23-24

MÁSTER UNIVERSITARIO EN
TECNOLOGÍAS DEL LENGUAJE

GUÍA DE ESTUDIO PÚBLICA

MINERÍA DE LA WEB

CÓDIGO 31101023

UNED

23-24

MINERÍA DE LA WEB

CÓDIGO 31101023

ÍNDICE

Nombre de la asignatura	MINERÍA DE LA WEB
Código	31101023
Curso académico	2023/2024
Título en que se imparte	MÁSTER UNIVERSITARIO EN TECNOLOGÍAS DEL LENGUAJE
Tipo	CONTENIDOS
Nº ETCS	6
Horas	150.0
Periodo	ANUAL
Idiomas en que se imparte	CASTELLANO

PRESENTACIÓN Y CONTEXTUALIZACIÓN

El curso se dirige a conocer las tecnologías existentes para extraer información de la web, tanto a partir de sus contenidos textuales (recuperación de información, extracción de información, creación de recursos lingüísticos, etc.) como de su estructura y su uso. Al finalizar el curso, el alumno deberá ser capaz de plantear la arquitectura completa de un sistema automático de acceso y extracción de información en la web.

Reseña del Profesorado:

Anselmo Peñas pertenece a la UNED (Universidad Nacional de Educación a Distancia) donde es Catedrático de Informática. Se incorporó al grupo investigación en Procesamiento del Lenguaje Natural de la UNED en 1998, obteniendo su doctorado con distinción especial y premio en 2002. También ha sido galardonado con el Premio de la Sociedad Española para el Procesamiento del Lenguaje Natural.

Más información: https://portalcientifico.uned.es/investigadores/183221/detalle

Laura Plaza forma parte del grupo NLP&IR de la UNED, donde es Profesora Titular de Universidad. Sus líneas de investigación se centran fundamentalmente en la generación automática de resúmenes y la clasificación de textos biomédicos, el estudio y detección de sexismo en redes sociales, la detección de desinformación en la web y el desarrollo de recursos no sesgados para tareas de Procesamiento de Lenguaje Natural. Ha realizado estancias de investigación en la Universidad de Sheffield y en el Royal Melbourne Institute of Technology. Ha trabajado como docente e investigadora en la Universidad Complutense de Madrid, en la Universidad Autónoma de Madrid y en la Universidad Nacional de Educación a Distancia.

REQUISITOS Y/O RECOMENDACIONES PARA CURSAR ESTA ASIGNATURA

Esta asignatura puede ser cursada aisladamente, aunque el estudiante se beneficiaría si hubiera cursado previamente o curse en paralelo la asignatura de *Fundamentos del procesamiento lingüístico,* además de *Diseño e implementación de Sistemas Informáticos*. Se requiere una lectura fluida de textos en inglés.

EQUIPO DOCENTE

Nombre y Apellidos	ANSELMO PEÑAS PADILLA (Coordinador de asignatura)
Correo Electrónico	anselmo@lsi.uned.es
Teléfono	91398-7750
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

Nombre y Apellidos	LAURA PLAZA MORALES
Correo Electrónico	lplaza@lsi.uned.es
Teléfono	
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

HORARIO DE ATENCIÓN AL ESTUDIANTE

La tutorización de los alumnos se llevará a cabo a través de la plataforma de e-Learning Alf y del correo electrónico de los profesores de la asignatura.

Información de contacto

Anselmo Peñas Padilla (Coordinador). Dpto. Lenguajes y Sistemas Informáticos (U.N.E.D)
e-mail: anselmo@lsi.uned.es
Horario de Asistencia al estudiante: Jueves de 9:30 a 13:30 horas

Laura Plaza. Dpto. Lenguajes y Sistemas Informáticos (U.N.E.D)
e-mail: lplaza@lsi.uned.es
Horario de Asistencia al estudiante: Jueves de 10 a 14 horas

COMPETENCIAS QUE ADQUIERE EL ESTUDIANTE

Competencias Básicas:

CB6 - Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación
CB7 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio
CB8 - Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB9 - Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades
CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

Competencias Generales:
CPG1 - Adquirir capacidad de abstracción, análisis, síntesis y relación de ideas.
CPG2 - Adquirir capacidad crítica y de decisión
CPG3 - Adquirir capacidad de estudio y autoaprendizaje
CPG4 - Adquirir capacidad creativa y de investigación
CPG5 - Adquirir habilidades sociales para el trabajo en equipo

