ASIGNATURA DE MÁSTER:

DESCUBRIMIENTO DE INFORMACIÓN EN TEXTOS

Curso 2017/2018

(Código:31101254)

1.PRESENTACIÓN

Ficha técnica:

- Tipo: Optativa
- Duración: Anual
- Créditos Totales y Horas: 6 / 150
- Horas de estudio teórico: 70
- Horas de trabajo práctico: 70
- Horas de actividades complementarias: 10

Reseña del Profesorado:

MARTÍNEZ UNANUE, RAQUEL:

Ha realizado la mayor parte de su actividad docente en el campo de la programación, la algoritmia, la documentación electrónica y la minería de textos. Su actividad investigadora reciente se centra en la minería de textos, especialmente en clustering de documentos tanto monolingües como multilingües, aplicada a diversos tipos de textos (páginas web, noticias, redes sociales, ...) y dominios, en particular el dominio médico.

Desde el año 2000 hasta la actualidad ha colaborado en programas de doctorado de tres universidades: la Universidad Complutense de Madrid, la Universidad Rey Juan Carlos y la UNED.

e.mail: raquel@lsi.uned.es

ARAUJO SERNA, LOURDES:

Desde 1990 ha desarrollado en universidades públicas diversa actividad docente relacionada con los lenguajes de programación y la algoritmia. Desde 1994 hasta la actualidad ha colaborado de forma continua en programa de doctorado de la Unvirsidad Complutense de Madrid y de la UNED. Su tema de tesis fue el estudio del paralelismo de Prolog y posteriormente ha trabajado en programación con restricciones sobre arquitecturas paralelas, ámbito en el que comenzó a aplicar técnicas de programación evolutiva. En la actualidad investiga en procesamiento del lenguaje natural, recuperación de información y en la aplicación de programación evolutiva a dichas áreas.

e.mail: lurdes@lsi.uned.es

FRESNO FERNÁNDEZ, VÍCTOR

Su actividad docente se ha centrado principalmente en el campo de la documentación electrónica y su investigación en el campo de la representación automática de textos, en especial de páginas web, así como en la clasificación y clustering de documentos HTML. Desde el año 2000 hasta la actualidad ha trabajado en el Instituto de Automática industrial (CSIC), la Universidad Rey Juan Carlos (URJC) y la Universidad Nacional de Educación a Distancia (UNED), colaborando en los programas de doctorado de dichas universidades.

e.mail: vfresno@lsi.uned.es

2.CONTEXTUALIZACIÓN

Esta asignatura de caracter optativo se imparte, tanto en el Máster Universitario en "Inteligencia Artificial Avanzada: Fundamentos, Métodos y Aplicaciones" como en el Máster Universitario en "Lenguajes y Sistemas Informáticos" de la ETSI Informática de la UNED. Esta asignatura es de carácter anual con una carga de 6 ECTS.

3.REQUISITOS PREVIOS RECOMENDABLES

Ninguno diferente de los generales de acceso a este programa de posgrado orientado a la investigación.

4.RESULTADOS DE APRENDIZAJE

El objetivo del curso es proporcionar al alumno una visión global de las técnicas y tecnologías involucradas en el descubrimiento de información en textos.

El aprendizaje está diseñado para permitir que el alumno adquiera una serie de *destrezas y competencias* que se enumeran a continuación:

- Saber lo que es un corpus y conocer los criterios por los que se clasifican, los tipos de anotaciones más comunes y los estándares utilizados.
- Conocer los modelos de representación comúnmente utilizados, así como los métodos de selección y reducción del número de rasgos.
- Saber distinguir los diversos niveles de información lingüística que se pueden utilizar en la representación de textos.
- Saber qué se entiende por minería de textos y conocer las principales técnicas y tecnologías implicadas.
- Saber qué es la clasificación automática de textos y sus características y tipos.
- Conocer diversos tipos de técnicas de aprendizaje automático que se pueden utilizar en la clasificación automática de textos.
- Conocer los modelos estadísticos más utilizados en el procesamiento del lenguaje.
- Saber utilizar las herramientas disponibles de clasificación automática de textos y tener criterios para seleccionar las más adecuadas.
- Saber qué es el clustering de textos y sus características y tipos.
- Conocer diversos tipos de algoritmos de clustering.
- Saber utilizar las herramientas disponibles de clustering de textos y tener criterios para seleccionar las más adecuadas.
- Conocer algoritmos de etiquetado léxico y análisis sintáctico.

