Lingüística computacional

Presentación

Escuela Nacional de Antropología e Historia (ENAH)
Agosto – diciembre de 2015

Profesor

- Dr. Carlos Méndez (cmendezc@iingen.unam.mx)
- Licenciado en informática, maestro en lingüística hispánica y doctor en lingüística (UNAM)
- Investigador posdoctoral
- Lingüística computacional: aprendizaje no supervisado de morfología
- Inferir, a partir de un corpus no etiquetado y de forma automática, una descripción morfológica de una lengua

Profesor

- Proyecto CONACyT *Caracterización de huellas textuales para análisis forense*
- Perfilamiento de autor: determinar automáticamente el género y edad del autor de un documento
- Análisis de rasgos estilísticos del autor: uso de signos de puntuación, léxico, cantidad de palabras por oración, riqueza léxica, etcétera

Profesor

- Proyecto OCRMX S. A. *Desarrollo de un sistema de publicación de una biblioteca de arte mexicano*
- 1) Publicar la biblioteca
- 2) Encontrara automáticamente libros que traten del mismo tema
- 3) Agrupar automáticamente libros por contenido

Materia

- Optativa
- Miércoles de 9 a 13
- Laboratorio de lingüística

Objetivos

- 1) El alumno reconocerá los fundamentos del análisis lingüístico desde la perspectiva computacional.
- 2) El alumno reconocerá las principales propuestas de la Lingüística computacional para el estudio de algunos fenómenos lingüísticos.
- 3) El alumno será capaz de proponer un análisis lingüístico con perspectiva computacional.
- NO es curso de programación
- Conocer y usar programas de análisis lingüístico

Contenido

- Lingüística computacional
  - Definición y alcance
  - Retos
  - Estadística y probabilidad
- Morfología
  - Aprendizaje no supervisado de morfología
  - Autómatas de estados finitos (expresiones regulares)
  - Morfotáctica de estados finitos
- Morfosintaxis
  - Etiquetado de partes de la oración
  - Modelo de n-gramas
  - Autómatas probabilísticos

Contenido

- Sintaxis
  - Tipos de gramáticas
  - Análisis sintáctico
  - Análisis de dependencias
- Semántica léxica
  - Modelo de bolsa de palabras
  - Modelo de espacio vectorial
  - Modelo de semántica distribucional

Estrategias de aprendizaje

Se proponen las siguientes estrategias de aprendizaje:

- Exposición por parte del profesor.
- Lectura de artículos académicos y capítulos de libros.
- Discusión en clase.
- Prácticas con corpus lingüísticos (tareas).

Criterios de evaluación

Se tomarán en cuenta los siguientes elementos de evaluación:

- 2 exámenes teóricos (50%). Uno a medio semestre y uno al final
- 1 trabajo final (en formato de artículo científico) en donde se analice un fenómeno lingüístico usando corpus y con perspectiva computacional (30%)
- Tareas (20%)
  - Prácticas
  - Lecturas (cuestionario)

Bibliografía básica (mínima)

- Jurafsky, D. y Martin, J. H. (2007). Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River, N.J.: Pearson Prentice Hall.
- Manning, C. D. y Schütze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, Mass.: The MIT Press.

Propuestas opcionales

- Corpus segmentado morfológicamente para español para la 10th Conference on Language Resources and Evaluation (LREC 2016) PORTOROŽ, SLOVENIA, 23-28 May 2016 (grupal)

- Artículo científico sobre análisis automático en corpus para congreso o revista (individual o grupal)