Estudos Avançados em *R*: *Text mining*

Alinne de Carvalho Veiga (Orientadora) *
Renata Souza Bueno (Coorientadora) †

junho/2017

Resumo

O objetivo deste projeto é aprofundar os estudos de ferramentas existentes no software *R* para a extração, sumarização e análise de informações textuais de páginas da internet. Atualmente muitas pessoas e empresas usam a internet para expressarem suas opiniões. Com isso a quantidade de dados textuais que existe na rede de computadores é enorme e esses dados podem ser usados a favor de melhorias em vários aspectos. Sendo assim, a informação textual se torna um caminho pelo qual se pode ter uma informação valiosa. O avanço do conhecimento de ferramentas de análise textuais permitirá uma ampliação da capacidade de análise dos dados desta natureza. Além disso, o conhecimento adquirido durante a implementação deste projeto permitirá a criação de repositórios de documentação para uso nas esferas de ensino e pesquisa da ENCE.

Palavras-chave: *text mining; software R; web scrapping* .

1 Introdução

O sofware *R* é uma linguagem e um ambiente de código aberto para computação estatística e gráfica que fornece aos seus usuários uma ampla variedade de técnicas para realização de cálculos, manipulação e visualização de dados, além de modelagem estatística (R Core Team, 2014). Ainda segundo seus fundadores, o *R* permite aos seus usuários criar novas funcionalidades por meio de uma linguagem simples e integrada com outras linguagens, como C, C++ e Fortran.

Uma das grandes vantagens do uso do *R* para análise de dados é a sua possibilidade de extensão via *pacotes*, grande parte desenvolvida por pesquisadores e usuários do *R*, os quais permitem um link direto entre os mais recentes avanços em tecnologia, modelagem e computação e os usuários desta plataforma. O grande repositório online de pacotes do *R* é conhecido como CRAN (*Comprehensive R Archive Network*) e é mantido por uma rede de servidores espalhados pelo mundo.

Este projeto tem o objetivo de aprofundar os estudos das ferramentas existentes *webs crapping* e *text mining* utilizando o software *R* através do *RStudio*. *Web scraping* é a

*Doutora em Estatística Social - Universidade de Southampton (alinne.veiga@ibge.gov.br)
†Doutora em Estatística - UFRJ (renata.bueno@ibge.gov.br)

técnica desenvolvida para extrair dados de páginas da internet. Essa técnica então permite a criação de arquivos de dados para que sejam analisados no *R*. É uma espécie de mineração de dados, sendo que esses dados estão em web-sites, blogs, etc.. No *R*, são utilizados pacotes como `rvest`. Acredita-se que essa técnica é de grande importância para a era que estamos vivendo - Big Data - onde a cada dia existem mais dados e informações disponíveis online. *Text mining* é uma técnica que permite análise de informações, quando não se parte de um arquivo de dados retangular, com linhas representando unidades de análise e colunas representando variáveis. É uma técnica que analisa metadados considerando textos como `data.frame` possibilitando então sua visualização e sumarização.

Este projeto faz parte do Núcleo de Estudos Avançados em *R* (NeaR) que consiste em um grupo de pesquisa recém formado na Escola Nacional de Ciências Estatísticas (ENCE). O NeaR tem como finalidade aprofundar o domínio das ferramentas mais atuais para manipulação, análise e visualização de dados utilizando o software *R*. Atualmente o NeaR é composto por 6 professores da graduação em Estatística e, a partir do segundo semestre de 2017, pretende envolver alunos de graduação por meio da participação no Programa Institucional de Bolsas de Iniciação Científica (PIBIC) da ENCE. Pela relevância e atualidade dos estudos a serem desenvolvidos no âmbito do NeaR, em médio e longo prazo espera-se uma melhor qualificação do corpo discente e o desenvolvimento paralelo de uma documentação destas ferramentas mais atuais de análise de dados para melhoria das atividades de ensino e pesquisa da ENCE.

Além da linha de pesquisa em *Text Mining*, o NeaR também atua no aprofundamento de conhecimentos nas áreas de integraçao de bases e sistemas e Visualização de Dados utilizando o software *R*.

2 Métodos

Para o alcance dos objetivos do projeto o aluno bolsista estudará os principais pacotes para *webscraping* `rvest`, e para *text mining* `tm` e `tidytext` (Silge e Robinson, 2017) através do *RStudio* (RStudio Team, 2015). Outros pacotes como `ggplot2`, `SnowballC`, `plyr`, `dplyr`, etc. também deverão ser estudados.

Adicionalmente, casos reais discutidos no âmbito do NeaR poderão ser utilizados para implementação de códigos e criação de exemplos para posterior documentação. O projeto prevê ainda a troca de conhecimentos entre participantes do NeaR, por meio da realização de Seminários e encontros periódicos na ENCE.

Tabela 1: Plano de Trabalho

Etapas	Prazos
Revisão e aprofundamento na linguagem *R*	Ago/2017 a Out/2017
Levantamento de pacotes para *web scraping e text mining* no *R*	Nov/2017 a Dez/2017
Implementação de *web scraping e text mining*	Jan/2018 a Mai/2018
Documentação dos avanços e técnicas estudadas	Mar/2018 a Jun/2018
Preparação de relatório de atividades	Jun/2018

3 Plano de trabalho e requisitos necessários

O plano de trabalho para o aluno bolsista do projeto está descrito na Tabela 1. Em relação aos requisitos do aluno bolsista, é necessário apenas que este possua interesse em programação, goste de trabalhar em equipe e que aceite desafios. Além das exigências ordinárias envolvidas na participação de um aluno no Projeto PIBIC (elaboração de relatório e apresentação oral na Jornada de Iniciação Científica), este projeto ainda pretende envolver o aluno bolsista na preparação de trabalho para submissão em Congresso(s) de Estatística.

Referências

[1] Silge, J. and Robinson, D. (2017) *Text Mining with R. A Tidy Approach.* http://tidytextmining.com/.

[2] R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/.

[3] RStudio Team (2015). *RStudio: Integrated Development for R*. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.