

UNIVERSIDAD NACIONAL DE CÓRDOBA
Facultad de Matemática, Astronomía y Física

PROGRAMA DE CURSO DE POSGRADO

TÍTULO: Minería de Datos para Texto	
AÑO: 2017	CUATRIMESTRE: 2
CARGA HORARIA: 60	No. DE CRÉDITOS:
CARRERA/S: Ciencias de la Computación	
DOCENTE ENCARGADO: Laura Alonso Alemany	

PROGRAMA

1. Introducción a la minería de datos, análisis exploratorio de datos, aprendizaje no supervisado
2. Aprendizaje semi-supervisado
3. Introducción al procesamiento del lenguaje natural
4. Principios de evaluación, concursos
5. Reglas de asociación, correlación, significatividad
6. Clustering
7. Embeddings: reducción de dimensionalidad y acercamiento a causas latentes mediante métodos proyectivos
8. Representation learning, Transfer learning
9. Caracterización de fenómenos lingüísticos basada en datos
 - 9.1. Delimitación del vocabulario mediante tests de hipótesis
 - 9.2. Descubrimiento de clases de palabras mediante clustering
 - 9.3. Caracterización de clases de palabras mediante combinaciones de clustering y clasificación
 - 9.3.1. Desambiguación de sentidos
 - 9.3.2. Adquisición de subcategorizaciones
 - 9.3.3. Traducción automática estadística
 - 9.3.4. Adquisición automática de paráfrasis

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 LA EVALUACIÓN

Se realizará un examen teórico (30% de la nota), 2 trabajos prácticos (15% de la nota respectivamente) y un desarrollo y presentación oral dentro de uno de los temas (40% de la nota).