Máster en Business Analytics y Big Data

Edición 2015 / 2016

Módulo: Procesamiento de Lenguaje Natural Y Minería de Textos
Coordinador: Manuel Lucania, manuellucania@campusciff.net

OBJETIVOS

El objetivo general del módulo es dotar al alumno de la capacidad para entender las teorías básicas y los algoritmos que sustentan los algoritmos de Procesamiento de Lenguaje Natural así como de asegurarle una capacidad de resolución práctica que le permita aplicar las herramientas más conocidas del área a posibles problemas que las empresas afrontan hoy en día.

El énfasis de la asignatura está en la capacidad del alumno para proponer e imaginar soluciones reales a problemas concretos de NLP. Por ello, la práctica individual que deberá desarrollar y presentar, tiene el mayor peso dentro de la asignatura.

Objetivos concretos del módulo:

1. Comprensión teórica de las técnicas, fundamentos y aplicaciones prácticas del PLN
2. Manejo práctico de las distintas librerías y herramientas de NLP
3. Capacidad creativa para imaginar un problema y proponer una solución práctica

METODOLOGÍA

La primera sesión comenzará con una introducción puramente teórica, para a continuación combinar prácticas y trabajo en grupo con teoría. Para la segunda sesión, se repetirá de nuevo el esquema de combinar prácticas con teoría.

En cada sesión el profesor explicará los requerimientos de instalación necesarios para la siguiente sesión. Básicamente se requerirá Anaconda, NLTK y Gate para ambas sesiones.

En las aulas se crearán unos grupos de trabajo, y en cada sesión se iniciarán las prácticas de grupo (prácticas que no tendrán mucha carga o complejidad). Pero el verdadero peso de la asignatura recaerá sobre la práctica individual que cada alumno escoja y desarrolle fuera de las aulas y presente a toda la clase en la última sesión impartida.

PROGRAMA

Sesión 1:

Introducción histórica y tecnológica del NLP
Actividades: Esta sesión será puramente teórica. Se dará una visión introductoria del NLP, se explicará la dinámica de la asignatura y se formarán equipos de trabajo para las siguientes sesiones.
Materiales: Material teórico proporcionado por el profesor.

Cadenas de procesamiento NLP I : Tokenizer, Sentence splitter, NER, POS tagger, Lemmatizer.
Actividades: Se explicará la importancia de las cadenas de procesamiento en los sistemas basados en NLP, sus distintos elementos y los principales algoritmos. Se mostrará y practicará la llamada a librerías NLP (NLTK).
Materiales: Material teórico proporcionado por el profesor. Necesario tener instalada la librería NLTK.

Herramientas NLP : NLTK, BRAT y Gate
Actividades: Primera actividad puramente práctica, en la que se profundizará fundamentalmente en 2 herramientas que se han usado anteriormente: la librería NLTK y el visualizador BRAT. Se verán todas las posibilidades que ofrece el Framework Gate, tanto para la programación como para la visualización de cadenas NLP.
Materiales: Material teórico proporcionado por el profesor.

Sesión 2:

Cadenas de procesamiento NLP II : Resolución de correferencias, WSD, Gramáticas.
Actividades: Se explicarán los elementos terminales y de mayor complejidad dentro de las cadenas de procesamiento: sistemas de resolución de correferencias, y sistemas de desambigüación semántica. Se explicarán las gramáticas (fundamentalmente las basadas en Autómatas de Estados Finitos) y se demostrará su utilidad mediante prácticas con gramáticas Jape de Gate.
Materiales: Material teórico proporcionado por el profesor. Necesario tener instalada la librería Gate.

Text Mining.
Actividades: Se detallarán las principales técnicas y se usarán las librerías de Anaconda para solucionar un caso práctico cuyos datos el profesor distribuirá en clase.
Materiales: Material teórico proporcionado por el profesor. Necesario tener instaladas las librerías de Machine Learning de Anaconda.

Aplicaciones del NLP: Information Extraction, Sentiment Analysis, Conversational Agents, Semantic Search
Actividades: Se detallarán ejemplos de aplicaciones y técnicas de NLP que resuelven problemas de nuestro mundo real. Se hará una práctica de Information Extraction. Además, al finalizar la sesión cada uno de los alumnos deberá realizar una brevísima presentación pública de la idea que está desarrollando en su práctica individual.
Materiales: Material teórico proporcionado por el profesor.

