UNIVERSIDADE FEDERAL DE GOIÁS
Instituto de Informática
Pós Graduação em Banco de Dados Para Big Data

GUILHERME BARBOSA GOMES
VITOR AUGUSTO DIB MARTINHO

PROCESSAMENTO E VISUALIZAÇÃO DE ACÓRDÃO JURÍDICOS: UMA ABORDAGEM BASEADA EM NLP E BUSINESS INTELLIGENCE

Trabalho de Conclusão de Curso

Goiânia
2025

UNIVERSIDADE FEDERAL DE GOIÁS
Instituto de Informática
Pós Graduação em Banco de Dados Para Big Data

GUILHERME BARBOSA GOMES
VITOR AUGUSTO DIB MARTINHO

PROCESSAMENTO E VISUALIZAÇÃO DE ACÓRDÃO JURÍDICOS: UMA ABORDAGEM BASEADA EM NLP E BUSINESS INTELLIGENCE

Artigo apresentado na disciplina de Conclusão de Curso da Pós Graduação em Banco de Dados Para Big Data do Instituto de Informática da Universidade Federal de Goiás

Orientador: Prof. Dr. Márcio de Souza Dias

Goiânia
2025

Resumo

A crescente digitalização do sistema judiciário brasileiro tem gerado um volume massivo de documentos jurídicos, dificultando a análise manual de decisões e acórdãos. Este trabalho apresenta uma abordagem baseada em Processamento de Linguagem Natural (NLP) e Business Intelligence (BI) para a extração, tratamento e visualização de dados textuais provenientes de acórdãos do Superior Tribunal de Justiça (STJ). Foram utilizados métodos de pré-processamento textual, identificação de entidades nomeadas e extração de métricas linguísticas com a biblioteca spaCy, aplicando o modelo pt_core_news_lg. A análise foi conduzida sobre um subconjunto de 10 mil documentos, parte de um corpus com mais de dois milhões. Os dados extraídos foram estruturados e visualizados em um painel interativo no Power BI, permitindo insights como a frequência de termos jurídicos, a distribuição geográfica dos processos, entidades recorrentes e a densidade lexical dos textos. Os resultados demonstram a viabilidade de integrar NLP e BI na exploração de dados jurídicos, oferecendo suporte à tomada de decisões e à pesquisa acadêmica no campo do Direito.

Palavras-chave: Processamento de Linguagem Natural. Acórdãos. Business Intelligence. Visualização de Dados. Direito Digital.

Abstract

The increasing digitization of the Brazilian judiciary system has generated a massive volume of legal documents, making manual analysis of court decisions and rulings unfeasible. This work presents an approach based on Natural Language Processing (NLP) and Business Intelligence (BI) for extracting, processing, and visualizing textual data from rulings issued by the Superior Court of Justice (STJ). Text preprocessing methods, named entity recognition, and linguistic metric extraction were performed using the spaCy library with the pt_core_news_lg model. The analysis focused on a subset of 10,000 documents, part of a broader corpus of over two million. The extracted data were structured and visualized through an interactive dashboard built in Power BI, enabling insights such as the frequency of legal terms, geographical distribution of cases, recurrent entities, and lexical density of the texts. The results demonstrate the feasibility of integrating NLP and BI for legal data exploration, supporting both decision-making and academic research in the legal domain.

Keywords: Natural Language Processing. Rulings. Business Intelligence. Data Visualization. Legal Analytics.

Estrutura do Trabalho

1. Introdução

1.1 Contextualização

A visualização de dados desempenha um papel essencial na ciência de dados, permitindo a interpretação eficiente de grandes volumes de informações. Edward Tufte (1983) destaca que "a excelência na visualização de dados consiste na apresentação clara e precisa da informação complexa", ressaltando a importância de gráficos e tabelas bem estruturados para facilitar a análise. No contexto jurídico, onde a interpretação de acórdãos envolve a leitura de extensos documentos textuais, a utilização de técnicas avançadas, como o Processamento de Linguagem Natural (NLP), torna-se indispensável para transformar esse material em conhecimento acionável.

Com a digitalização acelerada do sistema judiciário, a quantidade de documentos legais disponíveis cresceu exponencialmente, tornando inviável a análise manual de acórdãos (McKinney, 2021). A análise textual automatizada permite extrair informações relevantes e identificar padrões que, de outra forma, passariam despercebidos. Heer, Bostock e Ogievetsky (2010) afirmam que "a visualização interativa permite a exploração e o entendimento intuitivo dos dados, facilitando sua análise", o que é especialmente relevante para um domínio tão complexo e técnico como o jurídico.

Além disso, a adoção de ferramentas de Business Intelligence (BI) tem revolucionado a forma como as informações são processadas e apresentadas. Stephen Few (2012) ressalta que "a transformação de dados em gráficos eficazes é essencial para a tomada de decisões baseadas em evidências". O Power BI, uma das principais ferramentas de BI disponíveis, possibilita a criação de dashboards interativos que sintetizam grandes volumes de informações e revelam padrões ocultos nos dados, permitindo que advogados, juízes e pesquisadores compreendam melhor as tendências e decisões judiciais.

No campo do NLP, técnicas como Tokenização, Stemming, Lematização e Modelos de Word Embeddings têm sido amplamente empregadas para categorizar, resumir e analisar sentenças judiciais. Segundo Jurafsky e Martin (2021), "o Processamento de Linguagem Natural permite não apenas a extração de informações, mas também a compreensão contextual do conteúdo textual, facilitando a análise jurídica automatizada". Essas técnicas, aplicadas em conjunto com BI, permitem transformar decisões judiciais em dados estruturados e visualizáveis, otimizando pesquisas e tomadas de decisão dentro do sistema jurídico.

