# Trabalho 2 PLN 2025 - Representações de Texto

## Implementação e Análise de Técnicas de Vetorização de Texto

### João Martinho, Felipe Bona

**Implementação e Adaptação para os Dados Textuais**

Utilizamos o programa BibleScraper para extrair e analisar textos bíblicos, aplicando diferentes técnicas de vetorização e representação de texto. Abaixo estão os resultados para cada técnica:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Técnica / Modelo | Cálculo de similaridade | Busca semântica | Visualização (PCA + clusters) | Variação na tokenização | Pontos |
| TF-IDF | Sim (cosine similarity) | Sim | Sim (agrupamento K-means) | Stopwords em português | 4,0 |
| Word2Vec | Sim (cosine similarity) | Sim | Sim (projeção 2D) | Lemmatização | 3,5 |
| Vetores pré-treinados (SpaCy) | Sim (cosine similarity) | Sim | Sim (projeção 2D) | Nenhuma | 4,0 |

**Detalhes de Implementação:**

a) Cálculo de Similaridade:

* Implementamos cálculo de similaridade usando cosine similarity para todas as técnicas;
* Para TF-IDF, usamos *sklearn.metrics.pairwise.cosine\_similarity*;
* Para Word2Vec e SpaCy, calculamos a média dos vetores de palavras e depois a similaridade.

b) Busca Semântica:

* Implementada na função *search\_by\_subject()*;
* Usa TF-IDF para encontrar versículos relacionados a um assunto;
* Limite de similaridade mínimo configurável (0,35).

c) Visualizações:

* Clusterização com K-means mostrada na aba "Análise de Tópicos";
* Visualização dos clusters de versículos por similaridade;
* Exibição das palavras mais relevantes por TF-IDF.

d) Variações Testadas:

* TF-IDF: com e sem stopwords em português;
* Word2Vec: com lemmatização prévia;
* SpaCy: usando o modelo pré-treinado *pt\_core\_news\_lg.*

e) Limitações Identificadas:

* TF-IDF não captura relações semânticas entre palavras;
* Word2Vec requer treinamento adicional para domínio específico;
* SpaCy tem vocabulário fixo e pode não reconhecer peculiares à Bíblia.

**Análise da Viabilidade e Utilidade:**

A aplicação das técnicas de vetorização ajudou na análise de textos bíblicos. Os principais resultados foram:

1. TF-IDF:

* Excelente para identificar termos-chave e similaridade superficial;
* Eficaz para busca por assuntos específicos;
* No entanto, não captura relações semânticas profundas.

2. Word2Vec:

* Captura melhor relações semânticas entre conceitos;
* Requer ajuste fino para o assunto da religião;
* Performance inferior a vetores pré-treinados no nosso caso de uso.

3. Vetores Pré-treinados (SpaCy):

* Melhor performance geral;
* Captura relações semânticas complexas;
* No entanto, o tamanho do vocabulário é fixo.

**Qualidade dos Dados:**

Os textos bíblicos apresentam desafios específicos:

* Linguagem inusitada em algumas traduções;
* Nomes próprios e termos específicos do assunto;
* Estrutura versificada em alguns trechos.

A limpeza dos dados (remoção de números de versículos, pontuação, etc.) foi importante para melhorar os resultados. A adição de stopwords específicas do assunto (ex.: "Deus", "Senhor") pode melhorar ainda mais a análise.

**Conclusão:**

Para esta aplicação, os vetores pré-treinados do SpaCy apresentaram os melhores resultados, combinando boa performance com a capacidade de capturar a semântica. TF-IDF mostrou-se a melhor opção para tarefas mais simples, como busca por termos. Word2Vec exigiria mais ajustes para ser competitivo neste domínio.

A abordagem híbrida, usando TF-IDF para busca inicial e vetores do SpaCy para análise semântica mais profunda, parece ser a mais promissora para trabalhos futuros.