T1 NLP - Roberto Martins

Introdução

Objetivo:

Dado corpo de texto (30 documentos jurídicos em PT-BR), identificar grupos (clusters) e determinar rótulos para cada grupo.

Etapas:

1. Análise do Dataset
2. Vetorização do texto
3. Agrupamento
4. Análises finais
5. Limitações

Corpus

Dataset Card

Natureza dos documentos: Documentos Processuais
Domínio: Jurídico
Tarefa: Clustering
Idioma: Português
Formato: Arquivos .txt
Quantidade de textos: 30
Total de tokens (excluindo pontuação/espaços): 17560 - modelo Spacy: “en_core_web_md”
Total de types (tokens únicos, case-insensitive): 2482 - modelo Spacy: “en_core_web_md”
Comprimento médio dos textos em tokens: 585,33 - modelo Spacy: “en_core_web_md”
Comprimento em tokens do menor texto: 384 - modelo Spacy: “en_core_web_md”
Comprimento em tokens do maior texto: 864 - modelo Spacy: “en_core_web_md”
Tamanho em disco (arquivos .txt): 117,75 KB
Qualidade dos textos: Textos jurídicos revisados
Quantidade de classes: Baseado em decisão final - 6
Distribuição em classes: N/A

Pré-Processamento

Etapas de Pré-processamento:

1. Transformação para minúsculas para unificar as formas dos tokens.
2. Tokenização e lematização com SpaCy (en_core_web_md).
3. Remoção de stopwords para eliminar palavras comuns e de baixo valor informacional.
4. Remoção de pontuação e tokens de espaço em branco para focar em palavras de conteúdo.
5. Filtragem de tokens não alfabéticos para excluir números e símbolos.

Decisões e Justificativa:

1. Lematização e remoção de stopwords reduzem o tamanho do vocabulário e o ruído.
2. Filtragem de tokens não alfabéticos prioriza termos jurídicos (principalmente palavras).
3. Foi utilizado modelo em inglês por conveniência; idealmente um modelo em Português (pt_core_news_md) seria mais adequado ao corpus.

BOW

Nesta etapa, usamos a abordagem baseada em BOW/TF-IDF para representar os documentos.

Construção:
Entrada: textos pré-processados conforme mencionado no slide anterior.
Ferramenta: TfidfVectorizer com configurações padrão (norma L2, smooth_idf).
Saída: matriz TF-IDF esparsa (30 × 2.468).

Testes e Limitações:
Não realizamos busca em grade (grid search) para ajustar parâmetros como min_df, max_df ou ngram_range; optou-se pelas configurações padrão do TfidfVectorizer, motivado pela ausência de conjunto de validação, trabalhos futuros incluem utilização de métricas como silhouette score para avaliar.

Word Embeddings

Para capturar semântica mais profunda, recorremos a embeddings de palavras, avaliando diferentes modelos.

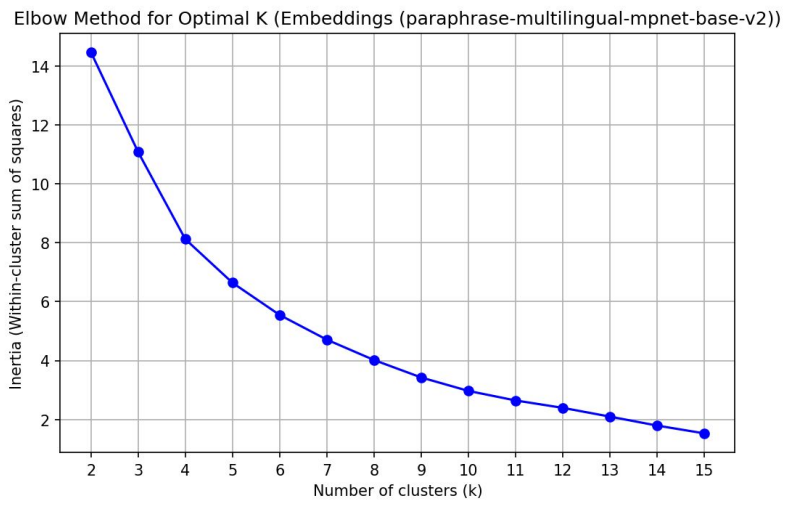
Modelos Avaliados:
SpaCy: en_core_web_md (96 d), pt_core_news_md (300 d)
SentenceTransformer: paraphrase-multilingual-mpnet-base-v2 (768 d), all-MiniLM-L6-v2 (384 d)
Agregação: média padrão dos embeddings de tokens (model.encode(raw_documents) ou doc.vector) - escolha devido a sua simplicidade, sem conjunto de validação para performar testes.

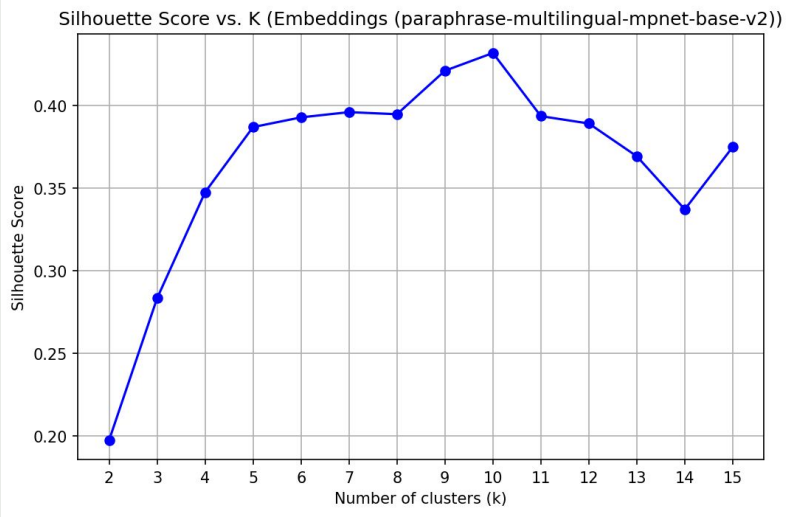
Testes e Seleção:
Sem ajuste de hiperparâmetros por modelo - sem conjunto de validação, trabalhos futuros indicam a possibilidade de fine-tuning dos modelos de forma semi-supervisionada.
Modelos comparados via desempenho de agrupamento (Silhouette para K-Means, DBCV para HDBSCAN).

