Aplicação de modelos para redes sociais em bases textuais obtidas via *Web Scraping*

Bolsista: Pedro Henrique Sodré Puntel
Orientador: Gustavo da Silva Ferreira

Rio de Janeiro,
Junho 2019

1 Objetivo do Projeto

Este projeto de iniciação científica analisa dados de artistas a fim de permitir uma visualização da sua estrutura de popularidade com base em um ranqueamento semanal produzido pela *Billboard*. Utilizando modelos de espaços latentes para redes sociais, investigou-se como se davam as associações entre os artistas que já ocuparam o referido ranqueamento, dentro de um período considerado. Para tal, dois modelos foram considerados, sendo o primeiro responsável por projetar tal estrutura de relações num sistema de coordenadas de baixa dimensionalidade (Modelo de Escalonamento Multidimensional) e um último cuja finalidade foi modelar a similaridade/dissimilaridade entre os artistas (Modelo de Distâncias Latentes). Os resultados obtidos neste projeto permitiram visualizar e interpretar relações de popularidade implícitas entre os artistas, identificando ainda claros agrupamentos dos mesmos a partir dos padrões latentes observados.

2 Relevância e Contextualização do Tema

A popularização das redes sociais traz consigo a necessidade de avanços na modelagem das relações entre os indivíduos que compõem uma rede. Não obstante, os objetivos das análises de redes sociais são hoje bastante diversos, abrangendo tópicos que variam desde a identificação de *nós* (indivíduos) com papel central, bem como a sumarização e identificação de

grupos subjacentes na rede.

Sob este prisma, o avanço dos modelos estatísticos tem permitido análises a cerca de diversos aspectos de uma única rede, como por exemplo: opiniões, hábitos, desejos e graus de relacionamento (intra e entre grupos) dos atores que a compõem. Tais aspectos são muitas vezes utilizados para a personalização da oferta de produtos e serviços sendo esta uma das possíveis aplicações destes modelos.

Como consequência do avanço computacional, grandes bases de dados textuais encontram-se agora disponíveis na Internet, sendo a maioria destas encontradas de forma não-estruturada. Segundo Hoff *et al.*, 2008, cerca de 90% de toda a informação contida no universo digital era composta por dados não estruturados como textos, imagens e vídeos. Neste sentido, a habilidade de coletar, estruturar e analisar parte dessa grande quantidade de dados é essencial, sendo o principal objeto de estudo durante o início deste projeto.

Para lidar com a não-estruturação das bases de dados textuais, diversas ferramentas de *Web Scraping* foram desenvolvidas nos últimos anos, permitindo novos níveis de integração dos modelos matemáticos e estatísticos para com as informações latentes disponíveis na Internet.

3 Metodologia

3.1 Web Scraping

Segundo Glez-Peña *et al.*, 2013, Web Scraping pode ser definido como um processo sistemático de extração de conteúdos da Internet. Neste processo, um agente de software, um robô, imita a interação de um ser humano com as páginas e servidores da Web. Esse agente é programado para acessar tantos sites da Internet quanto necessário, analisar seus conteúdos para encontrar e extrair dados de interesse e estruturar essa informação conforme desejado.

Para Giannini *et al.*, 2017, o Web Scraping pode ser implementado de três maneiras :

- Ferramentas *Ad-Hoc*: programas de uso geral, implementados em uma linguagem de programação, instruídos para navegar e extrair dados de páginas específicas.
- Automação de navegador: ferramentas que imitam interações sequenciais de um suário através de um navegador para com uma determinada página da Internet, reproduzindo-as de forma automática.
- Ferramentas *point-and-click*: dispositivos capazes de detectar as partes de uma página que contém os dados automaticamente.

As ferramentas *Ad-Hoc* e de automação de navegador constituem uma abordagem mais geral, podendo produzir melhores resultados, uma vez que as propriedades das linguagens de programação muitas vezes auxiliam na coleta e limpeza dos dados. Contudo, é importante mencionar que ambas estão sujeitas a invariável volatilidade de uma página da web, de forma que uma simples alteração da sua estruturação pode comprometer todo o processo.

Para tarefas mais simples e pontuais, as ferramentas *point-and-click* são bastante eficientes por requirirem do usuário um menor conhecimento sobre a estruturação de página. Nas abordagens anteriores, a compreensão de linguagens como *HTML* (*HyperText Markup Language*) e *XML (Extensible Markup Language)* bem como a de ferramentas mais avançadas como expressões regulares e *XPath (XML Path Language)* tornam-se indispensáveis MUNZERT *et al.*, 2014.

Entretanto, em alguns casos, o uso de Web Scraping não é necessário para a coleta dos dados. É prática comum de muitas empresas disponibilizam *APIs (Application Programming Interfaces)*, muitas vezes de forma gratuita, para que usuários ou programas possam acessar seus bancos de dados e informações de maneira muito mais simples, rápida e prática. Porém, usualmente, existe um limite de chamadas que podem ser feitas as *APIs* gratuitas para evitar um congestionamento destes serviços.

3.2 Modelos para Redes Sociais