ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

GUIA DOCENTE

CURSO 2022-23

MÁSTER EN INGENIERÍA INFORMÁTICA (PLAN 2018)

DATOS DE LA ASIGNATURA

Nombre:
EXTRACCIÓN DE DATOS MASIVOS DE INTERNET
Denominación en Inglés:
Extraction of Massive amounts of data from Internet

Código:	**Tipo Docencia:**	**Carácter:**
1180417	Semipresencial	Optativa

Horas:

	Totales	**Presenciales**	**No Presenciales**
Trabajo Estimado	75	15	60

Créditos:

Grupos Grandes	**Grupos Reducidos**			
	Aula estándar	**Laboratorio**	**Prácticas de campo**	**Aula de informática**
1.5	0	1.5	0	0

Departamentos:	**Áreas de Conocimiento:**
TECNOLOGIAS DE LA INFORMACION	LENGUAJES Y SISTEMA INFORMATICOS
Curso:	**Cuatrimestre**
1º - Primero	Segundo cuatrimestre

DATOS DEL PROFESORADO (*Profesorado coordinador de la asignatura)		
Nombre:	**E-mail:**	**Teléfono:**
* Jacinto Mata Vazquez	mata@dti.uhu.es	959 217 652
Docente por contratar (Departamento_TECNOLOGIAS DE LA	Docente_T135@uhu.es	

Datos adicionales del profesorado (Tutorías, Horarios, Despachos, etc...)

Los horarios de clases y de tutorías están disponibles en:

https://www.uhu.es/etsi/informacion-academica/informacion-comun-todos-los-titulos/horarios-2/

Despachos del profesorado

- Jacinto Mata Vázquez → Edificio de la ETSI (P162)

DATOS ESPECÍFICOS DE LA ASIGNATURA

1. Descripción de Contenidos:

1.1 Breve descripción (en Castellano):

El objetivo general de la asignatura es formar profesionales con destrezas para la obtención automatizada de información y conocimiento útil por cauces informales, principalmente de internet, teniendo en cuenta la innovación y el volumen de datos como los ejes estratégicos en el que sustentar la analítica de datos. Web crawling, búsqueda, análisis de redes sociales, extracción de datos estructurados, integración de información, minería de opinión y análisis de sentimientos, minería de uso de la Web, minería de registros de consulta, publicidad web o sistemas de recomendación serán algunos de los temas tratados.

Temas:

- Los Problemas de la Extracción de Conocimiento de Información No Estructurada.
- Extracción de datos estructurados a partir de APIs (Facebook, twitter, google). Formato JSON.
- Extracción de Conocimiento a partir de Información Semi-Estructurada. XML, DTDs y Consultas (SAX, XQuery, XPath).
- Extracción de Conocimiento a partir de Documentos No Estructurados.
- Web mining. Recuperación de información y búsquedas web. Análisis de redes sociales.
- Web crawling. Wrapper.
- Open Data. Linked (Open) Data. Calidad de datos. Conectividad.

1.2 Breve descripción (en Inglés):

Web Data Extraction aims to study automatic information extraction techniques: web crawling, web search techniques, analysis of social networks, extraction of structured data, information integration.

Topics:

- Extraction of Knowledge from Unstructured Information.
- Extraction of Structured data from APIs (Facebook, twitter, google). JSON format.
- Extraction of Knowledge from Semi-Structured Information. XML, DTDs and Queries (SAX, XQuery, XPath).
- Extraction of Knowledge from Unstructured Documents.
- Web Mining Information Retrieval and Web Searches.
- Web Crawling. Wrapper
- Open Data. Linked (Open) Data. Data quality.

2. Situación de la asignatura:

2.1 Contexto dentro de la titulación:

En las asignaturas de la especialidad de Big Data y Cloud Computing se presentan algoritmos e infraestructuras que tienen como objetivo soportar un volumen ingente de datos para obtener valor

empresarial. En esta asignatura, nos centramos en los mecanismos existentes para obtener grandes volúmenes de datos de internet.

