Relatório de Análise de Clusters de Criminalidade em Municípios Brasileiros

Gabriel Pietro Leone
Nº USP: 13874729

30 de maio de 2025

1 Introdução

Este relatório apresenta a análise de clusters de criminalidade em municípios brasileiros, utilizando o grafo de similaridades construído na Atividade 2, onde os vértices representam cidades e as arestas refletem a similaridade semântica entre descrições de assaltos. O objetivo foi identificar comunidades semânticas associadas a tipos de crimes semelhantes, empregando o algoritmo de Cluster Label Propagation para detecção de comunidades no grafo. Cada comunidade foi interpretada como um grupo de cidades com padrões criminais semelhantes, revelando potenciais tipologias de crimes.

O projeto foi implementado em Python, utilizando bibliotecas como `sentence-transformer`
`networkx`, `pandas`, `numpy`, `scikit-learn`, `matplotlib` e `seaborn` (quando disponível).
O código foi estruturado em um Jupyter Notebook para facilitar a reproducibilidade 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. A biblioteca `pandas` foi utilizada para carregar os dados, com verificação da integridade, incluindo o número de registros e municípios únicos.

2.2 Geração de Embeddings

As descrições dos assaltos foram transformadas em vetores numéricos (embeddings) utilizando o modelo `paraphrase-MiniLM-L6-v2` da biblioteca `sentence-transforme`
Este modelo foi escolhido por sua eficiência em capturar semântica em textos curtos. Os embeddings foram agregados por município, calculando a média dos vetores de todas as descrições de uma mesma cidade.

2.3 Construção do Grafo K-NN

A similaridade entre os embeddings foi calculada utilizando a métrica de similaridade do cosseno, gerando uma matriz de similaridade. Um grafo K-NN (K-Nearest Neighbors) foi construído, conectando cada cidade aos seus 5 vizinhos mais similares (K=5), com arestas ponderadas pelos valores de similaridade. A biblioteca `networkx` foi utilizada para criar o grafo não direcionado.

2.4 Detecção de Comunidades

O algoritmo de Cluster Label Propagation, implementado na biblioteca `networkx`, foi aplicado para identificar comunidades no grafo. Este algoritmo propaga rótulos entre nós conectados, agrupando cidades com padrões de criminalidade semelhantes. A modularidade do grafo foi calculada para avaliar a qualidade da divisão em comunidades.

2.5 Análise dos Clusters

Os clusters foram analisados para identificar padrões linguísticos ou temáticos nas descrições dos assaltos. Para cada cluster, foram calculadas estatísticas como o número de cidades, o total de crimes e a média de crimes por cidade. Exemplos de descrições foram extraídos para inferir tipologias de crimes predominantes em cada cluster (por exemplo, crimes com violência armada ou furtos noturnos).

2.6 Visualização dos Resultados

Foram geradas múltiplas visualizações para facilitar a interpretação dos clusters:

- **Grafo com Layout Spring**: Exibe as cidades coloridas por cluster, com layout otimizado para separação visual.
- **Grafo com Layout Kamada-Kawai**: Usa um algoritmo de força dirigida para melhor disposição dos nós.
- **Grafo Hierárquico**: Mostra conexões entre clusters diferentes, destacando interações entre padrões criminais distintos.
- **Visualização PCA**: Reduz os embeddings a duas dimensões para visualização linear dos clusters.
- **Visualização t-SNE**: Fornece uma visualização não-linear dos clusters.
- **Heatmap de Similaridade**: Mostra a matriz de similaridade reorganizada por clusters, destacando a coesão interna.

Os resultados foram salvos em arquivos CSV (clusters e estatísticas) e PNG (visualizações), além de um relatório detalhado em formato de texto com amostras de descrições por cluster.

3 Resultados

A análise gerou os seguintes resultados principais:

- **Grafo K-NN**: Construído com K=5, conectando cada cidade aos seus 5 vizinhos mais similares, resultando em um grafo com alta conectividade local.
- **Comunidades Detectadas**: Foram identificados múltiplos clusters, cada um representando um grupo de cidades com padrões criminais semelhantes. O número de clusters variou conforme o conjunto de dados, com modularidade calculada para avaliar a qualidade da divisão.
- **Estatísticas dos Clusters**: Incluem o número de cidades, total de crimes e média de crimes por cidade. Os cinco maiores clusters foram destacados no console.
- **Visualizações**: As visualizações geradas (spring layout, Kamada-Kawai, hierárquica, PCA, t-SNE e heatmap) permitiram uma análise visual clara da separação entre clusters e da estrutura do grafo.
- **Análise Qualitativa**: Amostras de descrições por cluster foram extraídas, permitindo inferências sobre padrões criminais, como ”crimes com uso de arma” ou ”furtos em estabelecimentos comerciais”.

4 Análise e Discussão

Os clusters identificados refletem padrões semânticos nas descrições dos assaltos, agrupando cidades com tipos de crimes semelhantes. Por exemplo, um cluster pode conter cidades com predominância de ”assaltos noturnos com violência”, enquanto outro pode agrupar ”furtos em residências”. A escolha de K=5 no grafo K-NN equilibrou a densidade do grafo, garantindo conexões significativas sem sobrecarregar a rede.

O algoritmo de Cluster Label Propagation foi eficaz para identificar comunidades coesas, com a modularidade indicando a qualidade da divisão. As visualizações PCA e t-SNE complementaram a análise, mostrando a separação entre clusters no espaço dos embeddings, enquanto o heatmap destacou a alta similaridade dentro dos clusters.

Limitações incluem a dependência do parâmetro K e do modelo de embeddings, que podem influenciar a estrutura do grafo e os clusters formados. Além disso, a análise qualitativa das descrições foi limitada a amostras, e uma análise mais profunda poderia requerer técnicas adicionais de processamento de linguagem natural, como extração de palavras-chave.

5 Conclusão

O projeto demonstrou a eficácia do algoritmo de Cluster Label Propagation para identificar comunidades de cidades com padrões criminais semelhantes, com

base em descrições textuais de assaltos. As visualizações geradas facilitaram a interpretação dos resultados, enquanto a análise qualitativa sugeriu tipologias de crimes predominantes em cada cluster. O código é modular e pode ser adaptado para outros conjuntos de dados ou valores de K.

Para trabalhos futuros, recomenda-se explorar outros algoritmos de detecção de comunidades (como Louvain) e incorporar técnicas de análise de texto, como modelagem de tópicos, para refinar a identificação de padrões criminais. A integração de dados socioeconômicos também poderia enriquecer a análise, fornecendo contexto adicional para os clusters identificados.