

2.1 Recursos para mi aprendizaje |

Corpus lingüístico



Contextualización

Con el fin de tener un conjunto de documentos estandarizado que pudieran ser usados para llevar a cabo los análisis y pruebas con datos de texto, la comunidad científica fue creando lo que se llaman “corpus lingüísticos”. Así, un corpus lingüístico es un conjunto de documentos de texto que pueden abarcar diferentes temáticas como notas periodísticas, novelas de diversos autores, documentos oficiales, etc. El poder trabajar con un conjunto de documentos estándar ayudará a los investigadores a comparar las técnicas y modelos de inteligencia artificial que se vayan proponiendo. Actualmente, existe una gran cantidad de ellos, agrupados ya no solo por temáticas de los documentos de texto, sino también con base a las técnicas y métodos de inteligencia artificial que se desean validar, como por ejemplo corpus lingüísticos para entrenar modelos que sean capaces de responder preguntas en un chatbot, o bien para generar un resumen de un documento, o bien para distinguir y clasificar un documento por el tipo de temática sobre la cual se está hablando, o bien para traducir un documento de un idioma a otro, etc.

En esta semana empezarás a trabajar con alguno de estos corpus lingüísticos y documentos en el proceso de manipulación y análisis de datos de texto. La manipulación de la información por caracteres, o palabras o enunciados dependerá del tipo de tarea que desees resolver, o del tipo de información que se desee extraer o generar.


Revisa a continuación el siguiente material de estudio, el cual te ayudará a realizar las actividades del curso y lograr los objetivos de aprendizaje 1.3 y 1.4.


Lectura

Lee las siguientes secciones de los **capítulos 2 y 3**:

- La **sección 1: Accessing Text Corpora**, del capítulo 2 “Accessing Text Corpora and Lexical Resources”.
- Las **secciones 3.1 a 3.4**, del capítulo 3: “Processing Raw Text”.

Del libro de texto:

Bird, S., Klein, E., y Loper, E. (s. f.). *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. <https://www.nltk.org/book> 
(<https://www.nltk.org/book>)

- Licencia: Este libro electrónico está disponible bajo los términos de la licencia [Creative Commons Attribution Noncommercial No-Derivative-Works 3.0 US](http://creativecommons.org/licenses/by-nc-nd/3.0/us)  (<http://creativecommons.org/licenses/by-nc-nd/3.0/us>).
- Accesibilidad: En el documento en HTML se puede ajustar el tamaño del texto con la herramienta de zoom del navegador web.

Puedes revisar la teoría también en la siguiente **presentación**:


Falcón Morales, L. E. (2023). *Tipos de datos en NLP* [PDF]. Maestría en Inteligencia Artificial Aplicada. ITESM.

[MNA_NLP_semana_02_corpus_texto_teoría.pdf](https://experiencia21.tec.mx/courses/575069/files/226236774?wrap=1)

(<https://experiencia21.tec.mx/courses/575069/files/226236774?wrap=1>) 

(https://experiencia21.tec.mx/courses/575069/files/226236774/download?download_frd=1)

Puedes ir estudiando el siguiente material adicional:

[Acceso al material \(https://experiencia21.tec.mx/courses/575069/files/226235331?wrap=1\)](https://experiencia21.tec.mx/courses/575069/files/226235331?wrap=1)  (https://experiencia21.tec.mx/courses/575069/files/226235331/download?download_frd=1)

- Licencia: D. R. Tecnológico de Monterrey, México, 2023. Prohibida la reproducción total o parcial de esta obra sin expresa autorización del Tecnológico de Monterrey.
- Accesibilidad: En el documento en PDF, se puede ajustar el tamaño de la página con el visor de archivos PDF.

Existe mucho material en la web sobre regex, en particular puedes consultar la siguiente tabla resumen: [python-regular-expressions-cheat-sheet.pdf](#)

(<https://experiencia21.tec.mx/courses/575069/files/226235231?wrap=1>)_ ↓

(https://experiencia21.tec.mx/courses/575069/files/226235231/download?download_frd=1)

Revisa el siguiente **Jupyter-Notebook** para repasar la parte aplicada de la teoría de esta semana:

- [MNA_NLP_semana_02_ejercicios_complementarios.ipynb](#)

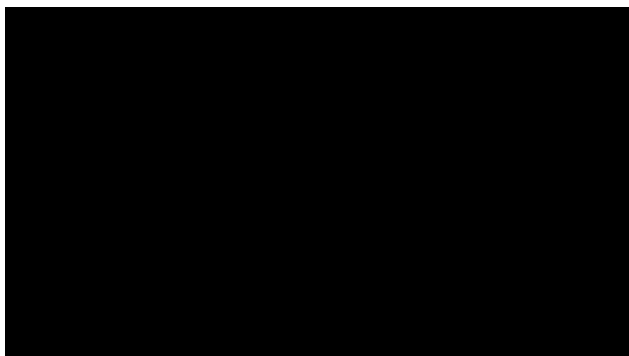
(<https://experiencia21.tec.mx/courses/575069/files/226236427?wrap=1>)_ ↓

(https://experiencia21.tec.mx/courses/575069/files/226236427/download?download_frd=1)



Videos Semana 2

Parte 1/2 - tipos de datos:



Parte 2/2 - regex:

