Universidad Nacional de Córdoba

FACULTAD DE MATEMATICA ASTRONOMÍA Y FÍSICA

Expte. 03-05-01535

RESOLUCION HCD N° 262/05

VISTO

La propuesta de la Comisión Asesora de Computación, para que se apruebe la materia "Minería de datos para texto" como Optativa de la Licenciatura en Ciencias de la Computación; y

CONSIDERANDO

Que es conveniente agregar a la nómina de materias optativas, aprobada por Res. HCD 207/02, la asignatura que se propone;

Que mediante Resolución HCS n° 122/02 se ha delegado en este Cuerpo la facultad de modificar la nómina de materias optativas del Plan de Estudios de la Licenciatura en Ciencias de la Computación;

EL HONORABLE CONSEJO DIRECTIVO DE LA
FACULTAD DE MATEMÁTICA, ASTRONOMÍA Y FÍSICA
R E S U E L V E :

ARTÍCULO 1°: Hacer lugar a lo solicitado por la Comisión Asesora de Computación de esta Facultad y, en consecuencia, modificar la nómina de materias optativas del Plan de Estudios de la Licenciatura en Ciencias de la Computación, incorporando a la misma la materia "Minería de datos para texto".

ARTÍCULO 2°: Fijar como programa, correlativas y carga horaria de la materia, los detallados en el Anexo que forma parte de la presente Resolución.

ARTÍCULO 3°: En cumplimiento de lo establecido en el artículo 2° de la Res. HCS n° 122/02, remítase a la Secretaría de Asuntos Académicos de la Universidad la presente resolución para su conocimiento y efectos.

ARTÍCULO 4°: Comuníquese y archívese.

DADA EN LA SALA DE SESIONES DEL HONORABLE CONSEJO DIRECTIVO DE LA FACULTAD DE MATEMÁTICA, ASTRONOMÍA Y FÍSICA, A VEINTIOCHO DIAS DEL MES DE NOVIEMBRE DE DOS MIL CINCO.

npk

Dr. WALTER N. DAL LAGO
Secretario General Fa.M.A.F.

Dr. DANIEL E. BARRACO DÍAZ
DECANO
Fa.M.A.F.

Universidad Nacional de Córdoba

FACULTAD DE MATEMATICA ASTRONOMÍA Y FÍSICA

Expte. 03-05-01535

ANEXO A RESOLUCIÓN HCD N° 262/05

MATERIA OPTATIVA	CORRELATIVAS			CARGA HORARIA
	PARA CURSAR		PARA RENDIR	
	REGULARIZADA	APROBADA	APROBADA	
Minería de datos para texto	Modelos y Simulación	Algoritmos y Estructuras de Datos II Probabilidad y Estadística	Modelos y Simulación Algoritmos y Estructuras de Datos II	120 hs.

Régimen de Cursado: Semestral.

INTRODUCCIÓN

La minería de texto consiste en descubrir información nueva y previamente desconocida mediante la extracción automática de información de varios recursos escritos. Un elemento clave es la relación entre las informaciones extraídas, de forma que se creen hechos o hipótesis nuevos que serán explorados en profundidad mediante métodos de experimentación más convencionales.

Este curso pretende ser una introducción al área de minería de datos aplicada a texto, desde una perspectiva de Procesamiento del Lenguaje Natural. Se describirá el área en relación a áreas bien establecidas como recuperación de información, procesamiento del lenguaje natural con métodos empíricos y descubrimiento de conocimiento en bases de datos.

Se trabajará mediante estudio de caso, presentando aproximaciones exitosas al descubrimiento de información en texto, para obtener una perspectiva general de:

- las necesidades de información que necesitan ser cubiertas,
- las propiedades de los textos que se pueden explotar,
- y cómo las intuiciones teóricas sobre propiedades textuales se pueden implementar en herramientas o procedimientos efectivos.

Al finalizar el curso, los estudiantes deberán haber adquirido

- una perspectiva general del área de minería de datos aplicada a texto,
- familiaridad (y capacidad operativa) con técnicas de aprendizaje automático no supervisado y semi-supervisado,
- madurez para hacer evaluaciones críticas del trabajo en el área.

PROGRAMA

1. Introducción a la minería de datos
2. Introducción al procesamiento del lenguaje natural
3. Principios de evaluación
4. Caracterización de fenómenos lingüísticos basada en datos
   (a) delimitación del vocabulario mediante tests de hipótesis
   (b) descubrimiento de clases de palabras mediante clustering y algoritmos genéticos
   (c) caracterización de clases de palabras mediante combinaciones de clustering y clasificación:
   - desambiguación de sentidos
   - adquisición de subcategorizaciones
   - traducción automática estadística
   - adquisición automática de paráfrasis

   (d) latent semantic analysis
5. técnicas de tratamiento de secuencias para lenguaje natural
   - modelos markovianos para modelar lenguaje
   - alineación múltiple de secuencias
   - identificación de discontinuidades
6. teoría de grafos aplicada a texto
   - identificación de nodos centrales
   - identificación de caminos relevantes

   técnicas de bootstrapping para aumentar recursos

BIBLIOGRAFÍA

J. Allen. 1987. *Natural Language Understanding*. The Benjamin/Cummings Publishin Company Inc.

R. Barzilay, K. McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. *Proceedings of the Meeting of the Association for Computational Linguistics 2001*

T. Briscoe, J. Carroll. 1997. Automatic extraction of subcategorization from corpora Proceedings of the 5th *Proceedings of the Meeting of the Association for Computational Linguistics 1997*

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-

M.L. Forcada. (2001) Corpus-based stochastic finite-state predictive text entry for reduced keyboards: application to Catalan. *Proceedings of the XVII Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural*

D. Koller., M. Sahami. 1996. Toward Optimal Feature Selection. *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 284-292, San Francisco, CA: Morgan Kaufmann

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. 1993. Automatic acquisition of a large subcategorisation dictionary from corpora *Proceedings of the Meeting of the Association for Computational Linguistics 1993*

C. Manning, H. SchÂ¨utze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

A. McCallum, A. Corrada-Emmanuel, X. Wang. 2004. *The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email.* Technical Report UM-CS-2004-096, 2004.

F.J. Och, H. Ney. (2000) Improved Statistical Alignment Models *Proceedings of the Meeting of the Association for Computational Linguistics 2000*

A. Venugopal, S. Vogel, A. Waibel. 2003 Effective Phrase Translation Extraction from Alignment Models *Proceedings of the Meeting of the Association for Computational Linguistics 2003*

D. Yarowsky. 1997. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *Proceedings of the Meeting of the Association for Computational Linguistics 1997*

Dr. WALTER N. DAL LAGO
Secretario General Fa. M. A. F.

Dr. DANIEL E. BARRACO DIAZ
DECANO
Fa.M.A.F.