CURSO: PROCESAMIENTO DEL LENGUAJE NATURAL

SIGLA: LET3220-1
CARÁCTER: OPTATIVO
CRÉDITOS:
MÓDULOS: 2
SEMESTRE: 2

I. Descripción

Este curso se centra en 5 aspectos: (i) explicar qué es, sus conceptos principales y sus áreas de estudio; (ii) aprender a programar algunas herramientas de análisis usando la librería NLTK (*Natural Language Tool-Kit*); (iii) conocer sus métodos básicos de análisis estadístico, sobre todo enfatizando el análisis de corpus textuales; (iv) identificar las principales líneas de investigación y; (v) reflexionar y desarrollar una propuesta de análisis basada en los métodos y herramientas vistas.

II. Objetivos

General:

Brindar a los estudiantes una panorámica general sobre lo que es el PLN, ponderando tanto una descripción teórica de sus conceptos y métodos principales, así como el aprendizaje de técnicas de programación para desarrollar herramientas de análisis.

Específicos:

- Definir en qué consiste el PLN.
- Definir sus conceptos y métodos básicos.
- Aprender a usar la librería NLTK, programada en lenguaje Python.
- Identificar sus principales líneas de investigación.
- Desarrollar un proyecto basado en alguna de las líneas de investigación revisadas.

III. Contenidos

- Nociones básicas: definición del PLN, gramáticas formales, gramáticas libres de contexto, *n*-gramas, corpus lingüísticos, etiquetado lingüístico, *chunking* y *parsing*, bases de conocimiento léxicas.
- Herramientas de análisis: módulos de procesamiento textual de NLTK.
- Técnicas de análisis: probabilidades condicionales, información mutua y entropía, *precisión & recall*.
- Líneas de investigación: recuperación y extracción de información, reconocimiento de entidades nombradas, clasificación de textos, minería de textos, desarrollo de ontologías.

IV. Metodología

- Clases expositivas
- Sesiones de taller con computadores para el aprendizaje en el manejo de herramientas
- Entrega de evaluaciones parciales
- Entrega de un reporte final de trabajo

Las actividades y los porcentajes considerados para dichas actividades son los siguientes:

- 2 Controles en el semestre: 20% (10% por cada control)

- Desarrollo de 4 trabajos parciales, para el desarrollo de una cadena de procesamiento para un corpus textual en inglés del área de medicina: 20%
- Reporte final del resultado obtenido tras desarrollar la cadena de procesamiento previamente descrita: 60%

Bibliografía mínima

Bird, Steven, Klein, Ewan y Loper, Edward (2010): *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit*, Sebastropol, Cal., O'Reilly Media.

Clark, Alexander, Fox, Chris and Lappin, Shalom (Eds.) (2010): *The Handbook of Computational Linguistics and Natural Language Processing*, Oxford, UK, Wiley-Blackwell.

Jurafsky, Daniel y Martin, James (2013): *Speech and Language Processing*, Third Edtion, New Jersey, Prentice Hall. Sitio electrónico: https://web.stanford.edu/~jurafsky/slp3/.

Bibliografía complementaria

Cole, Ronald A., et al. (1997): *Survey of the State of the Art in Human Language Technology*, Cambridge, Cambridge University Press.

Grishman, Ralph (1986): *Computational Linguistics*, Cambridge, Cambridge University Press.

Manning, Christopher y Schütze, Hinrich (2000): *Foundations of Statistical Natural Language*, Cambridge, Mass., MIT Press.

Mitkov, Ruslan (ed.) (2003): *The Oxford Handbook of Computational Linguistics*, Oxford, Oxford University Press.

Sitios electrónicos

Jurafsky, Daniel (2007): *Introduction to Computational Linguistics*, Stanford, Cal., Stanford University www.stanford.edu/~jurafsky.

Martin, James (2008): *Course of Natural Language Processing*, Boulder, Co., University of Colorado: www.cs.colorado.edu/~martin/csci5832/.

Bird, Steven, Klein, Ewan y Loper, Edward (2010): *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit*: www.nltk.org/book/.