

3.1 Recursos para mi aprendizaje |

Tokenización y pre-procesamiento de un texto



Contextualización

El proceso de tokenización nos permite determinar las unidades mínimas de información sobre las cuales un modelo de aprendizaje automático buscará extraer información que le permita entender los documentos. Es decir, el análisis morfológico, lexicográfico y semántico podrá empezar a automatizarse mediante las unidades de información mínimas, o tokens, que se estén empleando.

Sin embargo, sabemos tan solo una palabra puede generar una gran cantidad de pequeñas variaciones que incrementa la complejidad para entender el significado de un enunciado. Por ejemplo, la manera en que podemos conjugar un verbo regular o más aún un verbo irregular, es muy variable y eso incrementa en gran medida el proceso de análisis y automatización de los documentos. Esto nos llevará a técnicas que ayuden a simplificar los documentos a sus palabras raíces mediante técnicas como “stemming” y “lemmatization”.

Por otro lado, la frecuente aparición de una palabra o palabras en un documento puede ayudar a determinar el tipo de temática sobre la cual se está hablando y, por lo tanto, ayudar a clasificar documentos de acuerdo con su temática. Cuando en un documento, un conjunto de “n-palabras” aparecen usualmente juntas, es mejor considerarlas como “n-gramas”, es decir, asociarles un único identificador, independientemente de cada palabra por separado. Por ejemplo, “Tecnológico de Monterrey” se podría considerar como un “3-grama” o trigramas, en documentos donde se esté hablando de dicha institución y no cada una de estas tres palabras por separado.

Sin embargo, existe una gran cantidad de palabras de gran frecuencia que siempre aparecen en los documentos y que en general no proporcionar información relevante sobre el documento. Artículos, preposiciones, adverbios y conjunciones son este tipo de palabras, llamadas “stopwords”, que estarán apareciendo en cualquier tipo de documento, que tienen un gran porcentaje de participación, pero que no brindan información relevante sobre el documento en sí. Sin embargo, el gran porcentaje de frecuencia con el que aparecen en cualquier documento incrementa y dificulta en gran medida el proceso de análisis y automatización de un texto.

En esta tercera semana estudiarás algunas de estas técnicas que ayuden a continuar no solo simplificando un documento de texto, sino a empezar a prepararlo y extraer información de ellos generando el Vocabulario y Diccionario que nos proporcionen las palabras importantes de un documento, de acuerdo a su frecuencia de uso.

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.

Revisa a continuación el siguiente material de estudio, el cual te ayudará a realizar las actividades del curso y lograr los objetivos de aprendizaje 1.5 y 1.6.



Lectura

Lee el capítulo 1, Introduction to Regular Expressions, del libro de texto para seguir incrementando el conocimiento de regex:

Friedl, J. E. F. (2006). *Mastering Regular Expressions* (3.^a ed.). O'Reilly.

<https://learning.oreilly.com/library/view/mastering-regular-expressions/0596528124/>

 [\(https://learning.oreilly.com/library/view/mastering-regular-expressions/0596528124/\)](https://learning.oreilly.com/library/view/mastering-regular-expressions/0596528124/)


Lee el capítulo 2. NLP Pipeline, secciones 1 a 3 para conocer más acerca del procesamiento y limpieza de los datos:

Vajjala, S. (2020). *Practical Natural Language Processing*. O'Reilly.

<https://learning.oreilly.com/library/view/practical-natural-language/9781492054047/>

 [\(https://learning.oreilly.com/library/view/practical-natural-language/9781492054047/\)](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 © 2006, 2002, 1997 O'Reilly Media, Inc. 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) , de O'Reilly (en inglés).

Puedes revisar la teoría también en la siguiente presentación:

Falcón Morales, L. E. (2023). *Expresiones regulares (regex)* [PDF]. Maestría en Inteligencia Artificial Aplicada. ITESM. [Acceso al material](#)

<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](#)

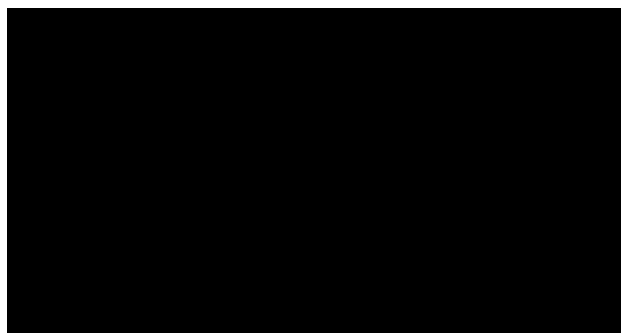
<https://experiencia21.tec.mx/courses/575069/files/226235231?wrap=1> 

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

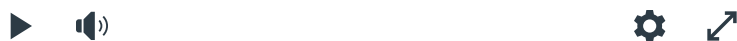
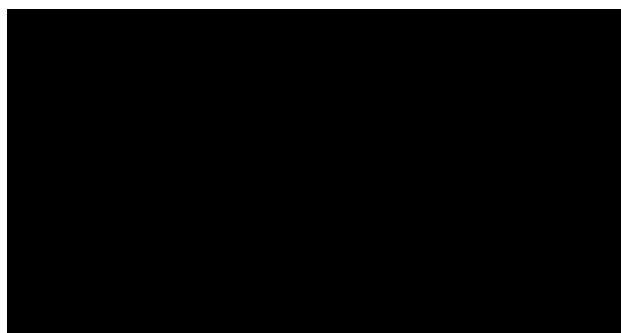
Igualmente revisa los siguientes Jupyter-Notebook para repasar la parte aplicada de la teoría de esta semana:

- [MNA_NLP_semana_03_Parte_1_ejercicios_complementarios.ipynb](#)
<https://experiencia21.tec.mx/courses/575069/files/226236590?wrap=1> 
https://experiencia21.tec.mx/courses/575069/files/226236590/download?download_frd=1
- [MNA_NLP_semana_03_Parte_2_ejercicios_complementarios.ipynb](#)
<https://experiencia21.tec.mx/courses/575069/files/226236594?wrap=1> 
https://experiencia21.tec.mx/courses/575069/files/226236594/download?download_frd=1

Video 1/3 - Diccionarios/Stemming/Lemmatization:



Video 2/3 - JupyterNotebook : Stemming/Lemmatization:



Video 3/3 - JupyterNotebook: Vectorización - matriz DTM

