UNIVERSIDADE DA CORUÑA

Guía Docente				
Datos Identificativos				2024/25
Asignatura (*)	Recuperación de Información e Minería Web		Código	614G03026
Titulación				
Descriptores				
Ciclo	Período	Curso	Tipo	Créditos
Grao	1º cuadrimestre	Terceiro	Optativa	6
Idioma	Castelán			
Modalidade docente	Presencial			
Prerrequisitos				
Departamento	Ciencias da Computación e Tecnoloxías da Información			
Coordinación	Barreiro Garcia, Álvaro	Correo electrónico	alvaro.barreiro@udc.es	
Profesorado	Barreiro Garcia, Álvaro Pérez Vila, Miguel Anxo	Correo electrónico	alvaro.barreiro@udc.es anxo.pvila@udc.es	
Web				
Descrición xeral	Esta materia aborda a recuperación de información en repositorios de documentos textuais e a web. Estúdanse modelos, técnicas e algoritmos actuais que permiten o crawling e minería web, procesamento, indexación e procura en coleccións de textos do rango de gigabytes, ata os terabytes de información que se manexan na web. Nesta materia o estudante comprenderá a arquitectura dos motores de procura de internet usados polas grandes compañías de Search Engines (Google, Bing, Yahoo, etc) e nas prácticas da mesma poderá desenvolver os módulos principais dun motor de procura. A Recuperación de Información e en particular na web expón extraordinarios retos debido ao volume e heteroxeneidade dos datos e fontes e ao amplo rango de intereses de usuarios privados e corporativos, por todo iso é un campo con amplas posibilidades de negocio e emprego nas TIC. Por outra parte a disciplina e básica para aumentar a fiabilidad e precisión da IA xenerativa e desenvolveu nos últimos anos os modelos neuronales de RI que supoñen unha das máis exitosas realizacións da IA.			

Competencias / Resultados do título	
Código	Competencias / Resultados do título

Resultados da aprendizaxe			
Resultados de aprendizaxe	Competencias / Resultados do título		
Coñecer, comprender e analizar os diferentes modelos de busca de información		B3 B4 B5 B8 B9 B10	C3
Coñecer, comprender e analizar as técnicas para unha implantación eficiente dos buscadores		B3 B4 B5 B8 B9 B10	C3
Coñecer, comprender e analizar as metodoloxías de avaliación dos sistemas de acceso á información		B3 B4 B5 B8 B9 B10	C3

Coñecer, comprender e saber utilizar tecnoloxías, marcos e bibliotecas para construír sistemas de recuperación de información		B3 B4 B5 B8 B9 B10	C3
Planificar e realizar a avaliación dos sistemas de recuperación de información		B3 B4 B5 B8 B9 B10	C3
Ser capaz de manexar correctamente os aspectos éticos, de privacidade, de confidencialidade e de seguridade destes sistemas.		B3 B4 B5 B8 B9 B10	C3

Contidos	
Temas	**Subtemas**
Introducción	Recuperación de Información e Search Engines. Arquitectura dun Search Engine. Grandes retos.
Recopilación de información.	Crawling, feeds, web scraping e minería web.
Procesamento de texto.	Preprocesamento. Parsing, documentos estructurados, anchor text e análisis de enlaces, internacionalización
Indices e procesado eficiente.	Indices Invertidos, compresión, construcción, procesado eficiente de consultas sobre índices invertidos
Formulación de consultas e presentación de resultados	Transformación de consultas, relevance feedback, pseudo-feedback, snippets e visualización de resultados.
Modelos de recuperación de información.	Booleano, espacio vectorial, probabilístico, BM25, Language Models, Relevance Models, modelos neuronales
Evaluación de sistemas de Recuperación de Información.	Datasets e iniciativas de evaluación. Métricas de eficacia e eficiencia. Training e test. Significancia estadística
Búsqueda distribuida e social.	Meta-buscadores, búsqueda distribuida e federada, redes sociais, sistemas de recomendación.

