Maestría en Inteligencia Artificial Aplicada

Proyecto Integrador

análisis exploratorio de datos

Alumno	Matrícula
Johanna Rodríguez Jaramillo	A01794010
Katherine Vanegas Salamanca	A01794113
Marcos Chávez Chávez	A01688507

Profesores Titulares

Dra. Grettel Barceló Alonso / Luis Eduardo Falcón Morales

Profesor Asistente

Horacio Martínez Alfaro

5 de mayo de 2024

Análisis exploratorio de datos

Contexto

Basados en el reto de Viakable se cuenta con aproximadamente 150 documentos sobre las normas para la fabricación de cables, por lo que es crucial identificar los patrones y estructuras en los documentos, especialmente con los detalles técnicos como títulos, listas y tablas.

Para el entrenamiento de modelo, consideraremos los siguientes enfoques:

Segmentación de Documentos: Mediante herramientas que reconozcan y separen diferentes secciones (títulos, listas, párrafos). Esto ayudará a identificar la estructura organizativa de cada documento.

Extracción de Entidades: Con el fin de extraer términos técnicos relevantes, pasos y regulaciones. Esto permitirá que el modelo reconozca y relacione requisitos específicos.

Análisis de Tablas: Se buscará extraer datos tabulados para comprender especificaciones o requisitos numéricos.

Clasificación de Texto: Clasificar las distintas partes en categorías relevantes como información de seguridad, necesidades de calidad y pasos de fabricación.

Resúmenes: Resumir las secciones individuales para tener una visión general, muy útil en documentos más largos.

Diccionario de Datos

UEN: Esta columna representa la Unidad Estratégica de Negocio, en este caso Industria.
Norma: El nombre o código que identifica el estándar, regulación o directriz de la industria. Ejemplo: "NMX-B-172-CANACERO-2018."
Fecha Actualización: El año en que se actualizó o revisó por última vez el estándar. Ejemplo: "2023."
Idioma: El idioma del documento del estándar. Todos parecen estar en "ESPAÑOL."
Nativo: Parece indicar si el documento está en su idioma original o nativo (aparece como "1" en todos los casos).
Megabytes: El tamaño del documento en megabytes.
Páginas: El número de páginas en el documento.
Letras: Número total de caracteres en el documento.
Palabras: La cantidad aproximada de palabras en el documento.
Imágenes: La cantidad de imágenes o elementos gráficos en el documento.

Observaciones Generales

Tamaños de Documentos: Los tamaños de los documentos varían considerablemente, desde tan solo 0.09 MB hasta más de 23.28 MB.

Páginas: Algunos documentos son muy cortos (5 páginas), mientras que otros son bastante completos (153 páginas).

Imágenes: El contenido gráfico varía, algunos documentos tienen solo una imagen, mientras que otros incluyen más de 150.

Características del Contenido: La cantidad de palabras varía desde poco más de mil hasta más de 100,000 palabras, lo que indica una diversidad en la longitud y profundidad de los documentos.

