CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

PLANIFICACIÓN DE LA DOCENCIA UNIVERSITARIA

GUÍA DOCENTE

Búsqueda y Análisis de la Información

1. DATOS DE IDENTIFICACIÓN DE LA ASIGNATURA.

Título:	Grado en Ingeniería del Software
Facultad:	Centro Universitario de Tecnología y Arte Digital (U-TAD)
Materia:	Ingeniería de datos
Denominación de la asignatura:	Búsqueda y Análisis de la Información
Curso:	3º
Cuatrimestre:	2
Carácter:	Obligatoria de Mención
Créditos ECTS:	6
Modalidad/es de enseñanza:	Híbrido Presencial
Idioma:	Castellano
Profesor/a:	Pedro Concejero Cerezo
E-mail:	pedro.concejero@u-tad.com
Teléfono:	

2. DESCRIPCIÓN DE LA ASIGNATURA

2.1 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 cantidades de datos.

2.2 Descripción de la asignatura

Comenzaremos con una nueva para muchos alumnos (o a lo mejor no; en ese caso la repasaremos) herramienta: el lenguaje de estadística R y el ecosistema (librerías, entornos), que nos permitirá utilizar un entorno muy potente tanto para los objetivos de análisis como de visualización.

A continuación veremos SNA, y esto se debe a que es una metodología potentísima en esta época de hiperconectividad. De hecho, comenzamos por este punto para permitirnos profundizar en el análisis de textos una vez conozcamos el contexto que nos aporta la SNA. Como parte importante de las técnicas de SNA veremos la visualización, con la ayuda de un software fascinante: gephi.

Continuaremos con unas herramientas y procedimientos esenciales: los de captura de información a través de webscrapping o de las APIs de plataformas bien conocidas, muchas de ellas integradas en librerías R que facilitan enormemente su uso.

Y procederemos a continuación a la minería de textos propiamente dicha, para lo cual será necesario conocer muy bien la matemática que subyace al procesamiento de textos en lenguajes de estadística, como R (también python y otros). Una vez dominado este requisito, pasaremos a la tarea más engorrosa pero siempre necesaria: la limpieza de datos, o pre-proceso, en este caso de textos. Una labor que es absolutamente esencial porque de su calidad dependerá la calidad de los objetivos finales, ya sean meramente descriptivos o de modelización estadística.

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.

3. COMPETENCIAS Y RESULTADOS DE APRENDIZAJE

3.1. 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.

4. CONTENIDOS

4.1. Temario de la asignatura

1. R ecosystem.

1.1. R programming environments, RStudio
1.2. R language.
1.3. R libraries, installation, maintenance.
1.4. The Tidyverse approach to data science.

2. Social Network Analysis (SNA).

2.1. The basics of SNA. Theory(-ies).
2.2. The igraph R library.
2.3. Visualizing networks: gephi.
2.4. Applications of SNA (academy, business).

3. Sources of text data.

3.1. Twitter etc. API's to these sources.
3.2. Web scrapping.
3.3. Other sources.

4. Text Mining.

4.1. The basics of text in R
- Character encoding
- Regular Expressions
- Managing strings. The stringr package

4.2. The tm and quanteda libraries.
4.3. The tidyverse approach to text mining.
4.4. Visualizing text: wordclouds.

5. Text Mining analyses.

5.1. Finding topics in texts: topicmodels.
5.2. Sentiment Analysis.

4.2. Desarrollo temporal

UNIDADES DIDÁCTICAS / TEMAS	PERÍODO TEMPORAL
1. Ecosistema R - Rstudio - igraph	10-25 febrero
2. SNA intro. + data Revisión plan trabajo final	3-4 marzo
SNA igraph SNA computing metrics SNA computing communities Visualizing networks Gephi	10-25 marzo
Semana Santa	29 marzo – 4 abril
3. Web scrapping y otras fuentes de datos	7-8 abril
4. Text Mining with R Texto en R: la matemática Quanteda Tidyverse	14 – 29 abril
Text Mining Visualización (wordclouds)	5 -6 Mayo
Text Mining Models Topic Modeling Sentiment Analysis	12 -27 Mayo

5. ACTIVIDADES FORMATIVAS Y MODALIDADES DE ENSEÑANZAS

5.1. Modalidades de enseñanza

La asignatura se desarrollará a través de los siguientes métodos y técnicas generales, que se aplicarán diferencialmente según las características propias de la asignatura:

- **Método expositivo/Lección magistral:** el profesor desarrollará, mediante clases magistrales y dinámicas los contenidos recogidos en el temario.
- **Estudio de casos:** análisis de casos reales relacionados con la asignatura.
- **Resolución de ejercicios y problemas:** los estudiantes desarrollarán las soluciones adecuadas aplicando procedimientos de transformación de la información disponible y la interpretación de los resultados.
- **Aprendizaje basado en problemas:** utilización de problemas como punto de partida para la adquisición de conocimientos nuevos.
- **Aprendizaje orientado a proyectos:** se pide a los alumnos que, en pequeños grupos, planifiquen, creen y evalúen un proyecto que responda a las necesidades planteadas en una determinada situación.
- **Aprendizaje cooperativo:** Los estudiantes compartirán con todos sus compañeros la información tanto de la resolución de ejecicios, casos y problemas, como del proyecto que se plantee para el fin de la asignatura.

