**Universidade de Brasília**
Programa de Pós-graduação em Computação Aplicada – PPCA
Fundamentos de Pesquisa Operacional

# Relatório - Sumarização extrativa da Constituição Federal

Aluno: João Robson Santos Martins

**Professor:**
Marcelo A. Marotta

**21 de fevereiro de 2025**

# 1 Motivação

A análise de normas, e em especial da Constituição Federal, é essencial para compreender a evolução do ordenamento jurídico e suas implicações na sociedade. Ao longo do tempo, diferentes versões da Constituição refletem mudanças nas prioridades políticas, econômicas e sociais de um país, influenciadas por contextos históricos e governamentais. O estudo comparativo dessas versões permite identificar padrões de transformação, continuidade ou ruptura em princípios fundamentais, direitos e estrutura do Estado. Além disso, a análise dos resumos dessas normas ao longo dos anos facilita a detecção de tendências legislativas e a adaptação do sistema jurídico às novas demandas da sociedade, contribuindo para uma interpretação mais aprofundada do processo constitucional.

Assim, a sumarização extrativa dessas propostas representa uma ferramenta valiosa para facilitar a compreensão das mudanças ao longo do tempo. Em particular, tal tarefa pode ser modelada como um problema de otimização combinatória, onde o objetivo é selecionar um subconjunto de unidades textuais (sentenças ou palavras) a partir de um grupo de documentos, de forma que cobertura informacional seja maximizada, garantindo coerência e respeitando restrições de tamanho. De forma mais detalhada, ao construir resumos de múltiplos documentos, geralmente tenta-se otimizar algumas propriedades [3], tais como:

- A maximização da cobertura da informação: o resumo deve conter unidades textuais relevantes.
- A minimização da redundância: o resumo não deve conter unidades textuais repetidas ou que possuam o mesmo significado.
- O respeito às restrições de comprimento: o tamanho do resumo deve permanecer dentro de um limite de palavras pré-definido.

Para atingir tal objetivo, diversos métodos de Processamento de Linguagem Natural (PLN) podem ser empregados, abrangendo desde abordagens estatísticas e baseadas em heurísticas ou otimização até os avanços mais recentes com LLMs [7]. Nesse contexto, a Programação Linear Inteira (Integer Linear Programming (ILP), em inglês) [3, 7] se destaca por oferecer uma formulação eficiente e interpretável, garantindo uma seleção ótima (dentro das restrições aplicadas) de sentenças que melhor representam o conteúdo original.

Portanto, este **experimento busca avaliar se há mudanças perceptíveis entre os temas abordados na versão original, de 1988, da Constituição Federal, e a última versão, de 2024** por meio dos resumos gerados pelo modelo de ILP construído e detalhado nas seções 2 e 3. Com isso, os resultados obtidos podem contribuir para o aprimoramento de métodos de sumarização de textos legislativos, permitindo uma análise mais eficiente de normas em geral e suas alterações ao longo do tempo. Além disso, a identificação de possíveis mudanças nos temas pode fornecer insights valiosos sobre a evolução das prioridades legislativas ao longo dos anos, auxiliando analistas e tomadores de decisão na formulação de políticas públicas mais informadas e alinhadas com as demandas da sociedade.

# 2 Descrição do modelo de otimização

O modelo de otimização utilizado no trabalho se baseia na formulação proposta por McDonald [3]:

$$\max \sum_{i=1}^{n} \alpha_i Rel(i) - \sum_{j=i+1}^{n} \alpha_{ij} Red(i,j)$$

**s.t.** $\forall i, j$ :

$$\alpha_i, \alpha_{ij} \in \{0, 1\} \tag{1}$$

$$\sum_i \alpha_i l(i) \leq K \tag{2}$$

$$\alpha_{ij} - \alpha_i \leq 0 \tag{3}$$

$$\alpha_{ij} - \alpha_j \leq 0 \tag{4}$$

$$\alpha_i + \alpha_j - \alpha_{ij} \leq 1 \tag{5}$$

onde:

- $n$ representa o número de sentenças nos documentos de entrada.
- $Rel(i)$ representa a relevância da sentença $i$.
- $l(i)$ é o comprimento da sentença $i$.
- $Red(i, j)$ representa a similaridade entre as sentenças $i$ e $j$.
- $K$ é o comprimento máximo permitido para o resumo.
- As variáveis $\alpha_i$, representadas coletivamente por $\alpha$, são binárias e indicam quais sentenças $i$ são incluídas no resumo.
- As variáveis $\alpha_{i,j}$ indicam se ambas as sentenças $i$ e $j$ são incluídas no resumo.
- A restrição (2) garante que o comprimento total máximo permitido não seja excedido.
- As restrições (3) a (5) garantem a consistência dos valores de $\alpha_i$, $\alpha_j$ e $\alpha_{i,j}$ (por exemplo, se $\alpha_{i,j} = 1$, então $\alpha_i = \alpha_j = 1$; e se $\alpha_{i,j} = 0$, então $\alpha_i = 0$ ou $\alpha_j = 0$).

Nos experimentos, a relevância da sentença ($Rel(i)$) é calculada por meio da soma do TF-Idf de seus termos, dado por:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

com:

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$\text{IDF}(t) = \log\left(\frac{N}{1 + |\{d \in D : t \in d\}|}\right)$$

onde:

- $\text{TF}(t, d)$ é a frequência do termo $t$ no documento $d$, calculada como a contagem $f_{t,d}$ do termo $t$ no documento $d$ dividida pelo número total de termos em $d$.
- $\text{IDF}(t)$ é o fator de inversão de frequência de documento, dado pelo logaritmo do número total de documentos $N$ dividido pelo número de documentos que contêm o termo $t$ (com um somador de 1 no denominador para evitar divisão por zero).
- $D$ representa o conjunto de todos os documentos.

Já a similaridade ($Red(i, j)$) é calculada com base na similaridade de cosseno entre a média dos vetores dos termos presentes nas sentenças. As representações ("*embeddings*") são geradas pelo modelo word2vec [4], mais precisamente uma versão CBOW (*continous bag-of-words*) treinada para Português [1].

# 3 Protótipo e Resultados

## 3.1 Dado

O dado utilizado para a realização dos experimentos consiste em duas versões da Constituição Federal: o conteúdo original, de 5 de outubro de 1988, e o texto atual, cuja última alteração foi feita em 20 de dezembro de 2024.

Os dados foram extraídos do portal normas.leg.br [5] e foram divididos em dois grupos de documentos: 1988 e 2024. Cada grupo contém 8 documentos, um para cada Título da norma:

- Título I: Dos princípios fundamentais
- Título II: Dos direitos e garantias fundamentais
- Título III: Da organização do Estado
- Título IV: Da organização dos Poderes
- Título V: Da defesa do Estado e das instituições democráticas
- Título VI: Da tributação e do orçamento
- Título VII: Da ordem econômica e financeira
- Título VIII: Da ordem social

Para cada um dos 16 títulos, foi realizado um pré-processamento do texto formado pelas seguintes etapas:

- Remoção de indicadores das partes das normas (”Art. 1º”, ”Parágrafo único”, ”§ 3º”, etc.).
- Transformação do texto para letras minúsculas.
- Remoção de pontuação e acentuação.
- Remoção de *”stopwords”* (palavras comuns do idioma), tais como artigos (”a”, ”o”), verbos (”ser”, ”estar”) e termos jurídicos frequentes (”lei”, ”salvo”, ”caput”, etc.).

Após essa processo, o valor da relevância (TF-Idf) foi calculado para um dos dois grupos de documentos e o texto pré-processado foi dividido em sentenças (frases terminadas em ”.”, ”;”ou ”:”). Além disso, foi feito o cálculo de similaridade entre os termos das sentenças geradas, possibilitando o uso como valores de redundância usados pelo modelo.

## 3.2 Implementação

O modelo descrito na seção 2 foi implementado por meio da biblioteca OR-Tools [6] como um problema de programação linear inteira mista (*Mixed Integer Programming*, em inglês). Sua implementação pode ser vista aqui.

Tais problemas requerem que algumas variáveis sejam inteiras (discretas) ou booleanas, como é o caso do problema da sumarização.

