IDENTIFICACIÓN DE LA UNIDAD DE APRENDIZAJE

<table>
<tr><td colspan="8">Unidad académica: Instituto de Investigación en Ciencias Básicas y Aplicadas</td></tr>
<tr><td colspan="8">Plan de estudios:
Licenciatura en Inteligencia Artificial</td></tr>
<tr><td colspan="4">Unidad de aprendizaje:

MINERÍA DE TEXTOS</td><td colspan="4">Ciclo de formación: Profesional
Eje general de formación: Teórico-Técnica
Área de conocimiento: Bases de la Inteligencia Artificial y la Ciencia de Datos
Semestre: 5º</td></tr>
<tr><td colspan="4">Elaborada por:
Dr. Jorge Hermosillo Valadez</td><td colspan="4">Fecha de elaboración: Abril, 2021</td></tr>
<tr><td>Clave:</td><td>Horas teóricas :</td><td>Horas prácticas :</td><td>Horas totales :</td><td>Créditos :</td><td>Tipo de unidad de aprendizaje :</td><td>Carácter de la unidad de aprendizaje :</td><td>Modalidad:</td></tr>
<tr><td>MT36CP030208</td><td>03</td><td>02</td><td>05</td><td>08</td><td>Obligatoria</td><td>Teórico - Práctica</td><td>Escolarizada</td></tr>
<tr><td colspan="8">Plan (es) de estudio en los que se imparte:
A partir de todos los programas impartidos por el Instituto de Investigación en Ciencias Básicas y Aplicadas.</td></tr>
</table>

ESTRUCTURA DE LA UNIDAD DE APRENDIZAJE

Presentación:

En la actualidad, la llamada *Sociedad del conocimiento* se refiere a la innovación de las tecnologías de la información y las comunicaciones en la que el incremento en las transferencias de la información modifica en muchos sentidos la forma en que se desarrollan muchas actividades en la sociedad moderna. En este contexto, la importancia de esta unidad de aprendizaje radica en conocer cómo la gestión de la información, la documentación y el conocimiento se perfilan como un componente estratégico de primer orden. Hoy en día, se maneja cada vez más información en formatos no estructurados o semiestructurados, como mensajes de correo electrónico, artículos, respuestas a preguntas abiertas, fuentes de noticias, formularios web, etc. Esta abundancia de información se presenta como un problema para muchas empresas e investigadores sobre cómo recopilar, explorar y aprovechar toda esta información; es aquí donde la Minería de Textos juega un papel fundamental.

Propósito: Conozca, describa y aplique los métodos y modelos de la *minería de textos,* como parte del procesamiento del lenguaje natural, para el análisis de sentimientos, la recuperación de información, el resumen automático de textos o la modelación del lenguaje en agentes artificiales, a fin de desarrollar sistemas inteligentes que puedan capturar temas y conceptos clave, y descubrir relaciones ocultas o tendencias existentes entre los datos, empleando criterios objetivos, comunicando resultados de manera efectiva, y profundizando en su campo de estudio profesional.
Competencias que contribuyen al perfil de egreso
Competencias genéricas: • CG8. Capacidad creativa. • CG14. Capacidad de aplicar los conocimientos en la práctica. • CG19. Conocimiento sobre el área de estudio y la profesión.
Competencias específicas: • CE11. Desarrolla sistemas computacionales inteligentes utilizando una computadora con la arquitectura y lenguaje de programación adecuados para la resolución de problemas con una actitud investigativa y socialmente responsable. • CE12. Implementa, prueba y mantiene proyectos de sistemas inteligentes empleando criterios de cumplimiento según estándares de calidad establecidos y aprovechando al máximo sus recursos, para resolver problemas científicos y tecnológicos y tomar decisiones que generen bienestar para la sociedad en su conjunto.

CONTENIDOS

Bloques	Temas
1. Representación de texto	1.1 El modelo Bolsa de Palabras 1.2 Características de documentos 1.3 Tokenización

	1.4 Filtrado de stop-words 1.5 Normalización 1.6 Árboles sintácticos 1.7 TF-IDF
2. Métodos de clasificación y agrupamiento de texto	2.1 Clasificación de texto 2.1.1 Árboles de decisión 2.1.2 Bosques aleatorios 2.1.2 Vecinos más cercanos 2.1.3 Clasificador Bayes ingenuo 2.1.4 Adaboost 2.2 Agrupamiento de texto 2.2.1 Métricas de similaridad 2.2.2 Agrupamiento por particiones 2.2.3 Agrupamiento jerárquico 2.2.4 Agrupamiento por grafos 2.3 Funciones de criterio de agrupamiento 2.3.1 Funciones de criterio interno 2.3.2 Funciones de criterio externo 2.3.3 Funciones de criterio híbrido 2.3.4 Funciones de criterio basado en grafos 2.4 Métricas de evaluación de clusters 2.4.1 Métricas basadas en conteo de pares 2.4.2 Pureza 2.4.3 Entropía, entropía condicional y conjunta 2.4.4 Medida F 2.4.5 Información Mutua Normalizada 2.4.6 Siluetas 2.4.7 Etiquetado de clusters
3. Procesamiento estadístico del lenguaje	3.1 Elementos básicos de Teoría de la Probabilidad 3.1.1 Espacios de probabilidad

