Universidade da Beira Interior

Prospeção em Redes Sociais e na Web

Relatório Projeto PRSW

Fábio Rico Maio
48286

Professor:
João Paulo da Costa Cordeiro

Junho 2024

Contents

1 Introdução

Neste relatório, apresento a análise de dois conjuntos de dados distintos: posts de um fórum sobre clima e dados de voos domésticos nos EUA durante 2013. O objetivo é aplicar métodos de data mining para gerar conhecimento e realizar visualizações inteligentes que facilitem a interpretação dos dados. Data mining, ou mineração de dados, é o processo de descobrir padrões, anomalias e correlações em grandes conjuntos de dados para prever resultados. Utilizando uma variedade de técnicas, tais como clustering, análise de grafos e visualização de dados, podemos extrair informações valiosas que ajudam na tomada de decisões.

Primeiro, realizei a análise dos posts sobre clima, onde foram aplicadas técnicas de limpeza e transformação de texto em vetores TF-IDF, seguidas da aplicação de clustering para identificar diferentes pontos de vista.

Em seguida, analisei os dados de voos domésticos, criando grafos ponderados para representar as ligações entre aeroportos, estudando os atrasos das companhias aéreas e investigando as possíveis causas dos atrasos. Utilizei algoritmos de centralidade e deteção de comunidades para identificar aeroportos chave e agrupamentos naturais de aeroportos dentro da rede de voos.

2 Análise dos Posts sobre Clima

2.1 Exercício 1: Limpeza e Preparação dos Dados

Neste exercício, realizei a limpeza e preparação dos dados dos posts sobre clima. Utilizei técnicas de limpeza de texto, como remoção de URLs, caracteres de pontuação e números, para preparar os dados dos posts para análise.

	datetime	user	text	senscore	viewpoint
0	2011-06-08 11:56:00	Victory Pete	Have you noticed how extreme the weather has become in the last 20 years Snow in Texas Tornados in Massachusetts Heat waves running rampant Drought and wildfires Just watch the news How about how windy it is at the flying field everyday Last year was my first year flying RC airplanes and everyday was windy I could only get a calm flight if I went at sunup And please dont say Al Gore made the whole thing up VP	0.937400	believer
1	2011-06-08 11:59:00	redhotpearl	I do think burning Fossil Fuels is helping at all if you believe the vast majority of scientists That being said good luck with this topic not to many science majors or even fans around here it seems	0.796400	believer
2	2011-06-08 12:04:00	debogus	Michael Mann made it up Egore is just a opportunist if you believe I believe a lot of scientist have jumped ship on that one East Anglia emails under the rug cause global warming not to many science majors or even fans around here it seems I not a big fan of scare tactic type manipulation for government funding and control	-0.676900	denier
3	2011-06-08 12:09:00	Victory Pete	I believe Egore is an opportunist but I dont believe anyone made it up except Mankind itself VP	0.925200	believer
4	2011-06-08 12:16:00	Looooeeeee!	Let just say I do think we can rule out entirely the influence that the extra CO2 has on recent global mean temperatures There seems to be a certain amount of fossil record supporting this trend	0.612400	believer
5	2011-06-08 12:18:00	ins4fun	Hi Pete If you look back in history Hurricanes Tornados or whatever bad weather we have has always been here Its just SO infrequent that when it does as much damage and gets as much attention as recently people think its something new Back in the 1930s there was a really ugly storm that wrecked lots of property and also in the 1950s i think Dont forget to factor in the fact that such tragedys are much more public thanks to TV news than they were in those years past and it only seems like this is a new trend I think a study of the weather patterns would indicate a more cyclic nature to the incidents of bad weather rather than an all of a sudden increase in one or another type of weather pattern eek EEK I sure know i 'd like to see a whole lot more of honest studys BEFORE another dime goes down the rathole known as Man Made Global Warming Thats for sure cool Cool	0.833500	unsure
6	2011-06-08 12:23:00	debogus	honest studys how would that get any funding	-0.510600	denier
7	2011-06-08 12:26:00	Park_Flyer	Last 20 years How about weather extremes over the last 50 years How 'bout 1,000 20 million years Weather extremes happen if they did the planet would be lifeless A deadly tornado killed nearly in Mass in the 50 that seems a bit more extreme than the recent one I answered no to your poll just in case your wondering	-0.771700	denier

Figure 1: Tabela com dados do climate.csv

Criei também outros 2 gráficos de modo a ter uma melhor compreensão dos dados em estudo:

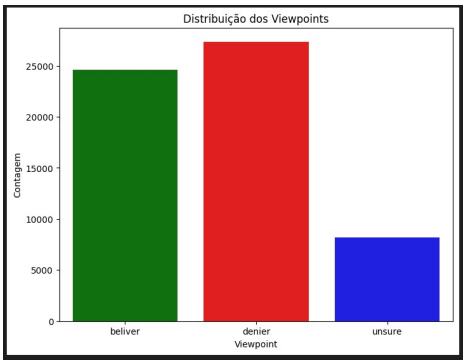

Figure 2: Gráfico de barras que divide entre as 3 emoções das pessoas

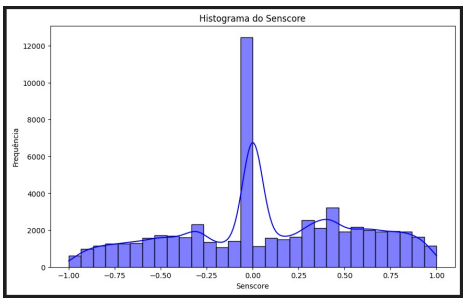

Figure 3: Histograma para vizualizar a distribuição dos dados "senscore"

2.2 Exercício 2: Clustering dos dados

Comecei por transformar os textos dos posts em vetores TF-IDF para análise subsequente. Utilizei a biblioteca 'TfidfVectorizer' do scikit-learn para transformar os textos em vetores TF-IDF, limitando o número de features para reduzir o vocabulário e simplificar os cálculos.

Depois apliquei o algoritmo K-Means para agrupar os posts em clusters, representando diferentes pontos de vista sobre o clima.

Acabei por reduzir a dimensionalidade dos vetores TF-IDF utilizando PCA e apliquei o K-Means para agrupar os posts em três clusters distintos.

A visualização dos clusters mostra três grupos distintos de posts, indicando diferentes pontos de vista sobre o clima.

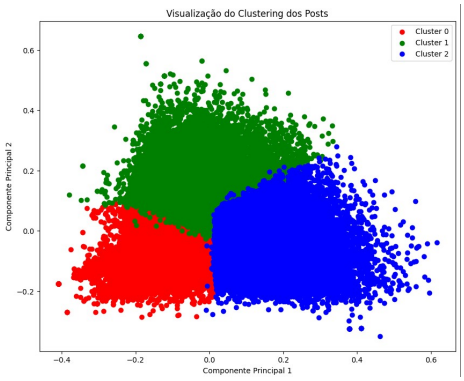

Figure 4: Visualização do Clustering dos Posts em 2D

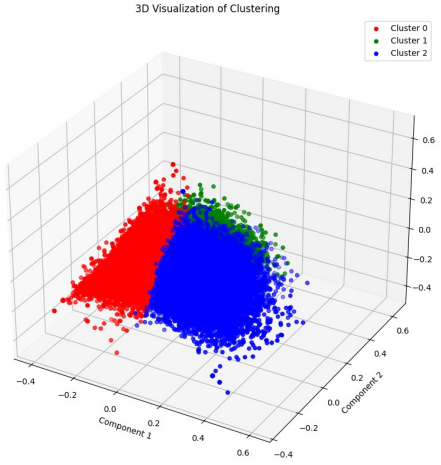

Figure 5: Visualização do Clustering dos Posts em 3D

3 Análise dos Voos Domésticos

3.1 Exercício 3: Grafo de Ligações entre Aeroportos

Neste exercício, representei as ligações entre aeroportos em um grafo, onde os vértices são os aeroportos e as arestas são a força das ligações, baseada no número de voos entre eles.

O grafo criado representa visualmente as conexões entre os aeroportos, tendo os 3 principais no centro e conforme mais grossa for a aresta, maior a força da ligação, destacando os principais hubs de voos.

