Programa de
Pós-Graduação em
Computação Aplicada

Modelo de programação linear para sumarização extrativa da Constituição Federal

João Robson Santos Martins

INTRODUÇÃO E MOTIVAÇÃO

- **Análise Comparativa da Constituição Federal**
  - Evolução das normas jurídicas reflete mudanças políticas, econômicas e sociais.
  - Comparação entre versões permite identificar padrões de transformação, continuidade ou ruptura.
  - Resumos das normas auxiliam na detecção de tendências legislativas.
- **Objetivo: Sumarização extrativa via otimização**
  - Problema modelado como otimização combinatória.
  - **Objetivos principais:**
    - Cobertura informacional: incluir sentenças relevantes.
    - Minimização da redundância: evitar informações repetidas.
    - Restrições de comprimento: respeitar limite de caracteres

DESCRIÇÃO DO MODELO

- **Baseado na formulação proposta por McDonald [3]:**

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

$$\textbf{s.t. } \forall i, j:$$

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

$$\sum_i \alpha_i l(i) \leq K$$

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

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

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

- n: número de sentenças nos documentos.
- α_i: variável binária, indica quais sentenças i são incluídas no resumo.
- α_{ij}: variável binária, indica se ambas as sentenças i e j são incluídas no resumo.
- $Rel(i)$: relevância da sentença i
- l(i): comprimento da sentença i.
- $Red(i,j)$: similaridade entre sentenças i e j
- K: comprimento máximo permitido para o resumo.

DESCRIÇÃO DO MODELO

- **Cálculo da relevância:**
  - A relevância *Rel(i)* de uma sentença é dado pela soma dos valores TF-Idf de seus termos
  - TF-Idf é dado por:

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

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

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

- *TF(t, d) :* 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*.
- *IDF(t):* 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.
- *D:* conjunto de todos os documentos.

DESCRIÇÃO DO MODELO

- **Cálculo da redundância:**
  - A redundância *Red(i, j)* de duas sentenças é calculada com base na **similaridade de cosseno** entre a média dos vetores dos termos presentes nas sentenças.
  - As representações ("*embeddings*") são geradas por um modelo word2vec [2] treinado em Português [3].

$$S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

DADOS

- **Versões utilizadas da Constituição Federal**
  - Original: 5 de outubro de 1988
  - Atual: 20 de dezembro de 2024
- **Extração e Estruturação dos Dados**
  - Fonte: Portal normas.leg.br [5]
  - Grupos de Documentos:
    - 1988: 8 documentos (um para cada Título)
    - 2024: 8 documentos (um para cada Título)
  - Títulos:
    - Dos princípios fundamentais
    - Dos direitos e garantias fundamentais
    - Da organização do Estado
    - Da organização dos Poderes
    - Da defesa do Estado e das instituições democráticas
    - Da tributação e do orçamento
    - Da ordem econômica e financeira
    - Da ordem social

PROCESSAMENTO E OTIMIZAÇÃO

- **Pré-processamento do texto:**
  - Remoção de indicadores das partes das normas (ex: "Art. 1º", "§ 3º").
  - Transformação para letras minúsculas.
  - Remoção de pontuação e acentuação.
  - Remoção de stopwords (ex: artigos, verbos comuns, termos jurídicos).
- **Cálculo da Relevância**
  - TF-Idf para cada grupo de documentos
- **Geração das sentenças**
  - Divisão do texto utilizando símbolos (., : e ;).
- **Cálculo da Redundância**
  - Média da similaridade de cosseno para os termos das sentenças geradas.
- **Otimização do modelo**
  - OR-Tools
  - Solver utilizado: SCIP
  - K (limite máximo de caracteres por resumo) = 100

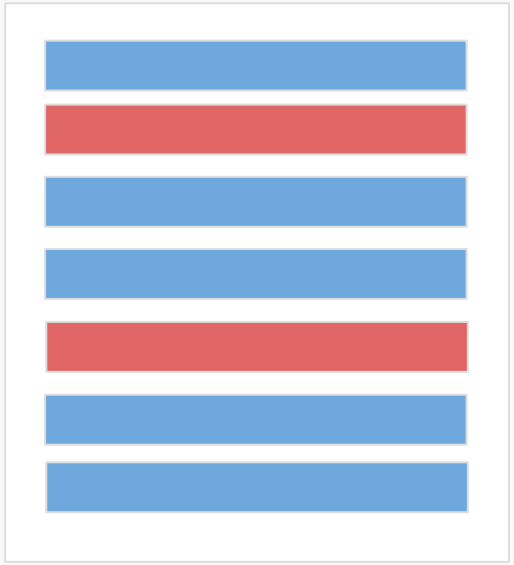
RESULTADOS - RESUMOS GERADOS

Título	1988	2024
Título I	"constituem objetivos fundamentais republica federativa brasil", "cooperacao povos progresso humanidade"	"autodeterminacao povos", "naointervencao", "repudio terrorismo racismo", "cooperacao povos progresso humanidade"
Título II	"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	"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	"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	"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	"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	"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	"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"

RESULTADOS - PROPORÇÃO DE SENTENÇAS ADICIONADAS EM RELAÇÃO AO TOTAL DE SENT.

Texto original: 5 sentenças

Texto atual: 7 sentenças
(2/7 foram adicionados)

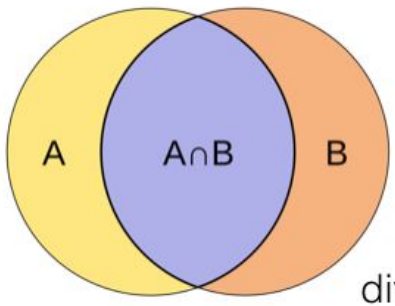


RESULTADOS - PROPORÇÃO DE SENTENÇAS ADICIONADAS EM RELAÇÃO AO TOTAL DE SENT.

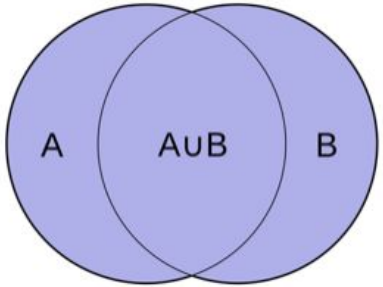
Título	Proporção de sentenças adicionadas (versão 2024)	Proporção dos resumos de 2024 contendo as sentenças adicionadas
Título I	0/26 (0 %)	0/4 (0 %)
Título II	33/250 (13 %)	1/5 (20 %)
Título III	182/419 (43 %)	1/3 (33 %)
Título IV	278/798 (34 %)	0/4 (0 %)
Título V	22/85 (26 %)	0/5 (0 %)
Título VI	346/529 (65 %)	3/9 (33 %)
Título VII	29/125 (23 %)	2/5 (40 %)
Título VIII	195/395 (49 %)	0/7 (0 %)

RESULTADOS - SIMILARIDADE (JACCARD) ENTRE SENTENÇAS (1988 VS 2024)

$$J(A,B) = \frac{\text{The intersect of A \& B}}{\text{The union of A \& B}}$$



division



RESULTADOS - SIMILARIDADE (JACCARD) ENTRE SENTENÇAS (1988 VS 2024)

Título	Textos completos (2024 vs 1988)	Resumos (2024 vs 1988)
Título I	1	0.2
Título II	0.81	0.43
Título III	0.49	0.17
Título IV	0.58	0.6
Título V	0.7	1
Título VI	0.32	0.42
Título VII	0.62	0.17
Título VIII	0.46	0.33

CONCLUSÕES

- Resumos de 2024 refletem algumas mudanças do texto original, mas com discrepâncias.
- Necessidade de ajuste no processo de geração dos resumos:
  - Aumento do limite máximo K
  - Adição de mais restrições.
    - Garantir que mudanças relevantes sejam melhor capturadas.
    - Garantir que determinados trechos tenha relevância maior (artigos > alíneas, por exemplo).

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. Constitui ¸ c˜ao da rep´ublica 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.