Text Mining

Laura Alonso i Alemany

Facultad de Matemática, Astronomía y Física
UNC, Córdoba (Argentina)
http://www.cs.famaf.unc.edu.ar/~laura

SADIO
7 de Mayo de 2010

qué vamos a ver hoy

introducción a extracción de información

extracción de información en texto semi-estructurado

opinion mining

presentación de proyectos

discusión de evaluación

introducción a extracción de información

filminas de Rob Gaizauskas
An Information Extraction Perspective on Text Mining: Tasks, Technologies and Prototype Applications

extracción de información de posts

- La sección de Wrapper Construction del tutorial de Craig Knoblock sobre Information Integration on the Web en AAAI-07.
- Matthew Michelson and Craig A. Knoblock. 2005. Semantic annotation of unstructured and ungrammatical text. IJCAI.
- Matthew Michelson and Craig A. Knoblock. 2009. Mining the Heterogeneous Transformations for Record Linkage. ICAI.

extracción de información de la wikipedia

y también la wikipedia

opinion mining

filminas de Lillian Lee
A Tempest. Or, on the flood of interest in: sentiment analysis, opinion mining, and the computational treatment of subjective language, de una charla en ICWSM'09.
filminas de Bing Liu sobre Opinion Mining de su curso Data and Text Mining.

presentación de proyectos

- Clasificación morfológica de lexicones frecuentes en fotolog
- Community Watch
- Sentiment Analysis
- Clasifiación de corrientes ideológicas de personas y organizaciones basada en noticias de los medios de difusión gráficos en línea
- Taggeado automático de programas

evaluación

- un proyecto más o menos grande
- tres proyectos chicos

posibles proyectos grandes

- armar un extractor de hechos a partir de corpus
- inducir o poblar una ontología
- extracción de información en la wikipedia
- obtener modelos de lenguaje de un corpus y detectar anomalías
  - utilizar abstracción por clases de palabra
  - aplicar para detección de autoría

posibles proyectos chicos

- clustering de palabras para identificar clases de palabras, en diferentes tipos de corpus
- juego de Shannon y su evaluación, en diferentes corpus
- inducción de temas de una colección documental
- clasificación temática de documentos, y análisis de error
- entrenamiento de un traductor automático estadístico, y análisis de error
- identificación de términos (incluyendo nombres propios) en un corpus, usando medidas de asociación entre palabras
- comparativa de performance de traductores automáticos

en el futuro...

lauraalonsoalemany@gmail.com

demoro en contestar pero termino haciéndolo!