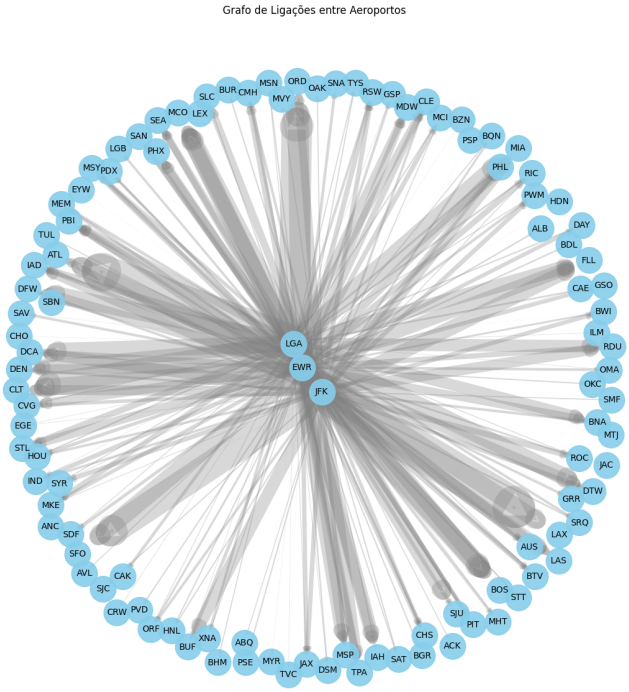

Figure 6: Grafo de Ligações entre Aeroportos

3.2 Exercício 4: Estudo dos Atrasos das Companhias Aéreas

Analisei os atrasos dos voos para diferentes companhias aéreas e gerei um ranking das companhias que mais se atrasam. Calculei o atraso médio por companhia aérea e apresentei os resultados em uma tabela.

Os resultados mostram quais companhias aéreas têm maiores problemas com atrasos, permitindo identificar áreas de melhoria.

	carrier	arr_delay	name
0	F9	21.920705	Frontier Airlines Inc.
1	FL	20.115906	AirTran Airways Corporation
2	EV	15.796431	ExpressJet Airlines Inc.
3	YV	15.556985	Mesa Airlines Inc.
4	OO	11.931034	SkyWest Airlines Inc.
5	MQ	10.774733	Envoy Air
6	WN	9.649120	Southwest Airlines Co.
7	B6	9.457973	JetBlue Airways
8	9E	7.379669	Endeavor Air Inc.
9	UA	3.558011	United Air Lines Inc.
10	US	2.129595	US Airways Inc.
11	VX	1.764464	Virgin America
12	DL	1.644341	Delta Air Lines Inc.
13	AA	0.364291	American Airlines Inc.
14	HA	-6.915205	Hawaiian Airlines Inc.
15	AS	-9.930889	Alaska Airlines Inc.

Figure 7: Lista de atrasos médios dos Aeroportos

3.3 Exercício 5: Representação Gráfica das Ligações Críticas em Termos de Atraso

Representei graficamente as ligações mais críticas em termos de atraso entre os diferentes aeroportos.

A visualização ajuda a identificar as rotas mais problemáticas em termos de atrasos.

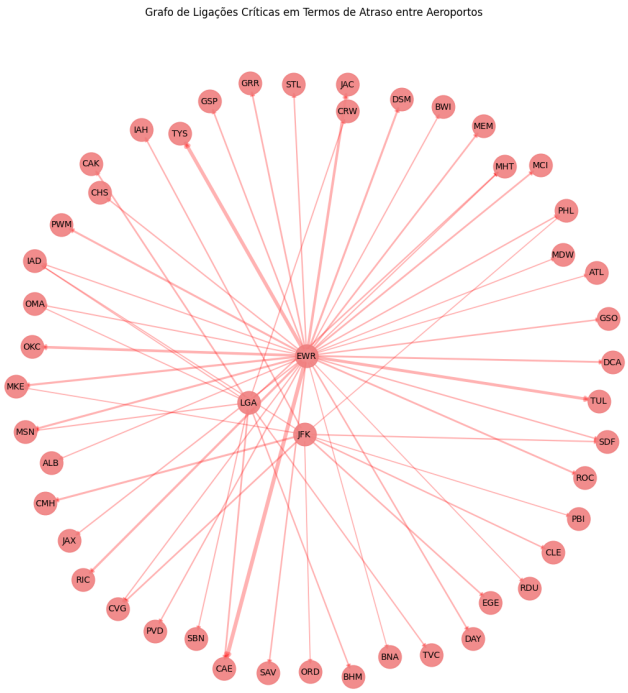

Figure 8: Ligações Críticas entre Aeroportos

3.4 Exercício 6: Estudo das Causas dos Atrasos na Chegada

Analisei possíveis causas para os atrasos na chegada, considerando fatores como a velocidade média do voo. Calculei a velocidade média dos voos e analisei a correlação com o atraso na chegada.

	speed	arr_delay
speed	1.000000	-0.135189
arr_delay	-0.135189	1.000000

Figure 9: Tabela de correlação entre a velocidade media e o atraso

Os resultados indicam uma correlação fraca entre a velocidade média do voo e o atraso na chegada, sugerindo que outros fatores podem ter maior impacto nos atrasos.

