GUÍA DOCENTE

BÚSQUEDA Y ANÁLISIS DE LA INFORMACIÓN

GRADO EN **INGENIERÍA DEL SOFTWARE**

MODALIDAD: A DISTANCIA

CURSO ACADÉMICO: 2023-2024

Denominación de la asignatura:	Búsqueda y Análisis de la Información
Titulación:	Ingeniería del Software
Facultad o Centro:	Centro Universitario de Tecnología y Arte Digital
Materia:	Ingeniería de Datos
Curso:	3º
Cuatrimestre:	2
Carácter:	OBM
Créditos ECTS:	6
Modalidad de enseñanza:	A distancia
Idioma:	Castellano
Profesor / Email:	Pedro Concejero Cerezo pedro.concejero@u-tad.com
Página Web:	http://www.u-tad.com/

DESCRIPCIÓN DE LA ASIGNATURA

Descripción de la materia

Los contenidos de la materia permiten a los alumnos comprender el flujo de búsqueda, ingesta, almacenamiento, procesamiento y análisis de información de datos y aproxima a los alumnos a las técnicas y tecnologías necesarias para la gestión de grandes cant

Descripción de la asignatura

El objetivo de esta asignatura es manejar con soltura las técnicas más modernas de captura de información de internet (conocido como “webscrapping”) además de las diferentes APIs ofrecidas por muchos servicios de internet para la captura de información disponible en internet. Se trata de una disciplina esencialmente práctica con infinidad de aplicaciones en muchas áreas de Ciencia de Datos, y para la que usaremos entornos opensource para los que hay multitud de librerías para los objetivos analíticos propuestos en la asignatura: R y python.

Pero qué duda cabe, como siempre en Ciencia de Datos, la preparación de datos (limpieza, consolidación), así como las técnicas básicas de manejo de textos (búsqueda, expresiones regulares) será el primer paso para poder abordar datasets reales de tamaño considerable.

Una vez obtenidos los datos, esta materia entra al detalle en dos áreas de aplicación de las técnicas de análisis y ciencia de datos que son muy especializadas por el tipo de datos en el que se basan:

Text Mining o minería de textos, para lo cual será necesario conocer la matemática que subyace al procesamiento de textos en lenguajes de estadística, como R (también python y otros). Veremos a continuación dos librerías o metodologías del entorno R para text mining: quanteda y el enfoque tidyverse. Y con ellas veremos cómo visualizar grandes cantidades de textos.

Por último veremos dos técnicas actuales de analítica de text mining: encontrar tópicos o temas en conjuntos de texto (Topic Mining) y análisis de sentimiento. Para abordar esta última parte de la asignatura contaremos con datasets resultado del procesamiento de todas las fases anteriores.

Y, en segundo lugar: Análisis de Redes Sociales (SNA en sus siglas inglesas) que utiliza toda la información disponible de nodos conectados entre sí (red social, o grafo conectado). Esta es una metodología potentísima en esta época de hiperconectividad. Como parte importante de las técnicas de SNA veremos la visualización, con la ayuda de un software fascinante: gephi.

Finalmente veremos aplicaciones de modelos Deep Learning en NLP, en concreto los llamados Large Language Models (LLM), algunos de los cuales (GPT-2 en concreto pero alguno más y seguro que irán saliendo según se desarrolle el curso) están disponibles en los frameworks más modernos de DL (tensorflow y pytorch

COMPETENCIAS Y RESULTADOS DE APRENDIZAJE

Competencias (genéricas, específicas y transversales)

COMPETENCIAS BÁSICAS Y GENERALES

CB1: Que los estudiantes hayan demostrado poseer y comprender conocimientos en un área de estudio que parte de la base de la educación secundaria general, y se suele encontrar a un nivel que, si bien se apoya en libros de texto avanzados, incluye también algunos aspectos que implican conocimientos procedentes de la vanguardia de su campo de estudio.

CB2: Que los estudiantes sepan aplicar sus conocimientos a su trabajo o vocación de una forma profesional y posean las competencias que suelen demostrarse por medio de la elaboración y defensa de argumentos y la resolución de problemas dentro de su área de estudio.

CB3: Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.

CB4: Que los estudiantes puedan transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado.

CB5: Que los estudiantes hayan desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía

CG1 - Capacidad para entender, planificar y resolver problemas a través del desarrollo de soluciones informáticas.

