**UNIVERSIDADE REGIONAL DE BLUMENAU**

**CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA**

**CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

**Disciplina: Processamento de Linguagem Natural**

**Unidade 2 – Web scraping e preparação de dados textuais**

**Data de entrega: 08/04**

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

|  |  |  |
| --- | --- | --- |
| **Dimensão** | **Critérios de avaliação** | **Peso** |
| **Base de dados textuais** | * Justificativa e explicação da escolha * Adequação dos dados às tarefas de PLN (volume, abrangência, variedade) * Organização e interpretabilidade da base de dados | 0,4 |
| **Scraping ou acesso via API (código)** | * Funcionamento e reprodutibilidade * Documentação dos procedimentos realizados (comentários ou texto markdown no notebook) * Criatividade / capacidade de resolução de problemas | 0,3 |
| **Limpeza e preparação dos dados** | * Dados foram tokenizados? * Dados foram normalizados? * Remoção de ruídos (ex.: tags sem uso) * Remoção de stopwords e pontuação | 0,3 |
| **“Bonus round”** | * Stemming e lematização * Bases adicionais (ex.: comparação entre diferentes fontes ou recortes amostrais) * Automatização na coleta (ex.: atualização periódica do scraping) | Até 0,3 |