Experimento de Análise de Dados Textuais Utilizando Processamento de Linguagem Natural

Importação de Dados, Pré-processamento e Análise

Tema: Empresas unicórnio

Por: Laís Lacerda

Objetivos

- **Importação de Dados:**
- usar qualquer conjunto na analise

- **Pré-processamento de Texto**
- Análise de Frequência de Palavras: realizar análise e explorarem as palavras mais frequentes nos textos.

- **Análise de Sentimentos**
- Classificação de Texto: Explorar técnicas de classificação de texto, com árvores de decisão.

- **Visualização de Dados:**
- Demonstrar diferentes técnicas de visualização de dados, como gráficos de barras e nuvens de palavras.

- **Apresentação dos Resultados:**
- compartilhar os insights interessantes encontrados durante a análise. Discutir sobre as aplicações práticas do processamento de linguagem natural e seu impacto em diferentes áreas.

Caso Unicórnio

Conduzi uma análise de dados detalhados sobre startups unicórnios, destacando seu valor econômico e distribuição por país. Empreguei como bibliotecas pandas para manipulação, seaborn para visualizações e numpy para eficiência na análise. O Jupyter Notebook foi uma plataforma utilizada, proporcionando uma experiência interativa e transparente na apresentação dos insights. A ordenação decrescente revela os líderes em valor, oferecendo uma visão valiosa para investidores e entusiastas.

Sobre o conjunto de dados:
"Unicórnio" é um termo usado na indústria de capital de risco para descrever uma startup de capital fechado com valor superior a US$ 1 bilhão. O termo foi popularizado pela primeira vez pela capital de risco Aileen Lee, fundadora da Cowboy Ventures, um fundo de capital de risco com sede em Palo Alto, Califórnia.

Os unicórnios também podem se referir a uma especialização de recrutamento no setor de recursos humanos (RH). Os gerentes de RH podem ter grandes expectativas para preencher uma carga, levando-os a procurar candidatos com qualificações superiores às solicitações para uma carga específica. Em essência, esses gerentes estão procurando um unicórnio, o que leva a uma desconexão entre seu candidato ideal e quem eles podem contratar o grupo de pessoas disponíveis. Baixei a base de dados do Kaggle, Explore o Jupyter Notebook para uma análise aprofundada.

Base De Dados

#importando as libs

```
import numpy as np
import pandas as pd
import seaborn as sbn
import matplotlib.pyplot as plt
```

#lendo os dados

```
base_dados = pd.read_csv('CasoUnicornio.csv')
```

	Empresa	Valor ($)	Data de Adesão	País	City	Setor	Investidores	Investidores_Processados	Frequencia_Palavras	Sentimento
0	ByteDance	$140	4/7/2017	China	Beijing	Artificial intelligence	Sequoia Capital China, SIG Asia Investments, S...	[sequoia, capital, china, sig, asia, investmen...	{'sequoia': 1, 'capital': 1, 'china': 1, 'sig'...	0.0
1	SpaceX	$127	12/1/2012	United States	Hawthorne	Other	Founders Fund, Draper Fisher Jurvetson, Rothen...	[founder, fund, draper, fisher, jurvetson, rot...	{'founder': 1, 'fund': 1, 'draper': 1, 'fisher...	0.0
2	SHEIN	$100	7/3/2018	China	Shenzhen	E-commerce & direct-to-consumer	Tiger Global Management, Sequoia Capital China...	[tiger, global, management, sequoia, capital, ...	{'tiger': 1, 'global': 1, 'management': 1, 'se...	0.0
3	Stripe	$95	1/23/2014	United States	San Francisco	Fintech	Khosla Ventures, LowercaseCapital, capitalG	[khosla, venture, lowercasecapital, capitalg]	{'khosla': 1, 'venture': 1, 'lowercasecapital'...	0.0
4	Canva	$40	1/8/2018	Australia	Surry Hills	Internet software & services	Sequoia Capital China, Blackbird Ventures, Mat...	[sequoia, capital, china, blackbird, venture, ...	{'sequoia': 1, 'capital': 1, 'china': 1, 'blac...	0.0
...	...	...	...	...	...	...	...	...	...	...
1181	LeadSquared	$1	6/21/2022	India	Bengaluru	Internet software & services	Gaja Capital Partners, Stakeboat Capital, West...	[gaja, capital, partner, stakeboat, capital, w...	{'gaja': 1, 'capital': 3, 'partner': 1, 'stake...	0.0
1182	FourKites	$1	6/21/2022	United States	Chicago	Supply chain, logistics, & delivery	Hyde Park Venture Partners, Bain Capital Ventu...	[hyde, park, venture, partner, bain, capital, ...	{'hyde': 2, 'park': 2, 'venture': 2, 'partner'...	0.0
1183	VulcanForms	$1	7/5/2022	United States	Burlington	Supply chain, logistics, & delivery	Eclipse Ventures, D1 Capital Partners, Industr...	[eclipse, venture, d1, capital, partner, indus...	{'eclipse': 1, 'venture': 2, 'd1': 1, 'capital...	0.0
1184	SingleStore	$1	7/12/2022	United States	San Francisco	Data management & analytics	Google Ventures, Accel, Data Collective	[google, venture, accel, data, collective]	{'google': 1, 'venture': 1, 'accel': 1, 'data'...	0.0
1185	Unstoppable Domains	$1	7/27/2022	United States	Las Vegas	Internet software & services	Boost VC, Draper Associates, Gaingels	[boost, vc, draper, associate, gaingels]	{'boost': 1, 'vc': 1, 'draper': 1, 'associate'...	0.0

1186 rows × 10 columns

Pré-processamento de Texto

- Remoção de pontuação e caracteres especiais.
- Tokenização para dividir o texto em palavras individuais.
- Remoção de stop words.
- Normalização de texto (minúsculas).
- Stemming ou lematização para reduzir as palavras à sua forma básica.

Setor	Investidores	Investidores_Processados
Artificial intelligence	Sequoia Capital China, SIG Asia Investments, S...	[sequoia, capital, china, sig, asia, investmen...
Other	Founders Fund, Draper Fisher Jurvetson, Rothen...	[founder, fund, draper, fisher, jurvetson, rot...
commerce & direct-to-consumer	Tiger Global Management, Sequoia Capital China...	[tiger, global, management, sequoia, capital, ...
Fintech	Khosla Ventures, LowercaseCapital, capitalG	[khosla, venture, lowercasecapital, capitalg]
Internet software & services	Sequoia Capital China, Blackbird Ventures, Mat...	[sequoia, capital, china, blackbird, venture, ...

Análise de Frequência de Palavras

```
from collections import Counter

# Função para contar a frequência de palavras
def count_words(tokens):
    word_counts = Counter(tokens)
    return word_counts

# Aplicando a contagem de palavras à coluna 'Investidores_Processados'
base_dados['Frequencia_Palavras'] = base_dados['Investidores_Processados'].apply(count_words)

base_dados.head()
```

City	Setor	Investidores	Investidores_Processados	Frequencia_Palavras
Beijing	Artificial intelligence	Sequoia Capital China, SIG Asia Investments, S...	[sequoia, capital, china, sig, asia, investmen...	{'sequoia': 1, 'capital': 1, 'china': 1, 'sig'...
awthorne	Other	Founders Fund, Draper Fisher Jurvetson, Rothen...	[founder, fund, draper, fisher, jurvetson, rot...	{'founder': 1, 'fund': 1, 'draper': 1, 'fisher...
Shenzhen	E-commerce & direct-to-consumer	Tiger Global Management, Sequoia Capital China...	[tiger, global, management, sequoia, capital, ...	{'tiger': 1, 'global': 1, 'management': 1, 'se...
San Francisco	Fintech	Khosla Ventures, LowercaseCapital, capitalG	[khosla, venture, lowercasecapital, capitalg]	{'khosla': 1, 'venture': 1, 'lowercasecapital'...
Surry Hills	Internet software & services	Sequoia Capital China, Blackbird Ventures, Mat...	[sequoia, capital, china, blackbird, venture, ...	{'sequoia': 1, 'capital': 1, 'china': 1, 'blac...