Competencias Específicas:
CE1 - Adquirir capacidad de comprender y manejar de forma básica los aspectos más importantes relacionados con los lenguajes y sistemas informáticos en general y, de manera especial, en los siguientes ámbitos: Tecnologías del lenguaje y de acceso a la información en web
CE3 - Adquirir capacidad de estudio de los sistemas y aproximaciones existentes y para distinguir las aproximaciones más efectivas.
CE4 - Adquirir capacidad para detectar carencias en el estado actual de la ciencia y la tecnología
CE5 - Adquirir capacidad para proponer nuevas aproximaciones que den solución a las carencias detectadas.
CE7 - Adquirir capacidad para proponer y llevar a cabo experimentos con la metodología adecuada como para poder extraer conclusiones y determinar nuevas líneas de actuación e investigación.

RESULTADOS DE APRENDIZAJE

1. Tener una visión amplia de las áreas relacionadas con la extracción de información en la web.
2. Hábito de lectura de artículos científicos.
3. Capacidad para 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.
4. Capacidad de reflexión sobre el material estudiado, necesaria para poder realizar una síntesis de calidad.
5. Capacidad para escribir textos con un formato de artículo científico, tanto en lo referente a la estructuración de contenidos, como de formato del propio artículo.

Objetivos por tema

Tema 1. Introducción

Objetivos:

O.1.1 Determinar los problemas que surgen al interactuar con la web.

O.1.2 Definir Minería de la web

O.1.3 Definir Crawling

O.1.4 Definir Búsqueda en web

O.1.5 Definir Minería de contenido de la web (minería de texto)

O.1.6 Definir Minería de uso de la web

O.1.7 Definir Minería de estructura de la web

O.1.8 Definir Dinámica de la web.

Tema 2. Consulta y búsqueda en web

Objetivos:

O.2.1 Determinar las características propias de la web que afectan a la búsqueda.

O.2.2 Caracterizar los tipos de información a considerar en la búsqueda en web (Contenido textual, Información en los enlaces, Estructura de enlace entre páginas, etc.).

O.2.3 Estudiar interfaces de exploración y visualización de la búsqueda.

O.2.4 Definir Metabúsqueda y Agentes web.

Tema 3. Minería de textos

Objetivos:

O.3.1 Definir corpus.

O.3.2 Comprender cómo se puede crear y usar un corpus a partir de la web.

O.3.3 Definir Extracción de Información textual.

O.3.4 Conocer la arquitectura de un sistema de Extracción de Información.

O.3.5 Definir Extracción de terminología.

O.3.6 Conocer alguna metodología de extracción de terminología a partir de la web.

O.3.7 Identificar la problemática asociada al lenguaje natural.

Tema 4. Minería de uso de la web

Objetivos:

O.4.1 Definir y establecer los objetivos de minería de uso de la web.

O.4.2 Determinar las etapas de procesamiento (Preprocesamiento, Inferencia de patrones, Análisis de patrones).

O.4.3 Conocer algunas herramientas existentes.

O.4.4 Identificar técnicas de aprendizaje aplicadas a minería de uso.

O.4.5 Saber qué son los sitios web adaptativos.

Tema 5. Minería de estructura de la web

Objetivos:
O.5.1 Definir y establecer los objetivos de la minería de estructura de la web.
O.5.2 Definir y modelar las nociones de Autoridad (authoritative page), prestigio, Centralidad y Co-cita.
O.5.3 Conocer cómo se realiza el ranking de páginas web basado en enlaces: PageRank y HITS.
O.5.4 Estudiar cómo se realiza el análisis de comunidades en la web.

Tema 6. Dinámica de la web

Objetivos:
O.6.1 Definir y establecer los objetivos del estudio de la dinámica de la web.
O.6.2 Determinar las características de la web susceptibles de estudio.
O.6.3 Estudiar la Ley de Zipf, "power laws" en la web así como sus aplicaciones.
O.6.4 Comprender cómo se determina el tamaño y tendencia de crecimiento de la web.
O.6.5 Comparar las web pública y web oculta.
O.6.6 Comprender cómo se determina la presencia de un idioma en la web.
O.6.7 Conocer estudios sobre la web española.

