Text Mining: Descrição da utilização do pacote Rfacebook

Carolina Peçaibes de Oliveira[1]

Guilherme Pumi[2]

Introdução

Com o advento da internet, tornou-se disponível uma grande quantidade de informações relevantes em forma de texto, e surgiu a demanda por processos e algoritmos capazes de obter, organizar, classificar, depurar e analisar esses dados não estruturados (sendo essas etapas que compõe o processo de *text mining*). Há interesse especial em entender os dados pertinentes ao comportamento do consumidor, sendo estes frequentemente expressos através das redes sociais. Com isso em mente, trazemos aqui uma introdução ao processo de obtenção de dados para aplicação posterior de *text mining*, utilizando o pacote *RFacebook* do software R.

Metodologia

Para a realização desse trabalho, executamos as etapas necessárias de autenticação para extrair dados utilizando o RFacebook. Posteriormente, utilizamos algumas funções do RFacebook aplicáveis à páginas públicas na página oficial do jornal Zero Hora (getPage e getPost) e outras aplicáveis à perfis pessoais na página pessoal da autora (getFriends, getLikes) no dia 05/10/2015, obtendo quatro base de dados com informações em forma de texto.

Desenvolvimento

Realizamos a instalação do pacote no software R usando os comandos apropriados. Em seguida, a documentação do pacote RFacebook apresenta como primeira função disponível o comando *fbOAuth* que exige as informações de *App ID* e *App Secret* para autenticação do acesso ao facebook através do R, informando que esses dados estão disponíveis no endereço www.developers.facebook.com/apps.Porém a partir dessa página não há instruções de como

[1] UFRGS - Universidade Federal do Rio Grande do Sul. Email: carolpecaibes@gmail.com

[2] UFRGS - Universidade Federal do Rio Grande do Sul. Email: guipumi@gmail.com

obter esses dados e o que eles são.

Verificamos através de pesquisa no FAQ do Facebook que esses dados provém da criação de um aplicativo, e que o usuário precisa registrar-se como desenvolvedor de apps utilizando sua conta pessoal do Facebook, registrar a criação de aplicativo. Isto feito, as informações de ID (*App ID*) e Senha de Acesso (*App Secret*) aprecem disponíveis, e podem ser inseridas no comando do R que deve ser rodado nesse momento.

Esse comando nos retorna uma URL que deve ser inserida nas informações de registro do aplicativo no Facebook. Com isso está completa a autenticação de acesso.

Para executar os demais comandos do pacote, é exigida a informação do *token*, que está disponível no aplicativo criado e autenticado.

Realizamos a extração de informações do feed de notícias a página oficial da *Zero Hora* usando a função **getPage**. Por se tratar de uma página pública, temos acesso aos seus dados completos. Informamos o número de postagens que queremos extrair e obtemos uma base de dados completa em forma de lista, incluindo o texto do post, o tipo de post, o link compartilhado (se houver), a data e hora da postagem, número de curtidas e número de comentários. Com a função **getPost** extraímos informações detalhadas de cada post, como quantidade de comentários e curtidas, nome do perfil dos usuários que comentaram ou curtiram o post, quantidade de curtidas de cada comentário, e data e hora dos mesmos.

Executamos alguns testes de extração e verificamos que não há um limite de quantidade de postagens e comentários que podemos obter. Além disso, por tratar-se de uma página pública, independentemente do tipo de configuração de privacidade do usuário do facebook, seus comentários e curtidas ficam disponíveis para extração e, posteriormente, análise.

A partir da página pessoal da autora, executamos a função **getFriends** para obter uma listagem do nome dos seus amigos na rede social, além das outras informações que o comando fornece como data de nascimento, gênero, profissão informada e escolaridade. Com a função **getLikes** obtemos a relação de páginas curtidas por aquele perfil. Verificamos que a primeira função retornou o nome apenas de dois amigos do perfil, o que é incorreto segundo a conferência da página original. De acordo com as configurações do facebook, apenas usuários que autorizam a visualização irrestrita de suas informações tem seus dados disponíveis por esse procedimento. Também não é possível extrair a lista de amigos de um perfil com visualização restrita, tornando as informações escassas para uma análise de mercado, por exemplo.

O output de cada um desses quatr comandos nos traz uma lista do R, que convertemos para o formato de matriz e exportamos para o excel, para fins de visualização. Cada tipo de variável de texto fica dividida por colunas, facilitando a análise.

Partindo para a etapa de conferência e limpeza da base de dados, focando na análise de *text mining* que queremos aplicar, verificamos na literatura que a exploração de dados desse assunto comumente parte da identificação de palavras-chave e listagem de termos mais frequentes. Dependendo do foco do trabalho, podemos querer que termos com grafias incorretas sejam computados na mesma contagem dos termos com grafia correta, ou podemos querer ignorar palavras de pouco interesse. Para esses casos, os dados não estruturados obtidos da rede social apresentam dificuldades, por ser um espaço em que a expressão de acordo com a língua culta não é mandatória.

Concluindo-se que queremos encontrar a frequência de palavras agrupando as diferentes grafias disponíveis, temos que propor um método para execução desse procedimento. Esse processo acaba dividido em duas abordagens aplicadas juntamente: a listagem de termos equivalentes, de acordo com o conhecimento da norma culta, da forma de escrita na rede social, e do conhecimento específico do assunto e do público de interesse; e a listagem de termos equivalentes de acordo com a análise exploratória de dados. Ambos exigem a conferência manual do texto, uma vez que um algoritmo padrão não é capaz de captar todas as nuances de variação dos dados possível, e apesar de dispendioso é necessário para termos uma base de dados de boa qualidadade.

Conclusões

O pacote*RFacebook*apresenta um conjunto de funções úteis para a extração de dados, embora todas as etapas necessárias para o funcionamento dos comandos do pacote não estejam descritas no documento que descreve sua utilização. Ele também se limita a extração de dados, não apresentando alternativas para análise dos mesmos, e sem corrigir qualquer dado necessário para correta análise posterior. Também as funções não podem contornar as configurações de confidencialidade da rede social, que não permite o acesso às informações de seus usuários sem autorização dos mesmos, o que limita a utilização das informações.

Referências

[1] BARBERA, P. *Documentação do pacote 'RFacebook'*. Disponível em: <cran.r-project.org/web/packages/Rfacebook> Acesso em: 10 de outubro de 2015.

[2] Francis, L; Flynn, M. Text Mining Handbook. *Casualty Actuarial Society E-Forum*, Spring 2010 61

[3] Vários autores. Seção de dúvidas frequentes para desenvolvedores de apps no Facebook.