Planificación				
Metodoloxías / probas	**Competencias / Resultados**	**Horas lectivas (presenciais e virtuais)**	**Horas traballo autónomo**	**Horas totais**
Prácticas de laboratorio	B3 B4 B5 B8 B9 B10 C3	14	21	35
Solución de problemas	B3 B4 B5 B8 B9 B10 C3	4	12	16
Proba mixta	B3 B4 B5 B8 B9 B10 C3	2	14	16
Traballos tutelados	B3 B4 B5 B8 B9 B10 C3	3	9	12

Sesión maxistral	B3 B4 B5 B8 B9 B10 C3	19	38	57
Lecturas	B3 B4 B5 B8 B9 B10 C3	2	12	14
Atención personalizada		0		0
***Os datos que aparecen na táboa de planificación son de carácter orientativo, considerando a heteroxeneidade do alumnado**				

Metodoloxías	
Metodoloxías	**Descrición**
Prácticas de laboratorio	Prácticas de laboratorio sobre plataformas de desenvolvemento de amplio uso na industria, nas compañias de Search Engines e nos grupos de investigación (Apache Lucene e outras librerías para outros módulos dos sistemas)
Solución de problemas	Problemas e cuestións para asentar e profundizar nos contidos expostos nas sesións maxistrais.
Proba mixta	Proba que versará sobre os contidos fundamentais da materia.
Traballos tutelados	Traballos complementarios plantexados polo profesor e realizados de forma autónoma
Sesión maxistral	O estudante asistirá ás explicacións dadas polo profesor sobre os distintos modelos, técnicas e algoritmos de Recuperación de Información. O profesor utilizará distintos niveis de abstracción-detalle e orientará ao estudante nas lecturas fundamentais e complementarias.
Lecturas	Lecturas para consolidar e complementar os coñecementos adquiridos. Temas: técnicas, aplicaciones, sistemas industriales.

Atención personalizada	
Metodoloxías	**Descrición**
Prácticas de laboratorio Solución de problemas Traballos tutelados	As lecturas complementarias poden requirir atención personalizada. Algúns problemas máis difíciles poden requirir unha atención personalizada. Ademais de avaliar o resultado da práctica de acordo cos requisitos esixidos, faise un seguimento do seu desenvolvemento. Debe respectarse a autonomía do alumno para que adquira unha maior destreza coas plataformas software utilizadas, pero o profesor pode resolver determinadas dificultades que poden bloquear o alumno durante un tempo excesivo dada a planificación da materia.

Avaliación			
Metodoloxías	**Competencias / Resultados**	**Descrición**	**Cualificación**
Prácticas de laboratorio	B3 B4 B5 B8 B9 B10 C3	Seguimento, defensa e avaliación dos resultados das prácticas realizados nas horas de clases prácticas de laboratorio.	30
Solución de problemas	B3 B4 B5 B8 B9 B10 C3	Resultados obtidos na realización de problemas, cuestións e cuestionarios realizados nas clases e controles levados a cabo.	20
Proba mixta	B3 B4 B5 B8 B9 B10 C3	Preguntas sobre os coñecementos adquiridos nas sesións maxistrais, lecturas, actividades prácticas e de problemas, cuestións e cuestionarios. É obrigatorio alcanzar un 40% da cualificación para superar a materia	50
Sesión maxistral	B3 B4 B5 B8 B9 B10 C3	Os contidos das sesións maxistrais serán avaliados na proba mixta e os estudiantes deben tamén saber aplicarlos a solución de problemas e cuestións, e na aplicación dos conceptos os traballos prácticos.	0

Observacións avaliación
Se non se obtén a puntuación mínima nas partes que o requiren, a nota máxima do alumno será 4'5. Para os alumnos a tempo parcial o baremo de cualificación e a avaliación continua son os mesmos que para os outros alumnos.

Fontes de información

Bibliografía básica	- W. Croft, D. Metzler, and T. Strohman. (2009). Search engines: Information retrieval in practice. Addison-Wesley - Manning, Christopher D. and Raghavan, Prabhakar and Schütze, Hinrich. 2008. (2008). . Introduction to Information Retrieval. Cambridge University Press https://ciir.cs.umass.edu/downloads/SEIRiP.pdfhttps://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdfhttps://ciir.cs.umass.edu/downloads/SEIRiP.pdfhttps://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf
Bibliografía complementaria	- Chengxiang Zhai, Sean Massung. (2016). Text Data Management and Analysis: A Practical Introduction to Information Retrieval. ACM Books - D. Jurafsky, JH Martin. (2009). Speech and Language Processing. Prentice Hall https://web.stanford.edu/~jurafsky/slp3/https://web.stanford.edu/~jurafsky/slp3/

Recomendacións
Materias que se recomenda ter cursado previamente
Materias que se recomenda cursar simultaneamente
Materias que continúan o temario
Observacións