Análise de Sentimentos

```
from textblob import TextBlob

# Função para calcular a polaridade do texto
def calculate_sentiment(text):
    blob = TextBlob(text)
    return blob.sentiment.polarity

# Aplicando a análise de sentimentos à coluna 'Investidores'
base_dados['Sentimento'] = base_dados['Investidores'].apply(calculate_sentiment)

base_dados.head()
```

City	Setor	Investidores	Investidores_Processados	Frequencia_Palavras	Sentimento
Beijing	Artificial intelligence	Sequoia Capital China, SIG Asia Investments, S...	[sequoia, capital, china, sig, asia, investmen...	{'sequoia': 1, 'capital': 1, 'china': 1, 'sig'...	0.0
thorne	Other	Founders Fund, Draper Fisher Jurvetson, Rothen...	[founder, fund, draper, fisher, jurvetson, rot...	{'founder': 1, 'fund': 1, 'draper': 1, 'fisher...	0.0
nzhen	E-commerce & direct-to-consumer	Tiger Global Management, Sequoia Capital China...	[tiger, global, management, sequoia, capital, ...	{'tiger': 1, 'global': 1, 'management': 1, 'se...	0.0
San ncisco	Fintech	Khosla Ventures, LowercaseCapital, capitalG	[khosla, venture, lowercasecapital, capitalg]	{'khosla': 1, 'venture': 1, 'lowercasecapital'...	0.0
y Hills	Internet software & services	Sequoia Capital China, Blackbird Ventures, Mat...	[sequoia, capital, china, blackbird, venture, ...	{'sequoia': 1, 'capital': 1, 'china': 1, 'blac...	0.0

