TÍTULO: Minería de datos para texto			
AÑO: 2024	**CUATRIMESTRE:** 2°	**N° DE CRÉDITOS:** 3	**VIGENCIA:** 3 años
CARGA HORARIA: 60 horas de teoría y 60 horas de práctica			
CARRERA/S: Doctorado en Ciencias de la Computación			

FUNDAMENTOS

En esta materia se presentan métodos y técnicas de análisis exploratorio de datos y aprendizaje no supervisado con especial atención a sus aplicaciones en procesamiento del lenguaje natural. Se proveen los fundamentos teóricos y metodológicos de aproximaciones como las reglas de asociación, clustering o embeddings, y se desarrollan aplicaciones prácticas de estas técnicas en diferentes problemas de procesamiento del lenguaje natural. Se aborda la cuestión de la evaluación en aprendizaje no supervisado, y se proveen diferentes métodos para ello.

OBJETIVOS

Al finalizar el curso los estudiantes habrán adquirido los fundamentos de las técnicas de aprendizaje no supervisado; habrán desarrollado o consolidado conocimientos sobre aprendizaje automático en general; se habrán familiarizado con el procesamiento del lenguaje natural, problemas clásicos del área y el amplio rango de soluciones que se pueden desplegar sobre esos problemas. Tendrán la capacidad de leer y entender artículos científicos del área, y de evaluar diferentes opciones para un problema y un contexto dados.

PROGRAMA

Unidad I: Unidad I: Introducción a la minería de datos, análisis exploratorio de datos, aprendizaje no supervisado
Inteligencia artificial, aprendizaje automático, aprendizaje supervisado y no supervisado. Minería de datos. Aprendizaje semi-supervisado.

Unidad II: Unidad II: Introducción al procesamiento del lenguaje natural
El lenguaje natural como objeto de estudio. Aplicaciones del tratamiento automático del lenguaje natural. Análisis por niveles del lenguaje natural. Generación. Métodos no supervisados para delimitar palabras, crear y enriquecer lexicones, análisis morfológico, análisis sintáctico y semántico.

Unidad III: Unidad III: Evaluación
Métricas. Limitaciones y fortalezas de las métricas. Concursos abiertos. Testbeds. Complemento entre análisis cualitativo y cuantitativo.

Unidad IV: Unidad IV: Reglas de asociación y Clustering
Correlación, significatividad. Diferentes formas de clustering: aglomerativo, jerárquico. Distancias. Combinaciones. Clustering como embedding.

Unidad V: Unidad VI: Modelos de Lenguaje y Embeddings
Reducción de dimensionalidad y acercamiento a causas latentes mediante métodos proyectivos. PCA. ICA. Autoencoders. Embeddings neuronales. Embeddings de modelos de lenguaje. Modelos de lenguaje,

Unidad VI: Unidad VI: El entorno de los aprendizajes automáticos

Representation learning. Transfer learning. Weak supervision.

Unidad VII: Unidad VII: Cuestiones éticas
Inteligencia artificial responsable. Ética de la inteligencia artificial. Impactos sociales y ambientales. Equidad y sesgo. Métricas de equidad. Exploraciones.

PRÁCTICAS

Se realizarán dos trabajos prácticos comunes a todos los estudiantes, uno sobre clustering y otro sobre embeddings, con consigna y objetivos comunes pero conjunto de datos a elección. Cada uno tendrá fecha de entrega 2 semanas después de que se comunica la consigna. En la segunda parte del curso se realizará un proyecto individual de investigación, con seguimiento semanal individualizado. Al final del curso se presentará oralmente el proyecto, un informe y el repositorio público.

BIBLIOGRAFÍA

R. Barzilay, K. McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. {\it Proceedings of the Meeting of the Association for Computational Linguistics 2001}
D. Brown et al. 1993. The Mathematics of Statistical Machine Translation. Computational Linguistics, 1993.
K. Church, P. Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics Vol. 16 (1), pp.22-29-
I. Goodfellow, Y. Bengio y A. Courville (2016). Deep Learning. MIT Press
T.K. Landauer, S.T. Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis: Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review
C. Manning, H. Schütze. 1999. Foundations of Statistical Natural Language Processing}. MIT Press.
NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook
D. Yarowsky. 1997. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. Proceedings of the Meeting of the Association for Computational Linguistics 1997

MODALIDAD DE EVALUACIÓN

Regularidad: entrega de dos informes de trabajos prácticos.
Aprobación: entrega de los tres informes (dos trabajos prácticos y un proyecto) y defensa oral del proyecto, que incluye examen oral con preguntas sobre el resto de la materia.

REQUERIMIENTOS PARA EL CURSADO

Conocimientos de probabilidad y estadística, conocimientos avanzados de programación