20-21

MÁSTER UNIVERSITARIO EN
TECNOLOGÍAS DEL LENGUAJE

GUÍA DE ESTUDIO PÚBLICA

MOTORES DE BÚSQUEDA WEB

CÓDIGO 31101042

UNED

20-21

MOTORES DE BÚSQUEDA WEB

CÓDIGO 31101042

ÍNDICE

Nombre de la asignatura	MOTORES DE BÚSQUEDA WEB
Código	31101042
Curso académico	2020/2021
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

Tipo	Optativa
	Cuatrimestre
Primero	Créditos/horas totales
6/150	Horas de estudio teórico
100	Horas de prácticas
50	Horas complementarias

Esta es la guía de la asignatura "Motores de búsqueda Web" que se imparte dentro del máster en Lenguajes y Sistemas Informáticos de la UNED. En esta guía se contextualiza la asignatura dentro del máster, se especifican los conocimientos previos necesarios para cursarla con éxito, sus objetivos de aprendizaje y contenidos, y la metodología con la que se estudiará.

Esta asignatura se encuadra en el módulo "ESP-LSI-1 Tecnologías del Lenguaje en la Web" dentro de la especialidad con el mismo nombre de la titulación de posgrado "Master en Lenguajes y Sistemas Informáticos". Dentro de esta especialidad, "Motores de búsqueda Web" aporta los fundamentos sobre los que aplicar tecnologías de procesamiento de textos más sofisticadas a gran escala.

REQUISITOS Y/O RECOMENDACIONES PARA CURSAR ESTA ASIGNATURA

Lectura fluida del inglés y conexión a Internet, además de los requisitos propios del máster.

EQUIPO DOCENTE

Nombre y Apellidos	JULIO ANTONIO GONZALO ARROYO (Coordinador de asignatura)
Correo Electrónico	julio@lsi.uned.es
Teléfono	91398-7922
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

Nombre y Apellidos	JUAN MARTINEZ ROMO
Correo Electrónico	juaner@lsi.uned.es
Teléfono	91398-9378
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

HORARIO DE ATENCIÓN AL ESTUDIANTE

Se realizará mediante la plataforma de posgrados de la UNED.

Julio Gonzalo Arroyo
email: julio@lsi.uned.es
Tfno: 913987922
Horario: Jueves de 16:00 a 20:00 horas

Juan Martínez Romo
email: juaner@lsi.uned.es
Tfno: 913989378
Horario: Jueves de 11:00 a 13:00 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.

CE6 - Adquirir capacidad de especificar, diseñar, implementar y evaluar tanto cualitativa como cuantitativamente los modelos y sistemas propuestos.

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

Objetivos generales de la materia

En este curso se estudian los aspectos esenciales para la recuperación de información en la Web: desde la naturaleza del problema (topología de la Web y características de los usuarios) hasta los retos tecnológicos planteados en la nueva generación de buscadores, pasando por los sistemas clásicos de recuperación de información, la arquitectura básica de un buscador Web, y los sistemas de recuperación basados en notoriedad, de los que Google es el ejemplo canónico.

Al finalizar el curso, el alumno debe ser capaz de plantear la arquitectura completa de un buscador Web, y debe ser capaz de diagnosticar las limitaciones de los sistemas actuales y proponer soluciones novedosas para superarlas.

Destrezas y competencias

El alumno adquirirá las siguientes destrezas y competencias:

Debe tener una visión de conjunto de las tecnologías relacionadas con la búsqueda Web, comprendiendo su evolución temporal y los retos de investigación que se plantean en la actualidad.

Debe ser capaz de realizar una lectura crítica de artículos científicos sobre el tema, de localizar y discriminar información bibliográfica relevante, y de sintetizar información de distintas fuentes.

Debe ser capaz de redactar con rigor científico y de comunicar y debatir con pares (en este caso, sus compañeros) sus análisis y opiniones en torno a los temas de la asignatura.

Debe ser capaz de diagnosticar las limitaciones del campo de investigación en motores de búsqueda Web y apuntar caminos para superarlas.

CONTENIDOS

Características de la búsqueda de información en la Web

1. Topología de la Web: Hubs, autoridades, islas, internet invisible, etc.
2. Necesidades de información y búsquedas web: perfil de usuarios.
3. Formas básicas de búsqueda: navegación y consulta. Directorios web versus motores de búsqueda.

Arquitectura básica de un motor de búsqueda

1. Crawling, Indexación, Procesado de la consulta, Recuperación,Presentación de resultados.
2. Arquitectura hardware/software.

Motores de búsqueda pre-Google: búsqueda basada en contenidos

1. Modelos tradicionales de recuperación de información (modelo booleano,modelo de espacio vectorial, modelos probabilísticos).
2. Limitaciones de los modelos RI en la web: pertinencia versus autoridad,vulnerabilidad a la manipulación externa (spamdexing).