Dessa forma, a presente pesquisa busca unir três pilares fundamentais: Visualização de Dados, Business Intelligence e Processamento de Linguagem Natural para oferecer uma solução inovadora na análise de acórdãos. A integração dessas áreas permitirá não apenas a

extração e categorização de informações relevantes, mas também a apresentação dessas informações de maneira intuitiva e interativa, facilitando sua interpretação e auxiliando na tomada de decisões estratégicas no setor jurídico.

1.2 Justificativa

A análise de decisões judiciais é um processo crítico para o funcionamento do sistema jurídico, pois permite identificar padrões em julgamentos, avaliar a coerência jurisprudencial e auxiliar advogados e magistrados na fundamentação de novas decisões. No entanto, o crescimento exponencial do volume de acórdãos dificulta uma análise manual eficiente. Segundo Lax & Cameron (2007), "o estudo sistemático de decisões judiciais pode revelar vieses institucionais, mudanças de interpretação ao longo do tempo e padrões de argumentação fundamentais para a evolução do direito". Assim, métodos automatizados e tecnologias avançadas tornam-se essenciais para otimizar essa tarefa.

A aplicação de Business Intelligence (BI) na área jurídica tem sido um diferencial significativo, pois possibilita a estruturação e a visualização de grandes volumes de dados de maneira clara e objetiva. Como argumenta Few (2012), "uma boa visualização de dados deve reduzir a complexidade sem perder a profundidade analítica", o que é crucial para permitir que operadores do direito compreendam rapidamente as tendências e insights derivados dos acórdãos.

Além disso, o Processamento de Linguagem Natural (NLP) tem demonstrado grande eficiência na extração de informações relevantes de documentos textuais extensos. Segundo Manning, Raghavan e Schütze (2008), "as técnicas de NLP podem não apenas identificar e classificar elementos textuais, mas também inferir significados contextuais e relações semânticas entre termos", tornando-se uma ferramenta valiosa na análise jurídica. Isso é particularmente relevante quando se considera que decisões judiciais contêm linguagens complexas, tecnicismos e uma estrutura argumentativa elaborada.

1.3 Objetivos

Objetivo Geral:

- Desenvolver um dashboard interativo no Power BI que permita a análise eficiente de acórdãos jurídicos por meio da integração de técnicas de Visualização de Dados, Business Intelligence e Processamento de Linguagem Natural.

Objetivos Específicos:

- Identificar padrões e tendências em acórdãos jurídicos a partir de análises estatísticas e textuais, utilizando NLP para extrair informações relevantes.
- Aplicar técnicas de NLP para segmentação e classificação dos textos jurídicos, facilitando a pesquisa e a compreensão das decisões.
- Construir indicadores visuais no Power BI que permitam a navegação intuitiva pelos dados extraídos dos acórdãos.

- Implementar filtros e funcionalidades interativas no dashboard para permitir análises personalizadas conforme o interesse dos usuários.
- Avaliar o impacto da ferramenta na otimização do tempo de pesquisa jurídica e na tomada de decisões baseadas em evidências.

2. Processamento e Tratamento dos Dados

2.1 Dados Utilizados

A base de dados utilizada nesta pesquisa foi extraída do corpus Ulysses Tesemô, disponibilizado publicamente no repositório oficial do projeto no GitHub. Trata-se de um dos maiores conjuntos de textos voltados ao domínio jurídico e governamental brasileiro, estruturado e publicado como parte do artigo "Ulysses Tesemô: a new large corpus for Brazilian legal and governmental domain" (Siqueira et al., 2024).

Dentre os diversos tipos de documentos disponíveis, optou-se por restringir a análise a 10.000 acórdãos cujo data_source é exclusivamente o Superior Tribunal de Justiça (STJ). Essa escolha teve como objetivo delimitar o escopo da análise, assegurando uma maior uniformidade textual e temática, o que facilita a aplicação das técnicas de Processamento de Linguagem Natural (NLP) e Business Intelligence (BI) adotadas neste trabalho. A seleção foi feita com base na estrutura previamente organizada no repositório, que contém metadados categorizando os documentos por origem, tipo e conteúdo.

Cabe ressaltar que a realização deste trabalho só foi possível graças à disponibilização aberta e estruturada do corpus Ulysses Tesemô pelos autores do projeto. A qualidade, a abrangência e a organização dos dados fornecidos foram fundamentais para viabilizar a aplicação das técnicas de Processamento de Linguagem Natural e Business Intelligence aqui propostas. Sem o acesso a essa base de dados pública e cuidadosamente curada, seria inviável reproduzir este estudo em escala, devido à dificuldade de acesso a documentos jurídicos em formato estruturado.

2.2 Tratamento dos Dados

Antes da construção do dashboard, foi necessário realizar um trabalho robusto de limpeza e tratamento dos dados textuais extraídos dos acórdãos. Esse processo foi essencial para garantir a qualidade da análise, removendo ruídos e estruturando as informações de maneira adequada. Para isso, utilizamos bibliotecas especializadas em manipulação e análise textual no Python, como spaCy, pandas, collections e re.

- **spaCy**: Utilizado para tokenização, lematização e reconhecimento de entidades nomeadas, permitindo categorizar termos jurídicos relevantes.
- **pandas**: Empregado para manipulação e estruturação dos dados, facilitando a organização e filtragem de grandes volumes de textos jurídicos.

- **collections**: Permitiu a contagem e análise de frequência de palavras e termos específicos, auxiliando na extração de insights estatísticos.
- **re (expressões regulares)**: Utilizado para padronização e limpeza textual, removendo caracteres especiais, números desnecessários e formatando os textos conforme as necessidades da análise.

