Minería de Texto

Curso Académico: (2024 / 2025) **Fecha de revisión: 30-04-2024**

Departamento asignado a la asignatura: Departamento de Estadística

Coordinador/a: UCAR MARQUES, IÑAKI

Tipo: Obligatoria Créditos ECTS : 3.0

Curso : 1 Cuatrimestre : 2

REQUISITOS (ASIGNATURAS O MATERIAS CUYO CONOCIMIENTO SE PRESUPONE)

Programación de Datos (19138)

OBJETIVOS

- Conocimiento de las estructuras y procedimientos propios de la minería de textos.
- Capacidad para usar métodos básicos de extracción de información de datos textuales.
- Capacidad de aplicar técnicas de procesamiento para preparar documentos para su modelado estadístico.
- Capacidad para evaluar y usar modelos básicos de predicción de información textual.

DESCRIPCIÓN DE CONTENIDOS: PROGRAMA

1. Introducción teórica al Procesamiento del Lenguaje Natural
1.1. Breve historia de la lingüística computacional y principales avances
1.2. Qué es el PLN y su papel en la Inteligencia Artificial
1.3. Estructura de un pipeline básico de PLN
1.4. Tareas y aplicaciones más comunes en la industria
1.5. Importancia actual en la sociedad digital, principales iniciativas

2. Introducción práctica al análisis automático del lenguaje con R
2.1. Importación de texto original, diseño de dataset y creación de estructuras de datos
2.2. Limpieza de texto, eliminación de stopwords y símbolos, valores ausentes y duplicados
2.3. Procesos de división y tokenización
2.4. Análisis básicos: conteo de palabras, extracción de ngramas, tablas de frecuencias
2.5. Análisis intermedios: análisis de distintividad, tf-idf, bag of words

3. Introducción al análisis de sentimiento
3.1. Qué es el análisis automático del sentimiento en un texto: la opinión, la emoción y la intención del emisor
3.2. Casos reales de análisis de sentimiento en la industria y limitaciones
3.3. Ejercicios prácticos de análisis automático de sentimiento: uso de lexicones y diccionarios, asignación automática de sentimiento, segmentación,
nubes de palabras
3.4. Creación de gráficos e informes de análisis de sentimiento

4. Introducción al modelado de tópicos
4.1. Qué es el modelado de tópicos, usos en la industria
4.2. Clasificación de textos en categorías: métodos supervisados y no supervisados
4.3. Ejercicios prácticos de modelado de tópicos: asociación de palabras y tópicos, identificación y caracterización de grupos naturales, términos comunes y solapamiento
4.4. Creación de gráficos e informes de modelado de tópicos para identificación de ideas representativas

5. Modelos de lenguaje
5.1. Qué son los modelos de lenguaje pre-entrenados y su impacto en el desarrollo del PLN y el

aprendizaje automático
5.2. Usos e implicaciones en la industria y situación actual, principales iniciativas
5.3. Ejercicios prácticos de uso y evaluación de modelos predictivos básicos con datos en texto

ACTIVIDADES FORMATIVAS, METODOLOGÍA A UTILIZAR Y RÉGIMEN DE TUTORÍAS

Actividades Formativas:

- Clases teórico-prácticas
- Tutorías
- Trabajo individual del estudiante
- Exámenes parciales y finales

Metodologías Docentes:

- Exposiciones en clase del profesor con soporte de medios informáticos y audiovisuales, en las que se desarrollan los conceptos principales de la materia y se proporciona la bibliografía para complementar el aprendizaje de los alumnos.
- Resolución de casos prácticos, problemas, etc., planteados por el profesor de manera individual o en grupo.
- Exposición y discusión en clase, bajo la moderación del profesor de temas relacionados con el contenido de la materia, así como de casos prácticos.
- Elaboración de trabajos e informes de manera individual o en grupo.

SISTEMA DE EVALUACIÓN

Peso porcentual del Examen Final:	40
Peso porcentual del resto de la evaluación:	60

- Participación en clase (10%)
- Trabajos individuales o en grupo realizados durante el curso (50%)
- Examen final (40%)

En la convocatoria extraordinaria, el sistema de evaluación será el siguiente:
1) Examen: 100%

BIBLIOGRAFÍA BÁSICA

- Gabe Ignatow and Rada F. Mihalcea An Introduction to Text Mining: Research Design, Data Collection, and Analysis., SAGE Publications, 2017

- Silge, J., & Robinson, D. Text mining with R: A tidy approach, O'Reilly Media, 2017

BIBLIOGRAFÍA COMPLEMENTARIA

- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed.), PEARSON, Prentice Hall, 2021

- Kumar, A., & Paul, A. Mastering text mining with r: Master text-taming techniques and build effective text-processing applications with R, Packt Publishing Limited, 2016

- Kwartler, T. Text mining in practice with R, Winley, 2017

- Marchette, D. J. Text data mining using R, Chapman & Hall Crc, 2018

- Ted Kwartler Text Mining in Practice with R, Wiley, 2017

RECURSOS ELECTRÓNICOS BÁSICOS

- Dan Jurafsky and James H. Martin . Speech and Language Processing (3rd ed.): http://https://web.stanford.edu/~jurafsky/slp3

- Julia Silge and David Robinson . Tex Mining with R: http://https://www.tidytextmining.com/