Análisis de las normas

UEN	Norma	Fecha actual	Idioma	Nativo	Megabytes	Páginas	Letras	Palabras	Imágenes
Industria	NMX-B-172-CANACERO-2018	2023	ESPAÑOL	1	1,638	62	110.112	20.088	59
Industria	NMX-CC-19011-IMNC-2012 DIRECTRICES PARA LA A	2021	ESPAÑOL	1	1,021	66	181.764	33.396	2
Industria	NMX-CH-036-SCFI-1994 INDICADORES DE CARATUL	2007	ESPAÑOL	1	0,257	13	17.556	4.264	8
Industria	NMX-CH-070-SCFI-1993 TERMOMETROS BIMETALIC	2007	ESPAÑOL	1	0,115	9	11.817	2.466	2
Industria	NMX-CH-099-IMNC-2005 MICROMETROS PARA ME	2008	ESPAÑOL	1	0,923	22	40.370	7.964	14
Industria	NMX-CH-110-1-SCFI-1993 INSTRUMENTOS DE MED	2007	ESPAÑOL	1	0,145	22	38.324	7.304	2
Industria	NMX-CH-115-1-SCFI-1993 ALTA TENSION - SISTEMA	2008	ESPAÑOL	1	0,144	17	31.824	5.696	2
Industria	NMX-CH-115-2-SCFI-1994	2008	ESPAÑOL	1	0,275	49	105.693	19.502	18
Industria	NMX-CH-131-1-SCFI-1993	2007	ESPAÑOL	1	0,13	20	31.340	6.520	2
Industria	NMX-CH-7500-1-IMNC-2006	2008	ESPAÑOL	1	0,464	28	32.872	6.300	13
Industria	NMX-EE-161-SCFI-1983	2006	ESPAÑOL	1	0,173	16	12.090	2.223	16
Industria	NMX-H-013-SCFI-1984	2005	ESPAÑOL	1	0,127	8	16.088	3.432	2
Industria	NMX-H-014-CANACERO-2021	2023	ESPAÑOL	1	0,477	17	43.707	9.197	4
Industria	NMX-H-014-SCFI-1984	2005	ESPAÑOL	1	0,119	11	18.931	3.949	2
Industria	NMX-J-002-ANCE-2018	2019	ESPAÑOL	1	6,337	12	25.884	5.304	4
Industria	NMX-J-010-1-ANCE-2018	2020	ESPAÑOL	1	1,036	19	41.648	6.745	23
Industria	NMX-J-010-ANCE-2018 Vigor AGO2019	2019	ESPAÑOL	1	57,521	110	283.470	64.460	13
Industria	NMX-J-012-ANCE-2019 Vigor 06OCT2019	2019	ESPAÑOL	1	1,478	22	51.128	10.252	26
Industria	NMX-J-014-ANCE-2016	2019	ESPAÑOL	1	3,012	16	32.992	6.656	19
Industria	NMX-J-027-ANCE-2020 Vigor 04JUL2021	2021	ESPAÑOL	1	1,913	11	19.492	3.905	13
Industria	NMX-J-030-ANCE-2021 Vigor 21DIC2022	2022	ESPAÑOL	1	1,295	23	50.899	10.442	32
Industria	NMX-J-032-ANCE-2022	2022	ESPAÑOL	1	0,885	26	66.144	13.468	29
Industria	NMX-J-035-ANCE-2018 Vigor 19MAR2020	2019	ESPAÑOL	1	1,702	12	24.132	5.172	14
Industria	NMX-J-036-ANCE-2018	2019	ESPAÑOL	1	0,973	12	18.552	3.060	17
Industria	NMX-J-037-ANCE-2019 Vigor 04JUL2021	2021	ESPAÑOL	1	1,307	14	26.138	5.292	16
Industria	NMX-J-040-ANCE-2020 Vigor 21DIC2022	2022	ESPAÑOL	1	0,361	11	15.136	3.311	3
Industria	NMX-J-041-SCFI-1965	2006	ESPAÑOL	1	0,174	12	8.616	1.800	5
Industria	NMX-J-043-ANCE-2015	2017	ESPAÑOL	1	5,884	13	20.085	3.757	3
Industria	NMX-J-054-ANCE-2015	2017	ESPAÑOL	1	5,176	12	17.064	2.892	3
Industria	NMX-J-058-ANCE-2019 Vigor 04JUL2021	2021	ESPAÑOL	1	1,514	20	23.860	4.740	22
Industria	NMX-J-059-1-ANCE-2020 Vigor 01NOV2021	2021	ESPAÑOL	1	0,765	14	15.638	3.514	17
Industria	NMX-J-061-ANCE-2015	2017	ESPAÑOL	1	7,86	16	27.510	3.675	3
Industria	NMX-J-062-ANCE-2014	2017	ESPAÑOL	1	7,825	18	35.456	7.305	5
Industria	NMX-J-066-ANCE-2017 Vigor MAY2018	2017	ESPAÑOL	1	5,284	14	15.078	2.996	5
Industria	NMX-J-091-ANCE-1982	2005	ESPAÑOL	1	0,156	14	11.898	2.151	7
