



EXTRACCIÓN Y PROCESAMIENTO AVANZADO DE LA INFORMACIÓN - CÓDIGO

Estudios: Grado en Información y Documentación

CURSO ACADÉMICO: 2026/2027

TIPO: Optativa

DEPARTAMENTO/S: BIBLIOTECONOMÍA Y DOCUMENTACIÓN

CRÉDITOS ECTS: 6 cts

CURSO: 4º

CUATRIMESTRE: 2º

PROFESOR/ES:

MANUEL BLÁZQUEZ OCHANDO

Correo electrónico: manublaz@ucm.es

Facultad de Ciencias de la Documentación (Despacho 103)

RECOMENDACIONES PARA CURSAR ESTA ASIGNATURA

Se recomienda cursar una vez se superen las siguientes asignaturas:

- Primer curso – Búsqueda y recuperación de la información
- Primer curso – Diseño de bases de datos para entornos documentales

La asignatura guarda relación con las siguientes materias:

- Primer curso – Búsqueda y recuperación de la información
- Primer curso – Diseño de bases de datos para entornos documentales
- Segundo curso – Bases de datos avanzadas para entornos documentales
- Tercer curso – Catalogación automatizada
- Tercer curso – Edición digital en el entorno web
- Tercer curso – Lenguajes de marcado, semántica y metadatos
- Tercer curso – Sistemas automatizados en unidades de información
- Cuarto curso – Datos e información para la empresa
- Cuarto curso – Búsqueda y posicionamiento en Internet
- Cuarto curso – Fuentes de información especializada y vigilancia tecnológica

COMPETENCIAS QUE ADQUIERE EL ESTUDIANTE Y RESULTADOS DEL APRENDIZAJE

1. Capacidad para aplicar técnicas de minería de textos y datos
2. Conocimientos técnicos para la recolección automatizada de información
3. Conocimientos en el tratamiento automatizado de la información
4. Desarrollar estrategias de clasificación de la información
5. Aplicar técnicas de big-data para el análisis automático de la información



DESCRIPCIÓN DE CONTENIDOS: PROGRAMA

Descripción y objetivos de la asignatura:

- Aprender los métodos de extracción de la información en Internet, para su reutilización, desarrollo de servicios de información y enriquecimiento documental automático.
- Aprender a procesar archivos XML de forma automática, para su posterior procesamiento en base de datos.
- Aprender los principios que sustentan la minería de datos y el big-data.
- Adquirir la habilidad para realizar migraciones de datos complejas, creando mapas de migración, planificando procedimientos de transformación, hasta su importación definitiva en la base de datos de destino.

Programa:

- Teoría básica de la extracción y procesamiento de datos en la Web.
- Tecnologías para la extracción de datos XML-XPath, funciones cURL, Objetos DOM, funciones file_get_contents, RESTful HTTP Get.
- Técnicas de parsing XML + Práctica.
- Técnicas de scraping orientadas a recursos Web + Práctica.
- Introducción a la minería de datos y el big-data.

METODOLOGÍA DOCENTE Y RÉGIMEN DE TUTORÍAS

El programa que se ofrece al estudiante para lograr los resultados previstos comprende las siguientes actividades:

1. Clases teóricas. El profesor expondrá y desarrollará los contenidos teóricos básicos sobre la extracción y procesamiento de información. Competencias tratadas: 1-5.
2. Clases prácticas. Desarrollo de actividades formativas prácticas con herramientas informáticas en las que el estudiante aplicará y pondrá en práctica los contenidos teóricos expuestos por el profesor. Competencias adquiridas: 1-5.
3. Proyectos grupales. Desarrollo de proyectos grupales relacionados con la extracción y procesamiento de información relacionada con el ámbito de la Documentación. Competencias adquiridas: 1-5.
4. Proyecto individual. Desarrollo de prácticas individuales sobre los contenidos de la asignatura, que muestre las competencias adquiridas a lo largo del curso. Competencias tratadas: 1-5.
5. Tutorías. Resolución de dudas que puedan surgir al estudiante y seguimiento de los trabajos prácticos individuales y grupales.



Tutorías y consultas:

- El horario de tutorías está establecido en el campus virtual de la asignatura. El alumno puede acercarse al despacho 103 en dicho horario para satisfacer cualquier duda o consulta, relativa con la asignatura y su contenido.
- También se pueden plantear las consultas o dudas a través de correo electrónico, dirigido a manublaz@ucm.es (*por favor, no usar la mensajería del campus virtual. Usar correo Gmail de la universidad para emitir cualquier comunicación*). Por otra parte, también se pueden concretar tutorías. Para ello indicar el rango horario más adecuado, para poder concretarlo.

SISTEMA DE EVALUACIÓN

- El estudiante deberá demostrar que ha alcanzado los resultados de aprendizaje previstos mediante la realización de ejercicios prácticos y un examen teórico. Los ejercicios prácticos supondrán el 50% de la nota y el examen el otro 50% restante.
- Para aprobar la asignatura es necesario aprobar tanto la teoría como la práctica.