2.2 Recomendaciones

Conocimientos de estándares de fuentes de información (XML, JSON, ...) y conocimientos básicos del lenguaje de programación Python

3. Objetivos (Expresados como resultado del aprendizaje):

Con la realización de esta asignatura, el estudiante tendrá capacidad para llevar a cabo tareas de recuperación de información estructurada, semiestructurada y no estructurada desde fuentes de datos heterogéneas, principalmente desde redes sociales y datos abiertos, para su posterior análisis y extracción de conocimiento.

Competencias Específicas:

- Conocimiento y capacidad para extraer información relevante incluida en páginas web y redes sociales
- Conocimiento de los principales formatos de intercambio de información, así como de los lenguajes de programación y librerías más utilizadas en tareas de extracción y análisis de información

4. Competencias a adquirir por los estudiantes

4.1 Competencias específicas:

-

4.2 Competencias básicas, generales o transversales:

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.

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

CG10: Capacidad para aplicar los principios de la economía y de la gestión de recursos humanos y proyectos, así como la legislación, regulación y normalización de la informática.

CG3: Dirigir, planificar y supervisar equipos multidisciplinares

CG9: Capacidad para comprender y aplicar la responsabilidad ética, la legislación y la deontología profesional de la actividad de la profesión de Ingeniero en Informática.

CG6 : Capacidad para la dirección general, dirección técnica y dirección de proyectos de investigación, desarrollo e innovación, en empresas y centros tecnológicos, en el ámbito de la Ingeniería Informática.

CG8 : Capacidad para la aplicación de los conocimientos adquiridos y de resolver problemas en entornos nuevos o poco conocidos dentro de contextos más amplios y mulitidisciplinares, siendo capaces de integrar estos conocimientos.

CG4: Capacidad para el modelado matemático, cálculo y simulación en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación, desarrollo e innovación en todos los ámbitos relacionados con la Ingeniería en Informática.

CT1 : Gestionar adecuadamente la información adquirida expresando conocimientos avanzados y demostrando, en un contexto de investigación científica y tecnológica o altamente especializado, una comprensión detallada y fundamentada de los aspectos teóricos y prácticos y de la metodología de trabajo en el campo de estudio.

CT2 : Dominar el proyecto académico y profesional, habiendo desarrollado la autonomía suficiente para participar en proyectos de investigación y colaboraciones científicas o tecnológicas dentro su ámbito temático, en contextos interdisciplinares y, en su caso, con un alto componente de transferencia del conocimiento.

CT5: Utilizar de manera avanzada las tecnologías de la información y la comunicación, desarrollando, al nivel requerido, las Competencias Informáticas e Informacionales (CI2).

CT4: Comprometerse con la ética y la responsabilidad social como ciudadano y como profesional, con objeto de saber actuar conforme a los principios de respeto a los derechos fundamentales y de igualdad entre hombres y mujeres y respeto y promoción de los Derechos Humanos, así como los de accesibilidad universal de las personas discapacitadas, de acuerdo con los principios de una cultura de paz, valores democráticos y sensibilización medioambiental.

CT3: Desarrollar una actitud y una aptitud de búsqueda permanente de la excelencia en el quehacer académico y en el ejercicio profesional futuro.

5. Actividades Formativas y Metodologías Docentes

5.1 Actividades formativas:

- Sesiones de teoría/problemas/casos prácticos sobre los contenidos del programa

- Sesiones prácticas en laboratorios especializados o en aulas de informática

- Actividades académicamente dirigidas por el profesorado: seminarios, conferencias, desarrollo de trabajos, debates, tutorías colectivas, ...

- Actividades de evaluación

- Lectura de los contenidos de los temas

- Entrega de ejercicios/prácticas/trabajos evaluables

- Actividades de autoevaluación

- Tutorías colectivas a través de plataformas de enseñanza virtual (foros, wikis, chats)

- Trabajo individual/autónomo del estudiante

- Actividades no presenciales con evaluación por pares

- Desarrollo cooperativo de trabajos utilizando herramientas de discusión asíncrona (foros, wikis, ...)

