UFRJ

cores
Laboratório de
Computação Social e
Análise de Redes Sociais

O Uso das Redes Sociais Digitais e as Eleições para o Cargo de Prefeito do Rio de Janeiro

Equipe: Ana Paula L. F. Vasconcelos - Ouvinte
Jivago Medeiros - Inscrito
Rafael Escalfoni - Ouvinte
Silas P. Lima Filho - Ouvinte
Sírius Thadeu Ferreira da Silva - Ouvinte

Problema

- Qual é o papel das redes sociais para as eleições municipais?

- Qual é a influência das redes sociais no resultado das eleições municipais de 2020 do Rio de Janeiro?

Objetivo Geral

- Identificar a relação entre as interações nas redes sociais (Twitter) e as eleições municipais de 2020 no Rio de Janeiro.

Objetivos Específicos

- Analisar:
  - Perfil dos usuários
  - Menções
  - Hashtags
  - Análise de Sentimentos
  - Análise Temporal

Dataset

- Tweets sobre candidatos à prefeitura da cidade do Rio de Janeiro em 2020
- Termos buscados: [{nomeCandidato} AND [{numeroPartido} OR {siglaPartido} OR $ANO]]
- Tamanho: ≃ 3.3GB (JSON) | ≃ 1.2GB (CSV)
- 580.354 tweets contendo um ou mais desses termos
- Período de Coleta: 29/09/2020 à 01/12/2020

Termos de Busca Utilizados na API do Twitter

'Benedita da Silva', 'Benedita13', 'BeneditaPT', 'BeneditadaSilva2021', 'Eduardo Bandeira de Mello', 'Bandeira18', 'BandeiraRede', 'BandeiradeMello2021', 'Clarissa Garotinho', 'Clarissa90', 'ClarissaPROS', 'ClarissaGarotinho2021', 'yro Garcia', 'Cyro16', 'CyroPST', 'Cyro2021', 'CyroGarcia2021', 'Eduardo Paes', 'Eduardo25', 'EduardoPL', 'EduardoPaes2021', 'Fred Luz', 'FredLuz30', 'redLuzNOVO', 'FredLuz2021', 'Glória Heloíza', 'GlóriaHeloíza20', 'GlóriaHeloízaPSC', 'GlóriaHeloíza2021', 'Henrique Simonard', 'HenriqueSimonard29', 'HenriqueSimonardPCO', 'Simonard2021', 'Luiz Lima', 'LuizLima17', 'LuizLimaPSL', 'LuizLima2021', 'Marcelo Crivella', 'Crivella10', 'CrivellaRepublicanos', 'Crivella2021', 'Martha Rocha', 'MarthaRocha12', 'MarthaRochaPDT', 'MarthaRocha2021', 'Paulo Messina', 'PauloMessina15', 'PauloMessinaMDB', 'PauloMessina2021', 'Renata Souza', 'RenataSouza50', 'RenataSouzaPSOL', 'RenataSouza2021', 'Suêd Haidar', 'SuêdHaidar35', 'SuêdHaidarPMB', 'SuêdHaidar2021'

Coleta de Dados

- Realizada via API oficial do Twitter
- Utilizou um script na linguagem Python com a biblioteca Tweepy
  - A biblioteca Tweepy simplifica o uso da API do Twitter encapsulando e abstraindo alguns dos passos necessários para a coleta de dados.
- Utilização da funcionalidade Sampled stream.
  - Funcionalidade que faz a API retornar uma amostra aleatória de 1% dos tweets públicos disponíveis em "tempo real" de acordo com os termos buscados.
- A persistência dos dados foi feita no SGBD MongoDB.
- Dados exportados para análise em formato CSV e JSON.