Como actividades formativas se tienen:

1. Actividades teóricas interaccionando con equipos docentes, tutores y compañeros.
   Resolución de dudas de contenido teórico de forma presencial, vía telefónica o en línea sobre la metodología, los contenidos o las actividades a realizar. Intercambio de información a través de un foro virtual.
2. Actividades prácticas interaccionando con equipos docentes, tutores y compañeros.

Resolución de dudas de contenido práctico de forma presencial, vía telefónica o en línea sobre la metodología, los contenidos o las actividades a realizar.Intercambio de información a través de un foro virtual.

3. Actividades teóricas desempeñadas autónomamente.
   Lectura reflexiva y crítica de las orientaciones metodológicas de la asignatura. Estudio de los materiales didácticos.
4. Actividades prácticas desempeñadas.
   Elaboración de trabajos individuales originales.

5.CONTENIDOS DE LA ASIGNATURA

Estructura y el contenido teórico de la asignatura se detalla a continuación:

Tema 1.- Introducción: interés y definiciones preliminares.
Tema 2.- Corpus: definiciones y tipología
Tema 3.- Estándares de anotaciones
Tema 4.- Modelos estadísticos para la caracterización de textos: Etiquetado léxico y sintáctico.
Tema 5.- Representación de textos: Modelos y funciones de pesado y reducción de rasgos.
Tema 6.- Técnicas de minería de textos. Clustering
Tema 7.- Técnicas de minería de textos. Clasificación automática.

Objetivos por tema y orientaciones breves:

Tema 1. Introducción: interés y definiciones preliminares.

- Objetivos: El objetivo global del tema es presentar al alumno aquellos conceptos y conocimientos preliminares sin los que no podría ubicar los contenidos de la asignatura. Se pretende, además, justificar el interés de la asignatura, motivar al alumno en su estudio y presentar las posibles aplicaciones de los contenidos.
- Orientaciones: Dentro de las actividades de aprendizaje se especifican las lecturas más adecuadas para cada uno de los objetivos del tema.

Tema 2. Corpus: definiciones y tipología.

- Objetivos: Se pretenden presentar las diversas definiciones de corpus existentes desde diversos puntos de vista, además de clasificarlos de acuerdo a diversos criterios comúnmente utilizados en la bibliografía.
- Orientaciones:Dentro de las actividades de aprendizaje se especifican las lecturas más adecuadas para cada uno de los objetivos del tema. Se presentarán y facilitará el acceso a numerosos ejemplos de tipos de corpus.

Tema 3. Estándares de anotaciones.

- Objetivos: Se pretende presentar el concepto de anotación, los tipos de anotaciones y los estándares de anotaciones en XML.
- Orientaciones: Dentro de las actividades de aprendizaje se especifican las lecturas más adecuadas para cada uno de los objetivos del tema. Se presentarán y facilitará el acceso a numerosos ejemplos de tipos de corpus con anotaciones, en particular XML.

Tema 4. Modelos estadísticos para la caracterización de textos: Etiquetado léxico y sintáctico.

- Objetivos: Dar a conocer al alumno los modelos estadísticos más utilizados en el procesamiento del lenguaje natural, tales como los Modelos de Markov Ocultos y las Gramáticas probabilísticas. También se darán a conocer algoritmos basados en estos

modelos para abordar problemas específicos de PLN.
- Orientaciones: Dentro de las actividades de representación y aprendizaje se especifican las lecturas más adecuadas para cada uno de los objetivos del tema. Se presentarán y facilitará el acceso a ejemplos y herramientas para el etiquetado léxico y el análisis sintáctico.

Tema 5. Representación de textos: Modelos y funciones de pesado y de reducción de rasgos.

- Objetivos: Se presentarán los modelos de representación más utilizados. Además se estudiarán los métodos de selección y reducción de rasgos más comunes en textos.
- Orientaciones: Dentro de las actividades de representación y aprendizaje se especifican las lecturas más adecuadas para cada uno de los objetivos del tema. Se presentarán y facilitará el acceso a ejemplos y herramientas para la selección de rasgos.

Tema 6. Técnicas de minería de textos. Clustering.

- Objetivos: Se presentará el campo de la minería de textos ubicando el clustering o agrupamiento automático en él. Se presentarán las principales técnicas y algoritmos de clustering, así como las técnicas que se suelen utilizar en su evaluación.
- Orientaciones: Dentro de las actividades de aprendizaje se especifican las lecturas más adecuadas para cada uno de los objetivos del tema. Se presentarán y facilitará el acceso a ejemplos y herramientas para clustering.

Tema 7. Técnicas de minería de textos. Clasificación automática.