CONTENIDOS

TEMA 1: INTRODUCCIÓN

El primer tema persigue introducir al alumno en los contenidos que serán tratados en la asignatura. El alumno adquirirá, en primer lugar, conocimiento acerca de los problemas habituales que los usuarios y sistemas experimentan en su interacción con la web. Seguidamente, se presentará al alumno una definición “inicial” de los principales conceptos relacionados con la extracción de información en la web, tanto a partir de sus contenidos como de su estructura y su uso. Estos conceptos serán desarrollados en detalle en los siguientes temas.

Contenido detallado:
1. Problemas que surgen al interactuar con la web.
2. Minería de la web.
3. Crawling.
4. Búsqueda en web.
5. Minería de contenido de la web (minería de texto).

6. Minería de uso de la web.
7. Minería de estructura de la web.
8. Dinámica de la web.

TEMA 2: CONSULTA Y BÚSQUEDA EN WEB

Este tema profundiza en el concepto de búsqueda en la web, ahondando fundamentalmente en los problemas que plantea tanto para usuarios como para desarrolladores de motores de búsqueda, así como en los diferentes tipos de contenidos presentes en la web.

Contenido detallado:

1. Características propias de la web que afectan a la búsqueda.
2. Tipos de información a considerar en la búsqueda en web.
1. Contenido textual.
2. Información en los enlaces.
3. Estructura de enlace entre páginas.
4. Otros tipos de información
3. Interfaces, browsing y visualización de la búsqueda.
4. Metabúsqueda.
5. Agentes web.

TEMA 3: MINERÍA DE TEXTOS

El tercer tema se desglosa en tres bloques principales. El primero desarrolla el concepto de corpus, desde su definición hasta los posibles usos y utilidades que proporciona en el contexto de la minería de la web. El segundo bloque trata los problemas de la extracción de información textual y de terminología en la web. Finalmente, se incluye un tercer bloque donde se presentan las tareas de clasificación, clustering y cálculo de la similitud textual.

Contenido detallado:

1. Qué es un corpus.
2. Creación de corpus.
3. Posibles usos y utilidad de un corpus.
4. Creación de corpus a partir de la web.
5. Ejemplos de algunos corpus y su finalidad.
6. Extracción de Información textual (Automatic Information Extraction).
7. Arquitectura de un sistema de EI.
8. Extracción de terminología.
9. Extracción de terminología a partir de la web.

10. Problemática asociada al lenguaje natural.

TEMA 4: MINERÍA DE USO DE LA WEB

El cuarto tema establece y desarrolla los objetivos de la minería de uso de la web, definiendo las etapas involucradas en el proceso de inferencia y análisis de patrones y presentando algunas de las técnicas de aprendizaje más utilizadas en el contexto de la minería de uso de la web.

Contenido detallado:

1. Definición y objetivos de minería de uso de la web.
2. Etapas de procesamiento.
1. Preprocesamiento.
2. Inferencia de patrones.
3. Análisis de patrones.
3. Herramientas existentes.
4. Técnicas de aprendizaje aplicadas a minería de uso.
5. Sitios web adaptativos.

TEMA 5: MINERÍA DE ESTRUCTURA DE LA WEB

El quinto tema establece y desarrolla los objetivos de la minería de estructura de la web, presentando y definiendo los conceptos de “autoridad”, “prestigio”, “centralidad” y “co-cita”, fundamentales en el estudio de la estructura de la web. Se introducen también algunos de los algoritmos más populares para el ranking de páginas web y el análisis de comunidades.

Contenido detallado:

1. Definición y objetivos de la minería de estructura de la web.
2. Definición, modelado y uso de las nociones de autoridad (authoritative page), prestigio, centralidad y co-cita.
3. Ranking de páginas web basado en enlaces: PageRank y HITS.
4. Análisis de comunidades en la web.
5. Otras aplicaciones de la minería de estructura.

TEMA 6: DINÁMICA DE LA WEB

El sexto tema establece y desarrolla los objetivos y fundamentos del estudio de la dinámica de la web, con el fin último de comprender cómo evoluciona la web (tanto en términos de tamaño como de otras características como el idioma, la distribución geográfica, etc.).

Contenido detallado:

1. Definición y objetivos del estudio de la dinámica de la web.
2. Características de la web susceptibles de estudio.
3. Ley de Zipf, "power laws" en la web.
4. Tamaño y tendencia de crecimiento de la web.
5. Web pública y web oculta.
6. Idiomas en la web.
7. Dominios en la web.
8. Estudios sobre la web española.

