

Processamento de Linguagem Natural Utilizando Dados Abertos Governamentais

William Santos Silva

05/08/2024

1 Introdução

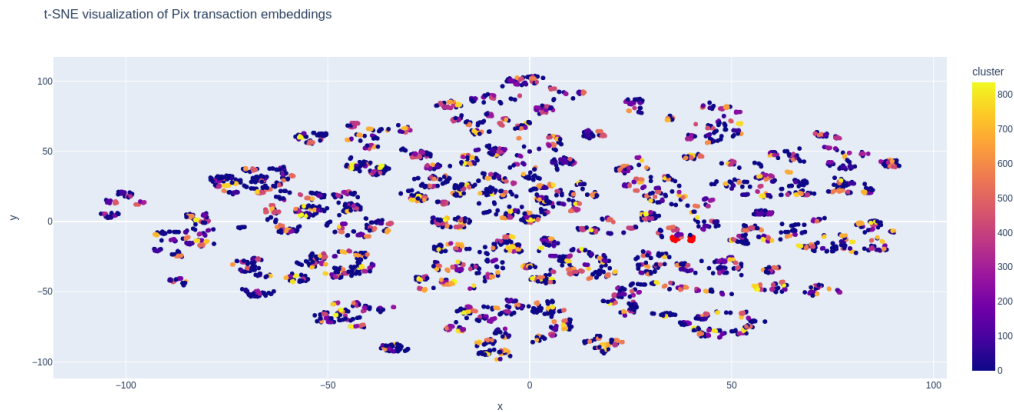


Figura 1: Visualização das transações Pix utilizando t-SNE (t-distributed Stochastic Neighbor Embedding).

Nesta atividade, exploramos a integração de técnicas avançadas de Processamento de Linguagem Natural (PLN) com dados abertos governamentais fornecidos pelo Banco Central do Brasil. O foco principal é na criação e utilização de embeddings por meio de modelos de transformers de última geração.

Para o armazenamento e consulta dos dados vetoriais, utilizamos o Milvus, uma ferramenta robusta para o gerenciamento de grandes volumes de dados vetoriais. A visualização dos dados é realizada com o algoritmo t-SNE

(t-distributed Stochastic Neighbor Embedding), que reduz a dimensionalidade dos dados e facilita a representação visual de suas relações em um espaço bidimensional.

O objetivo é utilizar essas técnicas para identificar transações Pix potencialmente fraudulentas. Partindo da hipótese de que uma transação é suspeita de fraude, realizamos uma busca semântica na base de dados para encontrar transações com características semelhantes, aumentando a probabilidade de identificar atividades fraudulentas.

2 Seleção do Conjunto de Dados

O conjunto de dados selecionado foi obtido do Banco Central do Brasil e contém estatísticas detalhadas sobre transações Pix. Este conjunto de dados inclui os seguintes parâmetros:

	A	B	C	D	E	F	G	H	I	J	K	L
	AnoMes	PAG_PFPJ	REC_PFPJ	PAG_REGIAO	REC_REGIAO	PAG_IDADE	REC_IDADE	FORMAINICIACAO	NATUREZA	FINALIDADE	VALOR	QUAN
1	202302	PF	PF	CENTRO-OESTE	NORDESTE	entre 30 e 39 anos	entre 30 e 39 anos	QRES	P2P	Pix	1148814.13	
2	202202	PF	PJ	Nao informado	CENTRO-OESTE	entre 20 e 29 anos	Nao se aplica	QRDN	P2G	Pix	21822	
3	202404	PJ	PF	NORDESTE	SUL	Nao se aplica	entre 50 e 59 anos	QRES	B2P	Pix	196557.95	
4	202307	PF	PF	Nao informado	NORDESTE	entre 50 e 59 anos	entre 20 e 29 anos	QRES	P2P	Pix	47498.03	
5	202402	PF	PF	SUL	Nao informado	entre 50 e 59 anos	até 19 anos	QRES	P2P	Pix	1431	
6	202305	PF	PF	NORDESTE	NORDESTE	entre 50 e 59 anos	mais de 60 anos	MANU	P2P	Pix	99224255.65	
7	202209	PJ	PJ	CENTRO-OESTE	CENTRO-OESTE	Nao se aplica	Nao se aplica	Nao disponivel	G2B	Nao disponivel	63.03	
8	202212	PF	PF	SUDESTE	CENTRO-OESTE	até 19 anos	entre 30 e 39 anos	QRDN	P2P	Pix	30316.61	
9	202203	PF	PF	SUL	NORDESTE	mais de 60 anos	até 19 anos	QRES	P2P	Pix	382.58	
10	202107	PF	PF	Nao informado	SUL	entre 20 e 29 anos	entre 20 e 29 anos	Nao disponivel	P2P	Nao disponivel	1366	
11	202406	PF	PF	SUL	NORDESTE	entre 40 e 49 anos	mais de 60 anos	MANU	P2P	Pix	1746678.27	
12	202110	PF	PF	Nao informado	SUL	entre 20 e 29 anos	entre 20 e 29 anos	QRDN	P2P	Pix	520.38	
13	202104	PJ	PJ	SUL	SUL	Nao se aplica	Nao se aplica	Nao disponivel	B2G	Nao disponivel	4346878.63	
14	202308	PF	PF	Nao informado	NORDESTE	até 19 anos	até 19 anos	QRES	P2P	Pix	407.42	
15	202401	PJ	PF	NORDESTE	Nao informado	Nao se aplica	entre 20 e 29 anos	MANU	B2P	Pix	504563.16	
16	202111	PF	PJ	SUDESTE	Nao informado	entre 40 e 49 anos	Nao se aplica	DICT	P2B	Pix	50027.75	
17	202211	PF	PF	CENTRO-OESTE	Nao informado	entre 20 e 29 anos	até 19 anos	Nao disponivel	P2P	Nao disponivel	1	
18	202407	PF	PJ	Nao informado	NORDESTE	entre 20 e 29 anos	Nao se aplica	QRDN	P2B	Pix Saque	5843.45	
19	202211	PF	PF	NORDESTE	Nao informado	até 19 anos	mais de 60 anos	QRES	P2P	Pix	60	
20	202106	PF	PF	NORDESTE	NORTE	entre 40 e 49 anos	até 19 anos	DICT	P2P	Pix	16643.83	
21	202402	PF	PF	SUDESTE	NORDESTE	entre 50 e 59 anos	até 19 anos	DICT	P2P	Pix	8839756.31	
22	202401	PF	PF	SUDESTE	NORTE	entre 30 e 39 anos	mais de 60 anos	QRES	P2P	Pix	248365.85	
23	202311	PF	PJ	SUDESTE	SUL	mais de 60 anos	Nao se aplica	QRES	P2B	Pix Saque	986.66	
24	202406	PF	PF	CENTRO-OESTE	SUDESTE	entre 50 e 59 anos	entre 30 e 39 anos	MANU	P2P	Pix	1323635.34	

Figura 2: Exemplo do conjunto de dados de transações Pix.

- **AnoMes:** Período da transação.
- **PAG_PFPJ:** Tipo de pagador (Pessoa Física ou Jurídica).
- **REC_PFPJ:** Tipo de receptor (Pessoa Física ou Jurídica).
- **PAG_REGIAO:** Região do pagador.
- **REC_REGIAO:** Região do receptor.
- **PAG_IDADE:** Faixa etária do pagador.
- **REC_IDADE:** Faixa etária do receptor.

- **FORMAINICIACAO:** Formação de iniciação da transação.
- **NATUREZA:** Natureza da transação.
- **FINALIDADE:** Finalidade da transação.
- **VALOR:** Valor da transação.
- **QUANTIDADE:** Quantidade de transações.

Este conjunto foi escolhido devido à sua amplitude, com mais de 400 mil registros, e à sua relevância para análises financeiras.

Para tornar o processamento viável, foi selecionada uma amostra de 10 mil registros. A amostra foi obtida utilizando a função `sample` do pandas com a semente (seed) 42, garantindo a reprodutibilidade dos dados amostrados por meio de um processo pseudoaleatório que pode ser consistentemente replicado.

Antes do processo de embedding, os dados de cada linha foram concatenados em uma única string. Essa abordagem permite representar os dados de uma transação como um vetor n-dimensional. Por exemplo, uma transação pode ser representada da seguinte forma:

```
"AnoMes: 202212, PAG_PFPJ: PF, REC_PFPJ: PF, PAG_REGIAO:
NORDESTE, REC_REGIAO: NORDESTE, PAG_IDADE: Nao informado,
REC_IDADE: entre 20 e 29 anos, FORMAINICIACAO: DICT, NATUREZA:
P2P, FINALIDADE: Pix, VALOR: 2963,41, QUANTIDADE: 21"
```

Essa concatenação facilita a criação de embeddings que capturam a complexidade dos dados transacionais em um espaço vetorial.

