Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Pedro Renato Ferreira da Silva

Criação de Coletor de Dados e Caracterização da Rede Social TikTok

MONOGRAFIA EM SISTEMAS DE INFORMAÇÃO I - PESQUISA TECNOLÓGICA

Jussara Almeida

RELATÓRIO FINAL

1 Introdução

As redes sociais online atualmente funcionam como uma extensão do mundo físico, replicando as dinâmicas sociais vigentes e, em certas situações, apresentando o seu próprio conjunto de regras e comportamentos peculiares.

A comunicação por meio da internet ocorre desde 1970, porém o modelo de redes sociais que conhecemos nos dias de hoje, como sendo um mecanismo coletivo e de colaboração massiva, evidenciou-se a partir de 2005, momento no qual houve uma mudança de paradigma na internet. As páginas na web passaram a ser orientadas aos usuários, ou seja, agora eles conseguiriam criar, compartilhar, colaborar e comunicar informações com outros usuários de forma simplificada, algo que não era possível em anos anteriores.

O advento das redes sociais online apresenta-se como um fenômeno relativamente recente, mas com um impacto marcante na sociedade, haja vista a notável influência nos âmbitos social, econômico, político e psicológico da população mundial. Sendo assim, surge a necessidade de estudos que buscam compreender o comportamento humano no ambiente virtual e os seus efeitos.

Dentre as principais redes sociais presentes nas plataformas digitais encontra-se o TikTok. Lançado em 2016 pela empresa chinesa ByteDance. em apenas um ano o aplicativo tornou-se o mais baixado do mundo e atualmente contém mais de 1 bilhão de usuários. Nesse sentido, o entendimento do comportamento dos usuários e criadores de conteúdo em tal rede social representa, consequentemente, a compreensão das dinâmicas sociais vigentes em um território bastante representativo do ambiente online.

A partir do cenário apresentado dois pontos se tornam evidentes: a necessidade de extração dos dados do TikTok e, posteriormente, a análise dos mesmos, a fim de gerar estudos, bem como validar hipóteses acerca do tema. Com base nisso, os objetivos de trabalho definidos serão, em um primeiro momento, o desenvolvimento de um programa capaz de coletar os dados presentes no TikTok.

2 Referencial Teórico

Determinado o problema desta pesquisa, este capítulo consiste em uma base teórica acerca dos assuntos que envolvem redes sociais e coleta de dados.

2.1 Redes Sociais

Definir o que são as redes sociais é um grande desafio dado o fato de ser um conceito repleto de nuances e em constante evolução. Por exemplo, Papacharissi, pesquisadora da área da comunicação, defende que a definição de rede social é dinâmica e específica de cada contexto.

Ao analisar a literatura acadêmica, entretanto, é possivel obter um consenso acerca da definição das redes sociais sob a ótica de três aspectos: (a) quais atividades a rede social permite, (b) como ela permite essas atividades e (c) o contéudo presente em tais redes sociais. Sendo assim, todo mecanismo que possui conteúdo gerado pelo usuário, permite a conexão, comunicação e interação com outros usuários, e ocorre mediante aplicações na internet com base nos princípios da Web 2.0, pode ser considerado como sendo uma rede social.

De modo abrangente, as redes sociais tornaram-se ubíquas na sociedade contemporânea, desempenhando papéis diversos que vão desde a comunicação pessoal até o marketing empresarial. Elas possibilitam a formação de comunidades virtuais, o fortalecimento de relações sociais existentes, a descoberta de novas conexões e o acesso a uma ampla gama de informações e conteúdos.

2.2 Coleta de Dados

Coleta de Dados de Conteúdo: Este tipo de coleta concentra-se na extração de informações presentes nos conteúdos gerados pelos usuários, como postagens, comentários, perfis, fotos e vídeos. Utilizando técnicas de web scraping ou APIs disponibilizadas pelas próprias plataformas, os pesquisadores podem coletar dados textuais, imagens e vídeos para análise.

Coleta de Dados de Comportamento do Usuário: A coleta de dados de comportamento visa capturar informações sobre como os usuários interagem e utilizam as redes sociais ao longo do tempo. Isso inclui dados sobre padrões de atividade, tempo de permanência, frequência de postagens, reações a conteúdos e interações com outros usuários. Esses dados são valiosos para entender tendências de uso e preferências dos usuários.

Coleta de Dados da Estrutura da Rede: Este tipo de coleta concentra-se na análise da topologia e da estrutura das redes sociais, incluindo informações sobre conexões entre usuários, grupos, comunidades e influenciadores. Através da coleta de dados sobre a rede, é possível identificar padrões de conexão, influenciadores chave e comunidades de interesse.

3 Desafios

A coleta de dados do TikTok para fins acadêmicos, analíticos, possui alguns desafios. A rede social chinesa não possui uma API pública disponível para uso irrestrito. No entanto, há uma API específica que o TikTok disponibiliza para fins de pesquisa acadêmica. Ainda assim, a maior limitação está no fato de que o acesso à esse recurso ocorre de maneira controlada. É necessário uma análise e aprovação pelos órgãos de controle do TikTok. Além disso, o Brasil não encontra-se na lista de países que possuem acesso permitido à API.

