Introducción al módulo 2. Procesamiento y modelado

Contextualización

En el módulo anterior obtuvimos diversas formas llevar a cabo una tokenización de las unidades mínimas de información (caracteres, palabras o enunciados) de un documento. Con dicha tokenización obtendremos diversas representaciones matriciales de cada documento, además de la construcción de los diccionarios (usualmente de palabras, pero puede incluir también caracteres ortográficos o caracteres especiales).

La bolsa de palabras, *bag-of-words* (BOW) en inglés, nos permite tener una primera representación para entrenar modelos y llevar a cabo técnicas de aprendizaje supervisado, como las de análisis de sentimiento y de aprendizaje no supervisado, como el modelado de temas (*topic modeling* en inglés).

A medida que crece la cantidad de documentos a analizar, las matrices asociadas empiezan a tener cientos de miles o millones de componentes, ya que para documentos muy grandes el diccionario generado llega a ser de decenas de miles o cientos de miles de palabras. Sin embargo, generalmente la información no cero llega a constituir un pequeño porcentaje de la matriz, lo cual nos llevará a la utilización de las llamadas matrices dispersas (*sparse matrices*, en inglés), un formato que permite guardar solamente la información no cero de una matriz.

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

• Aplicar herramientas computacionales y librerías adecuadas para el análisis de textos. • Evaluar los modelos de procesamiento de lenguaje natural.	2.1 Contrastar los conceptos que llevan a la automatización en el análisis de un texto. 2.2 Ilustrar las características principales de un texto. 2.3 Producir modelos de bolsa-de-palabras. 2.4 Examinar documentos mediante de cúmulos de palabras.	2.1 Bolsa de palabras y nube de palabras.	Consultar en **5.1 Recursos para mi aprendizaje Bolsa de palabras y nube de palabras (https://experiencia21.tec.mx/courses/575069/pages/dot-1-recursos-para-mi-aprendizaje-%7C-bolsa-de-palabras-y-nube-de-palabras)**
	2.5 Programar diferentes métodos para la normalización de un texto. 2.6 Proponer diversos modelos TF-IDF.	2.2 Matriz documento-término. 2.3 TF-IDF.	Consultar en **6.1 Recursos para mi aprendizaje Document Matrix (https://experiencia21.tec.mx/courses/575069/pages/dot-1-recursos-para-mi-aprendizaje-%7C-documen matrix)**
	2.7 Comparar los diferentes tipos de casos de matrices dispersas.	2.4 Matrices dispersas (CSR, CSC, COO).	Consultar en **7.1 Recursos para mi aprendizaje Modelos de clasificación I (https://experiencia21.tec.mx/courses/575069/pages/dot-1-recursos-para-mi-aprendizaje-%7C-modelos-clasificacion-i)**

	2.8 Argumentar las problemáticas de similaridad de documentos y palabras. 2.9 Experimentar con las principales métricas de similaridad.	2.5 Modelos de clasificación de documentos.	
	2.10 Argumentar los diferentes métodos del modelado de tópicos o temas. 2.11 Experimentar con el método de indexado semántico latente (LSI). 2.12 Resolver problemas con el método de asignación latente de Dirichlet (LDA).	2.6 Latent semantic indexing (Indexación semántica latente). 2.7 Latent Dirichlet allocation (Asignación latente de Dirichlet).	Consultar en **8.1 Recursos para mi aprendizaje Modelos de clasificación II (https://experiencia21.tec.mx/courses/575069/pages/dot-1-recursos-para-mi-aprendizaje-%7C-modelos-clasificacion-ii)**