Relatório final
Trabalho prático de Mineração de Dados

Ricardo Guimarães PG50858

Filipa Guimarães PG50853

João Lobo PG49366

Braga, 16 de Junho de 2023

Introdução

Como proposta de avaliação final para a Unidade Curricular de Mineração de Dados, foi nos solicitada a criação de um trabalho prático que constasse na implementação de um estudo que envolvesse a recolha, o processamento, a análise e subsequente mineração de dados. Estes dados teriam de ser recolhidos em fontes (preferencialmente múltiplas) publicamente e gratuitamente acessíveis e teriam como objetivo de extrair conhecimento útil e não óbvio.

Tendo em conta que o nosso grupo é composto exclusivamente por estudantes do mestrado em Humanidades Digitais, decidimos focar-nos numa vertente mais humanística, escolhendo um trabalho que tivesse como base de dados primária uma rede social e que necessitasse de pouco trabalho estatístico na sua execução.

Motivação

A motivação para o tema escolhido parte do facto de que nos encontramos num mundo intrinsecamente digital, onde todas as pessoas estão ligadas entre si pelas redes sociais, o que as torna uma importante fonte de informação (e desinformação). Por essa razão quisemos determinar o impacto das mesmas (mais especificamente do Twitter) na saúde das pessoas através da disseminação de afirmações falaciosas, usando como questão de investigação "Será que a afirmação pública de Donald Trump durante a pandemia de COVID19 sobre ingestão de lixívia gerou desinformação difundida?".

Recolha dos dados Twitter

Como fontes de dados da rede social decidiu-se que iria ser o Twitter, pois após uma breve pesquisa inicial encontramos dados da plataforma online de marketing e consumer data Statista.com que afirmavam que o Twitter é das principais redes sociais usadas pelos americanos para obtenção de notícias (cerca de 53% dos americanos usa), assim como a rede social preferida do sujeito em análise, Donald Trump.

Originalmente o plano seria usar o API gratuito do Twitter para realizar a extração do dataset contendo os tweets que correspondiam a nossa query, mas devido a profundas alterações na política de disponibilização do API que ocorreram recentemente, a extração por este meio ficou impossibilitada. Tendo já passado algum tempo desde a definição do tema do projeto, e não querendo desviar dos objetivos iniciais do trabalho, decidimos procurar soluções alternativas para a extração dos tweets.

Após dezenas de horas de tentativas de extração com dezenas de bibliotecas diferentes e aplicações diferentes testadas, finalmente encontramos um método que poderia resultar. A biblioteca Selenium permitiu-nos criar uma janela automatizável de browser, que automaticamente faz login no Twitter usando as credenciais de qualquer elemento do grupo, pesquisa o query e faz scroll na página para carregar mais tweets. Por sua vez, a biblioteca BeautifulSoup tem uma função de

webscraping e extrai o código html da página na qual os tweets estão localizados. Foi necessário um enorme esforço para conseguir que este script extraísse os tweets de forma correta, pois o Twitter apenas renderiza um pequeno número de tweets de cada vez, o que levou a que se tivesse de ajustar minuciosamente os parâmetros de scrape/scroll down/quantidade de extrações para que todos os tweets fossem extraídos sem que o script fosse desnecessariamente demorado. Após ter estudado o funcionamento das ferramentas e ter aprimoriado o script, a extração tornou-se incrivelmente fácil, e conseguimos com sucesso extrair 4 datasets:

- Dataset de tweets no mês anterior à declaração polémica, com um total de 0 tweets encontrados (dataset vazio).
- Dataset de tweets da semana imediatamente a seguir à declaração polémica, com um total de 3033 tweets encontrados e extraídos.
- Dataset de tweets da segunda semana a seguir à declaração polémica com um total de 71 tweets encontrados e extraídos.
- Dataset de tweets do mês a seguir à semana a seguir à declaração polémica com um total de 161 tweets encontrados e extraídos.

Usamos o seguinte prompt para extração dos datasets:

```
"inject disinfectant" "trump" OR "president" min_faves:1 since:2020-04-23_00:00:01_EST until:2020-04-30_11:59:01_EST -filter:replies
```

Achamos que seria importante restringir a extração a tweets que também contivessem “Trump” ou “Presidente” para reduzir o número total de tweets para que só ficassem aqueles que fariam a conexão a Donald Trump e os quais nos provam se foi ele a causa do aumento ou não. Também filtramos tweets com 0 likes por terem uma mais baixa probabilidade de serem difundidos e filtramos respostas para evitar threads irrelevantes.

Como a biblioteca BeautifulSoup recolhe todos os metadados de cada tweet de forma inteira, acabamos por fazer uma pré limpeza e pré-seleção ao apenas codificar no script a extração dos elementos username; usartag; timestamp e texto do tweet.

