1. Datos Generales de la asignatura

Nombre de la asignatura:	Procesamiento de lenguaje natural
Clave de la asignatura:	CDF-2419
SATCA[1]:	3-2-5
Carrera:	Ingeniería en Ciencia de Datos

2. Presentación

Caracterización de la asignatura
La asignatura le permitirá descubrir al alumno los aspectos relevantes de textos con base a la normalización, procesamiento semántico, clasificación y agrupamiento, el análisis de personalidad, el análisis de opinión y la relación con otros tipos de datos, además de la extracción de información.
Intención didáctica
La intención didáctica de esta asignatura es proporcionar a los estudiantes una comprensión sólida para el diseño, creación y modelado semántico y los tipos de datos aplicando metodologías de modelado de datos, clasificación entre otros. Asimismo, fomenta el trabajo en equipo, la comunicación efectiva, solución de problemas, creatividad e ingenio con un alto sentido ético. Esta asignatura se relaciona de manera antecedente con Aprendizaje Automático, Adquisición de Datos y Programación Avanzada para la Ciencia de Datos.

3. Participantes en el diseño y seguimiento curricular del programa

Lugar y fecha de elaboración o revisión	Participantes	Observaciones
Instituto Tecnológico Superior de Alvarado del 21 al 23 agosto de 2023.	Representante del Instituto Tecnológico Superior de Alvarado.	Propuesta inicial.
Tecnológico Nacional de México 30 octubre 2023	Representante del Instituto Tecnológico de: Querétaro y del Instituto Tecnológico Superior de Alvarado.	Presentación de la propuesta de la carrera de Ingeniería en Ciencia de Datos.

[1] Sistema de Asignación y Transferencia de Créditos Académicos

Instituto Tecnológico de Querétaro Campus Norte del 19 al 22 de marzo 2024.	Representantes de los Institutos Tecnológicos de: Morelia, Puebla, Querétaro, Tehuacán. Instituto Tecnológico Superior de Alvarado. CENIDET. Representante de Ciencias Básica de los Institutos de: Celaya, Morelia y CIIDET.	Diseño y/o desarrollo curricular de la carrera de Ingeniería en Ciencia de Datos.
Tecnológico Nacional de México del 22 al 24 de abril del 2024	Representante del Instituto Tecnológico de Querétaro e Instituto Tecnológico Superior de Alvarado.	Contraste y ajuste de las asignaturas de Ingeniería en Ciencia de Datos con respecto a las de Ing. en Inteligencia Artificial, Ing. en Desarrollo WEB e Ing. en Ciberseguridad
Tecnológico Nacional de México del 27 al 31 de mayo del 2024.	Representantes de los Institutos Tecnológicos de: Morelia, Querétaro. Instituto Tecnológico Superior de Alvarado. CENIDET.	Consolidación curricular de la carrera de Ingeniería Ciencia de Datos

4. Competencia(s) a desarrollar

Competencia(s) específica(s) de la asignatura
Analiza el procesamiento semántico, la clasificación y agrupamiento de datos para brindar soluciones al tratamiento de información, basándose en modelos de análisis de personalidad, el análisis de opinión y la extracción de la información. Habilidad para aplicar técnicas avanzadas de procesamiento de lenguaje natural. Capacidad para optimizar procesos de extracción de la información en la Contextualización y normalización de textos. Destreza en la aplicación de algoritmos para el procesamiento semántico de textos para la identificación de patrones y la realización de predicciones. Competencia en el uso de técnicas de clasificación y agrupamiento de textos para la programación de información relevante. Habilidad para interpretar los resultados obtenidos y aplicarlos en la toma de decisiones informadas, impulsando la innovación y el desarrollo en diversos ámbitos profesionales.

5. Competencias previas

Comprende y aplica los conceptos sobre el aprendizaje automático, la adquisición, conjunto de datos y la programación avanzada para la Ciencia de Datos para aplicarlos en modelos que resuelvan problemas computacionales. Analiza requerimientos definidos por el cliente para generar soluciones al tratamiento de información y programación avanzada para la Ciencia de Datos.

