**UNIVERSIDADE REGIONAL DE BLUMENAU**

**CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA**

**CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

**Disciplina: Processamento de Linguagem Natural**

**Unidade 3 – Representações de Texto**

**Data de entrega: 22/04**

**Equipe:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

O trabalho deverá apresentar, por meio da implementação real com a base de dados escolhida e de avaliação qualitativa, um “benchmark” comparando o uso de vetores/modelos/embeddings para diferentes tarefas. Serão avaliados:

**1) Implementação e adaptação para os dados textuais do grupo. Para cada técnica ou modelo serão empregados os seguintes critérios (4 pontos):**

a) Foi apresentado cálculo de similaridade entre os documentos?

b) Foi testada alguma forma de busca semântica baseada em vetorização? (importante: não é busca simples por correspondência entre strings!)

c) Foram geradas visualizações para avaliação e validação dos resultados?

d) Foi testada ao menos uma variação (ex. com texto limpo versus stemming ou lemmatização etc., uso de n-grams)?

e) Foi preenchida a tabela de síntese (modelo abaixo), avaliando os resultados (muito bom, bom, ruim, péssimo) e comentando brevemente as limitações?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Técnica / modelo** | **Cálculo de similaridade entre documentos** | **Busca semântica** | **Visualização (ex. com PCA + clusters, mas serão aceitas outras)** | **Variação na tokenização? Qual?** | **Pontos** |
| **TF-IDF** |  |  |  |  | 1,33 |
| **Word2Vec** |  |  |  |  | 1,33 |
| **Vetores pré-treinados (SpaCy,BERT OpenAI etc.)** |  |  |  |  | 1,33 |
| **Outras técnicas (opcional)** |  |  |  |  | Bônus |

**2) Texto analítico, resumindo e avaliando a viabilidade, utilidade e resultados da aplicação das tarefas estudadas (cálculo de similaridade, busca etc.) para os dados textuais empregados pelo grupo. Deverá incluir reflexão sobre a qualidade dos dados utilizados.** Tamanho mínimo: 1 página, Times New Roman/Arial 12, espaçamento 1,5. **(5 pontos)**.