A combinação dessas ferramentas possibilitou transformar documentos jurídicos não estruturados em um formato organizado e pronto para visualização no Power BI.

Será apresentado abaixo o código utilizado, bem como uma breve explicação de cada trecho:

2.3 Bibliotecas e Dependências Utilizadas

Para o desenvolvimento do pipeline de extração e análise textual dos acórdãos jurídicos, foram utilizadas bibliotecas consolidadas da linguagem Python, que oferecem suporte robusto ao processamento de linguagem natural e à manipulação de dados. A seguir, descrevem-se as principais dependências importadas:

```
import spacy

import pandas as pd

from collections import Counter

import re

import os
```

2.4 Leitura e Padronização do Texto Jurídico

A primeira etapa do processo de tratamento textual consiste na leitura dos arquivos de origem. Como os acórdãos jurídicos foram disponibilizados em formato *.txt*, foi necessário implementar uma função que realizasse a leitura segura e eficiente desses documentos, além de aplicar um pré-processamento inicial para uniformizar a estrutura textual.

```
def extract_text_from_txt(txt_path):
    """Lê o conteúdo de um arquivo TXT."""
    try:
        with open(txt_path, "r", encoding="utf-8") as file:
            text = file.read()
        text = text.replace('\n', ' ').replace('\r', ' ')  # Remove quebras de linha
        return text
    except Exception as e:
        print(f"Erro ao ler o arquivo: {e}")
        return ""
```

Essa função tem como objetivo ler o conteúdo de um arquivo *.txt* fornecido pelo caminho. Após a leitura, o texto é imediatamente normalizado por meio da substituição de quebras de linha *(\n e \r)* por espaços em branco, garantindo que o conteúdo textual seja tratado como um fluxo contínuo. Esse procedimento reduz o risco de segmentações artificiais durante as etapas subsequentes de tokenização e análise linguística. Além disso, a função está protegida com um bloco *try/except*, que captura eventuais erros de leitura e assegura que o processamento continue mesmo diante de falhas pontuais nos arquivos.

2.5 Extração do Título

```
def extract_title(text):
    """Extrai uma frase representativa do início do texto como título."""
    doc = nlp(text)
    for sent in doc.sents:
        return sent.text.strip()
    return "Título Desconhecido"
```

A função acima utiliza o *spaCy* para identificar e retornar a primeira sentença do texto como um título representativo do acórdão. Caso nenhuma sentença seja encontrada, retorna um valor padrão. Essa abordagem fornece uma identificação inicial útil para fins de organização e visualização.

2.6 Extração de Palavras-chave Frequentes

```
def extract_keywords(text, top_n=10):
    """Extrai palavras-chave do texto baseando-se na frequência e relevância."""
    doc = nlp(text)
    palavras_relevantes = [token.text.lower() for token in doc if token.is_alpha and token.pos_ in
["NOUN", "PROPN", "ADJ"]]
    palavras_mais_frequentes = Counter(palavras_relevantes).most_common(top_n)
    return palavras_mais_frequentes
```

A função identifica as 10 palavras-chave mais frequentes no texto, considerando apenas substantivos, nomes próprios e adjetivos. O uso do *Counter* permite ranquear os termos mais representativos do conteúdo jurídico analisado.

2.7 Filtragem de Entidades Nomeadas

```
def limpar_entidades(lista, tipo):
    """Remove entidades curtas, irrelevantes ou que parecem ruído com técnicas heurísticas."""
    resultado = []
    for item in lista:
```

```
        item_limpo = item.strip().replace("\n", " ").replace("\r", " ")
        if len(item_limpo) < 3:
            continue
        if tipo == "PESSOA":
            if len(item_limpo.split()) < 2:
                continue
            if not all(p[0].isupper() for p in item_limpo.split()):
                continue
        elif tipo == "ORGANIZACAO":
            if len(item_limpo.split()) < 2 and not any(kw.lower() in item_limpo.lower() for kw in
organizacoes_keywords):
                continue
        elif tipo == "LOCAL":
            if item_limpo.upper() not in siglas_estados and len(item_limpo.split()) < 2:
                continue
        elif tipo == "LEI":
            if item_limpo.lower().strip() == "lei":
                continue
        resultado.append(item_limpo)
    return resultado
```

Esta função aplica regras específicas para filtrar entidades nomeadas extraídas automaticamente, eliminando termos curtos, genéricos ou com estrutura inconsistente. Os critérios variam conforme o tipo da entidade (pessoa, organização, local ou lei), garantindo maior precisão nas análises.