5.2 Metodologías Docentes:

- Clase magistral participativa

- Desarrollo de prácticas en laboratorios especializados o en aulas de informática en grupos reducidos

- Resolución de problemas y ejercicios prácticos

- Tutorías individuales o colectivas. Interacción directa profesorado-estudiantes

- Planteamiento, realización, tutorización y presentación de trabajos

- Conferencias y seminarios

- Evaluaciones y exámenes

- Visualización y escuchas de sesiones grabadas de seminarios ad hoc con entrevistas a expertos en algunos temas claves de la materia o vídeos seleccionados que incentiven algunas competencias

- Tutorías en línea. Utilización de foros y otros medios de comunicación e interacción con el profesorado

- Trabajos colaborativos. Llevar a cabo una actividad basada en un objetivo común en el que el estudiante debe colaborar activamente para realizarla

- Metodologías basadas en la acción. Revisión, planificación de las mejoras de trabajos con la participación de los estudiantes y el profesor

5.3 Desarrollo y Justificación:

Las sesiones de teoría sobre los contenidos del programa, las sesiones de resolución de problemas y las sesiones prácticas se llevarán a cabo, conjuntamente, en un laboratirio del Departamento de Tecnologías de la Información, en el horario establecido por el Centro. En estas sesiones, el profesorado explicará conceptos teóricos y se realizarán los ejercicios y las prácticas propuestas.

Las Actividades Académicas Dirigidas por el profesorado complementarán las actividades formativas anteriores. Actividades como seminarios y conferencias se programarán durante el curso en función de la disponibilidad de los ponentes.

Las actividades formativas no presenciales complementan a las actividades que se realizarán en el aula y servirán para que los estudiantes puedan seguir el desarrollo completo de la asignatura.

Las metodologías docentes no presenciales propuestas servirán para llevar a cabo un correcto proceso de enseñanza-aprendizaje en esta titulación semipresencial.

6. Temario Desarrollado

Tema 1. Extracción de información mediante *Interfaz de Programación de Aplicaciones* (API)

- Introducción
- Autenticación y autorización
- Extracción de información estructurada, no estructurada y semiestructurada
- Análisis y extracción de conocimiento de la información recuperada
- Herramientas de recuperación y análisis de información

Tema 2. Open Data

- Introducción
- Gobierno Abierto.
- Open Goverment Data
- Open Data. Principios. Utilización. Proyectos. Formatos.
- Linked Open Data
- Bases de Datos RDF. Sparql
- Puntos de Acceso.
- Consultas Sparql: Estructura Básica.
- Consultando a la DBPedia.

7. Bibliografía

7.1 Bibliografía básica:

- Data analysis with open source tools / Philipp K. Janert Publicación Sebastopol: O'Reilly, 2011
- Open Data Structures: An Introduction. Pan Moris. 2013

7.2 Bibliografía complementaria:

Recursos en Internet:

https://www.programmableweb.com/apis/directory

8. Sistemas y criterios de evaluación

8.1 Sistemas de evaluación:

- Examen de teoría/problemas

- Defensa de trabajos e informes escritos

- Pruebas de evaluación mediante plataformas de enseñanza virtual

- Participación en las actividades propuestas

8.2 Criterios de evaluación relativos a cada convocatoria:

8.2.1 Convocatoria I:

Los principios de evaluación de la asignatura siguen unos criterios de **evaluación preferentemente continua**, entendiendo por tal la evaluación diversificada que se lleva a cabo en distintos momentos del curso académico en curso.

La calificación final mediante evaluación continua se calculará mediante la siguiente fórmula:

Nota final = 0.2 * Nota Examen Teoría/Problemas + 0.5 * Pruebas de evaluación mediante plataforma de enseñanza virtual + 0.2 * Defensa de trabajos e informes escritos + 0.1 * Participación en las actividades propuestas

---

El examen de teoría/problemas consistirá en un cuestionario de preguntas tipo test o respuesta corta. Tiene carácter individual y su duración máxima se notificará con antelación suficiente. La materia objeto de examen será toda la tratada en la asignatura. No se podrá utilizar ningún tipo de recurso didáctico o documentación además de la proporcionada por el equipo docente el día del examen.

La defensa de trabajos e informes escritos consistirá en la realización, entrega y defensa de un trabajo que el estudiante desarrollará durante el curso. Tiene un carácter individual y se podrá utilizar cualquier material que se considere siempre que se referencie adecuadamente.

Las pruebas de evaluación en plataforma de enseñanza virtual consistirán en la resolución de problemas teórico/prácticos que se subirán a la plataforma durante el desarrollo de cada tema.