METODOLOGÍA

La asignatura consta de seis 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 de libros 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 y de más difícil lectura. A partir de ellos, el alumno conocerá la estructura y formato que deben seguir los textos de estas características y que el tendrá que escribir más adelante.
- Enlaces web: enlaces que apuntan a sitios web donde encontrar nuevas referencias bibliográficas, enlaces a sitios web con recursos y herramientas relacionados con el tema, enlaces a otros cursos o tutoriales, etc.

A partir de este material y con la guía de un cuestionario, el alumno debe realizar un resumen sintetizando el conocimiento que ha adquirido. La elaboración del resumen 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.
- Aprender a escribir textos con un formato de artículo científico, tanto en lo referente a la estructuración de contenidos, como de formato del propio artículo. En especial,

contextualizar la síntesis referenciando correctamente las fuentes utilizadas.
Tras la elaboración del resumen, el alumno debe realizar una entrega electrónica de su resumen y de los nuevos enlaces y referencias más importantes que ha encontrado a lo largo de su trabajo. Esto servirá de material de evaluación para el equipo docente, que podrá valorar no sólo los conocimientos adquiridos, sino también la evolución y el progreso del alumno en la adquisición de la metodología y actitud necesaria para un investigador.
Los últimos meses del curso se dirigen a afianzar los conocimientos adquiridos mediante la elaboración de un trabajo final de carácter personal. El trabajo puede ser propuesto por el propio alumno y preferiblemente deberá tener un carácter de aplicación de los conocimientos adquiridos.

SISTEMA DE EVALUACIÓN

TIPO DE PRIMERA PRUEBA PRESENCIAL

Tipo de examen: No hay prueba presencial

TIPO DE SEGUNDA PRUEBA PRESENCIAL

Tipo de examen2: No hay prueba presencial

CARACTERÍSTICAS DE LA PRUEBA PRESENCIAL Y/O LOS TRABAJOS

Requiere Presencialidad: No

Descripción

La evaluación del curso se realizará sobre la base de 6 resúmenes y un trabajo práctico.

Respecto a los resúmenes de los 6 temas del curso, el alumno deberá:

Leer los artículos científicos y referencias bibliográficas indicadas como bibliografía básica del tema (ver contenidos).

Se recomienda explorar los enlaces y lecturas sugeridas como material complementario.

Opcionalmente, buscar información que complete el material propuesto inicialmente.

Realizar el resumen correspondiente cada tema a partir de las lecturas básicas y complementarias indicadas en el apartado de contenidos, y conforme a la estructura dada.

Respecto al trabajo práctico, el alumno deberá proponer al equipo docente un trabajo personal que sirva para poner en práctica los conocimientos adquiridos en alguno de los 6 temas.

Criterios de evaluación

Respecto a los resúmenes, se valorará:
La completitud al explorar las fuentes.
La capacidad de realizar una síntesis personal, estructurando adecuadamente el conocimiento adquirido en las lecturas.
La claridad de exposición.
La correcta referencia a las fuentes utilizadas.
Respecto al trabajo práctico, se valorará:
Originalidad de la propuesta
Motivación del trabajo y objetivos planteados
Descricpión de la propuesta
Implementación
Evaluación y conclusiones
Referencias utilizadas

Ponderación de la prueba presencial y/o los trabajos en la nota final	Cada resumen supone un 10% de la calificación final. El trabajo práctico supone el restante 40% de la calificación final.
Fecha aproximada de entrega	
Comentarios y observaciones	

PRUEBAS DE EVALUACIÓN CONTINUA (PEC)

¿Hay PEC?	Si,PEC no presencial
Descripción	
Ver apartado anterior	
Criterios de evaluación	
Ponderación de la PEC en la nota final	
Fecha aproximada de entrega	
Comentarios y observaciones	

OTRAS ACTIVIDADES EVALUABLES

¿Hay otra/s actividad/es evaluable/s?	No
Descripción	
Criterios de evaluación	
Ponderación en la nota final	
Fecha aproximada de entrega	
Comentarios y observaciones	

¿CÓMO SE OBTIENE LA NOTA FINAL?

El resumen de cada tema será evaluado sobre 10 puntos. Cada resumen supone un 10% de la calificación final. El trabajo práctico supone el restante 40% de la calificación final.