Análise dos dados Twitter

Tendo em conta as datas e os volumes dos datasets, podemos concluir que, tal como tínhamos previsto no início do projeto, a declaração polémica do ex-presidente Norte Americano Donald Trump levou a um aumento de informações sobre injeção de lixivia.

Recolha de dados Saúde

Na segunda fonte de dados, o plano era procurar na plataforma Kaggle.com, uma comunidade online de data scientists e machine learning practitioners, por datasets relativos a emergências medicas nos EUA que contivessem injeções de desinfetantes como variável. Como não

conseguimos encontrar nenhum dataset relevante nessa plataforma nem em outras plataformas de datasets, decidimos criar o nosso próprio dataset.

A fonte dos dados foi a WISQARS, uma base de dados online do CDC (Centro de Controlo e Prevenção de Doenças dos Estados Unidos) que fornece informações (sobre a forma de texto corrido/gráficos e ficheiros CSV) sobre ferimentos letais e não letais em solo americano. Decidimos extrair os dados relativos às idades de 15-24, 25-34, 35-44 e 45-54, pois de acordo com a plataforma online de markting e consumer data Statista.com, o maior grupo de utilizadores do Twitter pertencem ao intervalo de idade dos 25-34 anos (cerca de 40% de todos os utilizadores) e a mediana dos utilizadores adultos é de 40 anos.

Devido a constrangimentos de tempo decidimos não automatizar a extração dos dados (algo que seria possível usando as bibliotecas Selenium e BeautifulSoup como se fez para os datasets do Twitter), fazendo uma extração manual dos dados, com uma pré-seleção de variáveis, que depois foram organizados para facilitar a sua leitura (visto que não se utilizou os dados para mais nada).

O dataset final de saúde conta com as variáveis do ano da ocorrência, o intervalo de idades, o número total de mortes geral, a média do total de mortes geral, o número de mortes por envenenamento, a percentagem de mortes por envenenamento e média da percentagem de mortes por envenenamento.

Análise dos dados Saúde

Analisando o dataset, pode-se ver um claro aumento de mortes por envenenamento no ano de 2020, que se destaca dos valores dos anos anteriores por uma margem considerável. Nos anos de 2016 a 2019 as mortes por envenenamento em todos os grupos de idades rondam os 56% (um total de 12244 mortes por ano) de causas de morte na categoria de morte por ferimento não intencional, o que no ano de 2020 sobe para 62% (um total de 16906 mortes por ano).

Após uma muito extensiva pesquisa durante as semanas que antecederam a entrega final, chegamos à infeliz conclusão que não existe uma base de dados correspondente aos anos 2016-2020 que refira especificamente o tipo de mortes/ferimentos mensalmente. Pelo que conseguimos entender do nosso trabalho de pesquisa parece que a pandemia de COVID-19 veio mudar significativamente a forma como se geram/processam os dados de saúde, pois a maior parte das bases de dados encontradas ou acabam antes de 2020, ou começam em 2020, não havendo praticamente nada que englobe o ano da pandemia com (ao mesmo tempo) anos anteriores e seguintes. Também não obtivemos sucesso a encontrar uma outra base de dados além da usada que seja específica no que toca ao tipo de envenenamento, pelo que tivemos de assumir na análise que o aumento global se iria traduzir da mesma forma num aumento dos casos de envenenamento por injeção de desinfetante.

Conclusão

Em suma, e devido ao facto de a base de dados de saúde ser pouco especifica e não indicar nem o tipo de envenenamento nem a incidência de emergências por unidade de tempo, não pudemos extrair conclusões concretas. Poderá admitir-se a existência de uma correlação entre esta subida de envenenamentos e a subida de tweets que falam sobre envenenamento por injeção de desinfetante, mas sem uma base de dados mais adequada tal será impossível de comprovar.

Foi um trabalho bastante simples na sua apresentação, mas que tendo em conta o nosso background prova ser um trabalho importante para o nosso percurso académico pois conseguimos extrair tweets sem usar o API do Twitter (algo que se pensava impossível) e conseguimos tirar conclusões a partir de dados extraídos sem que fosse necessariamente óbvio.

Como propostas de trabalho futuras, gostaríamos de tentar realizar análise de sentimento aos datasets de tweets para possivelmente determinar se o volume de informações é positivo ou negativo;

Bibliografia

Centers for Disease Control and Prevention. Web-based Injury Statistics Query and Reporting System (WISQARS) [Online]. (2003). National Center for Injury Prevention and Control, Centers for Disease Control and Prevention (producer). Disponível em: URL: www.cdc.gov/injury/wisqars. [2023 05 (maio) 25].

OpenAI. (2023). ChatGPT (May 24 version) [Large language model]. https://chat.openai.com