Industria	NMX-J-102-ANCE-2015	2017	ESPAÑOL	1	5,843	13	13.328	2.832	4
Industria	NMX-J-120-SCFI-1970	2006	ESPAÑOL	1	0,104	7	8.280	1.775	2
Industria	NMX-J-129-ANCE-2019 Vigor ABR2020	2020	ESPAÑOL	1	1,415	12	13.632	3.252	18
Industria	NMX-J-142-1-ANCE-2019 Vigor 01OCT2019	2019	ESPAÑOL	1	1,69	63	105.210	15.314	75
Industria	NMX-J-177-ANCE-2018 Vigor 30ENE2020	2019	ESPAÑOL	1	7,827	18	11.865	2.373	14
Industria	NMX-J-178-ANCE-2020 Vigor 26AGO2022	2021	ESPAÑOL	1	1,002	19	29.520	7.452	27
Industria	NMX-J-180-ANCE-2011	2011	ESPAÑOL	1	0,217	8	9.093	1.771	1
Industria	NMX-J-183-ANCE-2021 Vigor 10ENE2023	2022	ESPAÑOL	1	0,632	10	15.993	3.060	12
Industria	NMX-J-184-ANCE-2021 Vigor 11FEB2023	2022	ESPAÑOL	1	0,66	10	13.851	2.259	13
Industria	NMX-J-186-ANCE-2018	2019	ESPAÑOL	1	1,773	12	24.804	4.980	15
Industria	NMX-J-190-ANCE-2018 Vigor ABR2019	2018	ESPAÑOL	1	6,098	12	14.710	2.680	2
Industria	NMX-J-191-ANCE-2018 Vigor 06AGO2020	2019	ESPAÑOL	1	1,369	13	28.418	5.916	16
Industria	NMX-J-193-ANCE-2020 Vigor 06ENE2022	2021	ESPAÑOL	1	1,22	12	13.530	2.761	14
Industria	NMX-J-194-ANCE-2022 Vigor 04AGO2023	2022	ESPAÑOL	1	0,768	14	22.316	3.626	16
Industria	NMX-J-200-ANCE-2021 Vigor 11FEB2023	2022	ESPAÑOL	1	0,849	20	28.740	5.600	23
Industria	NMX-J-204-ANCE-2021 Vigor 11FEB2023	2022	ESPAÑOL	1	0,907	13	17.016	3.096	17
Industria	NMX-J-205-ANCE-2020 Vigor 21DIC2022	2022	ESPAÑOL	1	0,762	14	15.134	2.730	20
Industria	NMX-J-212-ANCE-2017 Vigor 15SEP2018	2018	ESPAÑOL	1	8,719	18	19.242	3.960	2
Industria	NMX-J-215-ANCE-2019 Vigor 05OCT2019	2019	ESPAÑOL	1	1,47	18	22.860	3.924	24
Industria	NMX-J-216-ANCE-2018 Vigor 06OCT2019	2019	ESPAÑOL	1	0,711	11	13.960	2.740	3
Industria	NMX-J-218-ANCE-2019	2020	ESPAÑOL	1	0,822	11	11.363	2.046	3
Industria	NMX-J-271_1-ANCE-2007	2010	ESPAÑOL	1	2,476	74	163.984	31.755	40
Industria	NMX-J-271_2-ANCE-2002	2012	ESPAÑOL	1	3,799	86	165.155	31.110	15
Industria	NMX-J-271_3-ANCE-2009	2012	ESPAÑOL	1	2,563	46	75.825	14.715	14
Industria	NMX-J-292-ANCE-2021 Vigor 10AGO2022	2022	ESPAÑOL	1	0,915	23	33.580	4.758	29
Industria	NMX-J-297-ANCE-2017	2017	ESPAÑOL	1	5,597	12	14.750	2.330	4
Industria	NMX-J-298-ANCE-2018	2020	ESPAÑOL	1	1,15	13	24660	3480	16
Industria	NMX-J-299-SCFI-1993	2005	ESPAÑOL	1	0,135	14	19.530	3.133	2
Industria	NMX-J-300-ANCE-2020 Vigor 08MAR2022	2021	ESPAÑOL	1	1,122	19	52.751	11.832	33
Industria	NMX-J-312-ANCE-2017	2018	ESPAÑOL	1	5,309	10	9.816	1.792	3
Industria	NMX-J-417-ANCE-2021 Vigor 22JUN2022	2022	ESPAÑOL	1	0,737	26	39.250	7.625	33
Industria	NMX-J-426-ANCE-2013	2014	ESPAÑOL	1	6,093	16	15.720	2.955	12
Industria	NMX-J-429-ANCE-2009	2010	ESPAÑOL	1	0,311	12	17.743	2.816	1
Industria	NMX-J-431-ANCE-2011	2011	ESPAÑOL	1	0,264	10	13.518	2.736	1
Industria	NMX-J-432-ANCE-2021 Vigor 11FEB2023	2022	ESPAÑOL	1	0,694	14	15.392	3.575	18
Industria	NMX-J-436-ANCE-2021 Vigor 10JUL2022	2022	ESPAÑOL	1	2,316	123	292.556	62.098	147