2.8 Análise Linguística e Extração de Informações do Texto

```
def analyze_text(text):
    doc = nlp(text)
    titulo = extract_title(text)
    palavras_chave = extract_keywords(text)

    num_palavras = len([token.text for token in doc if token.is_alpha])
    num_caracteres = len(text)
    num_sentencas = len(list(doc.sents))
    num_palavras_unicas = len(set(token.text.lower() for token in doc if token.is_alpha))
    densidade_lexical = num_palavras_unicas / num_palavras if num_palavras > 0 else 0

    contagem_classes = {
        "Verbos": len([token for token in doc if token.pos_ == "VERB"]),
        "Substantivos": len([token for token in doc if token.pos_ == "NOUN"]),
        "Adjetivos": len([token for token in doc if token.pos_ == "ADJ"]),
        "Pronomes": len([token for token in doc if token.pos_ == "PRON"]),
```

```
        "Numerais": len([token for token in doc if token.pos_ == "NUM"]),
    }

    entidades = {"PESSOA": [], "ORGANIZACAO": [], "LOCAL": [], "DATA": [], "MONEY": [], "LEI":
[]}
    entidades_encontradas = []
    for ent in doc.ents:
        label = ent.label_
        if label == "PER":
            entidades["PESSOA"].append(ent.text)
        elif label == "ORG":
            entidades["ORGANIZACAO"].append(ent.text)
        elif label == "LOC":
            entidades["LOCAL"].append(ent.text)
        elif label == "DATE":
            entidades["DATA"].append(ent.text)
        elif label == "MONEY":
            entidades["MONEY"].append(ent.text)
        elif re.search(r"art(igo)?\.?(\s+\d+)?|lei\s+\d+", ent.text.lower()):
            entidades["LEI"].append(ent.text)
        entidades_encontradas.append((ent.text.strip(), label))

    for tipo in ["PESSOA", "ORGANIZACAO", "LOCAL", "LEI"]:
        entidades[tipo] = limpar_entidades(entidades[tipo], tipo)

    df_resumo = pd.DataFrame({
        "Métrica": ["Título do Texto", "Número de Palavras", "Número de Caracteres", "Número de
Sentenças", "Palavras Únicas", "Densidade Lexical", "Verbos", "Substantivos", "Adjetivos", "Pronomes",
"Numerais"],
        "Valor": [titulo, num_palavras, num_caracteres, num_sentencas, num_palavras_unicas,
densidade_lexical,
                  contagem_classes["Verbos"], contagem_classes["Substantivos"],
contagem_classes["Adjetivos"],
                  contagem_classes["Pronomes"], contagem_classes["Numerais"]]
    })

    df_palavras_chave = pd.DataFrame(palavras_chave, columns=["Palavra", "Frequência"])
    df_entidades_encontradas = pd.DataFrame(entidades_encontradas, columns=["Entidade", "Tipo"])
    df_entidades = {chave: pd.DataFrame(Counter(valores).items(), columns=["Entidade", "Frequência"])
for chave, valores in entidades.items() if valores}

    return df_resumo, df_entidades, df_entidades_encontradas, df_palavras_chave
```

Esta função executa a análise completa do texto jurídico. Utiliza o *spaCy* para segmentar sentenças, contar palavras e identificar classes gramaticais. Também extrai palavras-chave, calcula a densidade lexical e reconhece entidades nomeadas (como pessoas, locais e leis), aplicando filtros personalizados. Os resultados são organizados em *dataframes* para posterior exportação e visualização no Power BI.

2.9 Processamento em Lote e Geração do Arquivo Consolidado

```
def process_folder(folder_path, output_file):
    todos_resumos = []
    todas_entidades_encontradas = []
    todas_palavras_chave = []
    entidades_agrupadas = {}

    for nome_arquivo in os.listdir(folder_path):
        if nome_arquivo.lower().endswith(".txt"):
            caminho_arquivo = os.path.join(folder_path, nome_arquivo)
            print(f"Processando: {caminho_arquivo}")
            texto = extract_text_from_txt(caminho_arquivo)
            if not texto:
                continue
            df_resumo, df_entidades, df_entidades_encontradas, df_palavras_chave = analyze_text(texto)

            df_resumo.insert(0, "Arquivo", nome_arquivo)
            df_entidades_encontradas.insert(0, "Arquivo", nome_arquivo)
            df_palavras_chave.insert(0, "Arquivo", nome_arquivo)

            todos_resumos.append(df_resumo)
            todas_entidades_encontradas.append(df_entidades_encontradas)
            todas_palavras_chave.append(df_palavras_chave)

            for tipo, df in df_entidades.items():
                if tipo not in entidades_agrupadas:
                    entidades_agrupadas[tipo] = []
                df.insert(0, "Arquivo", nome_arquivo)
                entidades_agrupadas[tipo].append(df)

    with pd.ExcelWriter(output_file, engine="xlsxwriter") as writer:
        pd.concat(todos_resumos).to_excel(writer, sheet_name="Resumo", index=False)
        pd.concat(todas_palavras_chave).to_excel(writer, sheet_name="Palavras-chave", index=False)
```

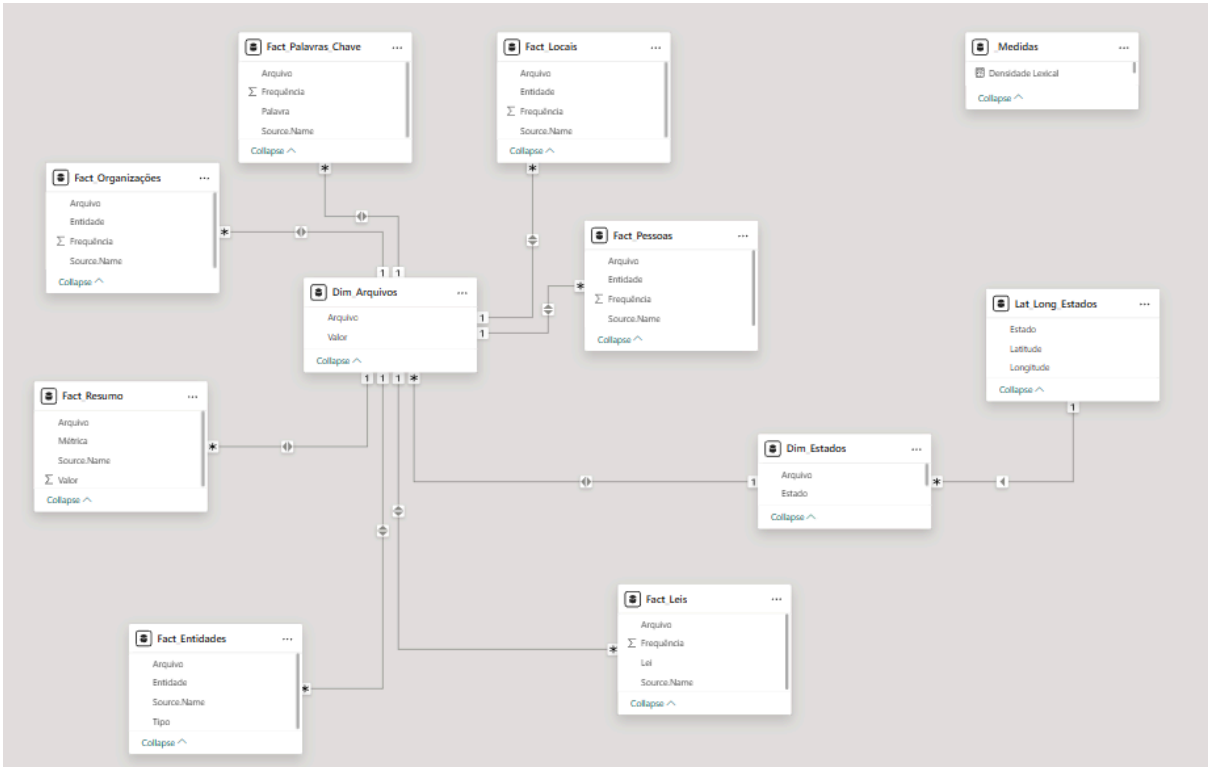
```
    pd.concat(todas_entidades_encontradas).to_excel(writer, sheet_name="Entidades Encontradas",
index=False)
    for tipo, lista_dfs in entidades_agrupadas.items():
        pd.concat(lista_dfs).to_excel(writer, sheet_name=tipo, index=False)

print(f"Consolidação concluída! Arquivo salvo em {output_file}")
```

