ASIGNATURA DE MÁSTER:

PROCESAMIENTO DEL LENGUAJE NATURAL

Curso 2016/2017

(Código:31101269)

1.PRESENTACIÓN

Este curso introductorio al procesamiento computacional del lenguaje natural aborda el diseño y la construcción de programas que pueden tratar, comprender y generar lenguaje natural. Se estudiarán los problemas y soluciones (modelos y técnicas) básicas en los niveles sintáctico, semántico y pragmático. Un capítulo de introducción y otro de áreas de aplicación, situarán la asignatura desde una perspectiva histórica, y permitirán conocer el estado actual de las realizaciones en este campo. La realización de un proyecto permitirá poner en práctica los conocimientos adquiridos.

Ficha técnica:

Tipo	Optativa
Cuatrimestre	Primero
Créditos/horas totales	6/150
Horas de estudio teórico	45
Horas de prácticas	45
Proyecto	60

Profesorado

Coordinadora: Dra. Felisa Verdejo Maillo
(Dpto. Lenguajes y Sistemas Informáticos, UNED)

http://nlp.uned.es/web-nlp/index.php?option=com_content&view=article&id=9

Profesor: Dr. Enrique Amigó Cabrera
(Dpto. Lenguajes y Sistemas Informáticos, UNED)

http://nlp.uned.es/~enrique/

2.CONTEXTUALIZACIÓN

Esta asignatura está encuadrada en

-el Módulo ESP-IA-1: Sistemas Inteligentes de diagnóstico, planificación y control (30 créditos, 1er. semestre), línea de especialización optativa que incluye materias que describen la aplicación de los fundamentos y métodos en diferentes áreas, tales como la Visión Artificial y la Robótica Perceptual y Autónoma, la Minería de Datos, el Descubrimiento de Información en Textos y el Procesamiento del Lenguaje Natural.

- en la linea de especialización de Tecnologías del lenguaje en la web, en el master de Lenguajes y Sistemas Informáticos (LSI)

Permitirá por tanto al alumno poner en práctica los fundamentos y métodos adquiridos tanto en IA como en LSI para el procesamiento del Lenguaje Natural. A su vez le proporcionará conocimiento y tecnología para valorar la incorporación del lenguaje natural en diferentes aplicaciones interactivas, o que hagan uso de de información textual.

3.REQUISITOS PREVIOS RECOMENDABLES

Es importante una lectura fluida del inglés y disponer de conexión a internet. En cuanto a contenidos, este curso tiene relación estrecha con las siguientes asignaturas de la carrera de Ingeniería Informática: Teoría de autómatas, Procesadores de Lenguaje, e Introducción a la Inteligencia Artificial, que proporcionan la base en cuanto a formalismos y técnicas computacionales. Así mismo las asignaturas de programación, y especialmente aquellas en que se estudian paradigmas declarativos, constituyen un complemento interesante para cursar Procesamiento de Lenguaje Natural. El lenguaje de programación que se utilizará es Python.

4.RESULTADOS DE APRENDIZAJE

En la primera parte del curso, mediante el estudio de la bibliografía el alumno adquirirá una visión amplia de las técnicas de procesamiento de lenguaje natural en los niveles léxico, sintáctico y semántico y sus aplicaciones. Los conocimientos adquiridos a nivel teórico se pondrán en práctica mediante mediante la realización de ejercicios con una herramienta, NLTK, de libre disposición que los alumnos utilizarán para la elaboración de un analizador morfológico y un analizador sintáctico y semántico sobre un subdominio abordable del lenguaje. Paralelamente, los conocimientos adquiridos a nivel global y la capacidad de síntesis se pondrán en práctica mediante el desarrollo de una serie de resúmenes (el primero de ellos guiado por preguntas) . En la segunda parte del curso el alumno adquirirá la destreza necesaria para elaborar un sistema de procesamiento de lenguaje orientado a una tarea específica. Con este curso el alumno asimilará tanto el potencial de las técnicas existentes de procesamiento de lenguaje como de sus limitaciones, siendo capaz de analizar en qué casos es factible aplicar estas técnicas en la resolución de un problema.

5.CONTENIDOS DE LA ASIGNATURA

Estructura y contenido teórico

Tema 0. Python y NLTK

Una introducción al lenguaje de programación Python para adquirir el nivel necesario para el uso de las herramientas disponibles en NLTK.

Tema 1. Introducción

Se identifican algunos de los problemas más importantes que se plantean en el estudio y tratamiento computacional del lenguaje natural, y se da una breve descripción histórica del desarrollo de esta disciplina.

Tema 2. Autómatas finitos, procesamiento de unidades morfológico-léxicas, N-gramas

Se fijan los conceptos de expresiones regulares y los operadores asociados además de autómatas finitos y lenguajes regulares. Se introduce además el concepto de morfología en inglés y, mediante lecturas complementarias, morfología castellana. El tema aborda a continuación las técnicas de procesamiento morfológico basadas en lexicones, transductores y la aproximación de stemming. Finalmente se estudian los N-gramas. Se proponen unos ejercicios prácticos a realizar con NLTK.

Tema 3. Etiquetado sintáctico

En este tema se establece un puente entre los niveles léxico y sintáctico. Se describe la taxonomía de palabras aplicables a diferentes lenguas, y las diferentes técnicas de etiquetado sintáctico existentes.

Tema 4. Gramáticas de contexto libre para el análisis de lenguaje natural

Se introducen las estructuras de la oración, incluyendo los conceptos de constituyente, sintagmas nominales y verbales, oraciones coordinadas, y su representación mediante gramáticas de contexto libre.

Tema 5. Parsing

Este tema se centra en las técnicas fundamentales de análisis sintáctico: descendente ("top-down") y ascendente ("bottom-up").

Tema 6. Unificación de rasgos

Se describe el análisis sintáctico mediante la unificación de rasgos, su implementación y el diseño de restricciones de unificación.

Tema 7. Semántica y análisis semántico

Este tema aborda en general las diferentes técnicas de procesamiento a nivel semántico del lenguaje. Se introducen los conceptos de nivel semántico, predicados de primer orden y análisis semántico dirigido por sintaxis, entre otros. Incluye además el nivel léxico semántico en el que se describen relaciones semánticas entre palabras, y bases de datos léxico semánticas. Se propone un ejercicio práctico con la herramienta NLTK.

Tema 8. Discurso, extracción de información y resúmenes

Este tema incluye el estudio de conceptos básicos de nivel de discurso como son la segmentación y resolución de correferencias. Finalmente nos centraremos en dos tipos de aplicaciones que son hoy día muy utilizadas: la extracción de información y los resúmenes, para estudiar la clase de problemas que se plantean y el alcance de las técnicas para tratarlos. A partir de esta base, se propone un trabajo personal de carácter teórico y práctico, que pone en juego los conocimientos adquiridos en la asignatura.

Objetivos por tema y orientaciones breves

Tema 0. Procesamiento del lenguaje y Python

Objetivo: Instalar el entorno del lenguaje y las herramientas para las practicas de la asignatura y adquirir un nivel de manejo de las mismas (NLTK y Python)

Orientación: Estudiar el capitulo 1 de libro indicado para las practicas en la bibliografía y hacer los ejercicios correspondientes. Previamente será necesario descargarse Python 3.5.1 y NLTK.

Tema 1. Introducción

Objetivos: Entender el procesamiento de lenguaje natural desde una perspectiva global.

Orientaciones:Lectura del capítulo 1 del libro base y bibliografía complementaria.

Tema 2:Autómatas finitos, procesamiento de unidades morfológico-léxicas, N-gramas

Objetivos: Refrescar los conocimientos sobre expresiones regulares y autómatas finitos. Aprender los conceptos fundamentales del análisis morfológico y las técnicas algorítmicas que permiten implementarlo.

Orientaciones:Lectura del capítulo 2 del libro base (debe suponer un refresco de conceptos conocidos) estudio del capítulo 3 (3.1 a 3.8, 3.9 solo la introducción), 4 (hasta el 4.8 inclusive), y lecturas complementarias. Para la puesta en práctica de estas técnicas se pondrá a disposición del alumno unos ejercicios prácticos a realizar con módulos de NLTK. Además, se pondrá a disposición del alumno sitios WEB en donde testear analizadores existentes.

Tema 3: Etiquetado sintáctico

Objetivos: Asimilar los conceptos de etiquetado sintáctico y las dos técnicas básicas de etiquetado: por reglas y técnicas estocásticas.