Para sua solução, foi utilzado o *solver* SCIP e as variáveis mostradas em 2 foram modeladas da seguinte forma:

- $Rel(i)$: dicionário i: Rel(i) representando a relevância de cada sentença de entrada.
- $l(i)$: Dicionário i: len(i) representando o comprimento de cada sentença de entrada.
- $Red(i, j)$: Dicionário (i, j): Red(i, j) representando a similaridade entre pares de sentenças.

- $K$: comprimento máximo permitido para o resumo, configurado como **100** para o experimento.
- $\alpha_i$: um *BoolVar* para cada sentença de entrada.
- $\alpha_{i,j}$: um *BoolVar* para cada sentença de entrada.

### 3.3 Resultados

O resultado da geração está consolidado na tabela 1.

| **Título** | **1988** | **2024** |
|---|---|---|
| Título I<br>Dos Princípios Fundamentais | "constituem objetivos fundamentais republica federativa brasil", "cooperacao povos progresso humanidade" | "autodeterminacao povos", "naointervencao", "repudio terrorismo racismo", "cooperacao povos progresso humanidade" |
| Título II<br>Dos Direitos e Garantias Fundamentais | "perda bens", "direitos sociais", "direitos politicos", "pleno exercicio direitos politicos", "dezoito anos vereador" | "direitos sociais", "revogado", "direitos politicos", "nacionalidade brasileira", "pleno exercicio direitos politicos" |
| Título III<br>Da Organização do Estado | "vedado uniao estados distrito federal municipios", "uniao", "distrito federal", "servidores publicos militares" | "uniao", "municipios", "vinte vereadores municipios cento sessenta mil habitantes trezentos mil habitantes" |
| Título IV<br>Da Organização dos Poderes | "presidente senado federal", "supremo tribunal federal", "tribunais juizes trabalho", "tribunais juizes militares" | "presidente senado federal", "supremo tribunal federal", "superior tribunal justica", "tribunais juizes militares" |
| Título V<br>Da Defesa do Estado e das Instituições Democráticas | "estado defesa estado sitio", "estado defesa", "vigencia estado defesa", "estado sitio", "policia ferroviaria federal" | "estado defesa estado sitio", "estado defesa", "vigencia estado defesa", "estado sitio", "policia ferroviaria federal" |
| Título VI<br>Da Tributação e do Orçamento | "impostos", "cabe complementar", "impostos uniao", "imposto iii", "imposto iv", "imposto i", "impostos municipios", "imposto ii" | "cabe complementar", "imposto iii", "revogado", "imposto iv", "imposto vi", "imposto viii deste", "imposto i", "imposto ii", "vedados" |
| Título VII<br>Da Ordem Econômica e Financeira | "politica agricola fundiaria reforma agraria", "compatibilizadas acoes politica agricola reforma agraria" | "funcao social propriedade", "revogado", "revogada", "politica agricola fundiaria reforma agraria", "seguro agricola" |
| Título VIII<br>Da Ordem Social | "ordem social", "seguridade social", "saude", "gestao democratica ensino publico forma", "melhoria qualidade ensino" | "ordem social", "saude", "previdencia social", "educacao cultura desporto", "educacao", "melhoria qualidade ensino", "cultura" |

Tabela 1: Comparação dos resumos dos títulos entre 1988 e 2024

A partir dos resumos apresentados, observa-se certa evolução nos temas abordados em 2024 em comparação a 1988, com a inclusão de novos conceitos e a retirada de tópicos anteriores.

Para avaliar a qualidade da geração, avaliando se as mudanças significativas foram capturadas pelos resumos de 2024, é possível utilizar duas métricas:

- Proporção de sentenças adicionadas à versão de 2024 que também foram adicionadas aos resumos de 2024
- Similaridade de Jaccard [2] entre textos originais e entre os resumos

Os resultados das duas avaliações são mostrados nas tabelas 2 e 3. A partir da Tabela 2, observa-se que a proporção de sentenças adicionadas ao texto original varia significativamente entre os títulos. Algumas seções, como "Da Tributação e do Orçamento"(65%) e "Da Ordem Social"(49%), passaram por uma reformulação mais profunda, enquanto outras, como "Dos Princípios Fundamentais"(0%) e "Da Defesa do Estado e das Instituições Democráticas"(26%), tiveram poucas adições. No entanto, a proporção dessas novas sentenças refletidas nos resumos de 2024 é relativamente baixa, com apenas alguns resumos incorporando parte dessas mudanças.

