UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CÂMPUS MEDIANEIRA

MINERAÇÃO DE DADOS APLICADO A REDES SOCIAIS UTILIZANDO A LINGUAGEM DE PROGRAMAÇÃO PYTHON

DATA MINING APPLIED TO SOCIAL NETWORK USING PYTHON AS A PROGRAMMING LANGUAGE

Resumo

O objetivo deste documento é apresentar conceitos sobre mineração de dados aplicado a redes sociais, assim como a utilização da linguagem de programação Python como ferramenta para os processos de data mining.

Palavras-chave: data mining, redes sociais, python.

1. Introdução

Redes sociais se tornaram um termo comum e uma chave fundamental para o estilo de vida moderno. Hoje em dia, a maioria das pessoas, independente da idade, sexo, crença, utilizam uma ou mais redes sociais. A princípio, esses ambientes *online* focavam-se na comunicação, por exemplo; a possibilidade de se comunicar com alguém distante e tornar esse diálogo pessoal, seguro e, de alguma forma, próximo, ajudando na popularização desse tipo de tecnologia. No decorrer dos anos e com o avanço tecnológico, diferentes tipos de redes sociais surgiram com ideias semelhantes ou, até mesmo, extremamente diferentes, não sendo apenas para a comunicação, mas para outros fins como o compartilhamento de mídias, localização, críticas, *mini-blogs*, perguntas e respostas, negócios, profissão, música, artes, venda e troca de produtos, entre outros.

Por causa do grande número de usuários e atividades que estes realizam, dados são gerados o tempo todo. Seja uma interação simples como a utilização de um botão de “curtir” ou então a publicação de um vídeo com os comentários a respeito do mesmo. A todo momento estas ações geram dados e em quantidades gigantescas.

Surge então a necessidade de compreender o que são esses dados, se possuem algum tipo de informação, se apresentam padrões, se é possível utilizá-los para predizer algum comportamento ou até mesmo para tomadas de decisões. Essas questões são respondidas através do uso de técnicas de mineração de dados (*data mining*), aonde conseguem coletar estes dados e aprimorá-los com o intuito de extrair o conhecimento necessário ou a descoberta de informações úteis.

Para este exercício de mineração é necessário então o uso de ferramentas para facilitar ou, até mesmo, automatizar todos os processos que *data mining* possui. A linguagem de programação Python tem se tornado uma excelente escolha nos últimos anos. Segundo Russel (2013), a linguagem Python apresenta uma coleção de bibliotecas que permite trabalhar com dados desde a sua coleta, limpeza e processamento, até a sua apresentação e visualização.

2. Redes sociais

Facebook, *Twitter*, *LinkedIn*, *Google+* e, muito comum entre desenvolvedores, o *GitHub* são exemplos populares de redes sociais. Logo, possuem grande número de usuários e diversas interações que estes realizam a cada momento, gerando uma quantidade gigantesca de dados. Esses dados são informações sobre pessoas, comportamentos, gostos, marcas e vários outros tipos de conteúdo. Devido a diversidade e a vasta quantidade desse tipo de informação, algumas redes sociais as utilizam para o aprimoramento de conteúdo ou, então, para o comércio de dados para empresas, por exemplo; de publicidade e marketing, que fazem a mineração desses dados para encontrar padrões de seus usuários e, assim, conseguir aumentar suas vendas, reduzir riscos e, até mesmo, gerar novas tendências.

Conforme comentado na introdução deste artigo, a evolução tecnológica permite a popularização e também melhores formas de interação e acesso às redes sociais. A grande maioria das empresas de redes sociais disponibilizam meios em que terceiros possam utilizar uma parte dos dados gerados para a construção ou integração de novos serviços e funcionalidades, por exemplo; aplicações que permitem um usuário ter acesso utilizando informações pessoais provenientes de alguma rede social.

Através da utilização de uma API (*Application Programming Interface*) é possível fazer a coleta desses dados. Imielinksi, Virmani, Abdulghani (1996) definem API como um conjunto de padrões de programação e instruções para acesso a um aplicativo de *software* que é baseado na *Web*. A disponibilidade de uma API é permitir que produtos sejam desenvolvidos utilizando os serviços da empresa em questão.