Las competencias básicas (CB7 y CB10) y la competencia general CG8 que los estudiantes deben adquirir en esta asignatura se evaluarán mediante el examen de teoría/problemas y las pruebas de evaluación mediante plataformas de enseñanza virtual. Por otro lado, los resultados de aprendizaje se evaluarán, además de con los sistemas de evaluación anteriores, mediante la defensa de trabajos e informes escritos y la participación en las actividades propuestas. Por último, las competencias básicas (CB6, CB8 y CB9), las competencias generales (CG3, CG4, CG6, CG9 y CG10) y las competencias transversales (CT1, CT2, CT3, CT4 y CT5) se evaluarán con la defensa de trabajos e informes escritos y la participación en las actividades propuestas.

Matrícula de Honor

Para la obtención de la matrícula de honor, el estudiante deberá obtener un 10 en su nota final. En el caso de que haya más estudiantes con esta calificación, y no sea posible otorgarlas todas por razón del número de estudiantes matriculados, ésta/s se le otorgará/n a aquellos que consigan mejor calificación en la resolución de una prueba adicional cuya fecha de celebración se acordará entre los estudiantes implicados.

8.2.2 Convocatoria II:

La calificación final en esta convocatoria se calculará con la misma fórmula de la convocatoria I.

Para esta convocatoria II se podrán conservar las calificaciones obtenidas en la convocatoria I en cualquiera de los criterios de evaluación.

8.2.3 Convocatoria III:

En esta convocatoria se aplicará la “evaluación única final” tal como se describe en el siguiente apartado.

8.2.4 Convocatoria extraordinaria:

En esta convocatoria se aplicará la “evaluación única final” tal como se describe en el siguiente apartado.

8.3 Evaluación única final:

8.3.1 Convocatoria I:

Aquellos estudiantes que quieran acogerse a la evaluación única final deberán comunicarlo en las dos primeras semanas de impartición de la asignatura, o en las dos semanas siguientes a su matriculación si ésta se ha producido con posterioridad al inicio de la asignatura. Para estos casos se aplicará la siguiente fórmula para su evaluación:

Nota final = 0.4* Examen de Teoría/Problemas + 0.4 * Examen de prácticas + 0.2 * Defensa de trabajo propuesto

Estas tres pruebas se realizarán el día fijado por el Centro. El examen de teoría consistirá en la resolución de problemas y preguntas teóricas relacionadas con el temario de teoría. El examen de prácticas consistirá en una actividad práctica en aula de informática relacionadas con los contenidos de la asignatura. La defensa de trabajo propuesto consistirá en la presentación de los resultados de un trabajo desarrollado a partir de un enunciado que se entregará con, al menos, una semana de antelación.

8.3.2 Convocatoria II:
La evaluación única final en esta convocatoria se calculará de la misma forma que en la convocatoria I.
8.3.3 Convocatoria III:
La evaluación única final en esta convocatoria se calculará de la misma forma que en la convocatoria I.
8.3.4 Convocatoria Extraordinaria:
La evaluación única final en esta convocatoria se calculará de la misma forma que en la convocatoria I.

9. Organización docente semanal orientativa:							
Fecha	Grupos Grandes	G. Reducidos				Pruebas y/o act. evaluables	Contenido desarrollado
		Aul. Est.	Lab.	P. Camp	Aul. Inf.		
20-02-2023	2	0	2	0	0		Presentación y Tema 1
27-02-2023	2	0	2	0	0		Tema 1
06-03-2023	2	0	2	0	0		Tema 1
13-03-2023	2	0	2	0	0		Tema 1
20-03-2023	2	0	2	0	0		Tema 1
27-03-2023	2	0	2	0	0		Tema 2
10-04-2023	2	0	2	0	0		Tema 2
17-04-2023	1	0	1	0	0	Pruebas de evaluación	
24-04-2023	0	0	0	0	0		
01-05-2023	0	0	0	0	0		
08-05-2023	0	0	0	0	0		
15-05-2023	0	0	0	0	0		
22-05-2023	0	0	0	0	0		
05-06-2023	0	0	0	0	0		
12-06-2023	0	0	0	0	0		
TOTAL	**15**	**0**	**15**	**0**	**0**		