UNIVERSIDAD HEMISFERIOS

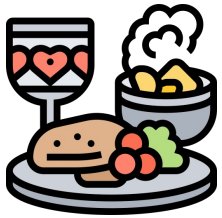
Procesamiento del lenguaje

Janneth Chicaiza Espinosa

Bienvenida y Unidad 1

Janneth Chicaiza Espinosa

- Máster en Ciencia de Datos por la Universidad Oberta de Catalunya
- Doctora en Software y Sistemas por la Universidad Politécnica de Madrid.
- Docente a tiempo completo en la Universidad Técnica Particular de Loja.
- Trabajo en líneas relacionadas con ciencia de datos, representación del conocimiento y procesamiento de lenguaje natural.



Procesamiento del lenguaje

La asignatura proporciona una comprensión fundamental, pero práctica de las técnicas y herramientas para la interacción entre las máquinas y el lenguaje humano.

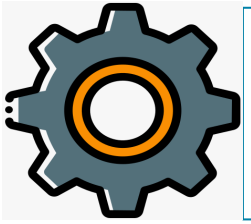
Escribe la primera estrofa del poema "LA HIERBA" de Violeta Luna

Cuántas cosas que pude haberlas dicho
y no las dije.
Cuántas horas que pude disfrutarlas
y no fueron.
Cuántas letras que se quedaron sueltas
sin remedio.
Cuánta vida que pudo ser raíz
y es hoy astilla.

Escuela de Educación Online

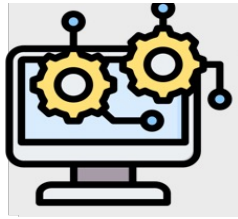
Bloques temáticos y actividades

1. Fundamentos y pre-procesamiento del texto



2. Modelado de lenguaje y extracción de características

3. Clasificación de texto y análisis de sentimientos



4. Generación de lenguaje natural

✓ Test de Unidad
✓ Caso práctico de Unidad

✓ Test Final
✓ Caso práctico final

Sistema de calificación (1)

Actividad	Ptos.	Fecha
PARCIAL 1 (sobre 30 puntos)		
Test de la Unidad 1: Fundamentos y preprocesamiento de texto	10	Domingo, 10 nov.
Caso práctico 2: Representación de texto	20	Hasta el martes, 12 nov.
PARCIAL 2 (sobre 30 puntos)		
Test de la Unidad 3: Clasificación de texto y análisis de sentimientos	10	Domingo, 17 nov.
Caso práctico 4: Aplicación de modelos generativos de lenguaje	20	Hasta el martes, 19 nov.

Sistema de calificación (2)

Actividad	Ptos.	Fecha
EVALUACIÓN FINAL (sobre 40 puntos)		
Test final: Unidades 1 a 4.	15	Jueves, 21 nov.
Caso práctico final: Preprocesamiento de texto y análisis de sentimientos.	25	Hasta el sábado, 23 nov.

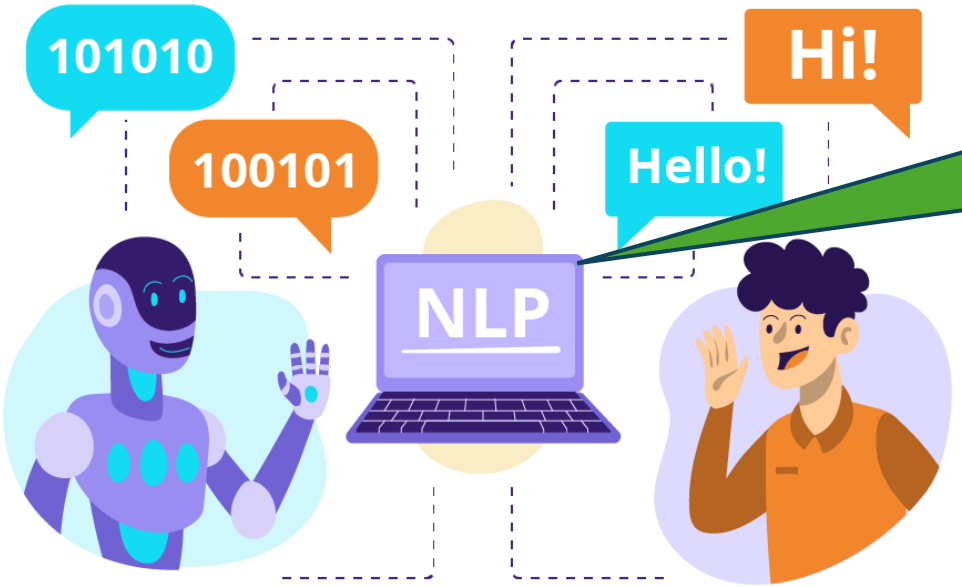
Unidad 1

Fundamentos y preprocesamiento de texto

Aplicar las técnicas de limpieza y normalización de datos de texto para prepararlos adecuadamente de tal manera que puedan ser usados en tareas de procesamiento del lenguaje natural

Motivación

- ✓ Crecimiento exponencial del volumen de la información textual -> NLP facilita la comprensión y análisis de este tipo de contenido.