CG3 - Conocimiento de los fundamentos científicos aplicables a la resolución de problemas informáticos

CG4 - Capacidad para simplificar y optimizar los sistemas informáticos atendiendo a la comprensión de su complejidad

CG9 - Capacidad para aprender, modificar y producir nuevas tecnologías informáticas

CG10 - Uso de técnicas creativas para la realización de proyectos informáticos

CG11 - Capacidad de buscar, analizar y gestionar la información para poder extraer conocimiento de la misma

COMPETENCIAS ESPECÍFICAS

CE3 - Conocimiento del álgebra relacional y realización de consultas en lenguajes procedurales para el diseño de esquemas de

bases de datos normalizados basados en modelos de entidad-relación

CE10 - Capacidad para manejar un gestor de versiones de código y generar la documentación de una aplicación de forma

automática.

COMPETENCIAS TRANSVERSALES

CT1 - Conocimiento de la definición, el alcance y la puesta en práctica de los fundamentos de las metodologías de gestión de proyectos de desarrollo tecnológico

CT2 - Conocimiento de los principales agentes del sector y del ciclo de vida completo de un proyecto de desarrollo y comercialización de contenidos digitales

CT4 - Capacidad de actualización del conocimiento adquirido en el manejo de herramientas y tecnologías digitales en función del estado actual del sector y de las tecnologías empleadas

CT5 - Desarrollo de las habilidades necesarias para el emprendimiento digital.

Resultados de aprendizaje

Al acabar la titulación, el graduado o graduada será capaz de:

- Comprender e implementar los métodos de almacenamiento y administración eficaz en entornos distribuidos de datos no estructurados.

- Conocer y saber aplicar las distintas técnicas de aprendizaje supervisado, semi-supervisado y no supervisado.

- Entender y aplicar las técnicas de Deep learning

- Ser capaz de recuperar información mediante técnicas de web scraping o APIs normalizadas

- Entender y aplicar las técnicas de análisis del lenguaje natural

- Ser capaz de analizar contenidos de redes sociales

- Entender la naturaleza y representación de las imágenes digitales.

- Conocer las aplicaciones de las redes neuronales al análisis y generación de sonido, imagen estática y video.

- Desarrollar soluciones informáticas aplicadas a la visión por computador.

- Desarrollar un proyecto completo de datos aplicando metodología iterativa, desde el diseño hasta el despliegue.

CONTENIDO

Información textual. Modelos de relevancia y similaridad.

Búsqueda de información textual y no textual.

Búsqueda en la web.

Análisis de redes sociales.

TEMARIO

1. El ecosistema R

1.1. Entornos (IDE) de programación R: RStudio

1.2. El lenguaje R

1.3. librerías R: instalación, mantenimiento

1.4. The Tidyverse approach to data science.

2. Minería de Texto (Text Mining) con R.

2.1. Aspectos básicos del texto en R

- Character encoding

- Expresiones regulares

- La librería stringr

2.2. Librerías tm y quanteda

2.3. The tidyverse approach to text mining.

2.4. Visualización de texto: wordclouds.

3. Fuentes de datos de tipo texto

3.1. Twitter y otras redes sociales. API's a estas fuentes

3.2. Web scrapping.

3.3. Otras fuentes de datos

4. Modelización y análisis en minería de texto

4.1. Encontrar los tópicos de un texto: topicmodels.

4.2. Análisis de sentimiento.

5. Análisis de Redes Sociales (Social Network Analysis -SNA).

5.1. Aspectos teóricos de SNA

5.2. El caso enron

5.3. La librería igraph R

5.4. Visualizar redes: gephi.

5.5. Aplicaciones de SNA

6. Aplicaciones de Deep Learning en NLP: LLMs

6.1. Transformers

6.2. NLP en entornos keras /TF y pytorch.org

6.3. LLMs en las APIS TF / Pytorch

6.4. Huggingface (https://huggingface.co/)

ACTIVIDADES FORMATIVAS Y METODOLOGÍAS DE APRENDIZAJE

Actividades formativas

Actividad Formativa	Horas totales	Horas síncronas
Sesiones teóricas virtuales síncronas	4,25	4
Sesiones teóricas virtuales asíncronas	22,50	0
Sesiones prácticas virtuales síncronas	2,25	2
Sesiones prácticas virtuales asíncronas	10,75	0
Debate y discusión oral y/o escrita.	8,50	0
Tutorías	4,00	4
Estudio independiente y trabajo autónomo del alumno	50,00	0
Elaboración de trabajos (en grupo o individuales	33,25	0
Actividades de Evaluación	3,75	0
Test de autoevaluación	5,00	0

Seguimiento de proyectos	5,75	6
TOTAL	150	16

Metodologías docentes

Método expositivo o lección magistral

Aprendizaje de casos

Aprendizaje basado en la resolución de problemas

Aprendizaje basado en proyectos

Aprendizaje cooperativo o colaborativo

Aprendizaje por indagación

Metodología Flipped classroom o aula invertida

Gamificación

Just in time Teaching (JITT) o aula a tiempo

Método expositivo o lección magistral

Método del caso

Aprendizaje basado en la resolución de problemas

Aprendizaje basado en proyectos

Aprendizaje cooperativo o colaborativo

Aprendizaje por indagación

Metodología flipped classroom o aula invertida

Gamificación

DESARROLLO TEMPORAL

Presentación - semana 1

Unidad 1 - semana 2-3

Unidad 2 - semana 4-5

Unidad 3 - semana 6-7

Unidad 4 - semana 7-8

Unidad 5 - semana 9-10

Unidad 6 - semana 11-12

Repaso - semana 13-14

Evaluación - semana 15

SISTEMA DE EVALUACIÓN

ACTIVIDAD DE EVALUACIÓN	VALORACIÓN MÍNIMA RESPECTO A LA CALIFICACIÓN FINAL (%)	VALORACIÓN MÁXIMA RESPECTO A LA CALIFICACIÓN FINAL (%)
Evaluación de la participación en clase, en prácticas o en proyectos de la asignatura	10	20
Evaluación de trabajos, proyectos, informes, memorias	10	20
Prueba Objetiva	60	70

CRITERIOS ESPECÍFICOS DE EVALUACIÓN

ACTIVIDAD DE EVALUACIÓN	CONVOCATORIA ORDINARIA	CONVOCATORIA EXTRAORDINARIA
Evaluación de la participación en clase, en prácticas o en proyectos de la asignatura	20	10
Evaluación de trabajos, proyectos, informes, memorias	20	20
Prueba Objetiva	60	70

Consideraciones específicas acerca de la evaluación

Será necesario que obtener una nota mínima de 4 puntos (sobre 10) en la prueba final presencial para que se realice la media con las actividades formativas.

BIBLIOGRAFÍA / WEBGRAFÍA

Bibliografía básica (Text Mining)

R para profesionales de los datos: una introducción

https://www.datanalytics.com/libro_r/

Text Mining with R- A Tidy Approach

https://www.tidytextmining.com/

Bibliografía adicional (Text Mining)

Hay literalmente centenares de tutoriales en línea para cada uno de los temas que vamos a tratar. Estos son los más esenciales, en mi opinión:

https://bookauthority.org/books/best-text-mining-books

https://en.wikibooks.org/wiki/R_Programming/Text_Processing

https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

https://quanteda.io/

https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf

https://rstudio-pubs-static.s3.amazonaws.com/266565_171416f6c4be464fb11f7d8200c0b8f7.html

https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html

Gayo-Abello, D. (2023): Recuperación de información y minería de textos. Disponible online: https://danigayo.prof/teaching/RI-SIW-WebSem/

Bibliografía básica (SNA)

Tutorial igraph:

https://kateto.net/networks-r-igraph

https://kateto.net/tutorials/

Robert A. Hanneman and Mark Riddle (2005): Introduction to social network methods https://faculty.ucr.edu/~hanneman/nettext/

Tutoriales de gephi

https://gephi.org/users/

Bibliografía adicional (SNA)

No puede haber listado más completo de bibliografía SNA que la del curso (coursera) de Lada Adamic:

https://github.com/ladamalina/coursera-sna/blob/master/Syllabus.pdf

Y sobre España, twitter, los trabajos de Mari Luz Congosto (quien fue también profe. en UTAD) son geniales:

https://www.researchgate.net/profile/Mariluz_Congosto

Deep Learning y LLMs

Jurafsky, D. ; Martin, J. H. (2023): Speech and Language Processing (3rd ed. draft). Disponible online: https://web.stanford.edu/~jurafsky/slp3/

https://keras.io/guides/keras_nlp/getting_started/

https://keras.io/guides/keras_nlp/transformer_pretraining/

https://keras.io/examples/nlp/

https://pytorch.org/tutorials/index.html

https://keras.io/api/keras_nlp/

MATERIALES, SOFTWARE Y HERRAMIENTAS NECESARIAS

Materiales:
Ordenador personal

Software:
R última versión (https://cran.r-project.org/), en el momento de escribir esta guía,v. 4.3.2 (en general será necesaria versión >= 4.3.x)

y RStudio Desktop:(https://posit.co/download/rstudio-desktop/)

+ librerías stringr, tm, quanteda, tidyverse, tidyte