URL de Acesso: https://olinda.bcb.gov.br/olinda/servico/Pix_DadosAbertos/versao/v1/aplicacao#!/recursos/EstatisticasTransacoesPix

2.1 Justificativa

O conjunto de dados das transações Pix foi selecionado devido à sua relevância no contexto econômico e financeiro do Brasil, além de atender aos critérios de volume de dados exigidos pela atividade.

3 Escolha do Modelo de Embeddings

Para a criação dos embeddings, foi escolhido o modelo `PORTULAN/serafim-335m-portuguese-pt` disponível no Hugging Face. Este modelo é um *sentence-transformer* especializado na língua portuguesa, projetado para mapear sentenças e parágrafos

em um espaço vetorial denso de 1024 dimensões. É particularmente adequado para tarefas como clustering e busca semântica.

URL do Modelo: <https://huggingface.co/PORTULAN/serafim-335m-portuguese-pt-se>

3.1 Justificativa

O modelo `PORTULAN/serafim-335m-portuguese-pt-sentence-encoder-ir` foi selecionado por sua eficácia na geração de embeddings de alta qualidade para textos em português. Sua capacidade de mapear sentenças e parágrafos em um espaço vetorial denso de 1024 dimensões é ideal para análises semânticas precisas. Além disso, o modelo é afinado para tarefas de *Information Retrieval* (IR), tornando-o particularmente útil para aplicações de busca semântica e clustering, atendendo aos requisitos da nossa análise.

O uso deste modelo permite capturar as nuances e a semântica dos dados textuais, contribuindo significativamente para a eficácia das análises realizadas com os embeddings gerados.

4 Criação dos Embeddings

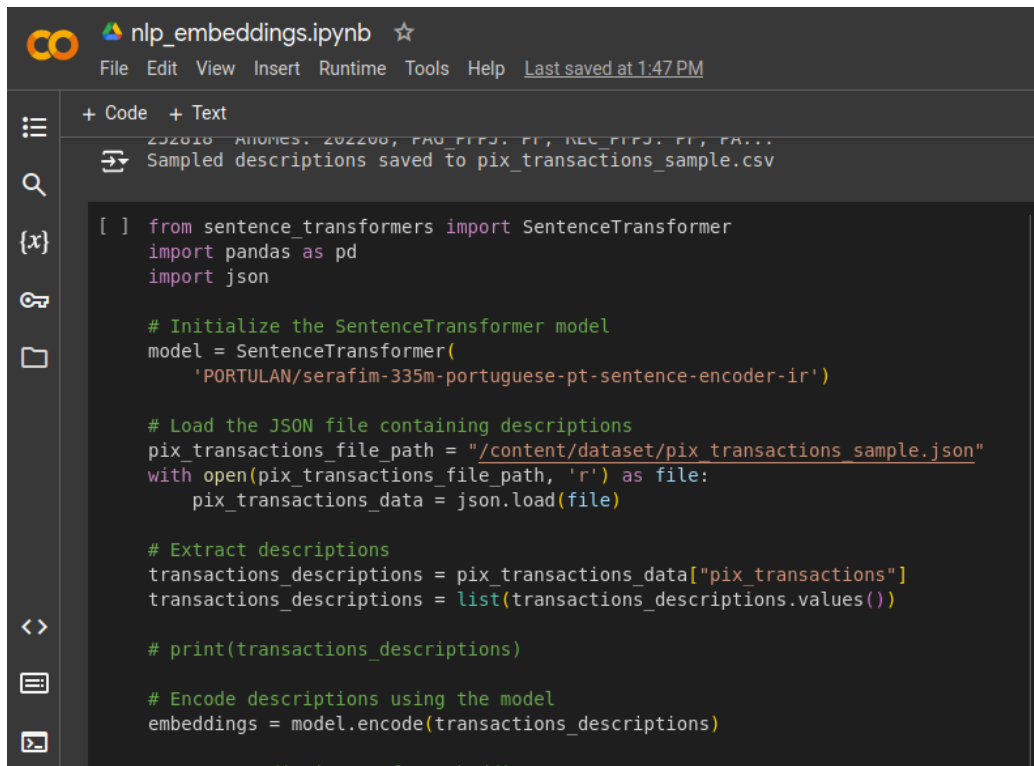
O processo de criação dos embeddings envolveu a utilização do modelo selecionado para converter as descrições das transações Pix em vetores de alta dimensão. Esses vetores capturam as características semânticas das descrições, permitindo a realização de análises posteriores.