Orientaciones:Capítulo 5 (hasta el 5.7) del libro base y lecturas complementarias para el etiquetado en castellano. Con caracter opcional se recomienda estudiar el capítulo 6

Tema 4: Gramáticas de contexto libre

Objetivos: Repaso de conceptos relativos a las gramáticas de contexto libre y estructuras de la oración.

Orientaciones:Capítulo 12 (hasta 12.7) del libro base y referencias a herramientas accesibles vía WEB o en NLTK.

Tema 5: Parsing

Objetivos: Estudio a fondo de las técnicas de análisis sintáctico.

Orientaciones:Capítulo 13 del libro y bibliografía complementaria

Tema 6: Estructuras de rasgos y unificación.

Objetivos: Comprender en profundidad el concepto de unificación y su aplicación en el procesamiento de lenguaje.

Orientaciones: Capítulo 15 (hasta 15.5) del libro base.

Tema 7: Semántica y análisis semántico

Objetivos: Conocer las diferentes técnicas de procesamiento a nivel semántico del lenguaje y los recursos léxico semánticos.

Orientaciones: Capítulos 17, 18, 19 (hasta 19,4) y 20.1 del libro base. Los conocimientos adquiridos en este tema serán también puestos en práctica utilizando modulos de NLTK.

Tema 8: Nivel de discurso, tareas de extracción de información y resúmenes

Objetivos: Estudio de conceptos básicos de nivel de discurso, y de las tareas mencionada

Orientaciones: Epígrafes de los capítulos 21.1, 21.4, 22.1, 22.2, 22.3.3, 23.3, 23.4,23.5, 23.6, y 23.7 del libro base.

Actividades prácticas programadas

Realización de tres trabajos de síntesis, a presentar en forma de artículos. Realización de ejercicios prácticos con módulos de NLTK y desarrollo de un proyecto que se definirá sobre la base de los conocimientos adquiridos en los temas teóricos. Muy importante: esta asignatura tiene evaluación continua, mediante la entrega de diciembre a mayo de los trabajos correspondientes en las fechas indicadas en el cronograma del curso.

Plan de trabajo

- Tema 0- Python y NLTK- 10 horas (semana 0)

- Tema 1. Introducción: 3 horas (semana 1)
- Tema 2: Autómatas finitos, procesamiento de unidades morfológico-léxicas: 6 horas (semanas 2, 3 y 4)
- Ejercicios de análisis morfológico: 10 horas.
- Temas 3: Etiquetado 3 horas (semana 5)
- Entrega del resumen correspondiente a temas 2 y 3, y ejercicios de morforlogía. El plazo de entrega se notificará en el entorno.
- Temas 4 y 5 : Gramáticas de contexto libre y parsing: 8 horas (semanas 6 y 7)
- Tema 6: Estructuras de rasgos y unificación: 5 horas (semana 8)
- Tema 7: Semántica y análisis semántico: 12 horas (semanas 9, 10 y 11)
- Ejercicios de análisis sintáctico y semántico: 25 horas. El plazo de realización de los ejercicios de sintaxis y semántica y la entrega del resumen correspondiente a los temas 4, 5, 6 y 7 se notificará en el entorno y en el cronograma de la asignatura.
- Tema 8: Nivel de discurso y aplicaciones de extracción y resumen: 8 horas (semana 12 y 13)
- Proyecto: 60 horas (semanas 14-22)

Se irán introduciendo en la plataforma alf de la asignatura orientaciones de manera dinámica según transcurra el curso.

6.EQUIPO DOCENTE

- MARIA FELISA VERDEJO MAILLO
- ENRIQUE AMIGO CABRERA

7.METODOLOGÍA

El curso consta de un conjunto de temas cuyo estudio se realiza con la siguiente metodología dentro de un paradigma de construcción de conocimiento:

Para cada tema, el alumno debe acceder al material propuesto por el equipo docente. Este material consta de:

- Bibliografía básica común a todos los temas. Se trata un libro con un conocimiento ya estructurado facilitando la introducción del alumno en la materia.
- Artículos científicos. Se propone la lectura de algunos artículos de carácter científico. Su contenido es más específico. Aparte de conocer su contenido, el alumno se familiarizará con la estructura y formato que deben seguir los textos de estas características.
- Enlaces web que apuntan a recursos y herramientas relacionados con el tema.