Campo de la IA que se enfoca en la interacción entre las computadoras y el lenguaje humano (voz y texto)

Source: link

Análisis léxico y morfológico

NLP concepts

NLP

Lenguaje natural

Vocabulario

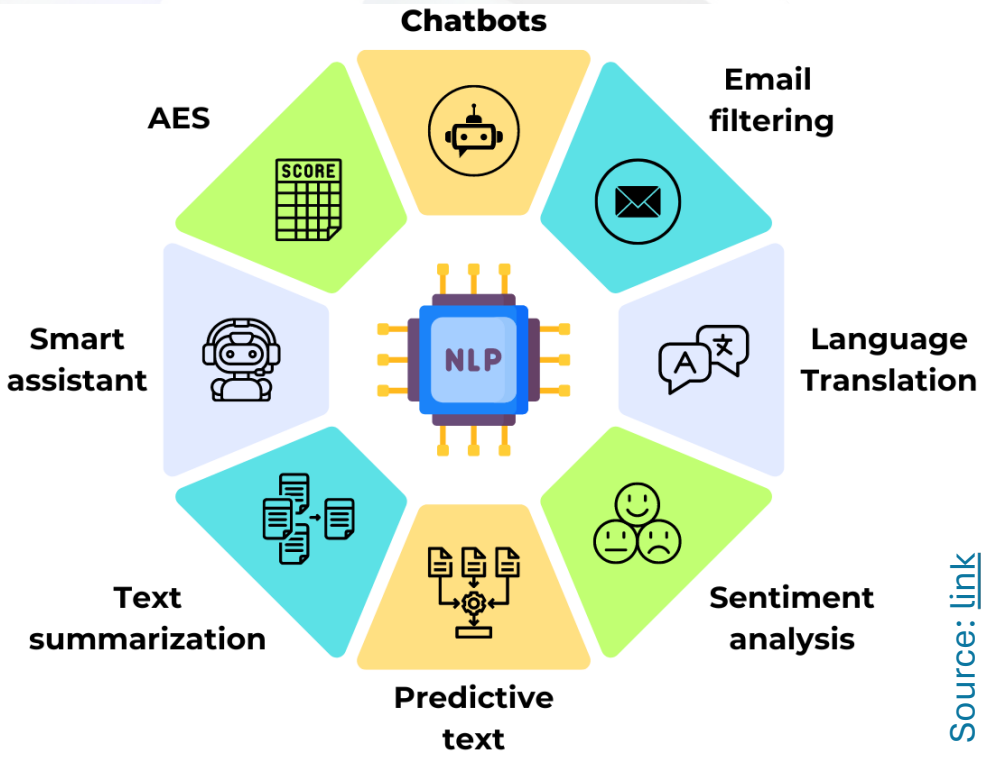
Words

Análisis semántico

Corpus

Análisis sintáctico

NLP applications



Source: link

NLP libraries & tools

python™

Desafíos ¿Lás máquinas comprenden el lenguaje natural?

1. Ambigüedad: El lenguaje natural es ambiguo: una misma palabra puede tener múltiples interpretaciones, dependiendo del contexto (polisemia).
2. Heterogeneidad lingüística o de vocabulario: La heterogeneidad (diferencias en términos) puede dificultar la comprensión automática del lenguaje.
3. Resolución de anáforas: La resolución de anáforas implica identificar a qué se refiere exactamente un pronombre o una frase de referencia.
4. Tratamiento de errores y ruido: Los errores tipográficos, las abreviaturas, las faltas de ortografía y otros tipos de ruido en los datos pueden dificultar el NLP, ya que pueden afectar la precisión de las tareas de análisis y generación de texto.
5. Entendimiento del contexto: Comprender el contexto en el que se utiliza el lenguaje es crucial para interpretar correctamente su significado (detectar sarcasmo, ironía o ambigüedad intencional).
6. Escasez de datos etiquetados: La obtención de datos etiquetados puede ser costosa y laboriosa.
7. Generalización: Los modelos de NLP pueden tener dificultades para generalizar el conocimiento aprendido en un conjunto de datos -> rendimiento deficiente.