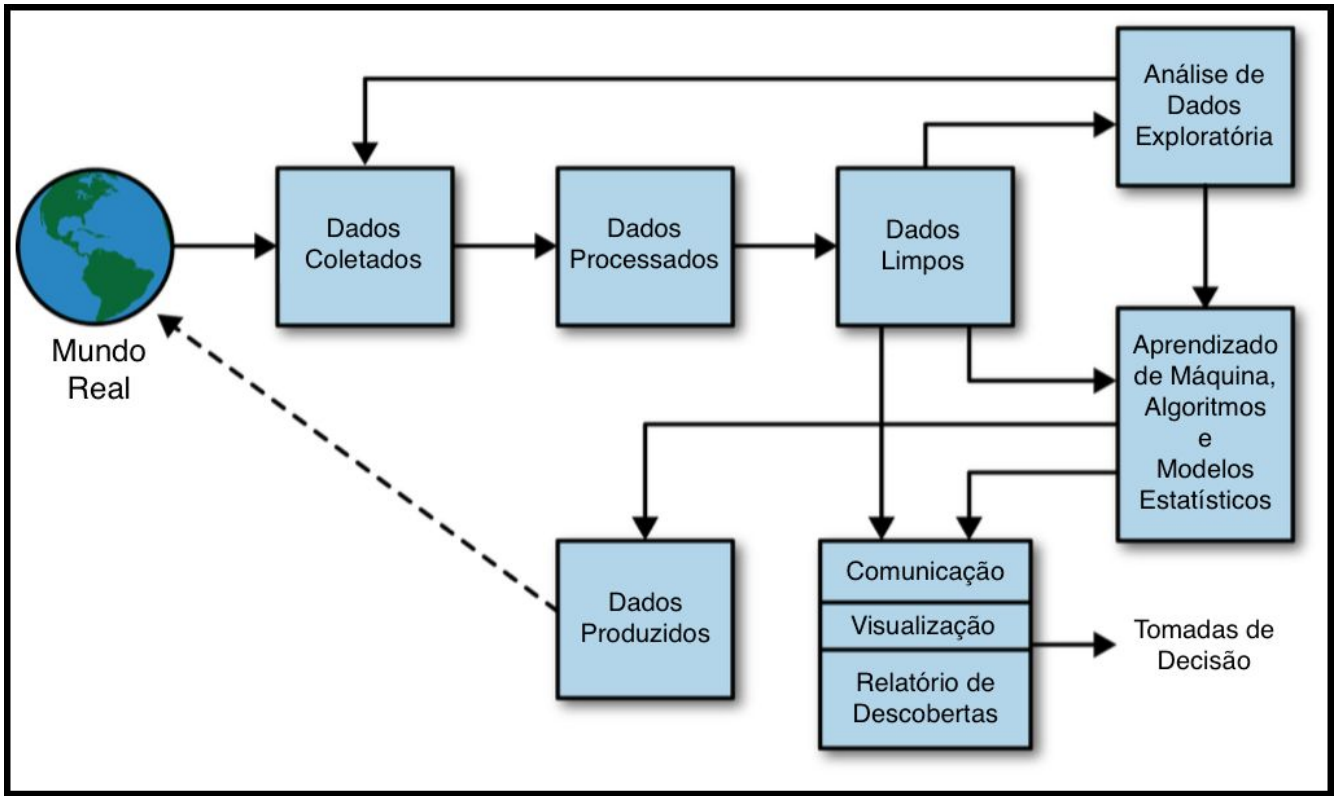
3. Data mining

Data mining consiste no processo de analisar dados em diferentes perspectivas e transformar em informação útil. Hoje em dia, o *data mining* é usado por companhias com grande foco em varejo, finanças, comunicação e marketing, para conseguirem determinar as relações de fatores internos como preço, posição de produto, ou habilidade de recurso humano, e fatores externos como indicadores econômicos, competições e população demográfica de clientes (HAN et al., 2012).

Essa análise de dados consiste em visualizar informações em diferentes maneiras e formas, plotando gráficos e planilhas. Através destes processos novas informações aprarecerão permitindo alguma previsão ou predição desse conteúdo. As observações dos dados levarão a uma reflexão que resultará em possibilidades ou probabilidades concretas para se exercer uma atividade (HAN et al., 2012).

Schutt e O'Neil (2014), explicam o processo de mineração de dados através da Figura 1, onde no primeiro momento existe um Mundo Real em que nele são produzido os dados através das atividades e interações de usuários em um sistema *Web*.
De alguma maneira, por exemplo, através de uma API, esses dados são coletados e então processados. O processo dos dados tem o objetivo de tornar os dados mais limpos e fáceis para serem trabalhados. Os dados, então, são apresentados como estruturas semelhantes a colunas e linhas.
Uma vez que se possui um conjunto de dados limpos e bem estruturados é preciso passar pela análise de dados exploratória para certificar a limpeza dos dados eliminando duplicações, valores em brancos, extremos absurdos e até mesmo dados que não foram catalogados ou foram catalogados incorretamente. Em alguns casos é necessário coletar mais dados, ou passar pelo processo de limpeza novamente.

