.br

os ideais das novas gerações (Y, milenials) e abrange, dentre outros, os setores de hospedagem e transportes. Em relação a esse último, o atual paradigma da mobilidade, baseado em combustíveis fosseis e no transporte individual, está atingindo seu limite, ambiental, econômico e social (Fournier, 2017). Deste modo, a posse e utilização de automóveis privados para satisfazer necessidades individuais se configuram como um problema cada vez mais desafiador para a mobilidade.

Uma das soluções na busca pela melhoria dos serviços de mobilidade é o compartilhamento de viagens (*carpooling*). Com grande potencial em termos de acessibilidade e sustentabilidade, tal compartilhamento pode reduzir o número total de quilômetros percorridos e mitigar o congestionamento, custos de viagens, consumo de combustível e emissões veiculares (Bischoff et al., 2018; Rayle et al., 2014; Santi et al., 2014). Além disso, esses serviços compartilhados geralmente oferecem tarifas mais baixas aos clientes, uma vez que os passageiros se comprometem a compartilhar o veículo com outros usuários e em troca, pagam menos pelas viagens (Farhan e Chen, 2018; Bischoff et al., 2018).

Neste sentido, plataformas de negócios que apresentam propostas de conectar motoristas dispostos a compartilhar viagens com passageiros cujas rotas são semelhantes, vêm ganhando destaque dentre os serviços de mobilidade. A pioneira nesta modalidade de negócios foi a francesa Blablacar; fundada em 2006 em Paris. A empresa conta hoje com uma base de 70 milhões de usuários em 22 países (Vleugels, 2019). A empresa não possui nenhum veículo, é simplesmente uma corretora que recebe uma comissão de 12% de cada reserva (Scott, 2016.). Ademais, outros modelos de negócios semelhantes vêm também surgindo de forma a ir de encontro a essas demandas de mobilidade compartilhada, tais como Waze Rider e Scoop (Bischoff et al., 2018; Clewlow e Mishra, 2017).

Esse estudo tem como principal objetivo realizar a coleta de dados da plataforma BlaBlaCar Brasil usando a técnica de *web scraping*. A coleta será feita considerando Lavras, MG como destino de viagem e em um período de tempo pré-determinado. Como objetivo secundário tem-se resumir, de forma didática, o procedimento utilizado e disponibilizar os códigos R utilizados nesse trabalho para que interessados possam utilizar essa técnica em outros contextos.

2. Metodologia

2.1 O objeto de estudo

Nesse trabalho a técnica *web scraping* foi aplicada ao site BlaBlaCar (https://www.blablacar.com.br/), uma plataforma de caronas de longa distância, criada em 2006 como uma startup francesa. A empresa está presente em diversos países, com usuários de diferentes faixas etárias. A proposta é conectar motoristas e passageiros por meio da plataforma onde os motoristas oferecem caronas para dividir o preço das viagens sem obter lucro. Cada usuário possui um perfil com informações de nome, idade, tempo de uso do aplicativo, avaliações recebidas e suas restrições como viajar com animais, com fumantes, entre outras. O passageiro faz a busca por data e trecho de interesse e todas as opções de viagens disponíveis são apresentadas, incluindo trajeto completo do motorista e preço cobrado.

O estudo foi feito considerando todas as viagens com destino a Lavras, MG, no período de 09 à 14 de março de 2019. Esse período foi escolhido por anteceder o início do semestre letivo na universidade pública situada na cidade. A escolha por essa cidade foi motivada pela forte cultura de

compartilhamento (carona) já existente entre estudantes dessa universidade, seja para deslocamento para suas cidades de origem ou internamente, até a instituição de ensino. Nesse último caso os usuários contam com o apoio da própria instituição, que estabelece pontos de caronas para regiões específicas da cidade. Vale destacar que o procedimento utilizado nesse trabalho pode ser feito para quaisquer outros trechos e, inclusive, ser adaptado para sites com padrão similar ao do BlaBlaCar.

2.2 Procedimento de coleta de dados

Para a coleta de dados do site BlaBlaCar foi utilizado o programa (R Core Team, 2018) e os pacotes xml2 (Wickham; Hester; Ooms, 2018) e rvest (Wickham, 2016). O primeiro tem a finalidade de estruturar arquivos HTML ou XML de forma eficiente, tornando possível a obtenção de *tags* e seus atributos dentro de um arquivo. O segundo é escrito sobre o pacote anterior e o httr (Wickham, 2018), responsável por realizar requisições web para obtenção das páginas de interesse, e por isso eleva ainda mais o nível de especialização para a "raspagem" de dados.