6. Temario

No.	Temas	Subtemas
1	Contextualización y normalización de textos.	1.1. Contextualización 1.1.1. Sistemas y aplicaciones de lenguaje natural. 1.1.2. Fundamentos lingüísticos y matemáticos para procesamiento de lenguaje. 1.1.3. Lenguajes de programación, librerías y software para procesamiento de lenguaje. 1.2. Normalización de textos. 1.2.1. Segmentación en palabras y oraciones. 1.2.2. Eliminación de elementos no relevantes. 1.2.3. Etiquetado con categorías gramaticales.
2	Procesamiento semántico de texto.	2.1. Representación vectorial de textos para su procesamiento semántico. 2.1.1. Modelo de bolsa de palabras. 2.1.2. Modelo de espacio vectorial para textos 2.2. Selección de características en vectores. 2.2.1. Frecuencia y probabilidad de palabras. 2.2.2. Funciones de frecuencia de término (TF) y frecuencia inversa (IDF). 2.2.3. Técnicas de mapeo de palabras a vectores numéricos (word embeddings) 2.3. Extracción de aspectos semánticos a partir de textos. 2.3.1. Palabras similares, asociaciones entre palabras, terminología y palabras clave. 2.4. Minería de tópicos. 2.4.1. Representación de tópicos mediante palabras clave. 2.4.2. Modelo generativo de tópicos. 2.4.3. Algoritmo de Asignación latente de Dirichlet (LDA). 2.5. Extracción de información a partir de textos. 2.5.1. Análisis sintáctico superficial y profundo para extracción de entidades e información.

		2.5.2. Expresiones regulares para extracción de información. 2.5.3. Resumen automático.
3	Clasificación y agrupamiento de textos.	3.1. Uso práctico de clasificación de textos. 3.2. Tipos de aprendizaje de máquina y clasificadores. 3.2.1. Clasificadores generativos y Clasificadores discriminativos. 3.3. Aprendizaje de máquina supervisado para clasificación de textos. 3.3.1. Clasificación de textos mediante Bayes ingenuo y Regresión logística. 3.3.2. Clasificación de textos mediante K vecinos más cercanos. 3.4. Aprendizaje de máquina no supervisado para agrupamiento de textos. 3.4.1. Métodos generativos para agrupamiento de textos 3.4.2. Agrupamiento jerárquico de textos. 3.4.3. Agrupamiento no jerárquico de textos. 3.4.4. Agrupamiento aglomerante de textos. 3.4.5. Agrupamiento de textos mediante Algoritmo de K promedios.
4	Análisis de personalidad y opinión.	4.1. Detección y análisis de aspectos de personalidad en textos. 4.2. Análisis de opinión sobre entidades. 4.2.1. Regresión logística ordinal para opinión. 4.2.2. Detección de polaridad de opinión con base en diccionarios de polaridad. 4.2.3. Detección de polaridad de opinión usando algoritmos de aprendizaje de máquina. 4.3. Análisis de opinión sobre aspectos y características específicas de entidades. 4.3.1. Extracción de aspectos y características. 4.3.2. Minería de opinión sobre aspectos y características.

		4.4. Análisis de sentimientos. 4.4.1. Análisis basado en diccionarios de sentimientos. 4.4.2. Análisis basado en aprendizaje de máquina.
5	Análisis de textos en relación con otros tipos de datos.	5.1. Textos como datos no estructurados y datos estructurados. 5.2. Análisis de textos en relación con datos de geolocalización. 5.3. Análisis de textos en relación con datos temporales. 5.4. Análisis de textos en series de tiempo. 5.5. Análisis de mensajes en redes sociales.