Motores de búsqueda post-Google

1. Autoridad absoluta: Algoritmos PageRank y HITS.
2. Autoridad relativa a un tema/consulta: Hilltop, Topic Distillation.- El motor de búsqueda Google: evolución de Pagerank (historia de URLs y enlaces, análisis de patentes de Google, Local Rank, Google Sandbox, etc), sistemas de publicidad contextual (adwords, adsense), señal de usuarios, vulnerabilidad.
3. Otros motores de búsqueda generalistas.

METODOLOGÍA

La general del programa de posgrado. En particular, el alumno realiza dos tipos de actividades en esta asignatura: las relacionadas con la consulta bibliográfica y las de implementación y experimentación. Las primeras son comunes a todos los alumnos y están fijadas dentro del material de estudio correspondiente a cada tema. En una segunda parte de la asignatura, cada alumno realiza un trabajo individual sobre un tema acordado con el equipo docente. Todo el material de estudio está disponible en el entorno virtual del posgrado, y toda la interacción entre profesores y alumnos se puede llevar a cabo en este entorno.

Las tareas que se asignan en esta asignatura tienen tanto que ver con la asimilación de los conocimientos propios de la materia, como con el desarrollo de la capacidad para investigar.Algunos de los tipos de tareas que se proponen son:

- Lectura y análisis de un artículo de investigación, contestando a preguntas como: ¿Se trata de un artículo de teoría, metodología, experimentación o aplicación? ¿Cuáles son sus aportaciones originales? ¿Cuáles son los argumentos/resultados esenciales que conducen a sus conclusiones?
- Evaluación simulada de un artículo, calificando de forma razonada su originalidad, su impacto potencial en el área, la pertinencia y completitud de las referencias bibliográficas, la calidad del trabajo (argumentos, metodología, diseño experimental, etc., la calidad de la presentación (organización, claridad expositiva, etc.). Discusión en grupo (tres alumnos) para alcanzar una única evaluación consensuada, estableciendo una figura de meta-revisor encargado de coordinar la discusión y redactar la evaluación final.
- Estudio del impacto de un artículo: ¿Cuáles son los aspectos del artículo por los que es referenciado? ¿Coinciden con los aspectos sobre los que los autores habían hecho énfasis, o son aspectos inicialmente marginales? ¿Se ha hecho algún avance sustancial respecto a las conclusiones del artículo? ¿Se han refutado las conclusiones del artículo, se han corroborado, se ha profundizado en ellas, se han propuesto vías alternativas?
- Actualización de un artículo de revisión del estado del arte, sintetizando los avances más significativos posteriores a la publicación de la revisión inicial.
- Propuesta de "lecturas recomendadas" para un tema, consensuando una lista razonada a partir del debate entre todos los alumnos de la asignatura.
- Evaluación comparada de servicios de búsqueda Web alternativos, utilizando tanto la revisión bibliográfica como la experimentación directa.
- Diseño e implementación de un servicio de búsqueda Web con algún componente novedoso, partiendo de herramientas de código abierto (como Lucene) o servicios Web (como las API de Google, Yahoo, etc).

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

En cada tema se proponen al estudiante varios ejercicios que deben cumplimentar y entregar a través del entorno virtual.

Criterios de evaluación

En la evaluación de los ejercicios se tiene en cuenta la exactitud y completitud de las respuestas, el tiempo de dedicación, la originalidad y la iniciativa personal.

Ponderación de la prueba presencial y/o los trabajos en la nota final	La calificación de los ejercicios representa un 50% de la nota final de la asignatura.
Fecha aproximada de entrega	23/12/2018
Comentarios y observaciones	

PRUEBAS DE EVALUACIÓN CONTINUA (PEC)

¿Hay PEC?	Si,PEC no presencial

Descripción

En el trabajo final de la asignatura, cada estudiante realiza un trabajo individual sobre el tema genérico propuesto por el equipo docente, o bien sobre un tema acordado entre el estudiante y el equipo docente. En el trabajo final se profundiza en algún aspecto de actualidad de los buscadores Web, realizando un resumen del estado del arte y sugiriendo ideas, preferiblemente originales, para avanzar el conocimiento en ese campo.

Criterios de evaluación

El trabajo final se evalúa teniendo en cuenta la capacidad del estudiante para

- localizar, manejar y citar adecuadamente fuentes bibliográficas relevantes

- analizar y sintetizar la información de esas fuentes

- realizar un diagnóstico adecuado del estado del arte en función de las fuentes consultadas

- proponer temas e ideas bien fundamentadas y preferiblemente novedosas

Ponderación de la PEC en la nota final	El trabajo final representa un 50% de la calificación de la asignatura.
Fecha aproximada de entrega	15/02/2019
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?