BIBLIOGRAFÍA BÁSICA

ISBN(13):null
Título:MINING THE WEB: DISCOVERING KNOWLEDGE FROM HYPERTEXT DATA2002
Autor/es:
Editorial:MORGAN KAUFMANN

ISBN(13):null
Título:WEB DATA MINING: EXPLORING HYPERLINKS, CONTENTS, AND USAGE DATA2007
Autor/es:
Editorial:Springer

Como bibliografía básica se aportarán, además, referencias dentro del curso virtual.

BIBLIOGRAFÍA COMPLEMENTARIA

Como bibliografía complementaria se aportarán referencias dentro del curso virtual.

RECURSOS DE APOYO Y WEBGRAFÍA

Material de estudio

Artículos (generalmente en inglés) disponibles en el sitio web de la asignatura:

Tema 1.

•Kosala, R. and Blockeel, H. Web Mining Research: A Survey. ACM SIGKDD Explorations, Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining. 2000.
•Chakrabarti, S. Data Mining for hypertext: a tutorial survey. ACM SIGKDD Explorations, Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining, 2000.
•Ricardo baeza-Yates. Excavando la web. El profesional de la información. v13, n1, 2004
•Bibliografía básica

Tema 2.

•Steve Lawrence and C. Lee Giles. Searching the World Wide Web. Science vol. 280, 1998.
•Nick Craswell, David Hawking and Stephen Robertson. Effective Site Finding using Link Anchor Information. Research and Development in Information Retrieval, SIGIR 2001.
•Dunja Mladenic. Text-Learning and Related Intelligent Agents: A survey. IEEE Intelligent Systems, 1999
•Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. ACM Press. Addison Wesley, 1999

•Bibliografía básica

Tema 3.

•Marti A. Hearst. Untangling Text Data Mining. Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999 (invited paper).
•Turmo, Jordi. Information Extraction, Multilinguality and Portability. Revista Iberoamericana de Inteligencia Artificial, N.22, vol. 5, Invierno 2003.
•Peñas, A., Verdejo, F. and Gonzalo, J. Terminology Retrieval: towards a synergy between thesaurus and free-text searching. In F.J. Garijo, J.C. Riquelme and M. Toro editors, Advances in Artificial Intelligence - IBERAMIA 2002, LNAI 2527, Lecture Notes in Computer Science. Springer-Verlag, 2002.

Tema 4.

•R. Cooley, B. Mobasher, and J. Srivastava. Web mining and Pattern Discovery on the World Wide Web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, ICTAI, 1997
•J. Srivastava, R. Cooley, M. Deshpande, P. Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, 2000.
•B. Mobasher. Web Usage Mining and Personalization. Chapter in Practical Handbook of Internet Computing, Munindar P. Singh (ed.), CRC Press, 2004
•Mike Perkowitz and Oren Etzioni. Adaptive Web Sites: an AI Challenge, IJCAI, 1997
•Thorsten Joachims, Dayne Freitag and Tom M. Mitchell.Web Watcher: A Tour Guide for the World Wide Web, IJCAI, 1997

Tema 5.

•Soumen Chakrabarti et al. Mining the Link Structure of the World Wide Web. Computer, volume 32, n.8, pp. 60-67, 1999.
•Ravi Kumar et al. The Web as a Graph. Proc. 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles of Database Systems, PODS, ACM Press, 2000.
•Broder et al. Graph Structure in the web. Proc.WWW9, 2000

Tema 6.

•M. Levene and A. Poulovassilis. Web Dynamics. Software Focus, 2, (2001), 60-67.
•Ricardo Baeza-Yates, Bárbara J. Poblete y Felipe Saint-Jean. Evolución de la Web Chilena . Centro de Investigación de la Web, 2003.
•Edward T. O'Neill, Brian F. Lavoie, Rick Bennett. Trends in the Evolution of the Public Web (1998 - 2002). D-Lib Magazine,Volume 9 Number 4, April 2003.

•Broder et al. Graph Structure in the web. Proc.WWW9, 2000.

IGUALDAD DE GÉNERO

En coherencia con el valor asumido de la igualdad de género, todas las denominaciones que en esta Guía hacen referencia a órganos de gobierno unipersonales, de representación, o miembros de la comunidad universitaria y se efectúan en género masculino, cuando no se hayan sustituido por términos genéricos, se entenderán hechas indistintamente en género femenino o masculino, según el sexo del titular que los desempeñe.