7. Actividades de aprendizaje de los temas

1. Contextualización y normalización de textos	
Competencias	Actividades de aprendizaje
Específica(s): Analiza la normalización de textos en lenguaje natural con base en fundamentos matemáticos, lingüísticos y normalización. *Genéricas:* ● Pensamiento crítico. ● Resolución de problemas. ● Comunicación oral y escrita.	● Realizar ejercicios de segmentación de textos en palabras y oraciones utilizando diferentes lenguajes de programación y librerías especializadas. ● Desarrollar scripts para eliminar elementos no relevantes de un conjunto de textos y aplicar etiquetado con categorías gramaticales. ● Analizar casos de estudio donde la normalización de textos mejora significativamente el procesamiento de lenguaje natural.
2. Procesamiento semántico de texto	
Competencias	Actividades de aprendizaje
Específica(s): Extrae aspectos semánticos e información a partir de la representación vectorial de textos, minería de tópicos, análisis sintáctico y expresiones regulares. *Genéricas:* ● Pensamiento crítico. ● Resolución de problemas. ● Comunicación oral y escrita.	● Implementar modelos de bolsa de palabras y espacio vectorial para la representación de textos. ● Utilizar técnicas de selección de características como TF (Frecuencia de Término) e IDF (Frecuencia Inversa de Documento) para mejorar la representación vectorial. ● Desarrollar una aplicación de minería de tópicos usando el algoritmo LDA y explorar diferentes métodos de extracción de información.

3. Clasificación y agrupamiento de textos	
Competencias	Actividades de aprendizaje
Específica(s): Aplica el algoritmo de clasificación o agrupamiento de textos con base a los clasificadores, aprendizaje de máquina supervisado y no supervisado *Genéricas:* ● Pensamiento crítico. ● Resolución de problemas. ● Comunicación oral y escrita.	● Aplicar clasificadores como Bayes ingenuo y regresión logística en tareas de clasificación de textos. ● Experimentar con métodos de aprendizaje de máquina no supervisado para el agrupamiento de textos, incluyendo algoritmos de K-medias. ● Realizar prácticas de clasificación y agrupamiento para distinguir entre diferentes tipos de documentos.
4. Análisis de personalidad y opinión	
Competencias	Actividades de aprendizaje
Específica(s): Examina aspectos de personalidad y sentimientos en textos a partir de rasgos y detección de polaridad. *Genéricas:* ● Pensamiento crítico. ● Resolución de problemas. ● Comunicación oral y escrita.	● Analizar textos para detectar rasgos de personalidad y opiniones utilizando técnicas de regresión logística y algoritmos de aprendizaje de máquina. ● Desarrollar un sistema para la minería de opiniones que identifique y clasifique opiniones sobre entidades específicas basadas en aspectos y características. ● Realizar un proyecto de análisis de sentimientos basado tanto en diccionarios como en aprendizaje automático.
5. Análisis de textos en relación con otros tipos de datos	
Competencias	Actividades de aprendizaje
Específica(s): Extraer información significativa y revelar patrones en conjuntos de datos heterogéneos. Además, aprenderán a aplicar técnicas de correlación y visualización para comprender mejor las relaciones entre los datos textuales y otros tipos de datos, lo que les permitirá abordar problemas complejos en campos como la inteligencia artificial, la ciencia de datos y la investigación interdisciplinaria. *Genéricas:* ● Pensamiento crítico. ● Resolución de problemas. ● Comunicación oral y escrita.	● Analiza conjuntos de datos textuales junto con datos visuales. Utilizando técnicas de visualización de datos adecuadas, creará representaciones gráficas que muestren las relaciones entre los datos textuales y otros tipos de datos. ● Desarrolla un sistema de clasificación integrado que combina datos textuales con otros tipos de datos. ● Analiza conjuntos de datos que contengan tanto texto como otros tipos de datos relacionados con las emociones, como imágenes o expresiones faciales. Implementarán técnicas de análisis de sentimientos para detectar y clasificar emociones en el texto y en los datos adicionales, explorando cómo se relacionan y cómo pueden complementarse mutuamente en la interpretación del sentimiento general de un mensaje o contenido.