La nota final se obtiene a partir de la calificación de los ejercicios de cada tema (50% de la nota final) y del trabajo final de la asignatura (50% de la calificación final). Es imprescindible entregar el trabajo final para poder superar la asignatura.

BIBLIOGRAFÍA BÁSICA

Arvind Arasu, Junghoo Cho, Hector García-Molina, Andreas Paepcke and Sriram Raghavan. Searching the Web. ACM Transactions on Internet Technology, vol. 1, n. 1, August 2001, pages 2-43.

BIBLIOGRAFÍA COMPLEMENTARIA

Tema 1. Características de la búsqueda de información en la WWW
Sobre estructura de la WWW:
- Kleinberg, JM. Hubs, authorities, and communities, ACM computing surveys 1999.
http://www.cs.brown.edu/memex/ACMCSHT/10/10.html
- A Borodin, GO Roberts, JS Rosenthal, P. Tsaparas. Finding authorities and hubs from link structures on the World Wide Web. Proc. WWW 2001.
http://www10.org/cdrom/papers/314/
Sobre tipología de búsquedas web:
- Rose, D. y Levinson, D. Understanding User Goals in Web Search. WWW 2004.
http://wwwconf.ecs.soton.ac.uk/archive/00000537/01/p13-rose.pdf
Sobre navegación versus consulta:
- Marti A. Hearst. Next Generation Web Search: Setting Our Sites In IEEE Data Engineering Bulletin, 2002.
http://www.sims.berkeley.edu/hearst/papers/data-engineering
- A. Peñas, F. Verdejo, J. Gonzalo, 2002. Terminology Retrieval: towards a synergy between thesaurus and free text searching. Advances in Artificial Intelligence - IBERAMIA 2002, LNAI 2527.
http://nlp.uned.es/pergamus/pubs/iberamia2002.pdf

Tema 2. Arquitectura básica de un motor de búsqueda.
Sobre crawling:
- J Cho, H Garcia-Molina, L Page. Efficient Crawling Through URL Ordering, WWW 1998.
- Allan Heydon and Marc Najork. Mercator: A Scalable, Extensible Web Crawler. In Proceedings of World Wide Web Conference, 1999, pages 219-229.
Sobre soporte hardware:

- L. A. Barroso, J. Dean, U. Hoelzle. Web search for a planet: the Google cluster architecture. IEEE 2003.

Tema 3. Motores de búsqueda pre-Google: recuperación basada en contenidos.

- D Hiemstra. Using Language Models for Information Retrieval. CTIT Ph.D. Thesis, 2001.
- G Salton, A Wong, CS Yang. A Vector Space Model for Automatic Indexing. Comm. ACM, 1975.
- N Fuhr. Probabilistic Models in Information Retrieval. The Computer Journal, 1992.

Tema 4. Motores de búsqueda actuales (generalistas): recuperación basada en autoridad.

Referencias:

- M Hollander. Google's PageRank Algorithm to Better Internet Searching. TR UMN.
- Brin, S. y Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW 1998.
- CHQ Ding, X He, P Husbands, H Zha, HD Simon. PageRank, HITS and a unified framework for link analysis. SIGIR 2002.
TH Haveliwala. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE T. on Knowledge and data engineering, 2003.

Tema 5. Temas avanzados.

- Guha, R. y Garg, A. Disambiguating People in Search. Proc. WWW 2004.
- S Lawrence, NJ Princeton. Context in Web Search, IEEE data engineering bulletin, 2000.
J Sivic, A Zisserman. Video google: A text retrieval approach to object matching in videos, ICCV 2003.
- SK Bhavnani, CK Bichakjian, TM Johnson, RJ Little. Strategy Hubs: Next-Generation Domain Portals with Search Procedures. Proc. ACM Conference on Human Factors in Computing Systems, 2003, ACM Press NY, USA.
- T Berners-Lee, J Hendler, O Lassila. The semantic Web. Scientific American, 2001.
- J Heflin, J Hendler. A Portrait of the Semantic Web in Action. IEEE Intelligent Systems, 2001.
- S Eissen, B Stein. Analysis of Clustering Algorithms for Web-Based Search. Springer-Verlag, 2002.
- J. Cigarrán, A. Peñas, J. Gonzalo, F. Verdejo, 2005. Automatic selection of noun phrases as document descriptors in an FCA-based Information Retrieval system. ICFCA 2005. Springer LNCS 3403.
Search Engines: Technology, Society, and Business. Materiales online del curso: http://www.sims.berkeley.edu/courses/is141/f05/schedule.html

RECURSOS DE APOYO Y WEBGRAFÍA

La plataforma de enseñanza virtual de posgrados de la UNED será la interfaz de interacción entre el alumno y sus profesores. Esta plataforma permite impartir y recibir formación, gestionar y compartir documentos, crear y participar en comunidades temáticas, así como realizar proyectos online.

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.