



Taller introductorio:

Procesamiento de Lenguaje Natural

Prof: Erasmo Gómez





NLP Fundamental

S

Taller WO1



AGENDA



01

PROPEDÉUTICA DEL CURSO

Presentación del curso y
del docente

03

LINGÜÍSTICA COMPUTACIONAL

Conceptos y técnicas básicas

02

INTRODUCCIÓN AL CURSO

Conceptos previos necesarios

04

ENTORNO DE PROGRAMACIÓN

Demostración y utilización de la
herramienta principal del curso





ERASMO G. MONTOYA

Software Engineer @ Airnguru
Research Leader @ Chana-PUCP
hector.gomez@pucp.edu.pe
NLP Specialist



Ciao!



Hola



¿En qué se ha trabajado?

- Modelo traducción automática Shipibo Konibo - Español
- Correctores ortográficos: Yanesha, Yine, Shipibo, Ashaninka.
- Síntesis de voz Shipibo Konibo
- Síntesis de voz Shiwilu
- Chana: Plataforma para aprender lenguas amazónicas.
- Iskonawa: Modelos generativos para la preservación.
- etc.





01

SOBRE EL CURSO

Metodología de trabajo



Hallo



CONTENIDO DEL CURSO

INTRODUCCIÓN

A PROCESAMIENTO DE
LENGUAJE NATURAL



REPRESENTACIÓN

VECTORIAL DE TEXTOS



MODELOS

DE LENGUAJE



TAREAS BÁSICAS

DEL PROCESAMIENTO DE
LENGUAJE NATURAL



DEEP LEARNING

EN PROCESAMIENTO DE
LENGUAJE NATURAL



TAREAS AVANZADAS

DEL PROCESAMIENTO DE
LENGUAJE NATURAL



привіт

02

INTRODUCCIÓ N AL CURSO




Conceptos matemáticos previos (revisión)





¿Cuáles son los objetivos del curso?



- 
- Introducir herramientas computacionales para el análisis lingüístico.
 - Aplicar Python al procesamiento de lenguas amerindias.
 - Familiarizarse con modelos de lenguaje, traducción automática y análisis de texto.
 - Desarrollar un proyecto final enfocado en lenguas de escasos recursos digitales.
- 
- 

¿Cuáles son las metodologías aplicadas en el curso?

- Clases teóricas y prácticas con ejemplos reales.
- Uso de Jupyter Notebooks y Google Colab.
- Ejercicios guiados para análisis de corpus y modelado computacional.
- Proyecto final basado en análisis de lenguas amerindias.

привіт

03

LINGÜÍSTICA COMPUTACIONA

L

Introducción



LINGÜÍSTICA COMPUTACIONAL

¿QUÉ ES?

- En general, es una rama de estudio entre la **Lingüística y la Computación**, enfocado al estudio del lenguaje humano con un enfoque computacional.
- Integra **teoría y modelos** de las ramas de la lingüística, matemática, inteligencia artificial e ingeniería de software.

TAREAS DE LA LINGÜÍSTICA COMPUTACIONAL

- Corpus lingüístico asistido por ordenador.
- Diseño de analizadores sintácticos (en inglés: parser), para lenguajes naturales.
- Diseño de etiquetadores o lematizadores (en inglés: tagger), tales como el POS-tagger.
- Apoyo en la generación de asistentes automáticos (chatbots).
- Estudio de la posible relación entre lenguajes formales y naturales.
- Modelos de traducción automática.

Fuentes:

http://www.coli.uni-saarland.de/~hansu/what_is_cl.html

<https://www.britac.ac.uk/blog/what-computational-linguistics>

Irake!

DEFINICIÓN DE CONCEPTOS

NLP

¿QUÉ ES?

- Disciplina que permite a las máquinas comprender, interpretar y generar lenguaje humano.
- Tareas clave:
 - Tokenización y segmentación.
 - Análisis morfológico y sintáctico.
 - Modelos de lenguaje y embeddings.
 - Traducción automática y análisis de sentimientos.

Modelos de Lenguaje

¿QUÉ ES?

- Algoritmos que asignan probabilidades a secuencias de palabras.
- Tipos:
 - Ngramas: Modelos estadísticos que predicen palabras basadas en contexto anterior.
 - Modelos neuronales: Utilizan redes neuronales para predecir y generar texto.
 - Ejemplo: GPT, BERT, FastText para lenguas de escasos recursos.

Traducción Automática

¿QUÉ ES?

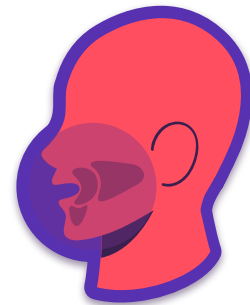
- Proceso de convertir texto de una lengua a otra usando modelos computacionales.
- Métodos:
 - Traducción basada en reglas.
 - Traducción estadística (SMT).
 - Traducción neuronal (NMT): basada en redes neuronales profundas.

Desafíos en lenguas amerindias: escasez de datos paralelos y baja representación digital.

Conversión de Voz a Texto (Speech To Text)

¿QUÉ ES?