	3.1.2 Probabilidad condicional e independencia 3.1.3 Teorema de bayes 3.1.4 Valor esperado y varianza 3.1.5 Distribuciones conjuntas y condicionales 3.1.6 Estadística Bayesiana 3.2 Fundamentos de Teoría de la Información 3.2.1 Entropía relativa o divergencia Kullback-Leibler 3.2.2 La relación con el lenguaje: la entropía cruzada 3.2.3 Perplejidad 3.3 Co-ocurrencia de palabras 3.3.1 Promedio y varianza 3.3.2 Pruebas de hipótesis 3.3.3 Información mutua 3.3.4 La noción de co-locación 3.4 Inferencia estadística 3.4.1 Modelos de n-gramas 3.4.2 Estimación por máximo de verosimilitud 3.4.3 Leyes de Laplace, Lidstone y Jeffreys-Perks 3.4.4 Validación cruzada 3.5 Incrustación de palabras 3.5.1 Determinando el contexto y la similaridad 3.5.2 Ventanas de contexto 3.5.3 Cálculo de incrustaciones 3.5.4 Modelos de atención con redes neuronales
4. Aplicaciones en Minería de textos	4.1 Recuperación de información 4.2 Resumen automático de textos 4.3 Análisis de sentimientos 4.4 Lingüística computacional

ESTRATEGIAS DE ENSEÑANZA - APRENDIZAJE

Estrategias de aprendizaje sugeridas (Marque X)			
Aprendizaje basado en problemas	(X)	Nemotecnia	( )
Estudios de caso	( )	Análisis de textos	( )
Trabajo colaborativo	(X)	Seminarios	(X)
Plenaria	( )	Debate	( )
Ensayo	( )	Taller	( )
Mapas conceptuales	( )	Ponencia científica	( )
Diseño de proyectos	(X)	Elaboración de síntesis	( )
Mapa mental	( )	Monografía	( )
Práctica reflexiva	( )	Reporte de lectura	( )
Trípticos	( )	Exposición oral	(X)
Otros			
Estrategias de enseñanza sugeridas (Marque X)			
Presentación oral (conferencia o exposición) por parte del docente	(X)	Experimentación (prácticas)	(X)
Debate o Panel	( )	Trabajos de investigación documental	( )
Lectura comentada	( )	Anteproyectos de investigación	( )
Seminario de investigación	( )	Discusión guiada	(X)
Estudio de Casos	(X)	Organizadores gráficos (Diagramas, etc.)	( )
Foro	( )	Actividad focal	( )
Demostraciones	( )	Analogías	( )
Ejercicios prácticos (series de problemas)	(X)	Método de proyectos	(X)

Interacción la realidad (a través de videos, fotografías, dibujos y software especialmente diseñado).	( )	Actividades generadoras de información previa	( )
Organizadores previos	( )	Exploración de la web	( )
Archivo	( )	Portafolio de evidencias	( )
Ambiente virtual (foros, chat, correos, ligas a otros sitios web, otros)	( )	Enunciado de objetivo o intenciones	(X)

CRITERIOS DE EVALUACIÓN

Criterios	Porcentaje
• Realización de prácticas • Búsqueda de información • Proyecto final • Tareas	30% 30% 30% 10%
Total	100 %

PERFIL DEL PROFESORADO

Licenciatura, Maestría o Doctorado en ciencias computacionales, matemáticas o ingeniería en áreas afines a las ciencias computacionales, con experiencia docente en el área.

REFERENCIAS

Básicas:

- Christopher D. Manning, Hinrich Schütze. (1999). *Statistical Natural language Processing*. The MIT Press, Cambridge, Massachusetts London England.
- Dipanjan Sarkar. (2016). *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data*. Apress.
- Michael W. Berry, Jacob Kogan (Editors). (2010). *Text Mining: Applications and Theory*. John Wiley & Sons Ltd.
- Steven Bird, Ewan Klein and Edward Looper. (2009). Natural Language Processing with Python. O'Reilly Media, Inc.

Complementarias:

- Jan Žižka, František Dařena, Arnošt Svoboda. (2020). *Text Mining with Machine Learning: Principles and Techniques*. CRC Press (Taylor & Francis Group).
- Uday Kamath, John Liu, and James Whitaker. (2019). *Deep Learning for NLP and Speech Recognition* (1st. ed.). Springer Publishing Company, Incorporated.