3. Apresentação do Dashboard

Após a etapa de extração, tratamento e estruturação dos dados em Python, os resultados foram exportados para um arquivo Excel contendo múltiplas abas, cada uma com um conjunto específico de informações: resumo estatístico, palavras-chave, entidades nomeadas por tipo (pessoas, organizações, locais, leis), entre outros.

Cada aba do arquivo Excel foi importada para o Power BI e tratada como uma **tabela fato**, representando eventos ou contagens associadas a cada documento analisado. Para organizar o modelo de dados, foi criada uma tabela **dimensional** chamada Dim_Arquivos, que contém o nome de cada arquivo analisado e serve como chave primária para filtrar e relacionar todas as demais tabelas fato. A estrutura relacional obedece a um tipo especial de Star-Schema Híbrido e é demonstrada na Figura abaixo:

Figura 01 – Modelo de Dados no Power BI com Tabelas Fato e Dimensão Arquivos



Fonte: Autoria Própria

A Dim_Arquivos está relacionada a diversas tabelas fato, tais como:

- Fact_Resumo: contém métricas como número de palavras, número de sentenças, densidade lexical, entre outras;
- Fact_Palavras_Chave: armazena as palavras mais frequentes e suas contagens por arquivo;
- Fact_Pessoas, Fact_Organizações, Fact_Locais, Fact_Leis: registram entidades nomeadas extraídas dos textos, com frequência de ocorrência;
- Fact_Entidades: consolida todas as entidades extraídas, categorizadas por tipo;
- Dim_Estados: associada ao estado identificado em cada documento;
- Lat_Long_Estados: usada para visualizações geográficas com latitude e longitude de cada UF.

Todas essas tabelas se conectam à Dim_Arquivos por meio do campo **Arquivo**, formando um modelo estrela. Além disso, a Dim_Estados está relacionada à tabela Lat_Long_Estados via campo **Estado**, o que permite a construção de mapas georreferenciados com a distribuição dos acórdãos por unidade federativa.

Essa modelagem permite filtrar e cruzar informações entre diferentes perspectivas, mantendo a integridade dos dados e promovendo análises comparativas, dinâmicas e interativas no painel visual.

Com os relacionamentos feitos, o BI foi dividido em 3 telas principais:

3.1 Tela 1 – Capa

A primeira tela (figura 02) funciona como uma introdução ao dashboard, apresentando o título do projeto, os objetivos principais e um panorama do conteúdo analisado. Esta tela tem o objetivo de contextualizar o usuário e prepará-lo para as análises subsequentes.

