UTILIZAÇÃO DE FERRAMENTAS DE WEB SCRAPING PARA MINERAÇÃO AUTOMATIZADA DE NOTÍCIAS

PE03190619/083

Gabriela Vieira Martins(IFSul Câmpus Charqueadas – Técnico em informática – gabimartins969@gmail.com)
Prof. Dr. Roberto Irajá Tavares da Costa Filho (Docente Orientador - IFSul Câmpus Charqueadas – robertocosta@charqueadas.ifsul.edu.br)
Câmpus Charqueadas

13° JIC IFSul JORNADA DE INICIAÇÃO CIENTÍFICA DO INSTITUTO FEDERAL SUL-RIO-GRANDENSE 2020

INSTITUTO FEDERAL Sul-rio-grandense

Introdução. A rápida propagação das notícias falsas consiste em um fenômeno de escala global que amplifica a desinformação e reduz o impacto de notícias verdadeiras. Uma vez que as pessoas estão cada vez mais polarizadas, e com alta rejeição a tudo que é contrário ao seu pensamento, a propagação de notícias falsas acaba se manifestando como uma consequência do desejo dos indivíduos de reforçar o seu ponto de vista**.**

Objetivo. Este trabalho está inserido em um projeto maior que objetiva empregar aprendizado de máquina para inferir a probabilidade de uma notícia ser falsa. Mais precisamente, o presente trabalho visa construir, por meio de um Web Scraping, um coletor automatizado de notícias previamente classificadas. De forma complementar ao cadastro manual de notícias, a coleta automatizada confere maior escalabilidade ao sistema de treinamento do aprendizado de máquina**.**

Metodologia. Para realizar esse projeto, foi construído um script na linguagem Python, empregando o framework Scrapy, muito utilizado para fazer o mapeamento de sites..

Esse script realiza o mapeamento de sites através das Tags HTML. Em um segundo momento, os testes foram realizados em sites selecionados de notícias, buscando sites que possuam um padrão bem definido de estrutura HTML para todas as notícias

```python
# -*- coding: utf-8 -*-
import scrapy


class TecnoblogSpider(scrapy.Spider):
    name = 'Tecnoblog'
    allowed_domains = ['tecnoblog.net']
    start_urls = ['http://tecnoblog.net/']

    def parse(self, response):
      for article in response.css("article"):
        link    = article.css("div.texts h2 a::attr(href)").extract_first()
        title   = article.css("div.texts h2 a::text").extract_first()
        author  = article.css("div.texts div.info a::text").extract_first()

        yield {'link': link, 'title': title, 'author': author}
```

Resultados. Como principais resultados é possível destacar o estudo do sistema operacional Linux; estudo do framework Scrapy; estudo e implementação de um script em Python, usando a framework Scrapy; além da realização de testes do script em sites de notícias.

Conclusão. A elaboração da ferramenta para coleta automatizada de notícias, apesar das dificuldades enfrentadas, permitiu um aprendizado significativo sobre tecnologias especializadas que não são abordadas ao longo do curso técnico em informática.

LAZER, D.M., et al., 2018. The science of fake news. Science, 359(6380), pp.1094-1096.PAL, S., et al., 2019. Applying Machine Learning to Detect Fake News. Indian Journal of Computer Science, 4(1), pp.7-12.RUCHANSKY, et al., 2017, November. Csi: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM.SHU, K., et al.. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), pp.22-36.VOSOUGHI, et al., 2018. The spread of true and false news online. Science, 359(6380), pp.1146-1151.ZHOU, X., et al., 2019, January. Fake News: Fundamental Theories, Detection Strategies and Challenges. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. ACM.

REALIZAÇÃO:
FAPERGS propesp INSTITUTO FEDERAL Sul-rio-grandense