4.1 Passos Envolvidos

1. **Carregamento das Descrições:** As descrições das transações foram carregadas a partir do arquivo JSON.
2. **Codificação das Descrições:** Utilizando o modelo de embeddings, as descrições foram convertidas em vetores. Cada descrição foi representada como uma string no formato:

“AnoMes: 202212, PAG_PFPJ: PF, REC_PFPJ: PF, PAG_REGIAO: NORDESTE, REC_REGIAO: NORDESTE, PAG_IDADE: Nao informado, REC_IDADE: entre 20 e 29 anos, FORMAI-NICIACAO: DICT, NATUREZA: P2P, FINALIDADE: Pix, VALOR: 2963,41, QUANTIDADE: 21”

Esta string foi então utilizada pelo modelo para gerar um vetor de 1024 dimensões. Para o processamento, foi utilizado o Google Colab que disponibiliza GPU para o processamento.



```
252018 - ANOMES: 202200, PAG_FPPJ: PT, REC_FPPJ: PT, PA...
Sampled descriptions saved to pix_transactions_sample.csv

[ ] from sentence_transformers import SentenceTransformer
import pandas as pd
import json

# Initialize the SentenceTransformer model
model = SentenceTransformer(
    'PORTULAN/serafim-335m-portuguese-pt-sentence-encoder-ir')

# Load the JSON file containing descriptions
pix_transactions_file_path = "/content/dataset/pix_transactions_sample.json"
with open(pix_transactions_file_path, 'r') as file:
    pix_transactions_data = json.load(file)

# Extract descriptions
transactions_descriptions = pix_transactions_data["pix_transactions"]
transactions_descriptions = list(transactions_descriptions.values())

# print(transactions_descriptions)

# Encode descriptions using the model
embeddings = model.encode(transactions_descriptions)

# Create a dictionary for embeddings
```

Figura 3: Ambiente Google Colab utilizado para processamento com GPU.

3. **Armazenamento dos Embeddings:** Os vetores resultantes foram armazenados em um arquivo JSON para uso posterior.

5 Armazenamento dos Embeddings no Banco de Dados Vetorial

Os embeddings foram armazenados no banco de dados vetorial Milvus, que permite consultas eficientes utilizando distâncias como a euclidiana e cosseno, entre outras, para similaridade semântica.

5.1 Passos Envolvidos

1. **Criação da Conexão com Milvus:** Estabeleceu-se uma conexão com o banco de dados Milvus.

```

dataset > {} pix_transaction_query.json > [ ] vector
1  {}
2  "vector": [
3  [
4      0.40860795974731445, -0.4091181457042694, 0.1914035975933075,
5      -0.2562173306941986, -0.4523962140083313, -0.9748979806900024,
6      0.20825627446174622, 1.2119090557098389, -0.10062997043132782,
7      1.0251961946487427, 0.13528741896152496, 0.26693856716156006,
8      -0.5651956796646118, 0.12902607023715973, -1.372467279434204,
9      -0.3929663896560669, -0.4253169894218445, 0.17604108154773712,
10     0.6958875060081482, -0.28480061888694763, -0.14502757787704468,
11     -0.1155146062374115, -1.0542749166488647, -0.5144249796867371,
12     -0.28820323944091797, 0.18217363953590393, 0.1245439425110817,
13     0.1238718181848526, -0.24676378071308136, -0.5888280272483826,
14     -0.30629998445510864, 0.01040477305650711, -0.4059048295021057,
15     0.1280483454465866, 0.2056107372045517, 0.4483802616596222,
16     0.3946329355239868, -0.17639392614364624, -0.5727613568305969,
17     -0.15239129960536957, 0.18291905522346497, -1.4511972665786743,
18     -0.7718914747238159, -0.4181630313396454, -0.13451875746250153,
19     -1.3311071395874023, 0.5914084911346436, -0.5682899355888367,
20     -0.3239600956439972, -0.07760492712259293, 0.24859662353992462,
21     0.8740231990814209, 0.4562850296497345, -0.1302027851343155,
22     0.09547830373048782, 0.24524912238121033, -0.22339226305484772,
23     -0.6423808932304382, -0.45504555106163025, 1.6877763271331787,
24     -0.6802427172660828, -0.1339883655309677, -1.049345850944519,
25     0.05901473015546799, -0.09719520807266235, -0.5684126019477844,
26     -0.5790218114852905, -0.7605175971984863, -1.2056747674942017,
27     0.9930568933486938, 0.7801121473312378, 0.6952186226844788,
28     0.7180768251419067, -0.037361208349466324, -0.004998489748686552,
29     -0.18447518348693848, -0.7278143763542175, -1.1075934171676636,
30     0.5259660482406616, 0.21022817492485046, 1.434178113937378,
31     -0.27715003490448, 0.6384868621826172, 0.8558654189109802,

```

Figura 4: Vetor de 1024 dimensões gerado a partir da string acima.

