FACULTAD DE INFORMÁTICA | UNIVERSIDAD NACIONAL DE LA PLATA

# MINERIA DE TEXTOS

**Año 2019**

Carrera: ***Maestría en Inteligencia de Datos orientada a Big Data***

Carga Horaria: 64 Hs.

Profesores a Cargo***:***
***Dr. Marcelo Errecalde***
***Dr. Alfredo Simón***
***Dr. Augusto Villa Monte***

## OBJETIVO

Introducir al alumno en el análisis de información no estructurada ingresada en formato texto con un enfoque teórico-práctico. El énfasis estará puesto en un análisis de los documentos que surge de la interacción del Procesamiento del Lenguaje Natural (PLN), la Recuperación de la Información (RI) y el Aprendizaje Automático (AA). Se revisarán distintas herramientas para la recolección, pre-procesamiento, representación, análisis y visualización de textos.

## MODALIDAD DE EVALUACION

Para aprobar el curso se requiere un 80% de asistencia y la realización de un trabajo final que se definirá una vez completada la exposición de los contenidos teóricos. Dicho proyecto podrá ser realizado en forma individual o grupal y tendrá por objetivo profundizar uno o varios de los conceptos vistos en clase. La entrega consistirá en un reporte técnico de calidad científica producto de un estudio experimental específico.

## PROGRAMA

- Introducción. Lenguaje, lingüística y computación. Minería de Texto (en contexto). Procesamiento de Lenguaje Natural (PLN). Dificultades del PLN. Tareas y Aplicaciones. El Proceso de KDD. KDD a partir de textos. Niveles del Lenguaje Natural

- Recuperación de Información en la Web. Modelos de búsqueda. Minería Web. Web scraping. Descarga. Recolección. Hipertexto. Repositorios digitales. Utilización de APIs. Navegador web, HTTP.

- Manipulación del texto a través de strings. Limpieza. Expresiones regulares. Partición del texto. Tokenización. Filtrado. “Stop-words”. Normalización. Truncado (“stemming”) y lematización (“lemmatization”). Etiquetado. Etiquetado de Partes de la Oración (Part of Speech (POS) Tagging). Desambiguación del Significado de las Palabras (WSD). Reconocimiento de Entidades Nombradas (NER).

- Modelos de representación de documentos. Los documentos. Características estáticas y dinámicas. El Modelo de Espacio Vectorial. Representación de bolsa de palabras (BoW). Representación distribucional de términos (Bag of Concepts). Document occurrence representation (DOR). Term co-occurrence representation (TCOR). Concise semantic analysis (CSA). Reducción de la dimensionalidad. Selección de términos. Transformación del espacio de términos. Indexado (análisis) de semántica latente

- Representaciones avanzadas de documentos. Cuantificando “estilo”. Enfoques basados en instancias versus enfoques basados en perfil. Modelos neuronales de textos. Embeddings. Modelos secuenciales de textos. Aprendizaje y clasificación incremental. Redes Neuronales Recurrentes. Clasificación anticipada de textos

- Análisis semántico. Diferencia con el sintáctico. Estructuras tales como glosarios, diccionarios, tesauros, ontologías, etc. Wordnet.

- Discusión de problemas: Agrupamiento, categorización, clasificación, obtención de resúmenes, topic detection, análisis de sentimientos.

- Aplicaciones: Atribución de Autoría. Determinación del perfil del autor (Author Profiling). Detección de Plagios. Análisis de Sentimientos/ Minería de Opiniones. Detección temprana de riesgos (pedófilos, rumores, depresión, anorexia). Calidad de Información en la Web. Recursos y herramientas: NLTK, gensym, LIWC, etc.

## BIBLIOGRAFIA


- Aggarwal, Charu C.; Zhai, ChengXiang. *Mining Text Data*. ISBN 978-1-4614-3222-7. Springer-Verlag New York, 2012.
- Bird, Steven; Klein, Ewan; Loper, Edward. *Natural Language Processing with Python*. ISBN: 978-0-596-51649-9. O'Reilly Media, Inc., 2009.
- Hernández Orallo, José; Ramírez Quintana, M.ª José; Ferri Ramírez, Cèsar. Capítulo: *Minería de web y textos*. En: *Introducción a la minería de datos*. ISBN 9788420540917. Pearson Educación, S.A., 2004.
- Martínez-Méndez, Francisco-Javier. *Recuperación de información: modelos, sistemas y evaluación*. ISBN 84-932537-7-4. El Kiosko JMC, 2004.
- Olivas Varela, José Angel. *Búsqueda eficaz de información en la web*. ISBN 978-950-34-0763-9. Edulp, 2011.
- Bosch, Mela. *La piel del jaguar: la escritura móvil. Heurística y hermenéutica en el tratamiento informático de los documentos*. ISBN 978-3-8473-6869-4. EAE, 2012.
- Peña, Rosalía; Baeza-Yates, Ricardo; Rodríguez Muñoz, José Vicente. *Gestión digital de la información: de bits a bibliotecas digitales y la Web*. ISBN: 9788478975143. RA-MA, 2002.
- Marimón Llorca, Carmen. *Análisis de textos en español : teoría y práctica*. ISBN 978-84-7908-994-8. Universidad de Alicante, 2008.
- Lebart, Ludovic; Salem, André; Bécue-Bertaut, Mónica. *Análisis estadístico de textos*. ISBN 84-89790-57-4. Milenio, 2000.
- Bécue-Bertaut, Mónica. *Minería de textos. Aplicación a preguntas abiertas en encuestas*. ISBN: 978-84-7133-793-1. LA MURALLA, S.A., 2010.
- Jurafsky, Dan; Martin, James H. *Speech and Language Processing* [online]. https://web.stanford.edu/~jurafsky/slp3/
- Torres-Moreno, Juan-Manuel. *Automatic Text Summarization*. ISBN 978-1-84821-668-6. ISTE Ltd and John Wiley & Sons, Inc., 2014.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval,* Cambridge University Press. 2008.
- Christopher D. Manning y Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press. Cambridge, Massachusetts. 1999.
- Charu C. Aggarwal. *Machine Learning for Text*, Springer, March 2018.
- Fabrizio Sebastiani. 2002. *Machine learning in automated text categorization*. ACM Computing Surveys. 34, 1, 1-47. 2002.