Agrupamento

Algoritmos:

K-Means: busca minimizar a soma das distâncias ao centróide e um algoritmo “clássico”, eficiente, é avaliado pelo coeficiente de silhueta e pela análise do cotovelo.

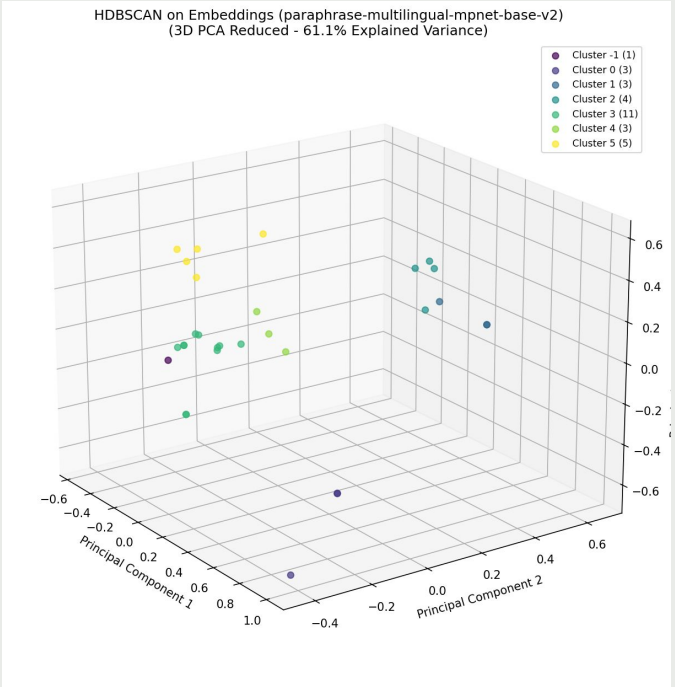




Agrupamento

Algoritmos:
HDBSCAN: agrupamento por densidade, capaz de destacar documentos atípicos (ruídos). Escolhido devido ao menor impacto da escolha dos parâmetros em comparação a outros modelos de clusterização por densidade como o DBSCAN.

Avaliado pelo **DBCV** (Density-Based Clustering Validation): avalia a qualidade de clusters baseados em densidade, considerando densidade intra-cluster e separação em relação ao ruído. Assim como pela análise visual dos gráficos.



Comparação e Análise de Resultados

Critérios de seleção do modelo de embedding:

- Silhouette Score + DBCV

Critérios de seleção do método de clusterização:

- Número de clusters que reflete pad encontrado
- Análises visuais

Escolha de clusterização final:
HDBSCAN + paraphrase-multilingual-mpnet-base-v2

Algorithm	Representation	#Clusters	Noise	Silhouette	DBCV
K-Means	TF-IDF	15	0	0.2627	–
HDBSCAN	TF-IDF	6	4	–	0.1554
K-Means	Embeddings (en_core_web_md)	13	0	0.2193	–
HDBSCAN	Embeddings (en_core_web_md)	2	4	–	0.1191
K-Means	Embeddings (paraphrase-multilingual-mpnet-base-v2)	10	0	0.4320	–
HDBSCAN	Embeddings (paraphrase-multilingual-mpnet-base-v2)	6	1	–	0.3846
K-Means	Embeddings (pt_core_news_md)	2	0	0.3211	–
HDBSCAN	Embeddings (pt_core_news_md)	4	2	–	0.2702
K-Means	Embeddings (all-MiniLM-L6-v2)	11	0	0.3162	–
HDBSCAN	Embeddings (all-MiniLM-L6-v2)	5	4	–	0.1865

Análise dos Clusters

Cluster 0 – Ambiental Paulista (3 docs): Principais termos: paulista, jurídicos, jurídica, brasileiro, legal, ilegal, brasileira, ambiental, ambientais, fiscalizem

Cluster 1 – Processos de Acusação Judicial (3 docs): Principais termos: brasileiro, judiciária, judicial, jurídica, advogado, jurídicas, acusado, tribunal, oficiada, ricardo

Cluster 2 – Casos Penais e Prisionais (4 docs): Principais termos: brasileiro, imputado, ricardo, julgado, acusado, prisional, caso, penal, fiscais, comarca

Cluster 3 – Decisões Judiciais Gerais (11 docs): Principais termos: brasileiro, brasileira, brasil, comarca, decretação, judiciais, juízo, procuradores, 3344ribeiro, luísa

Cluster 4 – Casos Financeiros (3 docs): Principais termos: paulista, brasileiro, brasil, autoriza, juros, copacabana, requerente, juízo, comarca, caso

Cluster 5 – Reclamações Trabalhistas (5 docs): Principais termos: paulista, brasileiro, brasileira, ricardo, laboradas, funcionário, comarca, juízo, reclamante, juros