8. Práctica(s)

- Normalización de textos.
- Generación de representación vectorial de textos.
- Extracción de palabras similares.
- Extracción de asociaciones entre palabras.
- Extracción de terminología y palabras clave.
- Minería de tópicos.
- Extracción de información.
- Generación de resumen.
- Clasificación de textos.
- Agrupamiento de textos.
- Análisis de aspectos de personalidad.
- Análisis de polaridad en opinión.
- Análisis de opinión sobre entidades.
- Análisis de opinión sobre aspectos y características específicas de entidades.
- Análisis de textos en relación con otros tipos de datos.

9. Proyecto de asignatura

El objetivo del proyecto que planteé el docente que imparta esta asignatura, es demostrar el desarrollo y alcance de la(s) competencia(s) de la asignatura, considerando las siguientes fases:

Fundamentación: marco referencial (teórico, conceptual, contextual, legal) en el cual se fundamenta el proyecto de acuerdo con un diagnóstico realizado, mismo que permite a los estudiantes lograr la comprensión de la realidad o situación objeto de estudio para definir un proceso de intervención o hacer el diseño de un modelo.

Planeación: con base en el diagnóstico en esta fase se realiza el diseño del proyecto por parte de los estudiantes con asesoría del docente; implica planificar un proceso: de intervención empresarial, social o comunitario, el diseño de un modelo, entre otros, según el tipo de proyecto, las actividades a realizar los recursos requeridos y el cronograma de trabajo.

Ejecución: consiste en el desarrollo de la planeación del proyecto realizada por parte de los estudiantes con asesoría del docente, es decir en la intervención (social, empresarial), o construcción del modelo propuesto según el tipo de proyecto, es la fase de mayor duración que implica el desempeño de las competencias genéricas y específicas a desarrollar.

Evaluación: es la fase final que aplica un juicio de valor en el contexto laboral-profesional, social e investigativo, ésta se debe realizar a través del reconocimiento de logros y aspectos a mejorar se estará promoviendo el concepto de “evaluación para la mejora continua”, la metacognición, el desarrollo del pensamiento crítico y reflexivo en los estudiantes.

10. Evaluación por competencias

Son las técnicas, instrumentos y herramientas sugeridas para constatar los desempeños académicos de las actividades de aprendizaje.

- Reporte de indagación.
- Presentación electrónica de resultados de indagación.
- Solución de casos.
- Reporte de proyecto.
- Presentación y reporte de proyecto.
- Reporte de uso de software.
- Reporte de prácticas.

11. Fuentes de información

1. Bengfort, B., Bilbro, R., & Ojeda, T. 2018 Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning O'Reilly Media, Inc./ 9781491963043.
2. Ganegedara, T. 2018 Natural Language Processing with TensorFlow: Teach Language to Machines Using Python's Deep Learning Library Packt Publishing Ltd./ 9781788478311.
3. Bird, S., Klein, E., Loper, E. 2009 Natural Language Processing with Python O'Reilly/ 9780596516499.
4. Eisenstein, J. 2019 Introduction to Natural Language Processing MIT press/ 9780262042840
5. Ghosh, S., & Gunning, D. 2019 Natural Language Processing Fundamentals: Build Intelligent Applications that can Interpret the Human Language to Deliver Impactful Results Packt Publishing Ltd./ 9781789954043.
6. Jurafsky, D., & Martin, J. 2008 Speech and Language Processing Pearson Prentice Hall/ 9780131873216.
7. Vajjala, S., Majumder, B., Gupta, A., & Surana, H. 2020 Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems O'Reilly Media/ 9781492054054.
8. Zhai, C., Massung, S. 2016 Text Data Management and Analysis ACM and Morgan & Claypool Publishers/ 9781970001167.