PRÁTICA 01 – SEQUÊNCIAS, PALAVRAS E TOKENS

PRÉ-REQUISITO: Possuir o ambiente configurado para a linguagem Python e ter instalado a biblioteca NLTK.

O pré-processamento é uma etapa fundamental em projetos de Processamento de Linguagem Natural (PLN), pois ajuda a preparar os dados para análise e modelagem. As principais atividades realizadas durante essa etapa podem ser resumidas em: Coleta de Dados, Limpeza de texto, Tokenização, Normalização, Remoção de stopwords e Extração de características.

Com base nisso, leia atentamente as instruções abaixo e realize cada uma das etapas conforme solicitado utilizando a biblioteca NLTK da linguagem Python. Lembre-se de registrar todo o processo em um relatório técnico de acordo com as normas da ABNT.

1 – COLETA DE DADOS

- Escolher uma empresa real que utilize um e-commerce ao qual seja possível coletar informações das avaliações reais realizadas por seus clientes.
- Criar a base de dados contendo no mínimo 20 avaliações (positivas, negativas e neutras). Esses dados podem estar armazenados em um array de strings. Observação: essa etapa já foi realizada na atividade “ESTUDO DE CASO: ANÁLISE DE SENTIMENTO DE AVALIAÇÕES DE PRODUTOS”.

2 – LIMPEZA DE TEXTO

- Deve ser realizada a remoção de caracteres especiais, números ou símbolos indesejados. É uma etapa opcional, portanto, se a base de dados não possuir essas informações, basta pular essa etapa.

3 – TOKENIZAÇÃO

- Dividir cada texto (avaliação) em palavras (tokens) ou frases (sentenças). Isso é essencial para as análises subsequentes.

4 – NORMALIZAÇÃO

- LOWERCASING: Converter todo o texto para minúsculas para evitar duplicação de palavras por causa de capitalização.
- LEMATIZAÇÃO: Reduzir palavras as suas formas base. Por exemplo, “correr” e “correndo” podem ser lematizadas para o termo “correr”.
- RADICALIZAÇÃO: Remover os sufixos de palavras para chegar a raiz, embora isso possa ser menos preciso que a lematização dependendo do contexto.

5 – REMOÇÃO DE STOPWORDS

- Stop words são palavras comuns (artigos, pronomes, advérbios, preposições, entre outras) que muitas vezes são removidas, pois podem não contribuir para a análise que está sendo realizada.

6 – EXTRAÇÃO DE CARACTERÍSTICAS

- Extraia as seguintes características da base de dados: quantidade de registros, quantidade de tokens e quantidade de types.
- Realize a operação de POS Tagging (Part-of-Speech Tagging). O POS Tagging é o processo de marcar as palavras em um texto com suas respectivas categorias gramaticais, como substantivos, verbos, adjetivos, entre outros.