

Propuesta para asignatura optativa de postgrado de computación:

Minería de datos para texto

Laura Alonso i Alemany

Facultad de Matemática, Astronomía y Física

Universidad Nacional de Córdoba

1 Introducción

La minería de texto consiste en descubrir información nueva y previamente desconocida mediante la extracción automática de información de various 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,
- capacidad para replicar y progresar en líneas de trabajo ya iniciadas en este área

2 Programa tentativo

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
7. técnicas de bootstrapping para aumentar recursos

3 Bibliografía

J. Allen. 1987. *Natural Language Understanding*. The Benjamin/Cummings Publishing 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. Schütze. 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*

Evaluación La materia se evaluará como sigue:

50% una tarea realizada individualmente, con tutoría

30% realización de un artículo de investigación y revisión crítica de los artículos de los demás compañeros

20% evaluaciones críticas (por escrito) de las lecturas presentadas en clase

Estimación del número de candidatos a inscribirse en el curso

Carga horaria 120 horas, con un mínimo de 60 horas presenciales, 10 horas de tutoría y 50 cubiertas por las dos tareas individuales.