Análise e tratamento dos dados

Foi realizado um pré-processamento, através do qual foram identificadas e resolvidas diversas situações, tais como:

- Problemas de recurso ao tentar ler arquivo JSON:
  - Estouro de memória.
- Tratamento dos dados JSON;
- Expansão das colunas JSON;
- Remoção das colunas segundo determinados critérios;
- Tratamento dos atributos nulos;
- Processamento do texto.

Exploração dos dados

Dificuldades encontradas:

- Tamanho do arquivo JSON (hospedagem, manipulação dos dados, recursos, etc.);
- Colunas com objetos JSON (múltiplos dados retratados em uma única coluna);
- Identificação/representação dos objetos JSON em String;
- Conversão de nulos em objetos JSON vazios;
- Conversão de listas para strings JSON.

Limpeza dos dados

- Eliminação de colunas com taxa de nulos acima de 80%;
- Eliminação de colunas com um único valor;
- Eliminação de colunas sem utilidade para nossa análise (replicadas, etc.);
- Tratamento das colunas JSON:
  - Transforma os objetos JSON de cada coluna em um DataFrame;
  - Repete o processo de limpeza acima para cada DataFrame;
  - Faz o Join desses DataFrames com o dataset principal.
- Preenchimento de nulos;
- Exportação do Dataset limpo.

Limpeza dos dados

Resultados Finais:

- Dataset Original:
  - 31 colunas;
  - 580354 linhas;
  - 1.29 GB.
- Dataset Limpo:
  - 25 colunas (excluídas as colunas JSON originalmente expandidas);
  - 580354 linhas;
  - 373 MB.

Ao exportar o dataset limpo para um novo arquivo CSV, tivemos que contornar um problema envolvendo quebras de linha nos campos de texto.

Em andamento...

Processamento de texto:

- Atomização;
- Normalização;
- Radicalização;
- Remoção de “Stop Words”.

Análise Exploratória de Dados

- Distribuição de tweets ao longo do tempo;
- Origens dos tweets;
- Proporção de tweets longos/curtos;
- Taxa de ReTweets;
- Linguagens mais utilizadas;
- Cluster de usuários;
- Cluster de localidades;
- Estatísticas de interações sociais;
- Correlação entre atividade e antiguidade;
- Quantidades de hashtags;
- Quantidades de menções;
- Perfil dos usuários:
  - Bag of Words.
- Análise de sentimento:
  - Word Cloud.