- Objetivos: Se presentará el campo de la minería de textos ubicando la clasificación automática en él. Se presentarán las principales técnicas y algoritmos de aprendizaje aplicados a clasificación, así como las técnicas que se suelen utilizar en su evaluación.
- Orientaciones: Dentro de las actividades de aprendizaje se especifican las lecturas más adecuadas para cada uno de los objetivos del tema. Se presentarán y facilitará el acceso a ejemplos y herramientas para la clasificación automática.

Las actividades prácticas programadas son:

- Corpus. Tipologías.
- Extracción de información lingüística a partir de anotaciones. Estándares de anotaciones. Salidas de herramientas de uso frecuente.
- Representaciones de diversos tipos de documentos. Uso de metainformación.
- Clustering: algoritmos partitivos y jerárquicos.
- Clasificación automática: aprendizaje supervisado y semisupervisado.

Otras actividades programadas se irán generando de forma dinámica en el curso virtual.

6.EQUIPO DOCENTE

- M. LOURDES ARAUJO SERNA
- RAQUEL MARTINEZ UNANUE
- VICTOR DIEGO FRESNO FERNANDEZ

7.METODOLOGÍA

La metodología es la general del programa de postgrado; junto a las actividades y enlaces con fuentes de información externas, existe material didáctico propio preparado por el

equipo docente. Se trata de una metodología adaptada a las directrices del EEES, de acuerdo con el documento del IUED.

La asignatura no tiene clases presenciales. Los contenidos teóricos se impartirán a distancia, de acuerdo con las normas y estructuras de soporte telemático de la enseñanza en la UNED.

El material docente incluye un resumen de los contenidos de cada tema y distintos tipos de actividades relacionadas con la consulta bibliográfica y la utilización de herramientas asociadas a las tecnologías y técnicas presentadas en el curso. Tratándose de un master orientado a la investigación, las actividades de aprendizaje se estructuran en torno al estado del arte en cada una de las materias del curso y a los problemas en los que se va a focalizar la práctica que el alumno deberá realizar.

8.BIBLIOGRAFÍA BÁSICA

Comentarios y anexos:

Como bibliografía de la asignatura se deberán estudiar capítulos seleccionados de las siguientes referencias:

- Gordon, A.D. Classification. 2nd Edition. Chapman & Hall/CRC, 1999.
- McEnery, T., Wilson, A. Corpus Linguistics. Edinburgh University Press, 1996.
- Manning, C.D., and Schütze, H. Foundations of Statistical Natural Language Processing. The MIT Press (2000).
- Mitchell, T. Machine Learning. McGraw Hill, 1997. (Nuevos capítulos creados en 2006 y disponibles en http://www.cs.cmu.edu/%7Etom/mlbook.html)
- S. Weiss; N. Indurkhya; T. Zhang; F. Damerau. Text Mining: Predictive Methods for Analyzing Unstructured Information, 2004.

9.BIBLIOGRAFÍA COMPLEMENTARIA

10.RECURSOS DE APOYO AL ESTUDIO

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

Se ofrecerán las herramientas necesarias para que, tanto el equipo docente como el alumnado, encuentren la manera de compaginar tanto el trabajo individual como el aprendizaje cooperativo.

11.TUTORIZACIÓN Y SEGUIMIENTO

La tutorización de los alumnos se llevará a cabo a través de la plataforma de e-Learning Alf, por teléfono y por correo electrónico:

- Raquel Martínez (coodinadora)
  email: raquel@lsi.uned.es
  Tfno: 913988725
  Horario guardias: Jueves de 11:30 a 13.30 y de 14.30 a 16:30

- Lourdes Araujo
  email: lurdes@lsi.uned.es
  Tfno: 913987318

Horario de guardias: Jueves de 11 a 13.30 y de 14.30 a 16:00

- Víctor Fresno
  email: vfresno@lsi.uned.es
  Tfno: 913988217
  Horario guardias: Jueves de 15:00 a 19:00

12.EVALUACIÓN DE LOS APRENDIZAJES

La evaluación final de la asignatura se realizará como:

1. Evaluación continua a través del seguimiento del alumno.
2. Evaluación continua a través de la realización de trabajos.
3. Evaluación global del proceso de aprendizaje y adquisición de competencias y conocimientos.
4. Calificación numérica de 1 a 10 según legislación vigente (RD 1125/2003)

A partir del tema 2, cada tema tiene asociada una tarea obligatoria cuya entrega en plazo es un requisito imprescindible para aprobar la asignatura. La realización correcta de todas las tareas obligatoria asegura una nota de aprobado.

Aquellos alumnos que deseen una calificación mayor pueden elegir uno de entre una serie de trabajos optativos que se proponen.

13.COLABORADORES DOCENTES

Véase equipo docente.