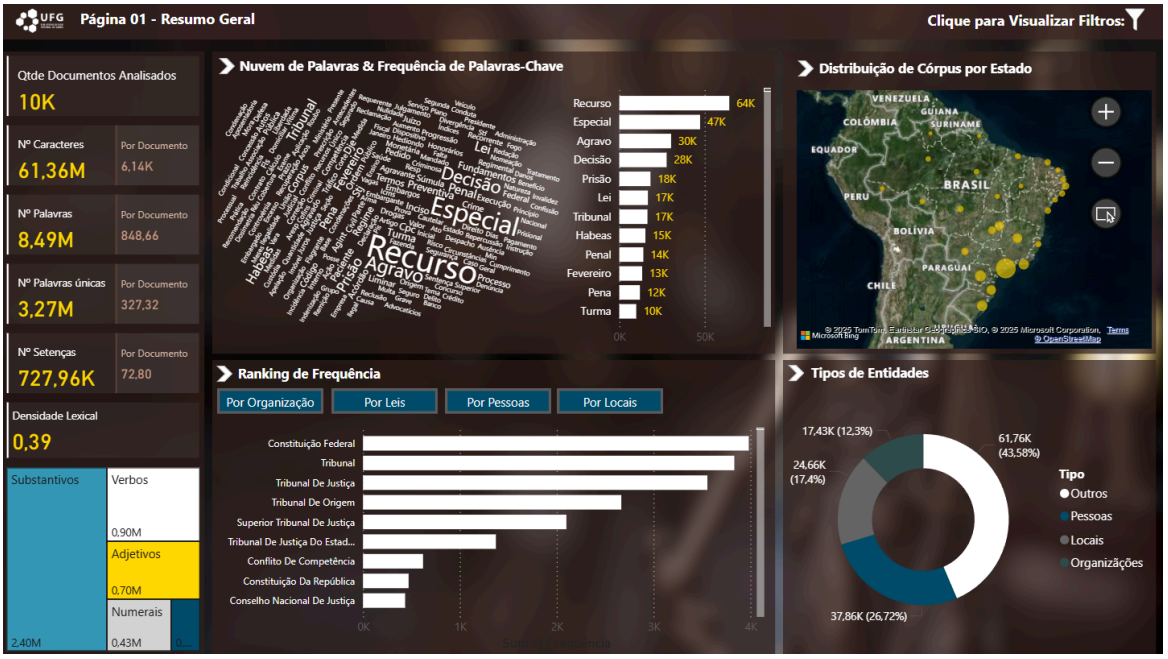
Figura 02 – Capa Introdutória do BI

Fonte: Autoria Própria

3.2 Tela 2 – Resumo Geral

A segunda tela (figura 03) concentra todas as análises realizadas sobre os acórdãos jurídicos, organizadas de forma interativa e intuitiva. Nessa tela, é possível observar os seguintes elementos e análises:

Figura 03 – Resumo Geral



Fonte: Autoria Própria

- **Indicadores Textuais Gerais:**
    - **Quantidade de documentos analisados:** número total de acórdãos considerados na análise.
    - **Quantidade de caracteres:** total de caracteres somados dos documentos.
    - **Número de palavras:** soma de todas as palavras presentes nos acórdãos.
    - **Número de sentenças:** quantificação das sentenças identificadas.
    - **Densidade lexical:** cálculo da proporção entre vocabulário total e vocabulário único, indicando a riqueza do texto.
- **Treemap por Classe Gramatical:**
    - Visualização gráfica que representa a distribuição das palavras nos acórdãos conforme sua classe gramatical. Os blocos indicam a frequência de:
        - Substantivos
        - Verbos
        - Adjetivos
        - Numerais
        - Pronomes
- **Nuvem de Palavras e Gráfico de Frequência de Palavras-Chave:**
    - A nuvem de palavras destaca os termos mais recorrentes nos documentos, permitindo uma leitura rápida dos temas centrais abordados.
    - O gráfico de barras complementa a visualização, listando as palavras com maior frequência de forma quantitativa.
- **Gráfico de Frequência por Tipo de Entidade:**
    - Apresenta as entidades nomeadas mais mencionadas, classificadas em:
        - **Organizações:** Constituição Federal, Tribunais, Tribunal Superior, Tribunal de Justiça Estadual, Conselho Nacional de Justiça.
        - **Leis:** dispositivos legais frequentemente referenciados.
        - **Pessoas:** nomes de autoridades, partes envolvidas ou relator.
        - **Locais:** unidades federativas, cidades, estados ou jurisdições mencionadas.
- **Distribuição Geográfica – Corpus por Estado:**
    - Mapa interativo ou gráfico de barras que exibe o número de acórdãos por estado brasileiro, facilitando a comparação regional.
- **Tipos de Entidades Reconhecidas:**
    - Resumo da quantidade e distribuição das entidades extraídas via NLP, segmentadas em três categorias principais:
        - Pessoas
        - Locais
        - Organizações

A página permite também ao usuário o uso completo dos filtros disponíveis, através de um botão no canto superior direito da página, que abre um menu auxiliar (Figura 04):