5.2. Actividades formativas

Actividad Formativa	Horas	Presencialidad
AF1 Clases teóricas / Expositivas	45	100%
AF2 Clases Prácticas	36	100%
AF3 Tutorías	9	50%
AF4 Estudio independiente y trabajo autónomo del alumno	57,5	0%
AF5 Elaboración de trabajos (en grupo o individuales)	28,5	0%
AF6: Actividades de Evaluación	9	100%

6. SISTEMA DE EVALUACIÓN

ACTIVIDAD DE EVALUACIÓN	VALORACIÓN MÍNIMA RESPECTO A LA CALIFICACIÓN FINAL (%)	VALORACIÓN MÁXIMA RESPECTO A LA CALIFICACIÓN FINAL (%)
SE1 Evaluación de la participación en clase, en prácticas o en proyectos de la asignatura	10%	30%
SE2 Evaluación de trabajos, proyectos, informes, memorias	40%	80%
SE3 Prueba Objetiva	10%	60%

6.1. Criterios de calificación

ACTIVIDAD DE EVALUACIÓN	VALORACIÓN RESPECTO A LA CALIFICACIÓN FINAL (%)
Evaluación de la participación en clase, en prácticas o en proyectos de la asignatura	10%
Evaluación de trabajos, proyectos, informes, memorias	45%
Prueba Objetiva	45%

Consideraciones generales acerca de la evaluación:

- La evaluación de la participación en clase, en prácticas o en proyectos de la asignatura se realizará a partir de la asistencia y la participación activa en clase y en el resto de las actividades desarrolladas durante el curso. Este aspecto representará el 10% de la calificación final de la asignatura en la convocatoria ordinaria.
- A lo largo del curso se plantearán actividades, ejercicios y problemas que deberán ser entregadas antes de la fecha indicada a través de la plataforma virtual. El trabajo final se plantea como de entrega obligatoria a través de la propia plataforma virtual y supondrá un 45% de la calificación final de la asignatura en la convocatoria ordinaria.
- A mitad de cuatrimestre se realizará un examen parcial, que será liberatorio si así lo desea el alumno con la condición de obtener al menos una calificación de 4.0 en dicho examen. Aquellos alumnos que no superen esa nota o que decidan descartarla voluntariamente, deberán realizar sendos exámenes correspondientes a los dos parciales en la fecha asignada para la convocatoria ordinaria de enero. **Los dos exámenes parciales representarán el 45% de la calificación final** en la convocatoria ordinaria (22,5% cada uno).

- En las primeras clases se propondrá un trabajo para desarrollar a lo largo de toda la asignatura, incorporando los conocimientos y capacidades que se vayan viendo en la misma. Habrá una gran libertad sobre el tema del que trate el trabajo o proyecto, así como de su formato de presentación. Pero será imprescindible que se comparta de forma periódica en primer lugar con el profesor, para ajustar tanto el alcance como la metodología usada, y en momentos concretos del desarrollo de la asignatura, con el resto de estudiantes. Esto se conoce como "peer review" o revisión por pares, y como tal, será parte de la evaluación tanto del proponente del trabajo como de los revisores.
  Este trabajo se presentará de forma inaplazable antes del 15 de enero de 2021, y su evaluación definitiva que tendrá en cuenta las revisiones por pares comentadas anteriormente, supondrá el 45% de la puntuación total de la asignatura, y para que esta calificación pueda promediar con el resto de calificaciones (exámenes) deberá superar un 4 (sobre 10).

- Para aprobar la asignatura en la convocatoria ordinaria la media aritmética de todos los elementos de calificación (exámenes parciales, trabajo o proyecto final y la participación o asistencia a clases) según la siguiente fórmula

  $calificaciones\_examen\_y\_trabajo = [(calif\_trabajo) + (control_1 + control_2)/2] /2$
  $Promedio\_final\_asignatura = 0.9*calificaciones + 0.1*asistencia$

  deberá ser igual o superior a 5 (siempre sobre 10). Para poder calcular este promedio, cada una de las calificaciones que entren en el cálculo deberá ser igual o superior a 4, de tal modo que para compensar puntuaciones menores que 5 en alguno de los componentes de la fórmula, las otras calificaciones deben ser claramente superiores a este valor. **En caso de no cumplirse alguno de estos requisitos, la asignatura se considerará automáticamente suspensa independientemente del resto de calificaciones.**

- En caso de no conseguir el aprobado en la convocatoria ordinaria de enero, el alumno podrá presentarse a la convocatoria extraordinaria de junio/julio 2021, donde realizará un examen final que representará el 50% de su calificación en dicha convocatoria, y en el que formará parte de la materia exigible al alumno todo el contenido de la asignatura visto en clase, y el restante 50% será el trabajo o proyecto que será en cualquier caso igualmente obligatorio para esta convocatoria.
- No se conservarán calificaciones de ningún tipo entre distintos cursos académicos, ni entre distintas convocatorias.

Consideraciones generales acerca del desarrollo de las clases:

- No está permitido consumir bebidas ni comidas en el aula. Tampoco está permitida la presencia de cualquier tipo de bebida en las mesas, incluso en envases cerrados.
- Se demandará del alumno una participación activa, necesaria para el desarrollo de las clases.
- Se exigirá al alumno un buen comportamiento en todo momento durante el desarrollo de las clases. El mal comportamiento que impida el normal desarrollo de la clase puede conllevar la expulsión del aula por un tiempo a determinar por el profesor.

7. 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

Bibliografía básica (SNA)

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

Robert A. Hanneman and Mark Riddle
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 (también profe. en UTAD) son geniales:

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

8.- MATERIAL, SOFTWARE Y HERRAMIENTAS NECESARIAS

TIPOLOGÍA DEL AULA:
Sala virtual del curso en Blackboard

MATERIALES DEL ALUMNO:
Ordenador personal, webcam y micrófono

SOFTWARE:
R 4.0 https://cran.r-project.org/
RStudio https://rstudio.com/products/rstudio/download/
librerías stringr, tm, quanteda, topicmodels
igraph for R: https://igraph.org/r/
Gephi https://gephi.org/