A partir de este material (y con la guía de unas preguntas para el primero), el alumno debe realizar tres síntesis (de 10-30 páginas), en forma de artículos, correspondiendo cada una a un bloque de temas, con el objetivo de sintetizar el conocimiento que ha adquirido. La elaboración del artículo se dirige a:

- Estimular la lectura detenida del material propuesto.
- Provocar la necesidad de buscar información que complete el material propuesto inicialmente. Esta búsqueda es un entrenamiento necesario en la formación del alumno como investigador. Con cada trabajo tendrá mayor capacidad para encontrar y discriminar fuentes de información relevantes, requisito para desarrollar cualquier trabajo de investigación posterior.
- Estimular una reflexión sobre el material estudiado, necesaria para poder realizar una síntesis de calidad.
- Adquirir destreza para la escritura de documentos técnicos.

Junto con la elaboración de artículos de síntesis, el alumno deberá realizar una serie de ejercicios para los que dispondrá de la herramienta NLTK. De esta forma, el alumno podrá centrarse en aspectos del lenguaje y diseño de gramáticas sin necesidad de implementar desde un principio autómatas y mecanismos de unificación. Para manejar esta herramienta es conveniente tener conocimientos del lenguaje de programación Python, por lo que la primera semana del curso se dedicará a este tema.

Los últimos meses del curso se dirigen a afianzar los conocimientos adquiridos mediante la elaboración de un proyecto en el que se pondrán en práctica las técnicas aprendidas con datos y herramientas similares a los utilizados en la practica profesional

8.BIBLIOGRAFÍA BÁSICA

ISBN(13): 9780135041963
Título: SPEECH AND LANGUAGE PROCESSING (segunda)
Autor/es: Jurafsky, Daniel ; Martin, James H. ;
Editorial: PEARSON EDUCATION

Buscarlo en libreria virtual UNED

Buscarlo en bibliotecas UNED

Buscarlo en la Biblioteca de Educación

Buscarlo en Catálogo del Patrimonio Bibliográfico

Comentarios y anexos:

El libro base (segunda edición) es imprescindible. La primera edición del libro de bibliografía básica no incluye algunos contenidos del curso.

Para los ejercicios prácticos:

Libro: Natural language processsing with Python, disponible en

http://victoria.lviv.ua/html/fl5/NaturalLanguageProcessingWithPython.pdf

Herramientas

Python (lenguaje de programación)

https://docs.python.org/3/

NLTK 3.0

http://www.nltk.org/

9.BIBLIOGRAFÍA COMPLEMENTARIA

Comentarios y anexos:
En el entorno virtual de la asignatura se pone a disposición de los alumnos material de estudio

complementario (artículos, recopilaciones y referencias a otro material disponible en la web)

10.RECURSOS DE APOYO AL ESTUDIO

La plataforma de aprendizaje en internet, proporcionará interfaz de interacción entre el alumno y sus profesores. Esta plataforma de e-Learning y colaboración permite impartir y recibir formación, gestionar y compartir documentos, crear y participar en comunidades temáticas, así como realizar proyectos online.

Además, los alumnos tendrán que descargarse la herramienta NLTK, para la realización de ejercicios prácticos

¿Qué es NLTK? Una presentación breve

http://desilinguist.org/pdf/crossroads.pdf

Para el proyecto se facilitará el acceso a los datos y herramientas necesarias para llevarlo a cabo.

11.TUTORIZACIÓN Y SEGUIMIENTO

La tutorización de los alumnos se llevará a cabo a través de una plataforma de aprendizaje en internet, que incluye comunicacion electrónica con el equipo docente, y áreas de compartición de trabajo. Los trabajos que se proponen se evaluan a lo largo del curso.

12.EVALUACIÓN DE LOS APRENDIZAJES

La evaluación se realizará sobre los trabajos: ejercicios prácticos, síntesis, así como del proyecto personal. Cada trabajo, definido como una tarea en el entorno de la asignatura, tiene un unico plazo de entrega. Estos plazos están también especificados en el cronograma del curso donde pueden visualizarse conjuntamente. Es necesario entregar en su plazo correspondiente tanto las síntesis como los ejercicios y el proyecto.

MUY IMPORTANTE : No hay mas que un único plazo de entrega para cada uno de los trabajos. No hay examen final.

13.COLABORADORES DOCENTES

Véase equipo docente.