Preprocesamiento de texto

"If 80 percent of our work is data preparation, then ensuring data quality is the important work of a machine learning team." – Andrew Ng

- Más del 80% de los datos no tienen estructura o son semiestructurados, el preprocesamiento de datos es un proceso esencial para realizar, previo al análisis de texto.
- Antes de analizar los datos de texto es importante prepararlos (Deepanshi, 2023), lo que implica **descomponer el lenguaje humano en sus partes más básicas** y luego comprender cómo estas partes se relacionan entre sí y trabajan juntas para crear significados en las oraciones.
- Las tareas de preprocesamiento constituyen una parte fundamental del NLP porque son **útiles para comprender la estructura y el significado de textos de manera automatizada**, lo que permite a las máquinas interpretar el lenguaje humano.
- Las técnicas de preprocesamiento buscan resolver las ambigüedades del lenguaje y crear representaciones del texto que facilite la comunicación humano-máquina.

Técnicas de preprocesamiento léxico y sintáctico

- Mejoran la estructura y comprensión de los textos.

Limpieza

- Remover: links, código HTML, signos o caracteres especiales, etc.

Tokenización

- Dividir el texto en unidades más pequeñas: palabras (tokens) / oraciones (sentences).

Stop words

- Eliminar stop words usando listas predefinidas y/o extendidas.

Normalización

- Convertir a minúsculas.
- Corregir errores ortográficos.
- Stemming/lematización.

Multipalabras

- Identificar secuencias de palabras adyacentes en una oración, como los n-gramas.

Limpieza de texto

Es el primer paso en NLP. Su objetivo es eliminar ruido y preparar el texto para el análisis mediante técnicas como tokenización y eliminación de stopwords.
Dependiendo de la naturaleza y calidad del texto a analizar, podemos aplicar diferentes **técnicas** como:

- Eliminación de URLs.
- Eliminación de etiquetas HTML.
- Eliminación de signos de puntuación o caracteres especiales.
- Remoción de números (si es necesario).
- Tokenización.

Además, luego de realizar la **tokenización** a nivel de palabras, podríamos aplicar otras técnicas de limpieza como:

- Eliminación de palabras frecuentes (Removal of frequent words) o de palabras raras (Removal of rare words).
- Eliminación de emojis (removal of emojis) o emoticones (removal of emoticons).
- Conversión de emoticones a palabras (Conversion of Emoticon to Words) o de emojis a palabras (conversion of emoji to words).

Tokenización

Dividir el texto en unidades más pequeñas, como palabras, para facilitar su procesamiento.

¿A qué nivel es adecuado dividir (tokenizar) el texto?

La sede de UHEM es Quito

- ¿Caracteres?

['L', 'a',' _', 's', 'e', de de UHEM es Quito

- ¿Palabras?

['La', 'sede', 'de' UHEM es Quito

- ¿Sub-palabras?: ['El', 'día', 'está', 'nub', '#lado']
- ¿n-gramas de palabras?: ['El día', 'día está', 'está nublado'] (n=2)

Tokenización a nivel de palabras o sentencias

- En este paso, el texto se divide el texto en unidades más pequeñas. Podemos utilizar la tokenización de oraciones (*sentences*) o la tokenización de palabras según el planteamiento de nuestro problema.

```
import nltk
nltk.download('punkt') # Paquete que contiene los tokenizadores
```

- Tokenización a nivel de sentencia

- Tokenización a nivel de palabra

```
texto = """El ceviche de pescado es bueno.
Pero, el servicio es malo."""”
```

```
from nltk.tokenize import sent_tokenize

sentences = sent_tokenize(texto)
```

```
['El ceviche de pescado es bueno.',
'Pero, el servicio es malo.']
```

```
from nltk.tokenize import word_tokenize

tokens = word_tokenize(texto)
```

```
['El', 'ceviche', 'de', 'pescado', 'es', 'bueno',
'.', 'Pero', ',', 'el', 'servicio', 'es', 'malo', '.']
```

Stop words

Quitar palabras comunes (como "el", "la", "y") que no aportan información relevante.