Execução

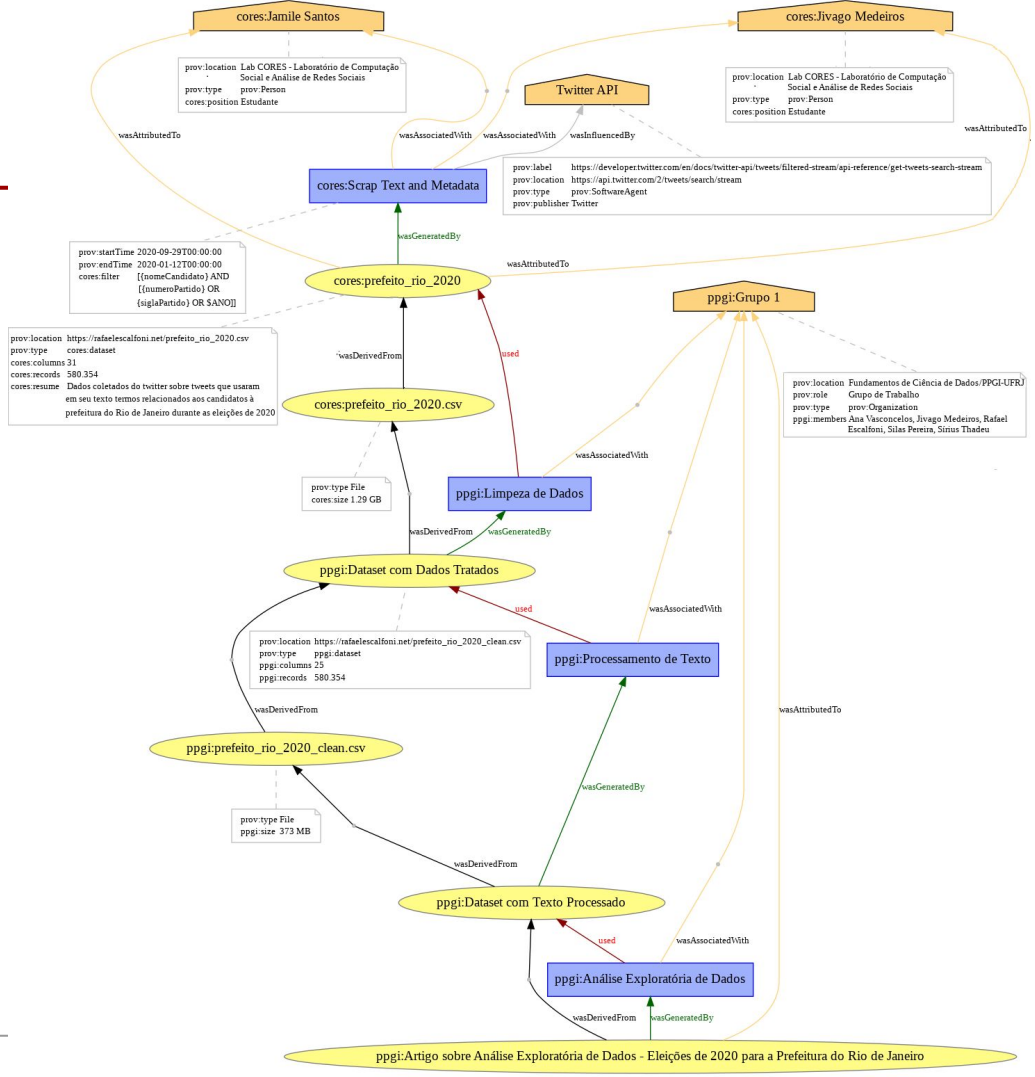
- Google Colaboratory ou Colab
  - Permite escrever código Python diretamente no navegador, com acesso gratuito a GPUs, TPUs e demais hardwares e softwares necessários, independentemente da potência do computador local.
  - Não é necessária configuração extra para começar a utilizar.
  - Permitem combinar código executável e rich text em um só documento, além de imagens, HTML, LaTeX e outros.
  - Os notebooks são armazenados no Google Drive, sendo possível compartilhar com outras pessoas que podem editar o documento.
  - Fornece completamentos automáticos para explorar atributos de objetos Python, bem como para visualizar rapidamente sua documentação.

