

4.1 Recursos para mi aprendizaje |

Matriz DTM, TF-IDF, diccionarios y etiquetado de palabras



Contextualización

Una vez tokenizado y estandarizado un documento de texto con las técnicas que has estudiado en las semanas previas, la vectorización de dichos documentos será parte esencial para el análisis y estudio posteriores.

Esto nos llevará a las matrices conocidas como Documento-Término, abreviada DTM, por sus siglas en inglés Document-Term-Matrix.

Igualmente estudiaremos la matriz conocida como TF-IDF que nos da información ponderada de los tokens más frecuentes en un corpus.

Un “diccionario” nos permitirá relacionar las palabras o tokens de un documento de texto mediante el formato “clave:valor” (“key:value” en inglés). Esta correspondencia nos permitirá asociar de manera única cada token de un documento de texto con un ID en un arreglo matricial, cuya información numérica estará asociada a la participación de cada token en dichos documentos. Dichas matrices serán a su vez la fuente de información para alimentar los modelos de aprendizaje automático o machine learning.

Finalmente, estudiarás el etiquetado gramatical o de palabras (POS-tag en inglés), técnica que asociará a cada token de un diccionario con su categoría gramatical, la cual permita obtener un mejor entendimiento semántico de un texto. Por ejemplo, será de gran utilidad el saber si una palabra es un verbo, o bien un sustantivo o un adjetivo, entre otros.

En esta cuarta semana, toma en cuenta que las técnicas de análisis y automatización de un documento de texto dependerán del idioma que se está analizando. Para cada idioma se deberá contar con la librería o paquete que lo soporte.

Revisa a continuación el siguiente material de estudio, el cual te ayudará a realizar las actividades del curso y lograr los objetivos de aprendizaje 1.7 y 1.8.





Lectura

Lee las siguientes secciones del **capítulo 3**:

- **3.4 Regular Expressions for Detecting Word Patterns.**
- **3.5 Useful Applications of Regular Expressions.**
- **3.6 Normalizing 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](https://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.


Lee también las siguientes **secciones** del **capítulo 2, NLP Pipeline**:

- **Pre-Processing**
 - **Preliminaries**
 - **Sentence segmentation**
 - **Word tokenization**
 - **Frequent Steps**
 - **Stemming and lemmatization**

Del libro del texto:

Vajjala, S., Majumder, B., Gupta, A., y Surana, H. (2020). *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly.
<https://learning.oreilly.com/library/view/practical-natural-language/9781492054047/>
 (<https://learning.oreilly.com/library/view/practical-natural-language/9781492054047/>)

- **Nota:** El URL de los ebooks que están en la base de datos de O'Reilly solo va a funcionar si se tiene una sesión activa en la plataforma. Es decir, primero deben entrar este link: <https://biblioteca.tec.mx/oreilly> (<https://biblioteca.tec.mx/oreilly>), autenticarse y posteriormente abrir los links directos a las lecturas.



- Licencia: Copyright © 2020 Anuj Gupta, Bodhisattwa Prasad Majumder, Sowmya Vajjala, y Harshit Surana. All rights reserved. Permisos del editor: se permite copiar/pegar.
- Accesibilidad: En la página de la obra dentro de la base de datos, se puede ajustar el tamaño de la letra, el color de fondo y el ancho del párrafo. Para mayor información, se puede consultar la página [Accessibility Guide](https://www.oreilly.com/online-learning/accessibility-guide.html)  (<https://www.oreilly.com/online-learning/accessibility-guide.html>), de O'Reilly (en inglés).

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

Falcón Morales, L. E. (2023). *matrices dtm y tfidf* [PDF]. Maestría en Inteligencia Artificial Aplicada. ITESM. [Acceso al material](#)

(<https://experiencia21.tec.mx/courses/575069/files/226235104?wrap=1>) 

(https://experiencia21.tec.mx/courses/575069/files/226235104/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.
- Puedes complementar la teoría de esta semana con el siguiente Jupyter Notebook:
[MNA_NLP_semana_04_ejercicios_complementarios.ipynb](#)
(<https://experiencia21.tec.mx/courses/575069/files/226236680?wrap=1>) 
(https://experiencia21.tec.mx/courses/575069/files/226236680/download?download_frd=1)
 - (Lectura Opcional) Si deseas profundizar en la teoría de las matrices dispersas (sparse matrices) y n-gramas, puedes consultar la siguiente presentación:
[MNA_NLP_semana_04_MatrizDispersa_Ngrama_teoría_Opcional.pdf](#)
(<https://experiencia21.tec.mx/courses/575069/files/226236685?wrap=1>) 
(https://experiencia21.tec.mx/courses/575069/files/226236685/download?download_frd=1)

El siguiente artículo es para irte preparando para los modelos llamados generadores de texto que estudiaremos en próximas semanas. Por el momento es opcional esta lectura.

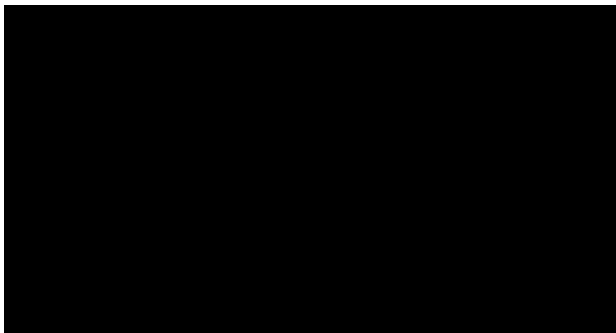
Wolfram, S. (2023, February 14). What Is ChatGPT Doing ... and Why Does It Work?

Stephen Wolfram Writings. <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/> 

[\(https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/\)](https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/)

- Licencia: © Stephen Wolfram, LLC. Except where otherwise noted, all materials, software, and information on the Site, as well as any unique visual or functional elements of the Site, are copyrighted and are protected by worldwide copyright laws and treaty provisions. They may not be copied, reproduced, modified, uploaded, posted, transmitted, or distributed in any way, in whole or in part, without WRI's prior written permission.
- Accesibilidad: En el documento en formato HTML, el tamaño del texto se puede ajustar con la herramienta zoom del navegador web.

Video semana 4 - parte 1/3 : Matrices DTM, TFIDF:



Video Semana 4 - parte 2/3 - JupyterNb_matrices dtm_tfidf_video:



Video Semana 4 -parte 3/3 - sparse matrix:

