Relatório de Análise de Similaridade de Crimes entre Municípios Brasileiros

Gabriel Pietro Leone
Nº USP: 13874729

30 de maio de 2025

1 Introdução

Este relatório apresenta a análise de similaridade de crimes entre municípios brasileiros, utilizando técnicas de processamento de linguagem natural e cadeias de Markov. O objetivo é mapear descrições textuais de assaltos em um grafo conexo, onde os vértices representam cidades e as arestas são baseadas na similaridade semântica das descrições. A partir do grafo, foi construída uma matriz de transição estocástica para estimar as probabilidades de ocorrência futura de assaltos em cada cidade, utilizando a distribuição estacionária de uma cadeia de Markov.

O projeto foi implementado em Python, utilizando bibliotecas como `sentence-transformer`
`networkx`, `pandas`, `numpy` e `scikit-learn`. O código foi estruturado em um Jupyter Notebook para facilitar a reprodutibilidade e visualização dos resultados.

2 Metodologia

2.1 Coleta e Pré-processamento dos Dados

Os dados foram obtidos de um arquivo Excel (`assaltos.xlsx`), contendo descrições textuais de assaltos e os respectivos municípios. O arquivo foi carregado utilizando a biblioteca `pandas`, e verificou-se a integridade dos dados, incluindo o número de registros e municípios únicos.

2.2 Geração de Embeddings

Para capturar o significado semântico das descrições dos assaltos, foi utilizado o modelo `paraphrase-MiniLM-L6-v2` da biblioteca `sentence-transformers`. Este modelo transforma cada descrição em um vetor numérico (embedding) no espaço vetorial. As descrições foram processadas, e os embeddings foram agregados por município, calculando-se a média dos vetores para cada cidade.

2.3 Construção do Grafo

A similaridade entre os embeddings foi calculada utilizando a métrica de similaridade do cosseno. Uma matriz de similaridade foi gerada, onde cada elemento representa a similaridade entre os embeddings de duas cidades. Um limiar de similaridade de 0.7 foi definido para determinar as conexões no grafo: cidades com similaridade igual ou superior a esse valor foram conectadas por uma aresta ponderada pelo valor da similaridade.

O grafo foi construído utilizando a biblioteca `networkx`, resultando em um grafo não direcionado onde os vértices são os municípios e as arestas representam a similaridade semântica entre as descrições dos assaltos.

2.4 Análise com Cadeias de Markov

A partir do grafo, foi gerada uma matriz de adjacência, que foi normalizada para formar uma matriz de transição estocástica. Cada elemento da matriz de transição representa a probabilidade de transição de uma cidade para outra com base na similaridade dos padrões de criminalidade.

A distribuição estacionária foi calculada resolvendo o sistema de equações associado ao autovalor 1 da matriz de transição transposta, utilizando a biblioteca `numpy`. Essa distribuição representa as probabilidades de longo prazo de ocorrência de assaltos em cada cidade, assumindo que o padrão de similaridade se mantém constante.

2.5 Salvamento e Visualização dos Resultados

Os resultados foram salvos em arquivos CSV, incluindo a matriz de similaridade, a matriz de transição e a distribuição estacionária. Um relatório resumido foi gerado em formato de texto, contendo as 10 cidades com maiores probabilidades estacionárias e estatísticas descritivas da distribuição. A visualização do grafo e da matriz de adjacência foi planejada utilizando `matplotlib` e `networkx`, embora a implementação específica da visualização dependa do conjunto de dados.

3 Resultados

A análise gerou os seguintes resultados principais:

- **Matriz de Similaridade**: Uma matriz simétrica foi gerada, indicando a similaridade semântica entre os padrões de criminalidade de diferentes cidades.
- **Grafo Conexo**: O grafo construído apresentou conexões entre cidades com padrões de assaltos semelhantes, com arestas ponderadas pela similaridade do cosseno.
- **Matriz de Transição**: A matriz de transição estocástica foi derivada da matriz de adjacência, garantindo que a soma das probabilidades em cada

linha fosse igual a 1.

- **Distribuição Estacionária**: A distribuição estacionária foi calculada, identificando as cidades com maior probabilidade de ocorrência futura de assaltos. As 10 cidades com maiores probabilidades foram listadas no relatório resumido.
- **Estatísticas**: Foram calculadas estatísticas descritivas da distribuição estacionária, incluindo média, desvio padrão, máximo, mínimo e mediana.

4 Análise e Discussão

A distribuição estacionária indica as cidades com maior probabilidade de ocorrência de assaltos com base na similaridade dos padrões criminais. Cidades com alta probabilidade estacionária podem ser interpretadas como aquelas com padrões de criminalidade mais consistentes ou influentes na rede, possivelmente devido a fatores socioeconômicos ou geográficos que favorecem a recorrência de assaltos semelhantes.

O limiar de similaridade (0.7) foi escolhido para equilibrar a conectividade do grafo, evitando um grafo excessivamente denso ou desconexo. A escolha do modelo `paraphrase-MiniLM-L6-v2` foi justificada por sua eficiência e capacidade de capturar nuances semânticas em textos curtos, como descrições de crimes.

A análise com cadeias de Markov assume que os padrões de criminalidade seguem uma dinâmica estacionária, o que pode ser uma simplificação. Fatores externos, como políticas públicas ou mudanças socioeconômicas, não foram incorporados no modelo, mas poderiam ser considerados em análises futuras.

5 Conclusão

O projeto demonstrou a viabilidade de utilizar embeddings de texto e cadeias de Markov para modelar padrões de criminalidade entre cidades brasileiras. As cidades com maiores probabilidades estacionárias foram identificadas como potenciais focos de atenção para políticas de segurança pública. O código implementado é modular e pode ser adaptado para outros conjuntos de dados ou limiares de similaridade.

Para trabalhos futuros, sugere-se incorporar dados adicionais, como variáveis socioeconômicas, e explorar outros modelos de embeddings ou métodos de análise de grafos, como comunidades ou centralidade, para enriquecer a análise.