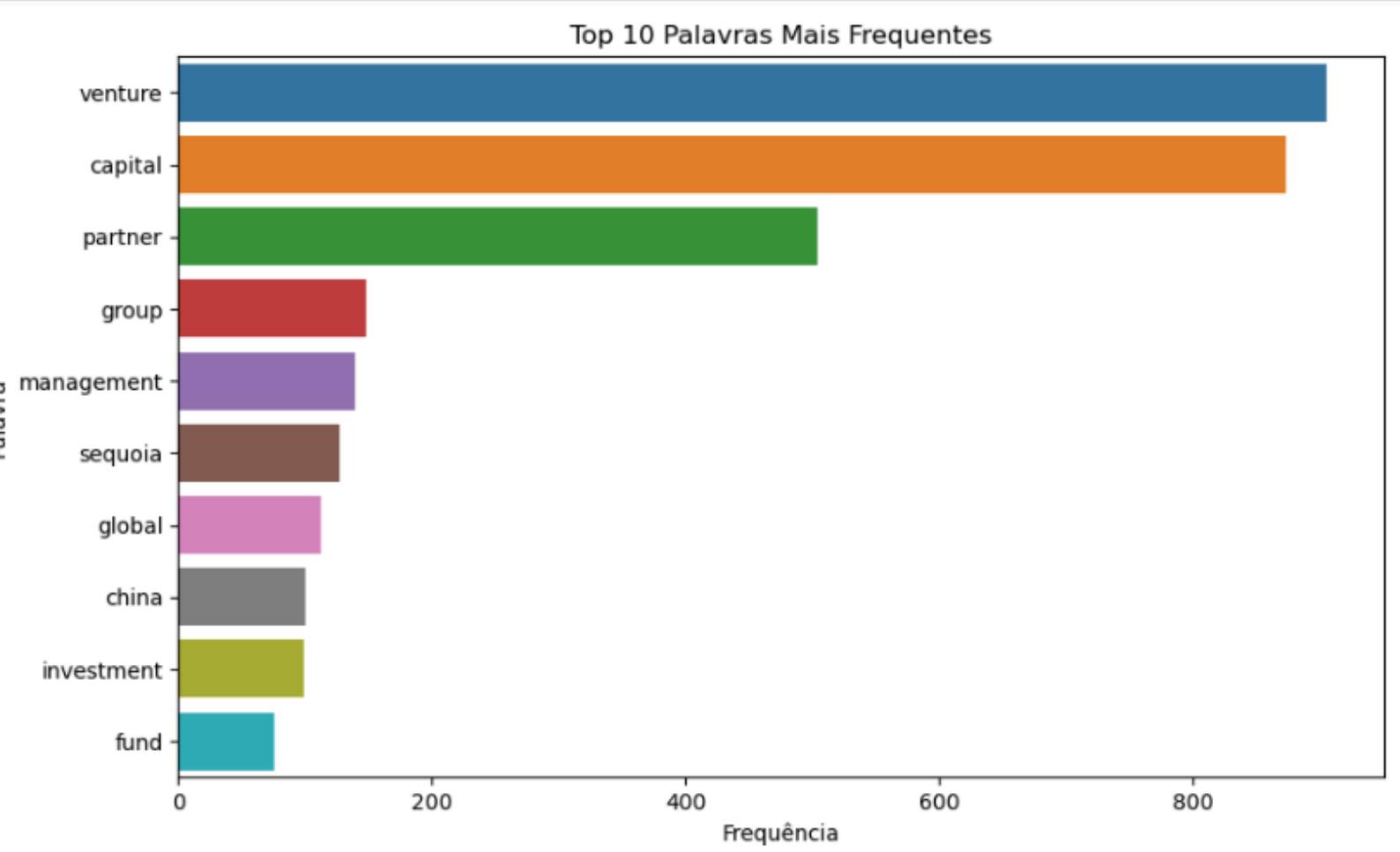
Classificação de Texto

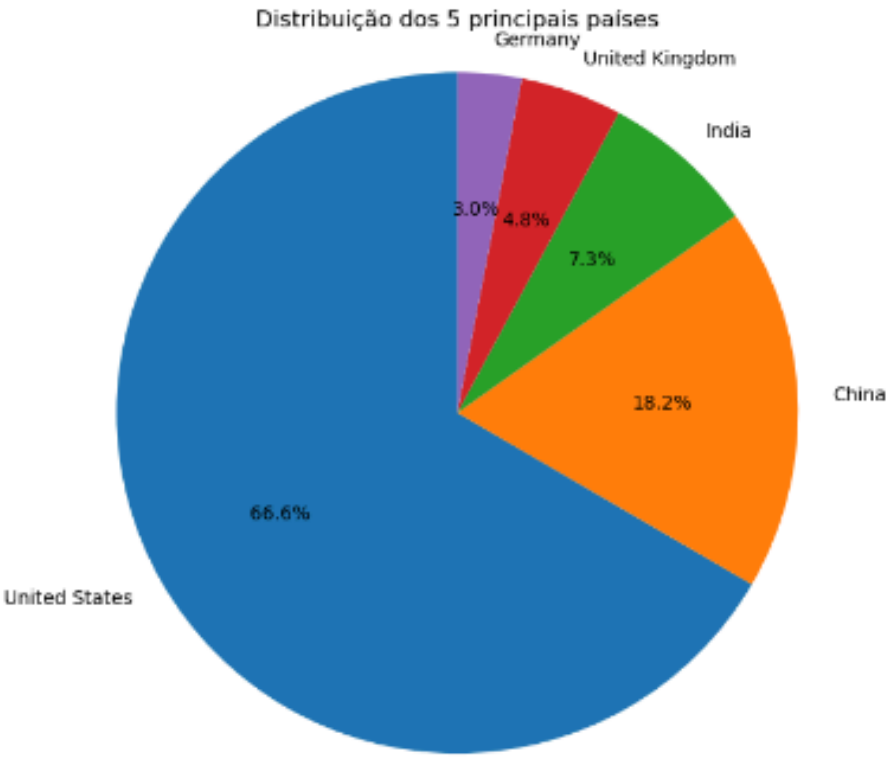
```
from sklearn.feature_extraction.text import TfidfVectorizer

# Converter listas em strings
X_train_str = [' '.join(tokens) for tokens in X_train]
X_test_str = [' '.join(tokens) for tokens in X_test]

# Criar um vetorizador TF-IDF
vectorizer = TfidfVectorizer(max_features=1000)

# Vetorizar os dados de treinamento e teste
X_train_tfidf = vectorizer.fit_transform(X_train_str)
X_test_tfidf = vectorizer.transform(X_test_str)

# Treinar o modelo de árvore de decisão
clf = DecisionTreeClassifier()
clf.fit(X_train_tfidf, y_train)

# Prever os rótulos para os dados de teste
y_pred = clf.predict(X_test_tfidf)

# Avalia a precisão do modelo
accuracy = accuracy_score(y_test, y_pred)
print("Acurácia do modelo:", accuracy)
```