Soma-se ao cenário descrito acima alguns desafios técnicos ao tentar obter dados por meio de web scraping (técnica de coleta de dados online), dentre eles:

Carregamento Dinâmico de Conteúdo. O TikTok é uma plataforma altamente dinâmica que depende de JavaScript para carregar conteúdo. Isso significa que simplesmente baixar o HTML de uma página não será suficiente para acessar os dados, já que o conteúdo é carregado de forma assíncrona. **Login e Autenticação.** Alguns dados no TikTok só são acessíveis após o login, o que significa o scraper precisará lidar com a autenticação. Isso é algo complexo, pois envolve o gerenciamento de cookies, tokens e autenticação multifator. **Técnicas Anti-Scraping**. O TikTok utiliza uma variedade de técnicas anti-scraping, como:

- **Captchas:** Para bloquear bots automatizados, o TikTok pode apresentar captchas que são difíceis para scripts contornarem.
- **Verificação de User-Agent:** O TikTok verifica a string User-Agent das requisições e bloquear aquelas que parecem vir de bots.

- **Tokens Dinâmicos:** O TikTok usa tokens dinâmicos que precisam ser incluídos nas requisições para que sejam válidos.
- **Limitação de Taxa de IP:** Fazer muitas requisições do mesmo endereço IP em um curto período pode levar a bloqueios temporários ou permanentes.

Mudanças Frequentes. O TikTok atualiza frequentemente seu site e APIs, o que pode quebrar scrapers que dependem de estruturas HTML ou endpoints específicos.

4 Abordagens adotadas

Em um primeiro momento foi realizada uma vasta pesquisa de soluções alternativas existentes capazes de fornecerem os dados do TikTok. A maioria dos resultados obtidos eram de ferramentas que não funcionavam devido à falta de atualizações e suporte. Contudo, dois resultados mostraram-se promissores.

4.1 API não oficial do TikTok

A API não oficial do TikTok possui diversas funcionalidades. No entanto, ao testá-la foi identificado que apenas uma função funciona corretamente. No caso, a função disponível para uso é a de pesquisa de um video por meio da URL (Uniform Resource Locator). Tal função retorna diversos campos de informação do vídeo como: id, descrição, duração, hashtags, número de visualizações e compartilhamentos, etc. Além disso, é possível obter o id dos vídeos relacionados.

Para utilizar a API não oficial de forma ideal para a coleta de dados, foi necessário desenvolver uma solução capaz de contornar os problemas e limitações inerentes de tal API. Ao modelarmos o problema na forma de grafo, sendo os vídeos os vértices, e as arestas as relações entre vídeos, conseguimos estabelecer o formato de uma árvore em que a URL inicial é a semente que origina os demais vertices por meio dos vídeos relacionados.

Foi identificado que o TikTok utiliza um determinado padrão de URL para acesso aos vídeos. Sendo assim, partindo de um vértice e em posse dos ids dos vídeos relacionados, foi possível recontruir as URLs de tais vídeos e acessá-los.

Com base na solução proposta acima, tornou-se possível coletar uma grande quantidade de vídeos de maneira escalável, partindo de apenas uma URL fornecida.

4.2 Plataforma Apify

A plataforma Apify é uma outra alternativa viável a ser utilizada. Trata-se de uma plataforma online que disponibiliza diversas APIs, dentre elas a API que dá acesso a dados de vídeos no TikTok. A busca por vídeos ocorre por palavra-chave ou por usuário. No caso de palavra-chave, é retornado um conjunto de vídeos que contém tal palavra em meio ao seu conteúdo, seja título do vídeo ou descrição. Ao utilizar a busca por usuário, é retornado o conjunto de vídeo produzido pelo usuário especificado.

A principal limitação da utilização da Apify está no fato de ser uma platforma freemium, ou seja, a coleta até uma determinado limite de vídeos é gratuita, após esse limite é necessário

pagamento. Contudo, um ponto a ser explorado é o fato da plataforma Apify renovar a quantidade de coleta gratuira mensalmente.

A solução proposta para esse cenário foi a criação de uma integração com a plataforma de forma a facilitar o acesso à API disponibilizada e a estruturação do armazenamento dos dados.

5 Resultados Obtidos

Ao final do MSI I foram obtidas duas ferramentas capazes de coletar dados do TikTok. A primeira ferramenta possibilita uma coleta mais geral dos vídeos, podendo ser analisados os aspectos relacionados ao algoritmo de recomendação do TikTok. A segunda ferramenta permite uma coleta de dados mais direcionada, possibilitando análises temáticas, separadas por assuntos. Ambas ferramentas serão utilizadas na segunda parte da monografia para a coleta em si de maneira sistemática dos dados.

6 Referencial Bibliográfico

Fast and reliable end-to-end testing for modern web apps. Disponível em: <https://playwright.dev/pytho Acesso em: 11 abr. 2024.

The Python standard library. Disponível em: <https://docs.python.org/3.12/library/index.html>. Acesso em: 11 abr. 2024.

VASCONCELLOS, Paulo H. S.; LARA, Pedro Diógenes A.; MARQUES-NETO, Humberto T. Analyzing Polarization And Toxicity On Political Debate In Brazilian TikTok Videos Transcriptions. ACM Web Science Conference 2023, [S. l.], p. 1-10, 30 abr. 2023.

McCay-Peet, L., Quan-Haase, A. (2017). What is social media and what questions can social media research help us answer. The SAGE handbook of social media research methods, 13-26.

Ackland, R. (2013). Web social science: Concepts, data and tools for social scientists in the digital age. London: SAGE.

Fennell, C. (2020). Big Data, Collection of (Social Media, Harvesting). In The International Encyclopedia of Media Psychology, J. Bulck (Ed.). https://doi.org/10.1002/9781119011071.iemp0006