La sede de UHEM es Quito

Tokenizar a nivel de sentencia

[“La”, “sede”, “de”, “UHEM”, “es”, “ Quito”]

Remover stop words

[“sede”, “UHEM”, “ Quito”]

Normalización de texto

- Se refiere al proceso de preparación de texto para su análisis o procesamiento. Involucra la eliminación de caracteres no alfanuméricos, como signos de puntuación, mayúsculas y minúsculas, tildes y acentos, y demás símbolos no relevantes.
- El **objetivo** es convertir el texto en una forma estándar, fácil de procesar y comparar. Esto ayuda a mejorar la calidad de los resultados de búsqueda, la clasificación de texto y otros algoritmos de NLP
- Tareas usuales:
  - Minúsculas (Lower casing)
  - Stemming y Lemmatization
  - Corrección ortográfica (si es relevante)
  - Normalización de espacios.

Comparativa de técnicas: ventajas y desventajas

Técnica	Ventajas	Desventajas
Tokenización	Divide el texto en unidades procesables	Puede perder información sobre relaciones entre palabras
Eliminación de Stopwords	Reduce dimensionalidad y ruido	Puede eliminar palabras importantes en ciertos contextos
Normalización	Mejora la consistencia y calidad de los datos	Requiere atención a errores y excepciones

Librerías de Python para preprocesamiento

NLTK (Natural Language Toolkit)	Amplio conjunto de herramientas para el procesamiento de lenguaje natural, incluyendo etiquetado PoS, WSD y NER.
spaCy	Biblioteca de alto rendimiento para NLP, con funcionalidades avanzadas para el análisis semántico de textos.

Ejemplos prácticos con NLTK

Tokenización con NLTK

Utilizar la función word_tokenize() para dividir el texto en palabras individuales.

Eliminación de Stopwords

Usar la lista de stopwords de NLTK y filtrar las palabras que coincidan.

Normalización

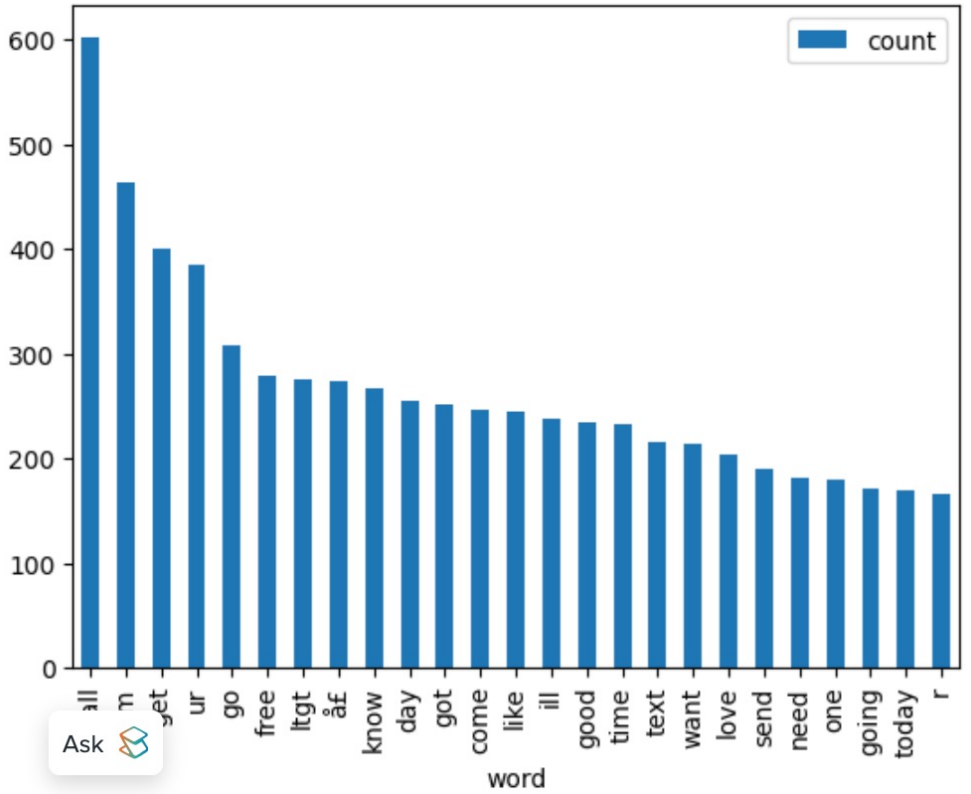
Aplicar funciones como lower() y SnowballStemmer() para convertir a minúsculas y lematizar.

Visualización

Generar nubes de palabras y gráficos de distribución de texto para explorar los datos.

Visualización: nube de palabras, distribución texto

Nube de Palabras
Visualización que muestra la frecuencia de palabras en un corpus de texto.
Distribución de Texto
Gráfico que ilustra la distribución de longitud de palabras, frases o documentos.



Gracias

uhemisferios uhe.oficial uhe_oficial Universidad Hemisferios

uhemisferios.edu.ec