3.5 Exercício 7: Algoritmos de Centralidade e Deteção de Comunidades

Utilizei algoritmos de centralidade e deteção de comunidades no grafo das ligações entre aeroportos.

Calculei a centralidade de grau dos nós e detetei comunidades utilizando o algoritmo de modularidade de greedy.

```
Top 5 Aeroportos por Centralidade de Grau:
EWR: 0.0113
JFK: 0.0604
LGA: 0.0509
IAH: 0.0283
MIA: 0.0283

Comunidades de Aeroportos:
Comunidade 1: ['ATL', 'AVL', 'BGR', 'BHM', 'BNA', 'BOS', 'BTV', 'BUF', 'BWI', 'CAE', 'CAK', 'CHO', 'CHS', 'CLE', 'CLT', 'CMH', 'CRW', 'CVG', 'DAY', 'DCA', 'DEN', 'DFW', 'DSM', 'DTW', 'EYW', 'FLL', 'GRR',
Comunidade 2: ['ABQ', 'ACK', 'AUS', 'BQN', 'BUR', 'EGE', 'HNL', 'JAC', 'JFK', 'LAS', 'LAX', 'LGB', 'MVY', 'OAK', 'PDX', 'PHX', 'PSE', 'PSP', 'SAN', 'SAT', 'SEA', 'SFO', 'SJC', 'SJU', 'SLC', 'SMF', 'STT']
Comunidade 3: ['ALB', 'ANC', 'BDL', 'BZN', 'EWR', 'HDN', 'MTJ', 'OKC', 'PVD', 'SNA', 'TUL']
```

Figure 10: Resultado dos graus de centralidade e comunidades.

Identifiquei os principais hubs e as comunidades de aeroportos, proporcionando insights sobre a estrutura da rede de voos.

3.6 Exercício Adicional: Criação de um Gráfico Interativo

3.6.1 Descrição do Exercício

Para enriquecer a análise dos dados de voos domésticos, criei uma visualização interativa do primeiro grafo usando a biblioteca Gravis. Esta visualização permite explorar dinamicamente as conexões entre aeroportos, proporcionando uma experiência mais envolvente.

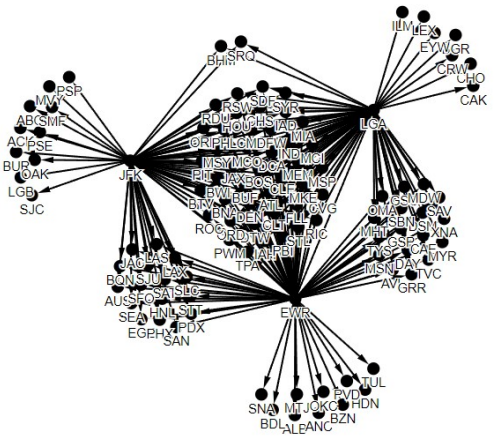

Figure 11: Resultado dos grafo interativo.

4 Conclusão

Neste relatório, utilizei técnicas avançadas de data mining para analisar dois conjuntos de dados distintos, extraindo informações valiosas e criando visualizações significativas. A análise dos posts sobre clima revelou três grupos distintos de discussões, identificados através de clustering, cada um representando diferentes pontos de vista sobre as mudanças climáticas. Esta abordagem ajudou a entender melhor as opiniões predominantes e a dinâmica das discussões online sobre este tema crucial.

Na análise dos voos domésticos, a criação de grafos ponderados permitiu visualizar claramente as conexões entre aeroportos e identificar os principais hubs de voos. O estudo dos atrasos das companhias aéreas revelou quais são as mais propensas a atrasos, proporcionando uma base para intervenções direcionadas. A análise das causas dos atrasos na chegada sugeriu que a velocidade média do voo tem uma correlação fraca com os atrasos, indicando que outros fatores devem ser considerados. Por fim, a aplicação de algoritmos de centralidade e deteção de comunidades no grafo das ligações entre aeroportos destacou os aeroportos mais centrais e os agrupamentos naturais de aeroportos, oferecendo insights sobre a estrutura da rede de voos.

A mineração de dados demonstrou ser uma ferramenta poderosa para extrair conhecimento e insights de grandes conjuntos de dados. As técnicas aplicadas neste relatório não apenas ajudaram a identificar padrões e anomalias, mas também forneceram uma base sólida para possíveis melhorias e intervenções tanto na gestão de discussões online quanto na operação de voos domésticos. A capacidade de transformar dados brutos em informações úteis é crucial para a tomada de decisões informadas e para a melhoria contínua de processos e serviços.