Industria	NMX-J-437-ANCE-2017 Vigor 15SEP2018	2018	ESPAÑOL	1	6,887	16	28.485	5.565	12
Industria	NMX-J-438-ANCE-2020 Vigor 26FEB2022	2021	ESPAÑOL	1	0,718	14	30.290	5.200	17
Industria	NMX-J-441-ANCE-2021 Vigor 11FEB2023	2022	ESPAÑOL	1	0,695	13	19.440	3.504	19
Industria	NMX-J-442-ANCE-2021 Vigor 04AGO2023	2022	ESPAÑOL	1	0,92	12	18.612	3.360	15
Industria	NMX-J-443-ANCE-2021 Vigor 10SEP2022	2022	ESPAÑOL	1	0,707	10	13.680	2.655	13
Industria	NMX-J-451-ANCE-2021 Vigor 22JUN2022	2022	ESPAÑOL	1	2,34	153	476.595	102.204	164
Industria	NMX-J-472-ANCE-2019	2019	ESPAÑOL	1	1,597	24	48.554	10.758	34
Industria	NMX-J-473-ANCE-2020 Vigor 22DIC2022	2022	ESPAÑOL	1	0,701	12	21.032	4.455	15
Industria	NMX-J-474-ANCE-2017	2018	ESPAÑOL	1	8,597	16	21.296	3.984	3
Industria	NMX-J-486-ANCE-2020 Vigor 08MAR2022	2021	ESPAÑOL	1	0,969	33	63.888	14.058	36
Industria	NMX-J-498-ANCE-2011	2012	ESPAÑOL	1	0,62	19	36.955	7.828	14
Industria	NMX-J-509-ANCE-2019	2021	ESPAÑOL	1	1,206	11	16.335	2.662	14
Industria	NMX-J-514-ANCE-2016	2019	ESPAÑOL	1	0,976	16	29.088	4.832	19
Industria	NMX-J-516-ANCE-2021 Vigor 05AGO2023	2022	ESPAÑOL	1	0,632	11	17.226	3.421	4
Industria	NMX-J-522-ANCE-2021 Vigor 05AGO2023	2022	ESPAÑOL	1	0,728	15	20.775	3.885	4
Industria	NMX-J-532-ANCE-2017	2018	ESPAÑOL	1	6,354	12	16272	2736	2
Industria	NMX-J-539-ANCE-2021	2022	ESPAÑOL	1	1,042	19	51.604	8.607	24
Industria	NMX-J-553-ANCE-2021 Vigor 22DIC2022	2022	ESPAÑOL	1	0,843	26	65.468	13.182	31
Industria	NMX-J-555-ANCE-2019	2020	ESPAÑOL	1	1,684	13	29.991	4.979	16
Industria	NMX-J-556-ANCE-2021 Vigor 14FEB2023	2022	ESPAÑOL	1	2,219	117	185.328	43.290	153
Industria	NMX-J-634-ANCE-2010	2011	ESPAÑOL	1	6,149	144	398.016	78.336	6
Industria	NMX-J-647-ANCE-2020	2021	ESPAÑOL	1	1,004	20	27.060	5.400	31
Industria	NMX-J-685-ANCE-2014	2018	ESPAÑOL	1	23,279	52	110.656	23.036	16
Industria	NMX-J-686-ANCE-2020	2021	ESPAÑOL	1	2,591	76	160.892	32.148	133
Industria	NMX-J-694-ANCE-2018	2020	ESPAÑOL	1	1,108	22	24.838	5.104	28
Industria	NMX-J-726-ANCE-2020 Vigor 25FEB2022	2021	ESPAÑOL	1	1,702	75	177.750	36.300	88
Industria	NMX-J-733-ANCE-2020	2021	ESPAÑOL	1	1,705	97	211.751	36.569	103
Industria	NMX-J-761-ANCE-2019	2020	ESPAÑOL	1	1,474	31	84.692	13.609	34
Industria	NMX-J-762 -ANCE-2020 VigorFEB2022	2021	ESPAÑOL	1	1,807	13	21.619	4.173	16
Industria	NMX-J-SAA-50001-ANCE-IMNC-2019	2021	ESPAÑOL	1	2,127	47	97.055	17.390	6
Industria	NMX-J-SAST-55001-ANCE-IMNC-2015	2021	ESPAÑOL	1	1	28	50.484	9.716	4
Industria	NMX-T-009-SCFI-1970	2009	ESPAÑOL	1	0,131	5	7.520	1.604	3
Industria	NMX-T-039-SCFI-1979	2006	ESPAÑOL	1	0,11	8	10.647	2.163	3
Industria	NMX-W-037-SCFI-1982	2009	ESPAÑOL	1	0,118	8	4.488	800	3
Industria	NMX-W-099-SCFI-1981	2009	ESPAÑOL	1	0,092	8	5.192	964	3
Industria	NMX-Z-012-1-SCFI-1987	2005	ESPAÑOL	1	0,744	89	179.568	34.104	22
Industria	NMX-Z-012-2-SCFI-1987	2005	ESPAÑOL	1	2,11	81	139.280	28.560	91
Industria	NMX-Z-012-3-SCFI-1987	2005	ESPAÑOL	1	0,211	19	36.309	8.151	5