BIBLIOGRAFÍA Y RECURSOS RECOMENDADOS

- Bharanipriya, V., & Prasad, V. K. (2011). Web content mining tools: a comparative study. *International Journal of Information Technology and Knowledge Management*, 4(1), 211-215.
- Bratko, I., Michalski, R. S., & Kubat, M. (1999). *Machine learning and data mining: methods and applications*.
- Corrales, J. S., & Lan, P. (2006). U.S. Patent Application No. 11/586,444.
- Cunningham, H. (2005). Information extraction, automatic. *Encyclopedia of language and linguistics*, 665-677.
- Fernández Villamor, J. I., Blasco Garcia, J., Iglesias Fernandez, C. A., & Garijo Ayestaran, M. (2011). A semantic scraping model for web resources-Appling linked data to web page screen scraping.
- Fry, C. (2005). U.S. Patent No. 6,880,125. Washington, DC: U.S. Patent and Trademark Office.
- Han, H., & Tokuda, T. (2008, July). A method for integration of Web applications based on information extraction. In *Web Engineering, 2008. ICWE'08. Eighth International Conference on* (pp. 189-195). IEEE.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- Heidorn, P. B. (2011). The emerging role of libraries in data curation and e-science. *Journal of Library Administration*, 51(7-8), 662-672.
- Johnson, F., & Gupta, S. K. (2012). *Web content mining techniques: A survey*.



International Journal of Computer Applications, 47(11).

- Kokkoras, F., Ntonas, K., & Bassiliades, N. (2013, September). DEiXTo: a web data extraction suite. In Proceedings of the 6th Balkan Conference in Informatics (pp. 9-12). ACM.
- Kostoulas, M. G., Matsa, M., Mendelsohn, N., Perkins, E., Heifets, A., & Mercaldi, M. (2006, May). XML screamer: an integrated approach to high performance XML parsing, validation and deserialization. In Proceedings of the 15th international conference on World Wide Web (pp. 93-102). ACM.
- Krzywdka, J. (2015). RESTful Web Services.
- Malik, S. K., & Rizvi, S. A. M. (2011, October). Information extraction using web usage mining, web scrapping and semantic annotation. In Computational Intelligence and Communication Networks (CICN), 2011 International Conference on (pp. 465-469). IEEE.
- Manber, U., & Lu, Q. (2005). U.S. Patent No. 6,920,609. Washington, DC: U.S. Patent and Trademark Office.
- Mayfield, J., & Finin, T. (2003, August). Information retrieval on the Semantic Web: Integrating inference and retrieval. In Proceedings of the SIGIR Workshop on the Semantic Web.
- Michael, K., & Miller, K. W. (2013). Big data: New opportunities and new challenges [guest editors' introduction]. Computer, 46(6), 22-24.
- Mitchell, R. (2015). Web Scraping with Python: Collecting Data from the Modern Web. " O'Reilly Media, Inc."
- Myllymaki, J. (2002). Effective web data extraction with standard XML technologies. Computer Networks, 39(5), 635-644.
- Nicola, M., & John, J. (2003, November). Xml parsing: a threat to database performance. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 175-178). ACM.
- Nikovski, D. N., & Esenther, A. W. (2008). U.S. Patent Application No. 12/239,859.
- Pautasso, C., Zimmermann, O., & Leymann, F. (2008, April). Restful web services vs. big'web services: making the right architectural decision. In Proceedings of the 17th international conference on World Wide Web (pp. 805-814). ACM.
- Richardson, L., & Ruby, S. (2008). RESTful web services. " O'Reilly Media, Inc."
- Rodriguez, A. (2008). Restful web services: The basics. IBM developerWorks.
- Russom, P. (2006). Best practices in data migration. Renton/USA.
- Scavuzzo, M., Di Nitto, E., & Ceri, S. (2014, September). Interoperable data migration between NoSQL columnar databases. In Enterprise Distributed Object Computing Conference Workshops and Demonstrations (EDOCW), 2014 IEEE 18th International



(pp. 154-162). IEEE.

- Singh, A., Korupolu, M., & Mohapatra, D. (2008, November). Server-storage virtualization: integration and load balancing in data centers. In Proceedings of the 2008 ACM/IEEE conference on Supercomputing (p. 53). IEEE Press.
- Taira, K., Kamakura, J., Hosoya, T., Suzuki, H., Haraguchi, T., Sasakawa, Y., & Enbutsu, H. (1999). U.S. Patent No. 5,930,806. Washington, DC: U.S. Patent and Trademark Office.
- Taylor, J. A., Merrell, A. R., Parks, R. L., & Taylor, A. (2003). U.S. Patent No. 6,654,830. Washington, DC: U.S. Patent and Trademark Office.
- Turland, M. (2010). php. Marco Tabini & Associates, Inc..
- Vargiu, E., & Urru, M. (2012). Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Artificial Intelligence Research*, 2(1), 44.