Figura 04 – Menu de Filtros

Fonte: Autoria Própria

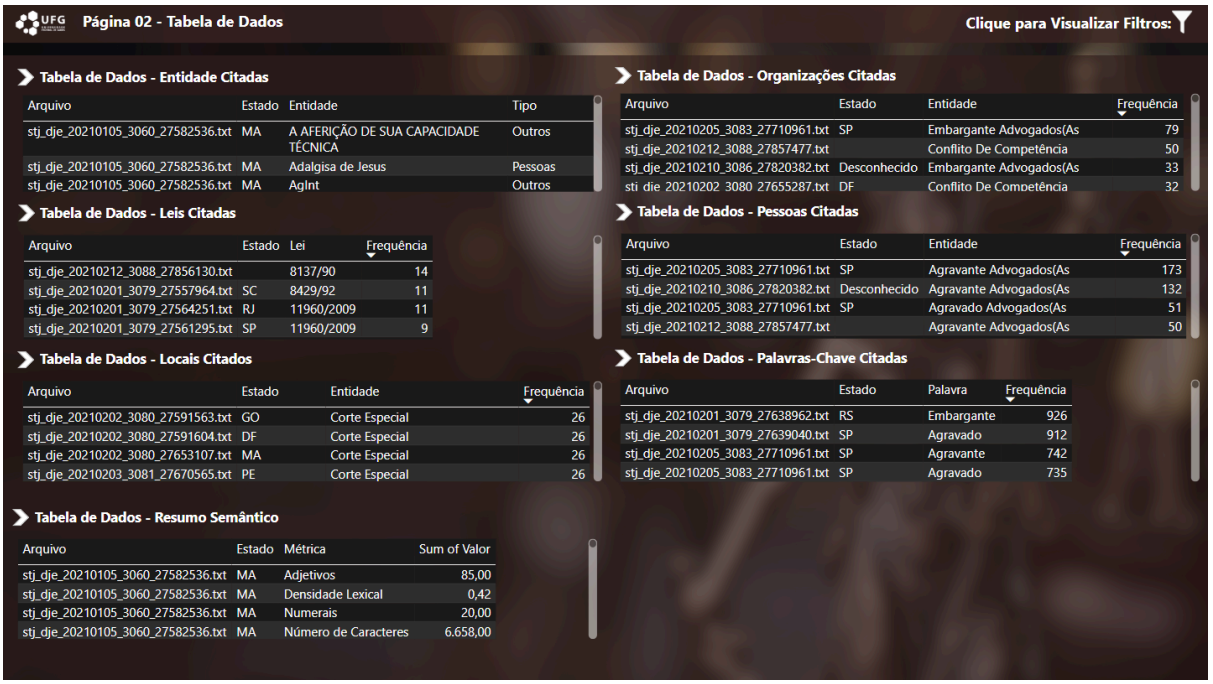
3.3 Tela 3 - Tabela de Dados

A página de Tabelas de Dados (Figura 05) do dashboard foi desenvolvida para apresentar, de forma detalhada e tabular, todas as informações extraídas dos acórdãos jurídicos. Ela contém seções específicas com as entidades identificadas nos textos, como **pessoas, organizações, locais e leis**, além das **palavras-chave** mais citadas.

Cada tabela exibe o nome do **arquivo analisado**, o **estado de origem do documento**, a **entidade ou palavra identificada** e sua respectiva **frequência de ocorrência**. Além disso, há uma tabela de **resumo semântico**, com as métricas linguísticas consolidadas, como número de caracteres, densidade lexical e quantidade de adjetivos e numerais.

Essa página permite uma análise minuciosa dos dados extraídos, facilitando a navegação e comparação entre documentos individuais e reforçando a rastreabilidade das informações apresentadas nos gráficos do painel.

Figura 05 – Página de Tabela de Dados



Fonte: Autoria Própria

3.4 Resumo das Possibilidades de Análise

O dashboard oferece uma ampla gama de análises, permitindo ao usuário:

- Avaliar a complexidade e extensão dos documentos jurídicos.
- Medir a densidade linguística do corpus.
- Identificar os principais temas e termos jurídicos recorrentes.
- Visualizar a estrutura gramatical predominante nos textos.
- Investigar a frequência e o tipo de entidades mencionadas (pessoas, leis, órgãos, locais).
- Comparar a origem geográfica dos acórdãos por estado.
- Explorar padrões e tendências que podem auxiliar em estudos jurídicos, decisões estratégicas ou pesquisa acadêmica.

Alguns Insights possíveis de tirarmos do BI são:

3.4.1 Volume e Complexidade Textual

- Foram analisados **10 mil acórdãos**, representando uma amostra significativa do corpus jurídico.

- O total de caracteres ultrapassa **61 milhões**, com uma média de **6.100 caracteres por documento**.

- A média de sentenças por acórdão é de **72,8**, o que confirma a complexidade textual dos documentos.

- A **densidade lexical média de 0,39** indica que, apesar da linguagem técnica, há certa repetição de termos — o que é esperado em textos jurídicos.

3.4.2 Palavras-Chave e Temas Dominantes

- Os termos mais frequentes foram **"Recurso" (64k)**, **"Especial" (47k)**, **"Agravo" (30k)**, **"Decisão" (28k)**, e **"Prisão" (18k)**.

- Isso evidencia que os temas centrais dos acórdãos giram em torno de **recursos processuais e decisões judiciais**, especialmente no contexto penal e recursal.

- A nuvem de palavras e os rankings mostram forte presença de conceitos como **"Tribunal", "Habeas", "Pena", "Turma"** — reforçando o foco em julgamentos colegiados e ações penais.

3.4.3 Entidades Mais Citadas

- Em **organizações**, destacam-se **Tribunal de Justiça**, **Superior Tribunal de Justiça**, **Conselho Nacional de Justiça**.

- Entre **pessoas**, há recorrência de partes processuais e advogados(as), como "Agravante Advogados(as)" — o que pode indicar o padrão de nomenclatura nos textos.

- As **leis** mais citadas incluem **8137/90**, **8429/92**, **11960/2009**, demonstrando o uso frequente de leis penais e de improbidade administrativa.

- Os **locais** mais mencionados incluem "Corte Especial", associado ao próprio STJ.

3.4.4 Distribuição Geográfica

- O mapa mostra ampla distribuição de documentos por todo o território brasileiro, com **maior concentração nos estados de SP, RJ, DF e RS**.

- Isso pode indicar onde há maior volume de processos remetidos ao STJ ou onde estão localizadas as instâncias de origem.

3.4.5 Estrutura Gramatical

- A contagem média de **substantivos (2,4M)** é bem superior à de **verbos (900k)** e **adjetivos (700k)** — coerente com a natureza descritiva e formal dos acórdãos.

- A análise por classe gramatical reforça o predomínio de linguagem técnica e nominal, típica de textos jurídicos.

4. Metodologia

A metodologia adotada neste trabalho foi estruturada em três etapas principais: coleta e preparação dos dados, processamento linguístico com técnicas de Processamento de Linguagem Natural (NLP) e visualização dos resultados utilizando Business Intelligence por meio da ferramenta Power BI. Esta abordagem permite transformar dados textuais não estruturados de acórdãos jurídicos em insights significativos, acessíveis e interativos para o usuário final.

4.1 Coleta de Dados

A base de dados utilizada para a análise é composta por acórdãos jurídicos disponíveis em formato digital, coletados de repositórios públicos. Os documentos foram inicialmente organizados em arquivos de texto bruto (.txt), preservando sua estrutura original. Essa fase exigiu atenção para garantir que os dados estivessem completos e representassem uma variedade de tribunais e jurisdições.

4.2 Pré-processamento Textual

O pré-processamento dos textos foi realizado com auxílio de bibliotecas Python, como re, pandas, spaCy e collections. As etapas envolvidas foram:

- **Limpeza textual:** remoção de símbolos, números isolados, pontuações e palavras irrelevantes (stopwords).
- **Tokenização:** segmentação dos textos em palavras e sentenças.
- **Lematização:** redução das palavras à sua forma canônica.
- **Contagem e extração de termos:** identificação das palavras mais frequentes e separação por classes gramaticais (substantivos, verbos, adjetivos etc.).
- **Reconhecimento de Entidades Nomeadas (NER):** com o uso do spaCy, foi realizada a extração automatizada de entidades do tipo Pessoas, Organizações e Locais.

Essas etapas foram fundamentais para padronizar e enriquecer os dados, viabilizando a análise quantitativa e qualitativa no ambiente do Power BI.

4.3 Estruturação e Modelagem dos Dados

Após o processamento, os dados foram organizados em DataFrames com estruturas tabulares, facilitando a manipulação e exportação para o Power BI. Cada linha representava um acórdão e suas respectivas métricas linguísticas (número de palavras, sentenças, densidade lexical etc.), além das entidades extraídas.

4.4 Construção do Dashboard

Utilizando o Power BI, foi possível construir um dashboard interativo com visualizações dinâmicas. Foram aplicadas medidas DAX para cálculo de indicadores, filtros para segmentação e gráficos como:

- Indicadores de volume textual
- Gráficos de barras, mapas e treemaps
- Nuvem de palavras
- Visualizações de frequência por entidade

A organização das visualizações foi orientada pela clareza e acessibilidade da informação, permitindo ao usuário navegar de forma fluida e extrair rapidamente os principais insights dos acórdãos jurídicos.

4.5 Validação e Testes

Por fim, a consistência dos dados foi validada por meio da comparação entre contagens manuais e automatizadas em amostras aleatórias. Também foram realizados testes no Power BI para garantir que os filtros e interações entre gráficos refletissem corretamente os dados subjacentes. O objetivo foi assegurar confiabilidade e utilidade prática à solução proposta.

5. Resultados e Discussão

A partir da análise dos acórdãos jurídicos e da aplicação das técnicas de NLP, observou-se a possibilidade de extrair e visualizar informações relevantes com maior agilidade e inteligibilidade. O dashboard desenvolvido permitiu sintetizar um grande volume de dados textuais em métricas interpretáveis e visualmente acessíveis.

A identificação de palavras-chave e entidades nomeadas mostrou-se eficaz, revelando padrões temáticos recorrentes nos acórdãos, como menções à Constituição Federal, tribunais superiores e órgãos reguladores. A distribuição gramatical dos termos revelou predomínio de substantivos e verbos, evidenciando a natureza descritiva e normativa dos textos.

Em termos quantitativos, foi possível medir a densidade lexical dos documentos, facilitando a avaliação da complexidade textual. O mapa interativo por estado permitiu identificar a origem regional dos documentos e comparar a atuação judicial entre as unidades federativas.

A ferramenta demonstrou potencial para auxiliar pesquisadores, operadores do direito e gestores na interpretação de dados jurídicos com base em evidências. A interatividade das visualizações no Power BI também ampliou as possibilidades de análise exploratória.

6. Conclusão

Este trabalho demonstrou a viabilidade e relevância de aplicar técnicas de Processamento de Linguagem Natural e Business Intelligence na análise de acórdãos jurídicos. A integração entre Python e Power BI permitiu transformar textos complexos em dados estruturados, passíveis de exploração visual.

O dashboard desenvolvido oferece um recurso inovador para a visualização de informações textuais, contribuindo para a transparência, eficiência e inteligibilidade dos dados judiciais. Entre os principais resultados estão a identificação de padrões temáticos, a análise da estrutura lexical e a classificação automática de entidades nomeadas.

Como sugestão para trabalhos futuros, recomenda-se a expansão da base de dados, a utilização de modelos linguísticos mais avançados (como transformers) e a inclusão de métricas qualitativas, como análise de sentimentos e detecção de similaridade entre acórdãos.

7. Referências Bibliográficas

SIQUEIRA, Felipe A.; VITÓRIO, Douglas; SOUZA, Ellen; SANTOS, José A. P.; ALBUQUERQUE, Hidelberg O.; DIAS, Márcio S.; SILVA, Nádia F. F.; CARVALHO, André C. P. L. F. de; OLIVEIRA, Adriano L. I.; BASTOS FILHO, Carmelo. Ulysses Tesemô: a new large corpus for Brazilian legal and governmental domain. Language Resources and Evaluation, 2024. DOI: 10.1007/s10579-024-09762-8. Disponível em: https://github.com/ulysses-camara/ulysses-tesemo

FEW, Stephen. Show Me the Numbers: Designing Tables and Graphs to Enlighten. Oakland: Analytics Press, 2012.

HEER, Jeffrey; BOSTOCK, Mike; OGIEVETSKY, Vadim. A Tour Through the Visualization Zoo. Communications of the ACM, v. 53, n. 6, p. 59-67, 2010.

JURAFSKY, Daniel; MARTIN, James H. Speech and Language Processing. 3. ed. [S. l.]: Pearson, 2021.

LAX, Jeffrey R.; CAMERON, Charles M. Bargaining and Opinion Assignment on the U.S. Supreme Court. Journal of Law, Economics, & Organization, v. 23, n. 2, p. 276-302, 2007.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

McKINNEY, Wes. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter. 3. ed. Beijing: O'Reilly Media, 2021.

TUFTE, Edward R. The Visual Display of Quantitative Information. Cheshire, CT: Graphics Press, 1983.