PROGRAMA CURSO ***INTRODUCCIÓN A LA LINGÜÍSTICA COMPUTACIONAL***

Profesores: César Aguilar
Curso semestral: Martes y jueves hrs.
(18 semanas de 3 horas / 36 sesiones de 1:30 hrs.)

DÍA 1 (01 de agosto)	**Sesión 1 Presentación general del curso** • Presentación • Dinámica del curso • Evaluación • Calendario • Lectura: Gelbukh y Sidorov (2010), Parte I: "Problemas generales del procesamiento de lenguaje natural"
DÍA 2 (06 de agosto)	**Sesión 2** • Continuación de la Sesión 1 • Visiones de mundo entre lingüística y ciencias de la computación • Métodos simbólicos, estadísticos e híbirdos • Lenguaje natural como un sistema combinatorio de símbolos • Lectura: NLTK site: Gelbukh y Sidorov (2010), Parte I: "Problemas generales del procesamiento de lenguaje natural"
DÍA 3 (08 de agosto)	**Sesión 3** • Aprendiendo a programar con Python y NLTK • Descripción general del lenguaje Python • Cómo instalarlo (sistemas operativos: Windows, Linux-Unix y Mac) • Breve descripción sobre lo que es programar • Instalando Python • Algunos procesos básicos • □ Lectura: NLTK site: https://sites.google.com/site/naturallanguagetoolkit/download
DÍA 4 (13 de agosto)	**Sesión 4** • Instalando NLTK • Paquetes anexos: Numpy y PyYaml • Recursos disponibles • Lectura: NLTK site: https://sites.google.com/site/naturallanguagetoolkit/download
15 de agosto	**Feriado legal: no hay clase**
DÍA 05 (20 de agosto)	**Sesión 5** • Definición de lenguajes formales • Morfología, sintaxis y semántica de los lenguajes formales • Jerarquía de lenguajes formales: lenguajes regulares, lenguajes libres de contexto y lenguajes sensibles al contexto. • Lectura: Bolshakov y Gelbukh (2004), Parte II: "A Historical Outline"

DÍA 6 (22 de agosto)	**Sesión 6** • Teoría de autómatas • Autómatas deterministas y no deterministas • Expresiones regulares • Lectura: Jurafsky y Martin (2010), Chapter II: “Regular Expression and Automata”
DÍA 7 (27 de agosto)	**Sesión 7** • Uso de expresiones regulares en Python • Expresiones regulares y patrones lingüísticos • Lectura: Jurafsky y Martin (2010), Chapter II: “Regular Expression and Automata”
DÍA 8 (29 de agosto)	**Sesión 8** • Expresiones regulares y patrones lingüísticos (continuación) • Análisis probabilístico en LC • Lectura: Dan Jurafsky: "Probabilistic Modeling of Language: An Introduction and Apologia", 2003 (presentación) • **Trabajo parcial I:** diseño de un autómata para la generación de correos electrónicos
DÍA 9 (03 de septiembre)	**Sesión 9** • Revisión y solución del Trabajo parcial I • Probabilidades independientes y condicionales • *N*-gramas • Entropía e información mutua • Lectura: Dan Jurafsky: "Probabilistic Modeling of Language: An Introduction and Apologia", 2003 (presentación)
DÍA 10 (05 de septiembre)	**Sesión 10** • **Aplicación del primer control parcial del curso.** Temas a tratar: I. Descripción general de la LC II. Lenguajes formales III. Teoría de autómatas IV. Expresiones regulares
DÍA 11 (10 de septiembre)	**Sesión 11** • Corpus lingüísticos • Definición y clasificación de corpus • Corpus anotados • Sistemas de anotación: Eagles y Penn TreeBank • Lectura: Torruela y Listerri (1999): “Diseño de corpus textuales y orales”
DÍA 12 (12 de septiembre)	**Sesión 12** • Corpus lingüísticos en NLTK: Brown Corpus, PennTreeBank, corpus en español • Trabajando con archivos de texto e Internet • Análisis de corpus: conteos de palabras, búsqueda de colocaciones, gráficos de dispersión • Lectura: NLTK Book: “Accessing Text Corpora and Lexical Resources”: http://nltk.org/book/ch02.html.

DÍA 13 (17 de septiembre)	**Sesión 13** • Reconocimiento y análisis de palabras • Distinciones entre *types* (tipos) y *tokens* (muestras) • El algortimo de Porter • Lectura: NLTK Book: "Categorizing and Tagging Words": http://nltk.org/book/ch05.html.
19 de septiembre	**Feriado legal: no hay clase**

DÍA 14 (24 de septiembre)	**Sesión 14** **Trabajo parcial II:** (i) Selección de 5 documentos en inglés, sin importar su extensión, y diseñar un corpus textual en formato plano (TXT). (ii) Implementar un programa para insertar etiquetas POST en el corpus elaborado. (iii) Generar los resultados obtenidos en un nuevo archivo TXT.
DÍA 15 (26 de septiembre)	**Sesión 15** • **Aplicación del segundo control parcial del curso.** Temas a tratar: I. Descripción de corpus lingüísticos II. Etiquetados sintácticos III. Reconocimiento de *tokens* y *tags* en corpus
DÍA 16 (01 de octubre)	**Sesión 16** • Análisis sintáctico: introducción • Descripción de gramáticas formales • Estructura de frase y categorías gramaticales • Lectura: Galicia y Gelbukh (2007), Capítulo II: "Formalismos gramaticales"
DÍA 17 (03 de octubre)	**Sesión 17** • Modelos de gramáticas formales • Gramáticas de adjunción de árboles sintácticos • Gramáticas léxicas • Ejemplo de gramática léxica en NLTK • Lectura: Galicia y Gelbukh (2007), Capítulo II: "Formalismos gramaticales
DÍA 18 (08 de octubre)	**Sesión 18** • Identificación de caracteres ASCII y Unicode • Análisis gramatical en NLTK • Chunking • Lectura: NLTK Book: "Extracting Information from Text": http://nltk.org/book/ch07.html.