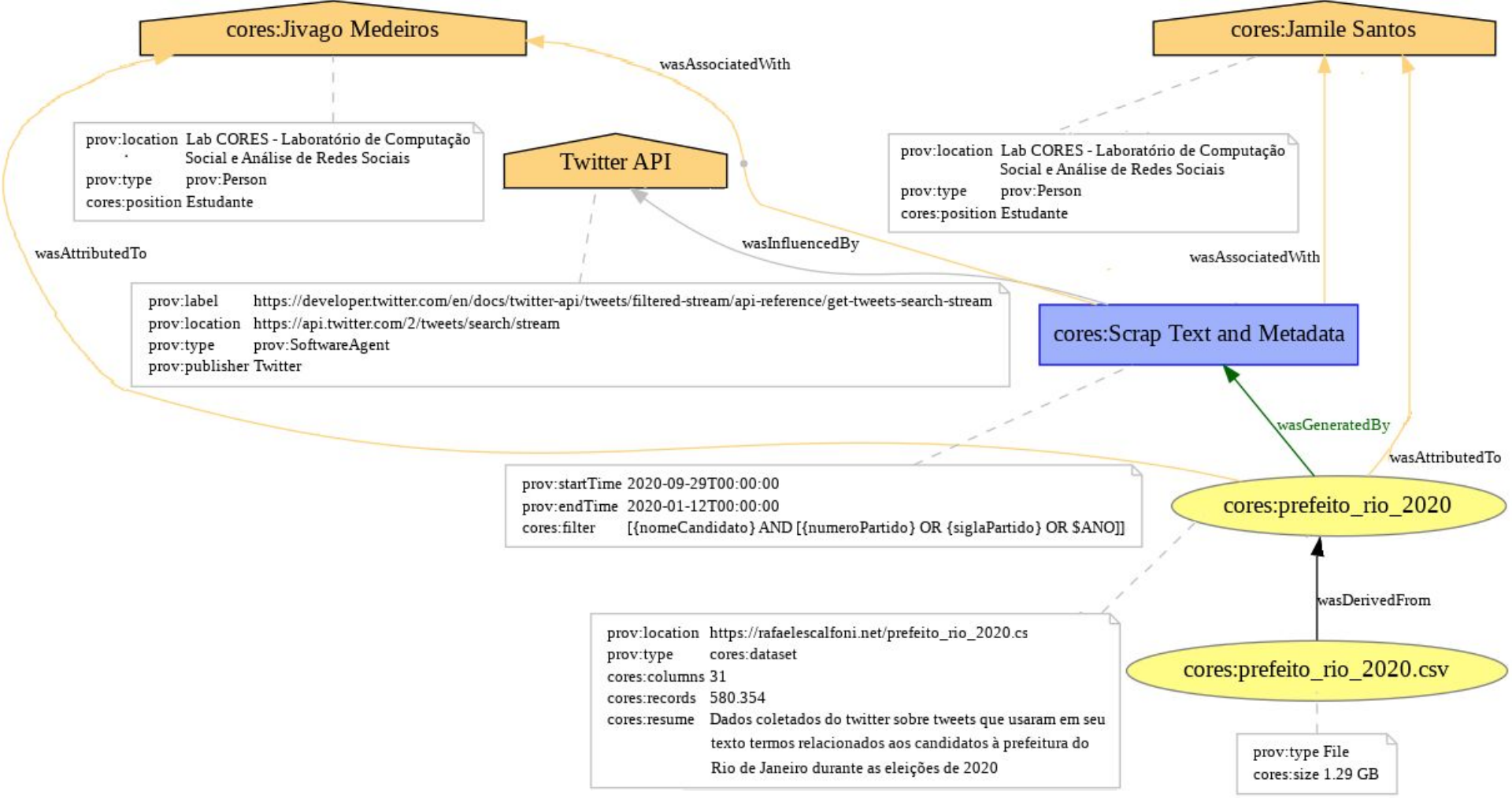
Execução

- Utilização de uma biblioteca de análise exploratória de dados, chamada Sweetviz.
  - Faz uso dos dataframes do pandas e cria um relatório HTML autocontido que pode ser visualizado em um navegador ou integrado em notebooks.
  - Cria visualizações perspicazes e bonitas com poucas linhas de código.
  - Fornece análises que levariam muito mais tempo para serem geradas manualmente, incluindo algumas que outras bibliotecas não fornecem tão rapidamente, como a análise de alvo, comparações de dataset, correlações, associações, etc.

Proveniência

Para coletar os metadados de proveniência do dataset e gerar o respectivo grafo de proveniência foi utilizada a biblioteca PROV do Python, em sua versão 2.0.0.




