TALENTO TECH

REGIÓN 3
CAUCA – NARIÑO
BOOTCAMP
ANÁLISIS DE DATOS

TIC

Universidad Tecnológica de Pereira

faceIT

TIC

GUÍA INICIAL – INTEGRADOR MÓDULO 3

TALENTO TECH

Universidad Tecnológica de Pereira

faceIT

Objetivo general

UNIDAD 1

Objetivo General: Al concluir esta lección, los estudiantes habrán adquirido las habilidades necesarias para:

- Comprender las técnicas fundamentales de análisis de textos, incluyendo la tokenización, el análisis de frecuencia de palabras y la estadística aplicada a textos. Esto les permitirá desglosar los textos en unidades manejables y comprender la importancia de la distribución de palabras en un corpus de texto.
- Reconocer la importancia de la tokenización en el modelado de textos, comprendiendo cómo esta técnica estructura los datos textuales y facilita su procesamiento en aplicaciones de procesamiento del lenguaje natural y análisis de textos.
- Familiarizarse con el concepto de lematización y su utilidad en la interpretación automática de textos. Los estudiantes aprenderán cómo la lematización simplifica el análisis de textos al reducir las palabras a su forma base, lo que facilita la identificación de relaciones semánticas y la realización de análisis precisos.

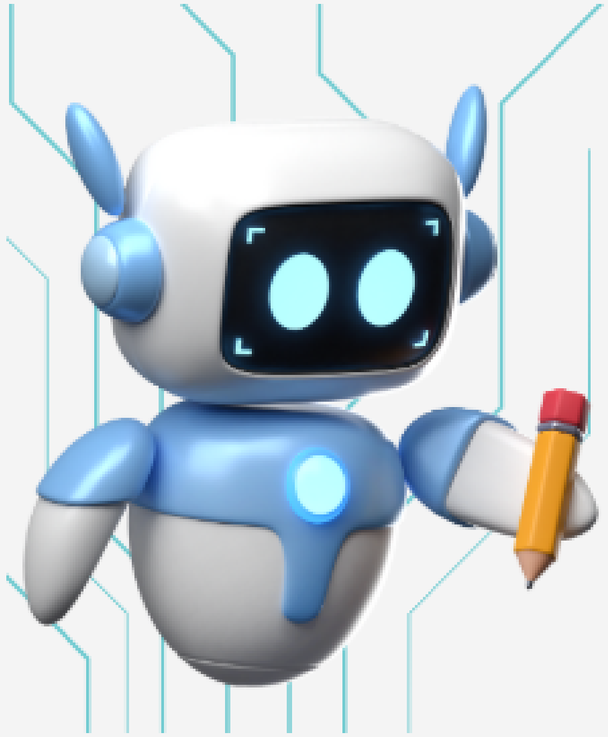
Competencias a desarrollar

- Análisis de textos
- Modelado de textos
- Interpretación automática de textos
- Prácticas con datos de textos

Analizar textos utilizando técnicas como la tokenización y el análisis de frecuencia de palabras para extraer información relevante y comprender la estructura y el contenido de los textos.

Aplicar la tokenización y la lematización para modelar textos de manera efectiva, representando datos textuales de manera estructurada y procesable para su análisis posterior.

Competencias a desarrollar



Habilidades en el uso de la lematización y otras técnicas de procesamiento del lenguaje natural para interpretar automáticamente textos, identificando relaciones semánticas y realizando análisis sintácticos y semánticos.

Aplicar los conocimientos teóricos adquiridos en un entorno práctico, utilizando herramientas y bibliotecas de análisis de textos para realizar tareas como la tokenización, lematización y modelización de temas en conjuntos de datos reales.

Activación de saberes previos

PLANTEAMIENTO DE LA SESIÓN

El análisis de texto es una disciplina fundamental en el procesamiento del lenguaje natural, y esta sesión práctica está diseñada para brindar a los participantes una comprensión sólida de las técnicas clave involucradas en este proceso. Se comenzará con una introducción al análisis de texto y su importancia en diversas aplicaciones, como la clasificación de documentos, la búsqueda de información y la generación de texto automático.

A continuación, se explorará la tokenización, que es el proceso de dividir un texto en unidades más pequeñas, como palabras o caracteres. Esta técnica es esencial para estructurar y procesar los datos textuales de forma adecuada. Los participantes realizarán una actividad práctica en la que tokenizarán diversos textos utilizando herramientas y técnicas específicas.

MATERIALES

PC con conexión a internet.

Activación de saberes previos

PLANTEAMIENTO DE LA SESIÓN

Después de comprender la importancia de la tokenización, se enfocará en el análisis de frecuencia de palabras en un corpus textual. Una vez que se haya dominado el análisis de frecuencia de palabras, se abordará la lematización, que consiste en reducir las palabras flexibles o derivadas a su forma básica (lema).

Se cubrirán las estadísticas de texto, incluyendo técnicas para analizar la distribución de palabras, extraer características y modelar temas. Comprender estas técnicas proporcionará una visión más profunda de la estructura y el contenido de los textos. Trabajar con herramientas y librerías populares de procesamiento del lenguaje natural, como NLTK (Natural Language Toolkit) en Python es necesario, ya que permite el entendimiento de diversos procesos en dicha librería.