Universidade Presbiteriana Mackenzie
Programa de Pós Graduação em Engenharia Elétrica e Computação (PPGEEC)

Influência de bots na Wikipedia

Aluno: Marcos Cordeiro de Brito Jr
Professor: Leandro Augusto da Silva

Março
2020

Conteúdo

1 Resumo

Esse trabalho tem como objetivo coletar informações das atualizações de artigos da "WIKIPEDIA"[3] e mensurar a influência de *bots* nos artigos publicados, baseado no estudo [6].

Através do uso de um *endpoint* disponibilizado pela "WIKITECH"[8], os dados serão coletados e armazenadas para serem analisados.

Com as informações coletadas, serão executados alguns algorítimos de *Machine Learning*, para mostrar a utilização de *bots* na atualização dos artigos da Wikipedia e tentar analisar a influência desse tipo de mecanismo relacionando com os temas dos artigos.

Ao final, a intenção é mostrar o resultado dos algorítimos em *dashboards* e relatórios sobre todo o estudo.

2 Objetivo

"A Wikipédia é um projeto de enciclopédia multilíngue de licença livre, baseado na web e escrito de maneira colaborativa.". Essa é a definição dada pelo próprio site[3].

O surgimento da Wikipedia mudou a forma de pesquisa sobre qualquer assunto. A utilização de livros e coleções de enciclopédias caíram muito após o avanço da internet e o surgimento da enciclopédia online. O grande problema desde seu surgimento é a qualidade das informações.

Por ser uma plataforma colaborativa onde qualquer um pode incluir ou alterar seu conteúdo, a qualidade como fonte segura de informações se torna questionável. Existem mecanismos que tentam mensurar a qualidade dos artigos [1] inclusive utilizando inteligência artificial [4].

Um artificio muito utilizado para verificação e atualização de artigos de forma automática são robôs conhecidos como *bots*. Esses *bots* podem ser usados tanto para benefício de colaboradores da plataforma, como de forma maliciosa por empresas, grupos políticos, *hackers* ou qualquer pessoa que queira postar inverdades sobre algum assunto. Em outras palavras, seria como cometer um ato de vandalísmo com as informações publicadas.

Esse termo foi utilizado em dois outros artigos que também servirão de inspiração para essa implementação[5][9]. Através da utilização de inteligência artificial, os estudos buscaram provar o vandalísmo e alterações maliciosas em artigos da Wikipedia.

O objetivo principal desse artigo é analisar a intenção desses *bots* através dos artigos que estão sendo atualizados. Através das informações coletadas, poderá ser analisado a quantidade de artigos alterados pelos *bots*, quais os temas mais alterados, quantidade de atualizações feitas pelos mesmos robôs, se os temas estão relacionados a grupos de grande influência como religião, política ou esporte, quantidade de atualizações comparadas com usuários comuns, países com maior número de alterações, entre outras possibilidades.

2.1 Técnicas e Tecnologias

A captura das informações será feita a partir de um *endpoint* disponibilizado pela "WIKITECH"[8], utilizando um script desenvolvido em **Python**. Esse script irá armazenar as informações no banco **MongoDB**, onde a *collection* principal irá seguir o *schema* dos próprios dados disponibilizados pelo endpoint[7].

Para a análise, a intenção é preparar os dados e utilizar de algorítimos de agrupamento[2] para criar *clustes* com as palavras encontradas nos títulos

e conseguir gerar uma análise para identificar os grupos e as intenções das alterações.

Esse desenvolvimento também será feito utilizando a linguagem **Python** e bibliotecas auxiliáres, assim como a geração dos *dashboards* e resultados das análises dos dados coletados. Como ferramentas de auxílio desse desenvolvimento, será utilizado o **Anaconda** e o **Jupyter Notebook**.

Toda infraestrutura do ambiente será feito na plataforma de *cloud* **AWS** da **Amazon**, utilizando como ferramenta de apoio os recursos de *containers* do **Docker** para provisionar todos recursos necessários de forma rápida.

Referências

[1] Quang-Vinh Dang and Claudia-Lavinia Ignat. Measuring quality of collaboratively edited documents: The case of wikipedia. *IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, 2016.

[2] DANIEL GOMES FERRARI and LEANDRO NUNES DE CASTRO SILVA. *Introdução a mineração de dados*. Editora Saraiva, 2017.

[3] Wikipedia Foundation. Wikipédia, 2006.

[4] Jun Liu and Sudha Ram. Using big data and network analysis to understand wikipedia article quality. *Elsevier*, May 2018.

[5] Martin Potthast, Benno Stein, and Robert Gerling. Automatic vandalism detection in wikipedia. In *European conference on information retrieval*, pages 663–668. Springer, 2008.

[6] Thomas Steiner. Bots vs. wikipedians, anons vs. logged-ins (redux): A global study of edit activity on wikipedia and wikidata. In *Proceedings of The International Symposium on Open Collaboration*, OpenSym '14, page 1–7, New York, NY, USA, 2014. Association for Computing Machinery.

[7] Wikimedia. mediawiki-event-schemas, 2019.

[8] Wikitech. Event platform/eventstreams, 2016.

[9] Qinyi Wu, Danesh Irani, Calton Pu, and Lakshmish Ramaswamy. Elusive vandalism detection in wikipedia: a text stability-based approach. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1797–1800, 2010.