Figura 1 – O processo de mineração de dados



Fonte: Adaptado de Schutt e O'Neil (2014)

Em seguida, são aplicados algoritmos de aprendizado de máquina ou também modelos estatísticos, como regressão linear, *k-Nearest Neighbors algorithm* (kNN), entre outros. O modelo escolhido nesta etapa depende do tipo de problema que está sendo solucionado, podendo ser um problema de classificação ou predição, até mesmo um simples problema de descrição.
É possível então interpretar, reportar e visualizar os resultados. Relatórios podem ser gerados com a finalidade de apresentar as informações obtidas pelo *data mining*. Segundo Schutt e O'Neil (2014), nesta etapa, além da visualização e apresentação, existem os "dados produzidos" que poderiam ser um classificador de *spam*, ou um algoritmo de busca ranqueada, ou um sistema de recomendações. O fundamental desta atividade é que estes "dados produzidos" podem ser incorporados novamente ao Mundo Real e outros usuários poderam interagir com este produto, gerando ainda mais dados resultando em um laço repetitivo de *feedback* (SCHUTT; O'NEIL, 2014).

4. Linguagem de programação Python

A linguagem Python se tornou a principal escolha para cientistas e aplicações científicas como linguagem de *script*, especialmente no campo de *machine learning*, ciências biológicas e ciências sociais (MCKINNEY, 2013).
A dificuldade de programação da linguagem Python é baixa, isso proporciona uma facilidade para o início de novos projetos. Segundo Janert (2011), existem duas situações que essa facilidade da linguagem proporciona: em um lado, existe uma abundância de projetos excelentes desenvolvidos com a linguagem Python; por outro lado, um número considerável desses projetos são abandonados ou descontinuados.
Dentro deste grande número de projetos pequenos ou mais especializados, existem cinco projetos que provêem uma ótima biblioteca de aplicações científicas para a linguagem:

- ***NumPy***: suporta *arrays* e matrizes multidimensionais, possuindo uma larga coleção de funções matemáticas para trabalhar com estas estruturas;
- ***SciPy*:** desenvolvida para trabalhar com *arrays NumPy*, fornece rotinas para integração numérica e otimização;
- ***pandas*:** manipular e analisar dados, oferece estrutura de dados e operações para manipulação de tabelas numéricas;
- ***matplotlib*:** permite plotar gráficos e planilhas;
- ***IPython*:** interpretador interativo da linguagem Python.

Em muitas organizações, é comum realizar pesquisas, prototipar e testar novas ideias utilizando mais de um domínio específico de linguagem computacional, como MATLAB ou R e, posteriormente, estas ideias viram parte de um sistema de produção maior, escrito, por exemplo, em Java, C#, ou C++.

Kaldero (2015), afirma que Python não é somente uma linguagem adequada para a pesquisa e prototipagem, mas também para o desenvolvimento de sistemas.
Devido a esta solução de apenas uma única linguagem, as organizações podem se beneficiar, tendo cientistas e tecnólogos usando o mesmo conjunto de ferramentas programáticas. Portanto, Python é a ferramenta escolhida pela maioria desses profissionais. Essa escolha se deve, não somente a alta produtividade que a linguagem fornece, mas também por ela ser uma ferramenta comum a diferentes times e organizações (KALDERO, 2015).

5. Conclusão

É muito comum se referir a época atual como "a era da informação", mas Han, et al. (2012) afirmam que a presente fase deveria se chamar "a era dos dados". Esta reputação é devido ao gigantesco volume de dados que é gerado a cada momento em virtude de atividade comercial, empresas de telecomunicações, buscas na *Web*, redes sociais, *blogs*, transmissão *online* de vídeos e outras atividades.
A necessidade, então, de meios eficientes de mineração de dados, como a utilização de ferramentas, métodos computacionais e algoritmos específicos para a extração de dados, permite a existência do *data mining* e é através dos seus processos que a informação útil é descoberta.
A linguagem Python está sendo utilizada como uma das principais ferramentas, dentre as diversas existentes, devido ao seu conjunto de bibliotecas extremamente ágeis para análise e manipulação de dados, assim como o desenvolvimento de *softwares* mais robustos com a finalidade de automatizar os processos de mineração de dados.

Abstract
The purpose of this document is to present concepts of data mining applied to social networks, as well as using the programming language Python as a tool for data mining processes.
Keywords: data mining, social network, python.

Referências

HAN, J. et al. **Data Mining: Concepts and Techniques.** Elsevier, 2012.

IMIELINSKI, Tomasz; VIRMANI, Aashu; ABDULGHANI, Amin. **DataMine: Application Programming Interface and Query Language for Database Mining.** Em: KDD. 1996. p. 256.

JANERT, P. K. **Data Analysis with Open Source Tools.** O'Reilly Media, 2011. ISBN 978-0-596-80235-6

KALDERO, N. **Why is Python a language of choice for data scientists?** 2015. Acesso em 15 de março de 2016. Disponível em: <http://qr.ae/RkIeiB>.

MCKINNEY, W. **Python for Data Analysis.** O'Reilly Media, 2013.

RUSSELL, M. A. **Mining the Social Web.** O'Reilly Media, 2013. ISBN 978-1- 449-36761-9.

SCHUTT, Rachel; O'NEIL, Cathy. **Doing Data Science.** O'Reilly Media, 2014. ISBN 978-1-449-35865-5.