cores:Jivago Medeiros
cores:Jamile Santos
wasAssociatedWith
prov:location Lab CORES - Laboratório de Computação Social e Análise de Redes Sociais
prov:type prov:Person
cores:position Estudante
Twitter API
prov:location Lab CORES - Laboratório de Computação Social e Análise de Redes Sociais
prov:type prov:Person
cores:position Estudante
wasAttributedTo
wasInfluencedBy
wasAssociatedWith
prov:label https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/api-reference/get-tweets-search-stream
prov:location https://api.twitter.com/2/tweets/search/stream
prov:type prov:SoftwareAgent
prov:publisher Twitter
cores:Scrap Text and Metadata
wasGeneratedBy
wasAttributedTo
prov:startTime 2020-09-29T00:00:00
prov:endTime 2020-01-12T00:00:00
cores:filter [{nomeCandidato} AND [{numeroPartido} OR {siglaPartido} OR $ANO]]
cores:prefeito_rio_2020
wasDerivedFrom
prov:location https://rafaelescalfoni.net/prefeito_rio_2020.cs
prov:type cores:dataset
cores:columns 31
cores:records 580.354
cores:resume Dados coletados do twitter sobre tweets que usaram em seu texto termos relacionados aos candidatos à prefeitura do Rio de Janeiro durante as eleições de 2020
cores:prefeito_rio_2020.csv
prov:type File
cores:size 1.29 GB

Proveniência (artefatos, agentes e processos)

- Responsáveis: Jamile Santos e Jivago Medeiros
- Modo de coleta: script em python + stream (real time) da API do twitter[1]
- *Script* executado por Jivago Medeiros
  - parâmetros de entrada na execução do *script:* - nenhum -
  - *script* executado em uma instância EC2 (máquina virtual) da AWS

[1] https://developer.twitter.com/en/docs/tutorials/stream-tweets-in-real-time

Proveniência (artefatos, agentes e processos)

- Tweets recebidos pela API foram armazenados no SGBD Mongodb versão 4.4.1
- Após a finalização da coleta foram gerados dois arquivos com o conteúdo dos dados utilizando o comando *mongoexport*; um arquivo JSON e um CSV.

Reprodutibilidade

- Seguiu-se os preceitos dispostos por Benureau e Rougier em “Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions”.
- Coletou-se todas as informações necessárias de ambiente.
  - Sistema Operacional e sua arquitetura; Linguagem de Programação e bibliotecas utilizadas.
- Criou-se scripts Python com o cuidado de documentar os passos executados durante todo o experimento, seguindo boas práticas de programação para tornar os códigos reusáveis.

Resultados Preliminares

- Uso de *dataset* com dados brutos (*raw data*), o que dificulta a definição de questões mais específicas;
- Processo de limpeza dos dados junto com análise exploratória auxilia no levantamento de questões a serem exploradas mais a fundo;
- Suficiência dos dados (após limpeza e tratamento).

Resultados Preliminares

- Usuários mais mencionados:
  - @**eduardopaes_**, @**felipeneto**, @**delmartharocha**, @**MCrivella**, @**OficialLuizLima**
- Usuários que mais postaram:
  - @**BelemJameson**, @**ESeCiroFossePr1**, @**meupaiseursal**, @**galloradical**
- Mídias externas mais postadas;
- Tweets com o maior número de RT.

Mídias externas mais postadas

https://gazetabrasil.com.br/politica/benedita-da-silva-usa-termo-racista-contra-sergio-camargo-capitao-do-mato/

https://www.jornaldacidadeonline.com.br/noticias/24720/nervosinho-eduardo-paes-tenta-se-eleger-em-cima-de-memoria-fraca-do-carioca

Tweets com mais RT

@romulocarvalho_: Não sei quem ganha, mas se Guilherme Boulos vencer em São Paulo e Eduardo Paes vencer no Rio de Janeiro, então o eixo...

@euupereira: Eduardo Paes provavelmente vai ser eleito prefeito do Rio de Janeiro esse ano e eu só penso nesse vídeo https://t.co/TzgQk0...

@dudunaweb: Eu no 1° turno: quem é LOUCO de votar no Eduardo Paes depois de tudo que ele fez no comando do Rio por 8 anos

@dudunaweb: Eu no 1° turno: quem é LOUCO de votar no Eduardo Paes depois de tudo que ele fez no comando do Rio por 8 anos