Dentre os procedimentos para a coleta, especifica-se a página da web a ser analisada e utiliza-se a função read_html para ler todo o texto do corpo do arquivo html. Dessa maneira, todos os dados da página são coletados. Para especificar os elementos mais relevantes é necessário usar as funções html_nodes e html_text que são responsáveis, respectivamente, por extrair partes de documentos HTML usando seletores XPath e css e extrair atributos, texto e nome de *tag* do html. Por fim, a saída é processada e convertida para o formato data frame.

Os dados coletados são, então, analisados utilizando desde estatísticas descritivas até modelos mais complexos ou técnicas de aprendizado de máquina.

3. Resultados parciais

Os primeiros resultados obtidos com a aplicação da técnica *web scraping* no site BlaBlaCar são apresentados na sequência. Pode-se observar que dos 258 perfis analisados somente 18,6% são de usuárias femininas enquanto 81,4% são de usuários masculinos. Pode-se observar, também, que as idades variam de 19 a 72 anos enquanto a idade média é de 34 anos.

Foi possível constatar também o fato de que, dentre os locais de origem das caronas, as mais frequentes são Belo Horizonte (19,5%) e São Paulo (15,2%), sendo que os preços médios das caronas nestes trajetos são de 43,00 e 63,43 reais, respectivamente. Além disto, as restrições mais comuns para passageiros são a proibição de fumar e de transportar animais durante as viagens.

4. Conclusões

A técnica de *web scraping* utilizando o programa R foi aplicada com sucesso ao site de compartilhamento de viagens BlaBlaCar. Pela característica do estudo, o banco de dados obtido não possui grande dimensão, mas essa técnica também seria bastante útil nessa situação. Na presente aplicação, foi possível coletar informações importantes sobre o perfil dos condutores cadastrados no aplicativo e das viagens para o trecho estudado.

Agradecimentos

A autora Paula R. Santos agradece a CAPES pela bolsa de mestrado. Os autores agradecem a FAPEMIG pelo apoio financeiro para participação no congresso.

Referências Bibliográficas

AGGARWAL, C. Data Mining: The Textbook. Softcover reprint of the original 1st ed. 2015. Springer, 2016

BISCHOFF, J.; KADDOURA, I.; MACIEJEWSKI, M.; NAGEL, K. Simulation-based optimization of service areas for pooled ride-hailing operators. *Procedia Computer Science*, *130*, 816-823. 2018.

CLEWLOW, R. R.; MISHRA, G. S. *Disruptive transportation: The adoption, utilization, and impacts of ride-hailing in the United States*. (UCD - Institute of Transportation Studies research report). 2017. Retrieved on August 29, 2018, from: <http://www.trb.org/Main/Blurbs/176762.aspx>.

FARHAN, J.; CHEN, T. D. Impact of ridesharing on operational efficiency of shared autonomous electric vehicle fleet. *TransportationResearchPart C: Emerging Technologies*, *93*, 310-321. 2018.

FOURNIER G. The new mobility paradigm. Transformation of value chain and value proposition through innovations. In: ATTIAS, D. (Org.). *The Automobile Revolution: Towards a New Electro-Mobility Paradigm.* (1st ed.). Gewerbestrasse (Switzerland): Springer International Publishing, p. 21-47. 2017.

WICKHAM, H.; HESTER, J.; OOMS, J. xml2: Parse XML. 2018. URL https://CRAN.R-project.org/package=xml2.

WICKHAM, H. httr: Tools for Working with URLs and HTTP. 2018. URL https://CRAN.R-project.org/package=httr.

WICKHAM, H. rvest: Easily Harvest (Scrape) Web Pages. 2016. URL https://CRAN.R-project.org/package=rvest.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An introduction to Statistical Learning with applications in R. Springer. 2013. 426p.

OLIVEIRA, R. A.; PORTO, R. M. A. B. Extração de dados do site Tripadvisor como suporte na elaboração de indicadores do turismo de Minas Gerais: uma iniciativa em Big Data. *Pesquisa Brasileira em Ciência da Informação e Biblioteconomia*, João Pessoa, v.11, n.2, p. 026-037, 2016.

RAYLE, L., DAI, D., CHAN, N., CERVERO, R., SHAHEEN, S. Just a better taxi? A survey-based comparison of taxis, transit, and Ridesourcing services in San Francisco. *TransportPolicy*, *45*, 168-178. 2016.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018. URL http://www.R-project.org/.

SANTI, P.; RESTA, G.; SZELL, M.; SOBOLEVSKY, S.; STROGATZ, S. H.; RATTI, C. Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(37), 13290- 13294. 2014.

VLEUGELS, A. "Why French unicorn BlaBlaCar still believes in 'done is better than perfect'". Fundamentals | The Next Web. 2019. Acessado em 11 de março de 2019, disponível em; <https://thenextweb.com/adobe-fundamentals/2019/02/19/why-french-unicorn-blablacar-still-believes-in-done-is-better-than-perfect/.