MATERIALES

No es necesario adquirir ningún libro para el seguimiento del programa. Como libro de consulta para aquellos alumnos que quieran profundizar en algún aspecto de la asignatura, se propone:
Handbook of Natural Language Processing, Second Edition (Nitin Indurkhya, Fred J. Damerau)

También se recomienda la consulta de los artículos académicos publicados en la Revista de la SEPLN (Sociedad Española de Procesamiento de Lenguaje Natural), cuya versión online es gratuita:
http://www.sepln.org/

EVALUACIÓN

Niveles de consecución de los objetivos

Objetivo específico	*Nivel alto*	*Nivel medio*	*Nivel bajo*
O1 – Comprensión teórica de las técnicas, fundamentos y aplicaciones prácticas del PLN	El alumno es capaz de responder en clase a las preguntas que el profesor le formule acerca de sesiones anteriores. Participa activamente, hace preguntas en clase.	El alumno muestra interés, sin participar demasiado, y no respondiendo correctamente siempre a las preguntas del profesor.	Falta de interés por la asignatura, incapacidad para responder a las preguntas del profesor, falta de participación.
O2 - Manejo práctico de las distintas librerías y herramientas de NLP	El alumno lidera los equipos de trabajo en los que participa. Es capaz de usar con soltura las herramientas que el profesor ha enseñado en clase, y de aplicarlas en su práctica individual.	El alumno participa en los equipos y es capaz de aprender y usar tanto en su práctica individual como en las prácticas colectivas los conocimientos de librerías y herramientas.	El alumno a duras penas es capaz de usar las herramientas que se muestran en clase, y no colabora en los equipos de trabajo.
O3 - Capacidad creativa para imaginar un problema y proponer una solución práctica	El alumno, en su práctica individual hace una propuesta original, propia para resolver un problema distinto a los sugeridos por el profesor. Es capaz de implementar la solución, y su programa funciona según lo esperado.	La propuesta del alumno para su práctica individual se enmarca dentro de las sugerencias del profesor. El sistema programado funciona bien, aunque sin mucho esfuerzo en la parte de programación.	El alumno escoge las propuestas más sencillas del profesor, y su programa no es capaz de ejecutarse con éxito.

Modelo de evaluación

Elemento	*Peso*
Práctica individual	90,00%
Práctica grupal y participación en clase	10,00%

PROFESORADO

Manuel Lucania es un Ingeniero Informático experto en diseño de soluciones software de PLN (Procesamiento de Lenguaje Natural) e Inteligencia Artificial. A lo largo de su carrera profesional ha alternado la dirección de proyectos de Investigación en Lingüística Computacional y Text Mining con el desarrollo de productos con un fuerte componente de I+D.

Actualmente desempeña el cargo de CTO dentro una empresa cuyo lanzamiento se hará público en 2016. Anteriormente, como Ingeniero NLP ha pasado por diversas start-ups, la mayor parte de las veces creando productos desde cero, y en otras ocasiones agregando nuevos módulos y funcionalidades a productos ya existentes:

TAIGER
Se ha encargado de añadir nuevas funcionalidades a los productos de búsqueda semántica de Taiger, desarrollar ontologías y crear asistentes virtuales orientados a la industria bancaria. Ha desarrollado sistemas de NER/Information Extraction para la identificación automática de nombres de empresas, puestos de trabajo y lugares, con el objeto de normalizar la información procedente de curriculums y ofertas laborales publicadas en la web.

CogniCor Technologies
Ha liderado el equipo NLP de una innovadora herramienta de resolución automática de quejas de usuario desarrollando modelos basados en el Spanish Framenet y en la definición de Frames, Lexical Units, Frame Elements y el ASRL (Automatic Semantic Role Labelling). Además ha creado Agentes Virtuales (chat bots) para el sector bancario de habla inglesa.

Oteara (anteriormente Ximetrix)
Como socio fundador, ha diseñado Oraquo, una herramienta de rastreo de opiniones en Internet acerca de marcas, productos y personas. Ha dirigido proyectos de I+D con la Universidad Pompeu Fabra y la Universidad del País Vasco, donde se empleaban técnicas de Opinion Mining y se usaba el analizador multi-idiomas Freeling.

ASOMO
Ha liderado proyectos de Text Mining, Sentiment Analysis y Reputation Monitoring. web crawling y scraping, tokenization, stemming y chunking. Ha programado un corrector ortográfico avanzado basado en la distancia fonética. Ha liderado colaboraciones con el Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla (TALP) de la Universidad Politécnica de Cataluña y con el Intelligent Systems Group (ISG) de la Universidad del País Vasco.