TÍTULO: Procesamiento del Lenguaje Natural		
AÑO: 2019	**CUATRIMESTRE:** primero	**N° DE CRÉDITOS:**
CARGA HORARIA: 60 horas de teoría y 60 horas de práctica.		
CARRERA/S: Doctorado en Ciencias de la Computación		

FUNDAMENTOS

El Procesamiento de Lenguaje Natural (PLN) estudia el uso de algoritmos y estructuras de datos para el procesamiento automático del lenguaje humano. Es una rama de las Ciencias de la Computación, la Inteligencia Artificial y la Lingüística Computacional, que sirve tanto para el desarrollo de aplicaciones prácticas que utilicen tecnología basada en lenguaje humano, como para el estudio de los problemas fundamentales de la lingüística teórica y las ciencias cognitivas.

En este curso daremos una introducción a las principales tareas que componen el PLN, y los diferentes enfoques computacionales para encararlas. Haremos énfasis especialmente en el trabajo basado en corpus y en el uso de algoritmos de aprendizaje automático (Machine Learning). Repasaremos métodos clásicos de aprendizaje automático así como también enfoques modernos basados en redes neuronales profundas.

OBJETIVOS

El objetivo del curso es dar a las y los estudiantes un conocimiento general del campo de PLN. Serán capaces de identificar y comprender problemas concretos de PLN, y proponer soluciones para ellos.

PROGRAMA

Unidad 1: Procesamiento básico de texto

Expresiones regulares, tokenización, segmentación, normalización, lematización y stemming.

Unidad 2: Modelado de lenguaje

N-gramas, suavizado add-one y por interpolación, back-off. Evaluación con métricas de teoría la información (entropía y perplejidad). Aplicaciones: Generación de lenguaje y atribución de autoría.

Unidad 3: Etiquetado de secuencias

Etiquetado morfosintáctico (PoS tagging) y Reconocimiento de Entidades Nombradas (NER). Aprendizaje supervisado. Clasificadores: árboles de decision, regresiones logísticas y SVMs. Modelos Ocultos de Markov (HMMs), de Máxima Entropía (MEMMs) y Conditional Random Fields (CRFs). Algoritmo de Viterbi y beam search. Ingeniería de features, evaluación y análisis de error.

Unidad 4: Representación de palabras y modelos neuronales

Representación vectorial de palabras (word embeddings): word2vec, fasttext y GloVe. Aprendizaje y evaluación. Representación de oraciones y documentos.
Modelos de lenguaje neuronales: ULMFiT, OpenAI, ELMo, BERT.

Unidad 5: Temas complementarios

Análisis de sentimiento (sentiment analysis), análisis sintáctico (parsing), extracción de

información (information extraction), traducción automática (machine translation), recuperación de información (information retrieval) y búsqueda de respuestas (question answering).

PRÁCTICAS

Se realizarán cuatro Trabajos Prácticos (TPs). Los primeros tres serán realizados en torno a tres temas principales (modelado de lenguaje, etiquetado de secuencias y representaciones de palabras). En cada uno de ellos se implementarán sistemas completos, y se realizarán experimentos que permitan evaluar y comparar los diferentes modelos. Los TPs serán guiados a través de ejercicios con objetivos claros y medibles.

La evaluación será realizada a través de una entrega de código fuente y de un informe de resultados. Además de la resolución de los ejercicios, se evaluarán aspectos cualitativos como el uso de buenas prácticas de programación (versionado, testing, coding style, documentación, etc.).

El cuarto y último TP será de tema libre y tratará sobre el estudio y la replicación de resultados obtenidos en publicaciones científicas de conferencias o revistas del área. La evaluación será a través de la entrega de un informe y de una defensa oral.

BIBLIOGRAFÍA

BÁSICA

[1] Daniel Jurafsky and James H. Martin. Speech and Language Processing, 2nd Edition . Prentice Hall, 2nd edition, May 2008.
[2] Christopher D. Manning and Hinrich Schtze. Foundations of statistical natural language processing. Hardcover, June 1999.
[3] Bird, S., Klein, E., and Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, 1 edition.
[4] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825– 2830.

COMPLEMENTARIA

Proceedings de las conferencias más importantes relacionadas con el PLN:

- Association of Computational Linguistics (ACL)
- North American Chapter of the ACL (NAACL)
- European Chapter of the ACL (EACL)
- COLING (International Committee of Computational Linguistics)
- EMNLP (Empirical Methods in Natural Language Processing)

MODALIDAD DE EVALUACIÓN

Para la regularización del curso, deben aprobarse los primeros tres TPs. Para la aprobación, deben aprobarse todos los TPs, y se debe realizar una defensa oral del cuarto TP.

REQUERIMIENTOS PARA EL CURSADO

Se requieren conocimientos previos de Algoritmos y Estructuras de Datos, y de Probabilidad y Estadística.