2. **Preparação dos Dados:** Os dados foram preparados para serem inseridos no banco de dados, garantindo a correspondência entre os textos originais e seus embeddings.

```

# Create the dataset list
self.dataset = [
    {
        "id": int(key),
        "vector": self.vector_transactions["pix_transactions_embeddings"][text_keys.index(key)],
        "text": self.text_transactions["pix_transactions"][key]
    }
    for key in text_keys
]

```

Figura 5: Schema do banco de dados contendo ID, vetor e texto original.

3. **Inserção dos Dados:** Os embeddings foram inseridos no banco de dados Milvus.

6 Consultas de Similaridade Semântica

Utilizando a funcionalidade de busca por similaridade do Milvus, foram realizadas consultas para identificar transações Pix semelhantes com base nos embeddings gerados. No exemplo em questão, foi utilizada a string:

“AnoMes: 202212, PAG_PFPJ: PF, REC_PFPJ: PF, PAG_REGIAO: NORDESTE, REC_REGIAO: NORDESTE, PAG_IDADE: Nao informado, REC_IDADE: entre 20 e 29 anos, FORMAINICIA-CAO: DICT, NATUREZA: P2P, FINALIDADE: Pix, VALOR: 2963,41, QUANTIDADE: 21”

Como hipótese de uma transação potencialmente fraudulenta, com o objetivo de encontrar outras transações semelhantes.

6.1 Passos Envolvidos

1. **Formulação da Consulta:** Consultas foram formuladas para identificar descrições de transações semelhantes, possivelmente fraudulentas.
2. **Execução da Busca:** A busca foi executada no banco de dados Milvus utilizando a distância cosseno.
3. **Interpretação dos Resultados:** Os resultados foram interpretados para identificar padrões e insights. Por exemplo, ao buscar por 20 transações semelhantes, obteve-se o gráfico mostrado abaixo:

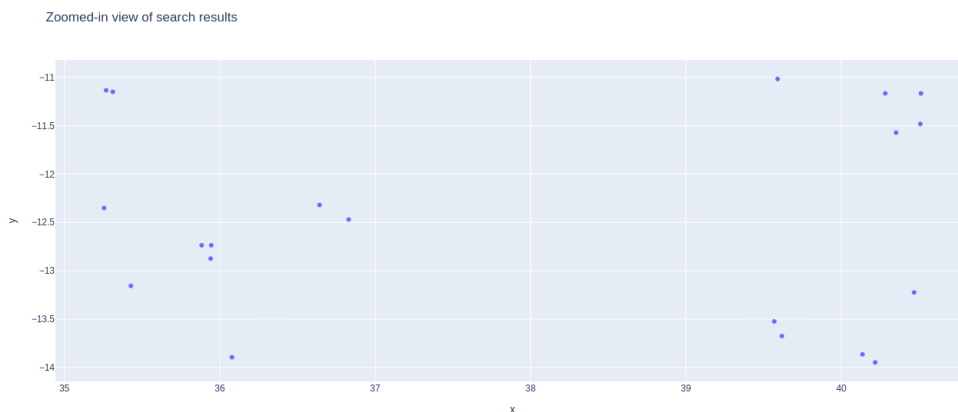


Figura 6: 20 transações semelhantes à string pesquisada.

7 Visualização dos Dados

Para a visualização dos dados, foram utilizadas técnicas de redução de dimensionalidade e clustering, combinadas com ferramentas de visualização interativa.

7.1 Passos Envolvidos

1. **Redução de Dimensionalidade:** Utilizou-se t-SNE para reduzir a dimensionalidade dos embeddings.
2. **Clustering:** Aplicou-se DBSCAN para identificar clusters nos dados. e tbm o T SNE
3. **Visualização:** Os dados foram visualizados utilizando Plotly para visualização interativa e Matplotlib para visualizações estáticas.

8 Conclusão

A atividade demonstrou como técnicas avançadas de PLN e ferramentas modernas de armazenamento e consulta de dados vetoriais podem ser aplicadas para analisar grandes volumes de dados governamentais. A integração dessas ferramentas permite realizar análises sofisticadas, oferecendo novos insights sobre os dados.