Observaciones Generales

Tamaños de Documentos: Los tamaños de los documentos varían considerablemente, desde tan solo 0.09 MB hasta más de 23.28 MB.

Páginas: Algunos documentos son muy cortos (5 páginas), mientras que otros son bastante completos (153 páginas).

Imágenes: El contenido gráfico varía, algunos documentos tienen solo una imagen, mientras que otros incluyen más de 150.

Características del Contenido: La cantidad de palabras varía desde poco más de mil hasta más de 100,000 palabras, lo que indica una diversidad en la longitud y profundidad de los documentos.

Análisis de los datos

Con el fin de realizar un acercamiento al preprocesamiento de los datos, realizamos una prueba aleatoria, para lo cual se definió un archivo de prueba procesado en Colab, con el fin de cumplir las siguientes validaciones.

Normas seleccionadas

- NMX-CC-19011-IMNC-2012 DIRECTRICES PARA LA AUDITORIA DE LOS SISTEMAS DE GESTION

- NMX-CH-036-SCFI-1994 INDICADORES DE CARATULA (con imágenes)

Proceso aplicado

Extracción de Texto desde un Archivo PDF:
Herramienta Usada: PyPDF2.
Proceso: Se abre el archivo PDF en modo de lectura binaria y, utilizando PdfReader, se extraen todas las páginas para obtener su texto.
Resultado: Se obtiene una cadena de texto con todo el contenido del archivo PDF.

Limpieza y Tokenización de Texto:
Stopwords: Se descargan las stopwords en español usando nltk para eliminarlas posteriormente.

Tokenización:
Herramienta Usada: nltk.word_tokenize.
Proceso: Se convierte el texto a tokens o palabras individuales.

Lematización:
Lematizador Usado: WordNetLemmatizer de nltk.
Proceso: Cada palabra se lematiza en función de su categoría gramatical:
Verbos: Lematización de verbos (por ejemplo, "hablar" en vez de "hablamos").

Pronombres: Lematización de sustantivos.
Adverbios: Lematización de adverbios.
Resultado: Las palabras se reducen a su forma base, facilitando el análisis posterior.

Eliminación de Duplicados: Se elimina cualquier palabra duplicada.
Filtrado por Longitud: Se eliminan tokens cuya longitud es menor a 1 caracteres.

Algunos hallazgos con respecto a las pruebas que se realizaron de relacionan a continuación:

- Con respecto a la norma NMX-CC-19011-IMNC-2012 DIRECTRICES PARA LA AUDITORIA DE LOS SISTEMAS DE GESTION la cual cuenta con un total de 33.396 palabras, con el proceso de limpieza se logra reducir a 2591 palabras
- La norma NMX-CH-036-SCFI-1994 INDICADORES DE CARATULA cuenta con 4.264, con la aplicación de la limpieza de los datos se reduce a 643 palabras

Dentro de la limpieza se optó por no borrar los números de las muestras ya que los mismos pueden denotar medidas de cables, intervalos que son importantes para el tipo de información que estamos manejando. Por otro lado, las imagenes que se encuentran dentro del documento no son tenidas en cuenta dentro de los datos preprocesados ya su información también se encuentra contenida en el texto.

Los resultados de esta prueba están publicados en el repositorio de Github https://github.com/Katty62870/Equipo-34

Conclusiones

El método utilizado emplea técnicas normales de procesamiento de lenguaje para extraer, limpiar y simplificar el contenido de un archivo PDF. Esto permite obtener una versión procesada del texto, más limpia y representativa, para un análisis y modelado posterior.

Bibliografía

Visengeriyeva, L., Kammer, A., Bär, I., Kniesz, A., y Plöd, M. (2023). CRISP-ML(Q). The ML Lifecycle Process. MLOps. INNOQ.

Kumar Mukhiya, S., y Ahmed, U. (2020). Hands-On Exploratory Data Analysis with Python. Packt Publishing.