Por exemplo, "Da Ordem Econômica e Financeira"apresenta 40% das sentenças nos resumos, enquanto resumos de títulos como "Da Ordem Social"e "Da Organização dos Poderes"não capturaram nenhuma das adições.

A Tabela 3 reforça essa análise ao medir a similaridade de Jaccard entre as versões de 1988 e 2024. No nível dos textos originais, os títulos com menor similaridade (como "Da Tributação e do Orçamento"com 0.32 e "Da Organização do Estado"com 0.49) indicam uma grande quantidade de mudanças. Já os resumos apresentam variação ainda maior, sugerindo que a condensação das informações pode não ter capturado todas as mudanças estruturais. Em especial, o título "Da Defesa do Estado e das Instituições Democráticas"apresenta uma similaridade máxima (1.0) entre resumos, sugerindo que essa seção permaneceu essencialmente inalterada no nível de síntese.

Esses resultados indicam que, embora os resumos de 2024 reflitam algumas mudanças do texto original, há uma discrepância na incorporação das alterações significativas. Algumas seções reformuladas em 2024 não foram devidamente representadas nos resumos, o que pode impactar a qualidade da extração de informações. Isso sugere a necessidade de um ajuste no processo de geração dos resumos, como aumento do limite máximo $K$ e adição de mais restrições, o que pode garantir que as mudanças mais relevantes sejam melhor capturadas.

| **Título** | **Proporção de sentenças adicionadas à versão de 2024** | **Proporção do resumo de 2024 contendo sentenças adicionadas** |
|---|---|---|
| Título I<br>Dos Princípios Fundamentais | 0/26 (0 %) | 0/4 (0 %) |
| Título II<br>Dos Direitos e Garantias Fundamentais | 33/250 (13 %) | 1/5 (20 %) |
| Título III<br>Da Organização do Estado | 182/419 (43 %) | 1/3 (33 %) |
| Título IV<br>Da Organização dos Poderes | 278/798 (34 %) | 0/4 (0 %) |
| Título V<br>Da Defesa do Estado e das Instituições Democráticas | 22/85 (26 %) | 0/5 (0%) |
| Título VI<br>Da Tributação e do Orçamento | 346/529 (65 %) | 3/9 (33 %) |
| Título VII<br>Da Ordem Econômica e Financeira | 29/125 (23 %) | 2/5 (40 %) |
| Título VIII<br>Da Ordem Social | 195/395 (49 %) | 0/7 (0 %) |

Tabela 2: Proporção de sentenças adicionadas ao texto original e aos resumos (2024)

| **Título** | **Textos completos (2024 vs 1988)** | **Resumos (2024 vs 1988)** |
|---|---|---|
| Título I<br>Dos Princípios Fundamentais | 1 | 0.2 |
| Título II<br>Dos Direitos e Garantias Fundamentais | 0.81 | 0.43 |
| Título III<br>Da Organização do Estado | 0.49 | 0.17 |
| Título IV<br>Da Organização dos Poderes | 0.58 | 0.6 |
| Título V<br>Da Defesa do Estado e das Instituições Democráticas | 0.7 | 1 |
| Título VI<br>Da Tributação e do Orçamento | 0.32 | 0.42 |
| Título VII<br>Da Ordem Econômica e Financeira | 0.62 | 0.17 |
| Título VIII<br>Da Ordem Social | 0.46 | 0.33 |

Tabela 3: Similaridade de Jaccard entre sentenças de 1988 e 2024

# Referências


[1] Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues, and Sandra Aluisio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks, 2017.

[2] Paul Jaccard. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44:223–270, 1908.

[3] Ryan McDonald. A study of global inference algorithms in multi-document summarization. In Giambattista Amati, Claudio Carpineto, and Giovanni Romano, editors, *Advances in Information Retrieval*, pages 557–564, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[5] normas.leg.br. Constituição da república federativa do brasil, 2024. Accessed: 2024-11-21.

[6] Laurent Perron and Vincent Furnon. Or-tools.

[7] Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. A systematic survey of text summarization: From statistical methods to large language models, 2024.