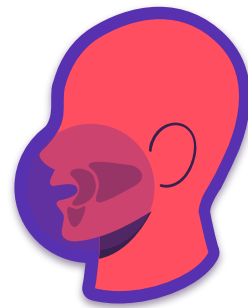
- Tecnología que convierte el habla humana en texto escrito.
- Etapas clave:
 - Extracción de características acústicas.
 - Decodificación y reconocimiento del lenguaje.
 - Ajuste de modelos para lenguas de escasos recursos.
- Aplicaciones: sistemas de reconocimiento de voz y transcripción automática.



Conversión de Texto a Voz (Text To Speech)

¿QUÉ ES?

- Tecnología que convierte texto escrito en habla sintetizada.
- Etapas clave:
 - Análisis del texto y normalización.
 - Conversión fonética (TexttoPhoneme).
 - Síntesis del habla usando modelos de voz.
- Aplicaciones:
 - Asistentes virtuales y lectores de pantalla.
 - Uso en lenguas amerindias para preservación y educación.



Fuentes:

http://www.coli.uni-saarland.de/~hansu/what_is_cl.html

<https://www.britac.ac.uk/blog/what-computational-linguistics>

CORPUS DE TEXTO

¿QUÉ ES?

- En Lingüística Computacional un corpus es una colección electrónica de texto, que se encuentra indexado de alguna manera y que incluye metadatos.
- Ejemplos:
 - Brown (1MM)
 - British National Corpus (100MM)
 - CORDE, CREA (250MM)
 - CODICACH (900MM)

Fuentes:

http://www.coli.uni-saarland.de/~hansu/what_is_cl.html

<https://www.britac.ac.uk/blog/what-computational-linguistics>

WORDNET

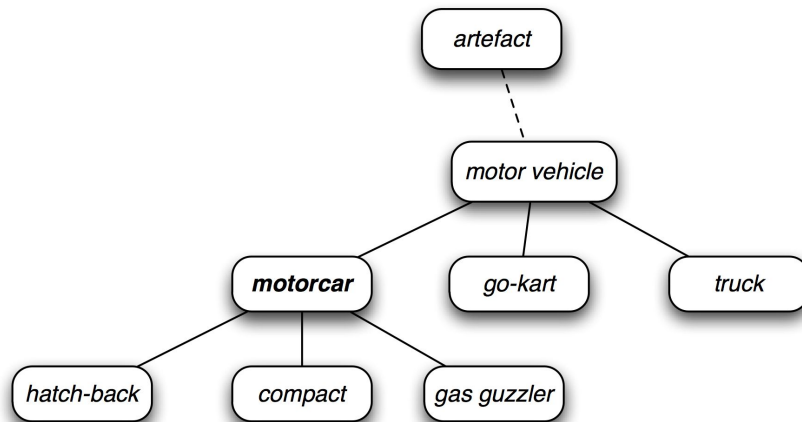
¿QUÉ ES?

- Base de datos léxico-conceptual en un idioma predeterminado estructurado en forma de red semántica, es decir, compuesta de unidades léxicas y relaciones entre ellas, que pretende ser un modelo del conocimiento léxico-conceptual de los hablantes de la lengua en cuestión.

- Ejemplo;

WordNet

WordNet Spanish



Fuentes:

http://www.coli.uni-saarland.de/~hansu/what_is_cl.html

<https://www.britac.ac.uk/blog/what-computational-linguistics>

TOKENIZACIÓN

¿QUÉ ES?

- La tokenización es un proceso utilizado en el campo del procesamiento del lenguaje natural (PLN) que consiste en dividir un texto en unidades más pequeñas llamadas “tokens”. Estos tokens pueden ser palabras individuales, frases, símbolos o cualquier otra unidad significativa para el análisis lingüístico.
- Ejemplo;
 - A nivel de palabras : Ese plato tiene buena pinta.
 - A nivel de sílabas: E se pla to tie ne bue na pin ta.
 - A nivel de stem: ese: ese plato: plat tiene: tien buena: buen pinta: pint
 - A nivel de lemma: ese: ese plato: plato tiene: tener buena: bueno pinta: pintar
 - A nivel de caracteres: E s e p l a t o t i e n e b u e n a p i n t a

Fuentes:

http://www.coli.uni-saarland.de/~hansu/what_is_cl.html

<https://www.britac.ac.uk/blog/what-computational-linguistics>

LENGUAJE FORMAL - REGEX

¿QUÉ ES?

- Las expresiones regulares, comúnmente abreviadas como "regex", son patrones de búsqueda utilizados para encontrar secuencias de caracteres dentro de cadenas de texto. Estos patrones pueden incluir caracteres literales (como letras o números), así como caracteres especiales que representan clases de caracteres, repeticiones, alternativas, entre otros.

SPECIAL			LITERAL		
.	^	\$	\.	\^	\\$
*	+	?	*	\+	\?
	[]	\\	\[\]
()	\	\\	\\	\\

Practica aqui!



Bibliografía

- Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing.*
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.*
- Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.*
- Sproat, R. (2003). *Computational Morphology: Finite-State Methods and Beyond.*

Bye!

Adieu

¡Gracias!