Seminario de Métodos Computacionales:

Estudio de Lenguas Amerindias

Prof: Erasmo Gómez

PUCP

PROCESAMIENTO de textos

Estudio de
lenguas Amerindias

AGENDA

Pre procesamiento

Limpieza Textual
Stopwords

Tokenización

Lematización
Vocabulario

Representación

Visualización
Representación Vectorial
Actividad

Procesamiento de Textos

¿QUÉ ES?

Definición del procesamiento de textos

El procesamiento de textos se refiere a las técnicas utilizadas para transformar texto crudo en datos que pueden ser analizados por computadora. Este proceso es fundamental en diversas aplicaciones, como la **clasificación**, que organiza textos en categorías, y la **traducción**, que facilita la traducción automática de lenguas.

Importancia del procesamiento de textos

El procesamiento de textos permite a los lingüistas extraer información valiosa y realizar análisis profundos de los datos textuales. También incluye el **modelado de lenguas**, que estudia patrones y estructuras lingüísticas, lo que es crucial para entender mejor el lenguaje.

Limpieza Textual

La limpieza textual es el primer paso en el procesamiento de textos. Este proceso incluye:

- **Eliminación de puntuación:** Quitar signos de puntuación que no aportan información.
- **Eliminación de símbolos:** Deshacerse de caracteres especiales que pueden interferir con el análisis.

Importancia: La limpieza de datos es crucial para asegurar que el análisis posterior sea preciso y significativo.

Tokenización

La tokenización es la técnica que consiste en dividir el texto en tokens, palabras u oraciones. Este proceso es esencial para el análisis de texto, ya que permite:

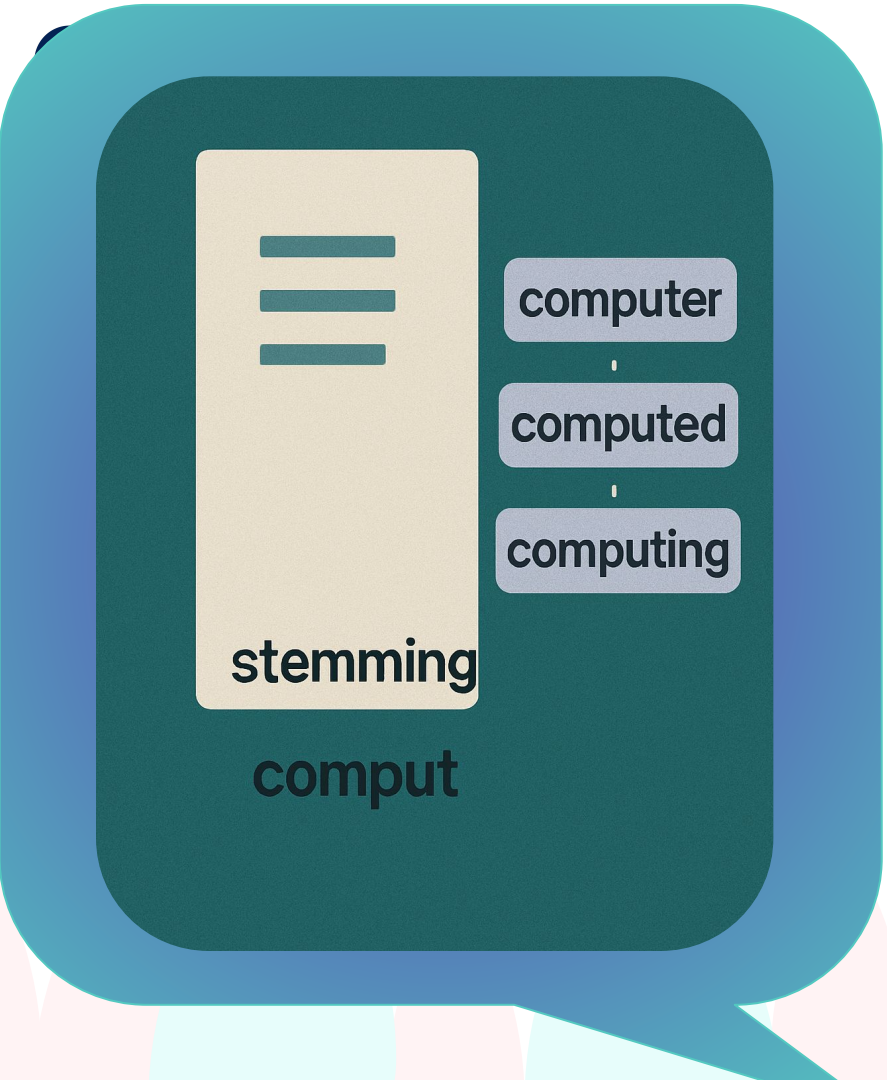
- **Identificar unidades significativas:** Facilita el análisis de la frecuencia de palabras y la estructura del texto.
- **Preparar datos para análisis posteriores:** Los tokens se utilizan en diversas aplicaciones de procesamiento de lenguaje natural.

Stopwords y Lematización

Definiciones Clave:

- **Stopwords:** Son palabras comunes (como ”y”, ”el”, ”de”) que se eliminan del texto porque no aportan valor analítico.
- **Lematización:** Es el proceso de reducir una palabra a su forma base o raíz, con un significado en la lengua.

Relevancia: Estas técnicas son fundamentales para mejorar la calidad del análisis de texto, permitiendo que los lingüistas se concentren en las palabras más significativas.



Stemming

El stemming es el proceso de reducir las palabras a su raíz morfológica. A diferencia de la lematización, que busca la forma base, el stemming corta las palabras a su raíz.

- **Lematización:** Considera el contexto y la gramática.
- **Stemming:** Se basa en reglas más simples y puede no producir palabras reales.

Uso en Procesamiento de Textos: El stemming es útil para reducir la variabilidad de las palabras y facilitar el análisis.

Visualización: Nube de Palabras

La visualización de datos a través de nubes de palabras es una técnica efectiva para interpretar los resultados del análisis textual.

- **Cómo funciona:** Las palabras más frecuentes aparecen en un tamaño mayor, lo que permite identificar rápidamente los temas principales.
- **Beneficios:** Facilita la comprensión de patrones y tendencias en los datos textuales.

Representación Vectorial

La representación vectorial de texto implica convertir texto en números para su análisis computacional. Este proceso es esencial para aplicar algoritmos de aprendizaje automático.

- **Métodos comunes:** TF-IDF (Term Frequency-Inverse Document Frequency) y Word Embeddings.
- **Importancia:** Permite a las computadoras procesar y analizar texto de manera eficiente.

Conclusión y Preguntas

En conclusión, hemos explorado los conceptos clave del procesamiento de textos y su aplicación en el estudio de lenguas amerindias. Los puntos clave incluyen:

- Importancia del **procesamiento de textos**.
- Técnicas de **limpieza, tokenización, lematización y visualización**.

Ahora abrimos el espacio para preguntas y discusión sobre el procesamiento de textos.

Actividad con Corpus

Ejercicio Práctico: Durante esta sesión, aplicarán lo aprendido en un ejercicio práctico utilizando un corpus.

- **Objetivo**: Procesar un conjunto de textos, aplicando técnicas de limpieza, tokenización y lematización.
- **Resultados esperados**: Los estudiantes obtendrán experiencia práctica en el manejo de datos textuales y su análisis.

¡Gracias!