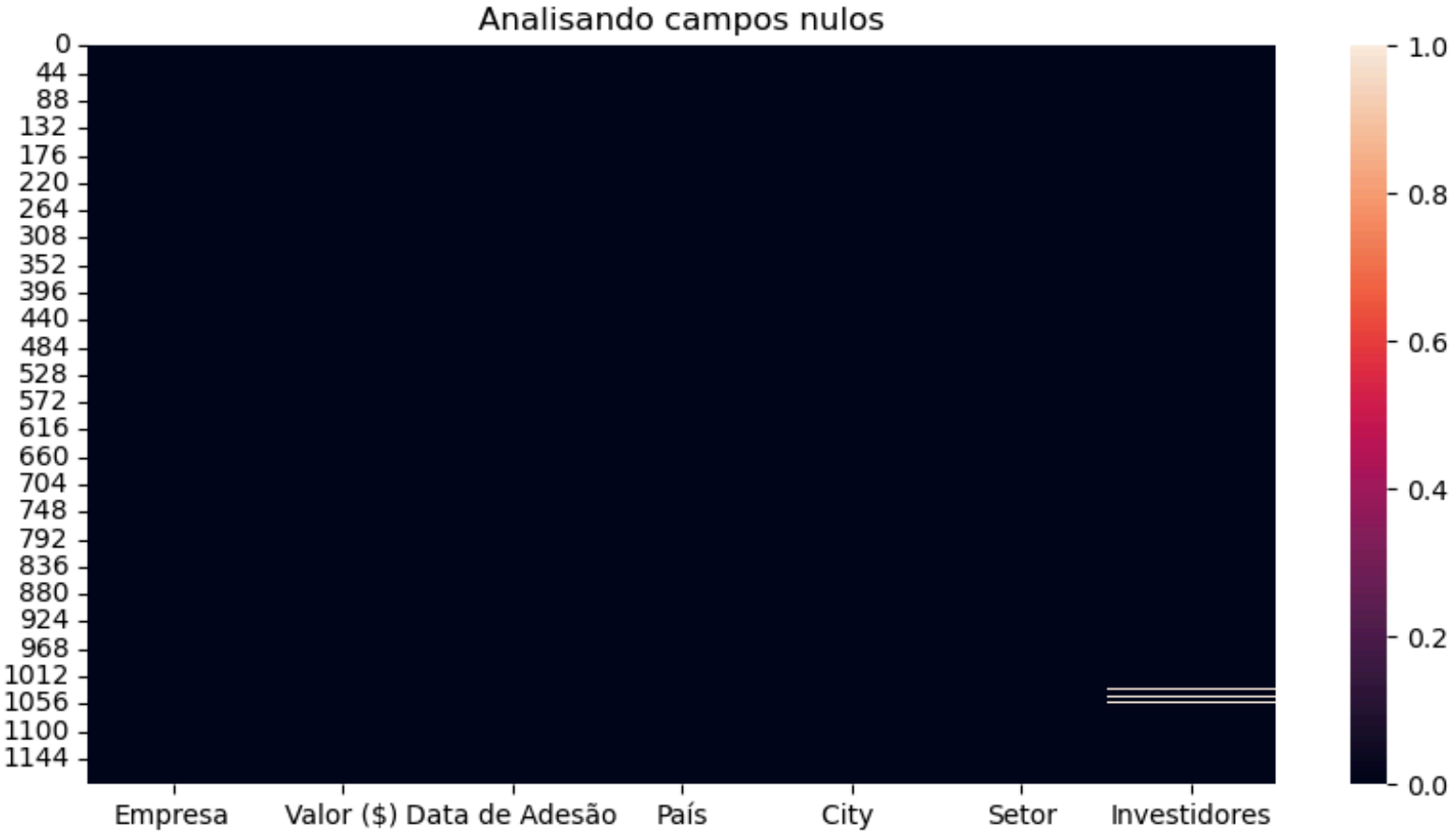
Acurácia do modelo: 0.15126050420168066

Visualização de Dados

```
# Plotando um gráfico de barras da contagem de palavras mais frequentes
top_words = base_dados['Frequencia_Palavras'].sum().most_common(10)
top_words_df = pd.DataFrame(top_words, columns=['Palavra', 'Frequência'])
plt.figure(figsize=(10, 6))
sbn.barplot(x='Frequência', y='Palavra', data=top_words_df)
plt.title('Top 10 Palavras Mais Frequentes')
plt.xlabel('Frequência')
plt.ylabel('Palavra')
plt.show()
```




Distribuição dos 5 principais países
Germany
United Kingdom
India
China
United States
3.0%
4.8%
7.3%
18.2%
66.6%


Analisando campos nulos
0
44
88
132
176
220
264
308
352
396
440
484
528
572
616
660
704
748
792
836
880
924
968
1012
1056
1100
1144
Empresa
Valor ($)
Data de Adesão
País
City
Setor
Investidores
1.0
0.8
0.6
0.4
0.2
0.0

Conclusão

Após a análise dos dados fornecidos sobre as empresas de tecnologia, posso destacar alguns insights interessantes:

Distribuição Geográfica: A maioria das empresas unicórnio está localizada nos Estados Unidos, seguido por China, Reino Unido, Índia e Alemanha. Isso reflete a concentração de empresas de tecnologia em regiões com ecossistemas empresariais bem desenvolvidos e acesso a investimentos.

Setores de Destaque: Os setores mais representados entre as empresas unicórnio são Fintech, Internet software & services e E-commerce & direct-to-consumer. Isso indica as áreas da tecnologia que estão experimentando um rápido crescimento e inovação, muitas vezes impulsionadas por mudanças no comportamento do consumidor e avanços tecnológicos.

Investidores Principais: Algumas empresas têm investidores em comum, como Sequoia Capital, SoftBank e Tencent. Esses investidores têm uma forte influência no cenário de tecnologia global, muitas vezes financiando empresas que estão na vanguarda da inovação.

