Introducción al Módulo 1. Corpus y expresiones regulares

Contextualización

En este primer módulo sabemos que los modelos de aprendizaje automático, como las redes neuronales, requieren de información numérica para procesar la información de sus datos de entrada. El análisis de documentos de texto es uno de los tipos de datos no estructurados que mayor investigación ha generado en las últimas décadas y al mismo tiempo, es el tipo de problemas que mayores retos ha planteado a la comunidad de inteligencia artificial.

Los idiomas en general y la estructura gramatical de una lengua en particular, son por demás muy complejos y más aún cuando se desean encontrar y sintetizar en un conjunto de reglas lógicas la manera en que podamos comunicarnos con una computadora en lenguaje ordinario o natural.

Desde los inicios de la inteligencia artificial a finales de la década de los años 50, un conjunto de científicos, liderados por el informático estadounidense John McCarthy, se dio a la tarea de automatizar la manipulación y entendimiento del lenguaje natural por parte de una computadora. Conceptos como “tokenización” y “expresiones regulares”, entre muchos otros, ayudarán en este sentido y serán los temas de estudio de este módulo.

Las expresiones regulares son una herramienta poderosa que se utiliza ampliamente en la programación y la informática en general. En este módulo aprenderás a utilizar expresiones regulares para buscar y manipular cadenas de texto en diferentes idiomas y contextos.

Por su parte, el proceso de tokenización te permitirá construir las unidades mínimas de información con las cuales llevarás a cabo la manipulación de los enunciados y documentos de texto. Y más importante aún, te llevará a construir un “diccionario” mediante el cual se transforma cada palabra o token en un vector numérico de información, que serán posteriormente los vectores de entrada de los algoritmos de aprendizaje automático o machine learning.

Plan del módulo

En la siguiente tabla, encuentra lo que aprenderás en este módulo y los medios para lograrlo.

Objetivo general de aprendizaje	Objetivo específico de aprendizaje	Tema	Materiales didácticos
Utilizar técnicas para la clasificación de textos.	1.1 Describir la evolución del análisis del lenguaje escrito. 1.2 Identificar la estructura y sintaxis de un lenguaje.	1.1 Antecedentes. 1.2 Estructura del lenguaje.	Consultar en **1.1 Recursos para mi apre Antecedentes y estructura del lenguaje (https://experiencia21.tec.mx/courses/5750 dot-1-recursos-para-mi-aprendizaje-%7C- antecedentes-y-estructura-del-lenguaje)** .
	1.3 Diferenciar la importancia de las palabras, las frases y la gramática en un texto. 1.4 Aplicar las representaciones semánticas en un texto.	1.3 Corpus lingüístico (text corpus)	Consultar en **2.1 Recursos para mi apre Corpus lingüístico (https://experiencia21.tec.mx/courses/5750 dot-1-recursos-para-mi-aprendizaje-%7C-co linguistico)**
	1.5 Ejemplificar los conceptos de tokenización. 1.6 Explicar los conceptos que involucran el pre-procesamiento de un texto: limpieza de un texto, identificación de las palabras de	1.4 Tokenización y normalización (stemming/lemmatization) 1.5 Expresiones regulares (regex) 1.6 N-grams y stopwords.	Consultar en **3.1 Recursos para mi apre Tokenización y pre-procesamiento de u (https://experiencia21.tec.mx/courses/5750 dot-1-recursos-para-mi-aprendizaje-%7C- tokenizacion-y-pre-procesamiento-de-un-te**

	enlace o stopwords, aplicar criterios de normalización.		
	1.7 Reconocer la sintaxis y estructura de un texto. 1.8 Construir los conceptos para la detección y separación de palabras.	1.7 Creación de diccionario 1.8 Etiquetado de palabras (POS-tag)s	Consultar en **4.1 Recursos para mi apre Creación de diccionario y etiquetado d palabras (https://experiencia21.tec.mx/courses/5750 dot-1-recursos-para-mi-aprendizaje-%7C-cr diccionario-y-etiquetado-de-palabras)**