DÍA 19 (10 de octubre)	**Sesión 19** • Evaluación estadística del análisis sintáctico • Precisión y cobertura • Chunking y N-gramas • Lectura: NLTK Book: "Extracting Information from Text": http://nltk.org/book/ch07.html.
DÍA 20 (15 de octubre)	**Sesión 20** **Trabajo parcial III:** (i) Retomar el corpus etiquetado previamente en formato plano (TXT). (ii) Elaborar una gramática mínima en NLTK para detectar patrones de frases en el corpus Brown. (iii) Repetir la misma búsqueda de patrones en su corpus anotado, y hacer comparaciones con los resultados del Brown.
DÍA 21 (17 de octubre)	**Sesión 21** • **Aplicación del tercer control parcial del curso.** Temas a tratar: I. Definición de gramáticas formales II. Diferencias entre TAG, LFG y HPSG III. Medición de resultados (precisión y cobertura)
DÍA 22 (22 de octubre)	**Sesión 22** • Semántica y lingüística computacional • Composicionalidad semántica • Representación formal del significado • Lectura: Jurafsky y Martin (2010), Chapter 17: "The Representation of Meaning".
DÍA 23 (24 de octubre)	**Sesión 23** • Semántica léxica • Bancos de conocimiento léxico • WordNet • EuroWordNet • Framenet • Lectura: Jurafsky y Martin (2010), Chapter 19: "Lexical Semantics".
DÍA 24 (29 de octubre)	**Sesión 24** • Etiquetado semántico • VerbNet: reconociendo estructuras argumentales • Resolviendo problemas de ambigüedad semántica • Lectura: Jurafsky y Martin (2010), Chapter 20: "Computational Lexical Semantics".
DÍA 25 (31 de octubre)	**Sesión 25** • Recursos léxicos en NLTK • Wordnet • Generando relaciones léxicas • Lectura: NLTK Book: "Accessing Text Corpora and Lexical Resources": http://nltk.org/book/ch02.html

DÍA 26 (29 de octubre)	**Sesión 26** **Trabajo parcial IV:** (i) En equipos, consultar WordNet y Framenet, y buscar información semántica sobre los siguientes verbos: *define*, *understand*, *comprehend*, *believe* y *conceive*. (ii) Enlistar todas las relaciones léxicas y semánticas que proponen WordNet y FrameNet. (iii) Señalar similitudes y diferencias sobre la información semántica recopilada.
DÍA 27 (31 de octubre)	**Sesión 27** • **Aplicación del cuarto control parcial del curso.** Temas a tratar: I. Semántica léxica II. Composicionalidad III. Descripción de recursos léxicos: WordNet, Framenet, VerbNet
DÍA 28 (05 de noviembre)	**Sesión 28** • **Líneas de investigación aplicadas:** recuperación y extracción de información • Definición • Diferencias entre recuperación y extracción • Técnicas y recursos disponibles • Lectura: Martínez, P. & García, A. (2002): "Utilizando recursos lingüísticos para mejora de la recuperación de información en la Web". *Revista Iberoamericana de Inteligencia*, No. 16: 55-64.
DÍA 29 (07 de noviembre)	**Sesión 29** • **Líneas de investigación aplicadas:** reconocimiento de entidades nombradas • Definición • Técnicas y recursos disponibles • Lectura: Ladino, C. (2012): "Reconocimiento de Entidades y Extracción de Relaciones de Textos en la Web: Estado del Arte. Sitio electrónico: http://201.234.78.173:8081/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000774782.
DÍA 30 (12 de noviembre)	**Sesión 30** • **Líneas de investigación aplicadas:** resumen automático • Definición • Técnicas y recursos disponibles • Lectura: Mateo, P., et al (2003): "Un sistema para resumen automático de textos en castellano." *Revista de Procesamiento del lenguaje natural*, No. 31 (2003): 29-36.
DÍA 31 (14 de noviembre)	**Sesión 31** • **Líneas de investigación aplicadas:** clasificación de textos • Definición • Técnicas y recursos disponibles • Lectura: Echegoyen, C. (2013): "Clasificación de textos". Sitio electrónico: www.sc.ehu.es/ccwbayes/members/carlos/home/index.html.

DÍA 32 (19 de noviembre)	**Sesión 32** • **Líneas de investigación aplicadas:** clasificación de textos • Definición • Técnicas y recursos disponibles • Lectura: Barrón, A., Vila, M. & Rosso, P. (2010): "Detección automática de plagio: De la copia exacta a la paráfrasis". En *Panorama actual de la lingüística forense en el ámbito legal y policial: Teoría y práctica. (Jornadas (in)formativas de lingüística forense)*. Madrid. Euphonia Ediciones: 76-96.
DÍA 33 (21 de noviembre)	**Sesión 33** • **Líneas de investigación aplicadas:** minería de textos • Definición • Técnicas y recursos disponibles • Lectura: Brun, R. E., & Senso, J. A. (2004). "Minería textual", *El profesional de la información*, 13(1): 11-27.