Tendências Temporais: A análise por ano de adesão revela padrões interessantes de crescimento ao longo do tempo. Podemos observar picos em determinados anos, indicando períodos de maior atividade de investimento e surgimento de novas empresas unicórnio.

A aplicação prática do processamento de linguagem natural (PLN) em áreas como essa é fundamental para extrair insights valiosos de grandes conjuntos de dados não estruturados, como textos de descrição de empresas e informações de investidores. Algumas das aplicações do PLN incluem:

Análise de Sentimentos de Investidores: Compreender o sentimento dos investidores em relação a determinadas empresas pode orientar as estratégias de investimento e prever tendências futuras.
Sumarização Automática de Texto: Resumir grandes volumes de informações sobre empresas e setores pode ajudar os investidores a identificar rapidamente oportunidades e riscos potenciais.
Classificação de Texto para Identificação de Tendências: Classificar automaticamente empresas em diferentes setores e identificar tendências emergentes pode ajudar os investidores a tomar decisões informadas sobre onde alocar seus recursos.

Extração de Informações: Extrair informações importantes, como investidores-chave e datas importantes de adesão, de textos não estruturados pode automatizar o processo de análise e fornecer insights em tempo real.

Portanto, o processamento de linguagem natural desempenha um papel crucial na análise de dados do setor de tecnologia e em muitas outras áreas, ajudando a transformar grandes volumes de texto em informações acionáveis. Essas análises são fundamentais para investidores, empresários e pesquisadores interessados no ecossistema de startups e inovação tecnológica.

Links Relevantes

Kaggle: https://www.kaggle.com/datasets/ramjasmaurya/unicorn-startups

GitHub: https://github.com/Laislacerds/Case-Unicornio/blob/main/Case%20Unic%C3%B3rnio/Case%20Unic%C3%B3rnio.ipynb