CARRERA: Licenciatura en sociología

ASIGNATURA:Taller de investigación IV (Minería de texto)

DOCENTE (s) A CARGO: Lic. Ariana Bardauil

AÑO: 2025

I. FINALIDAD Y/U OBJETIVOS

La asignatura Taller de investigación IV tiene como propósito introducir al estudiante en el análisis de datos no estructurados en formato textual, articulando un enfoque teórico-práctico. Mediante la integración de metodologías de investigación cualitativa y análisis estadístico, se pondrá especial énfasis en el procesamiento de grandes volúmenes de texto utilizando herramientas de Procesamiento de Lenguaje Natural, Recuperación de Información y Aprendizaje Automático con énfasis en su aplicación en proyectos de investigación sociológica. Todas las actividades prácticas y los análisis serán realizados utilizando el lenguaje de programación **R**

I.1. Objetivo general y específicos (obligatorio)

Objetivo general

Que los/las estudiantes se introduzcan en técnicas y herramientas para extraer conocimiento relevante a partir de grandes volúmenes de datos textuales no estructurados, con un enfoque en su aplicación a investigaciones en ciencias sociales.

Objetivos específicos

1. Comprender los principios teóricos y metodológicos del procesamiento de lenguaje natural (PLN) y el text-mining.
2. Desarrollar habilidades para la extracción, limpieza, preprocesamiento y análisis de datos textuales provenientes de diferentes fuentes.

3. Explorar técnicas de análisis exploratorio de datos (EDA) aplicadas a textos
4. Implementar técnicas avanzadas como modelado de tópicos, análisis de redes aplicado a textos y vectorización/embeddings.
5. Diseñar e implementar análisis predictivos y descriptivos usando machine learning para textos.
6. Reflexionar y evaluar el uso de técnicas de text-mining para la investigación.

II. CONTENIDOS

Unidad 1. La minería de texto (text-mining) y la investigación social.

Fundamentos y conceptos de Text Mining. Datos estructurados y no estructurados. Análisis textual y documental. Flujo de trabajo en un proyecto de text mining. Text mining en el contexto de la investigación cualitativa y cuantitativa. Usos y aplicaciones en ciencias sociales.

Unidad 2. Técnicas de recolección y manipulación de datos textuales.

Manipulación del texto a través de herramientas específicas del lenguaje R, como el manejo de strings y expresiones regulares, para realizar transformaciones y análisis iniciales. Exploración de fuentes de datos textuales relevantes para la investigación. Conexión a APIs, utilización de CURL. Subtítulos de youtube. Web scraping para recolectar datos de la web y redes sociales.

Unidad 3. Preparación de Datos Textuales y Análisis exploratorio

Aplicación de técnicas de preprocesamiento como tokenización, lematización y eliminación de ruido textual, junto con técnicas avanzadas como Partes de la Oración (POS Tagging), Desambiguación del Significado de las Palabras (WSD) y Reconocimiento de Entidades Nombradas (NER). Medidas de similitud, distancia y diversidad de léxico en textos. Sentiment Analysis.

Unidad 4. Modelos de clasificación y representación de textos

Modelos de representación vectorial de textos, como TF-IDF, word2vec y GloVe, para descubrir patrones semánticos y contextuales. Exploración de técnicas avanzadas como modelado de tópicos (LDA, STM) para analizar discursos e identificar temas latentes. Implementación de algoritmos para clasificación de documentos con paquetes específicos de R como topicmodels y text2vec

Unidad 5. Análisis predictivo y Network Analysis

Implementación de técnicas de Machine Learning en R para la predicción y clasificación de datos textuales: Modelos lineales (regresión logística y Naive Bayes), no lineales (Support Vector Machines - SVM y Random Forest), clustering (k-means, DBSCAN) y modelos avanzados como redes neuronales y transformers (BERT). Network Analysis: creación de grafos, cálculo de métricas de red (centralidad, modularidad), y visualización para interpretar relaciones en corpus textuales

III. ACTIVIDADES y METODOLOGÍA

El curso se desarrollará a través de una combinación de actividades sincrónicas y asincrónicas, diseñadas para garantizar un aprendizaje dinámico y flexible. Las actividades asincrónicas incluirán exposiciones breves a cargo del docente, enfocadas en los contenidos de cada unidad, junto con debates teóricos y demostraciones prácticas. Estas se complementarán con ejercicios prácticos, foros de discusión, y la resolución de consignas breves basadas en materiales específicos y guías de lectura. Por su parte, las clases sincrónicas estarán dedicadas a la resolución de problemas, la atención de consultas y el acompañamiento personalizado, permitiendo un seguimiento cercano de los avances y necesidades de aprendizaje de los/las estudiantes.

Programación Didáctica

Clase N°	Unidad	Objetivos	Contenidos	Actividades a desarrollar	Material didáctico	Criterios de Evaluación	Recursos (tipo de soporte)	Bibliografía
1	**1**	Presentación de la materia. Introducir el concepto de minería de texto. Comprender la diferencia entre datos estructurados y no estructurados. Explorar aplicaciones prácticas del Text Mining.	Concepto y aplicaciones del Text Mining. Introducción a datos no estructurados. Ejemplos de minería de texto aplicada a problemas sociales.	Exposición teórica introductoria con ejemplos. Análisis de un caso práctico con: Ejercitación de repaso de R y R studio	Presentación	Participación en clase		Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). Fundamentals of Predictive Text Mining. Capítulo 1. Friedl, J. E. F. (2006). Mastering regular expressions: Understand your data and be more productive (3.ª ed.). O'Reilly Media.

Clase N°	Unidad	Objetivos	Contenidos	Actividades a desarrollar	Material didáctico	Criterios de Evaluación	Recursos (tipo de soporte)	Bibliografía
2	**1**	Familiarizarse con herramientas en R para manipular texto. Aprender sobre strings y expresiones regulares	Taller práctico: Introducción a R y manipulación básica de texto. Resolución de ejercicios prácticos usando expresiones regulares.	Introducción a R y manipulación básica de texto. Utilización de expresiones regulares.	Archivo .qmd con el contenido de la clase.	Entrega de ejercitación práctica Participación en clase		Documentación: https://www.rdocumentation.org/packages/stringr/versions/1.5.1
3	1	Text Mining y Sociología						
4	2	Introducción a fuentes de datos textuales, APIs y CURL	Introducir las principales fuentes de datos textuales relevantes para la investigación en ciencias sociales. Enseñar cómo	Fuentes de datos textuales (APIs públicas, redes sociales, documentos digitales). Introducción a CURL para la	Cómo configurar y utilizar CURL para realizar peticiones a una API (por ejemplo, Twitter o YouTube). Descarga de	Archivo .qmd con ejemplos y ejercicios prácticos. Documentación de APIs de ejemplo.		Documentación: https://cran.r-project.org/web/packages/curl/curl.pdf

Clase N°	Unidad	Objetivos	Contenidos	Actividades a desarrollar	Material didáctico	Criterios de Evaluación	Recursos (tipo de soporte)	Bibliografía
			consumir APIs utilizando CURL y R. Aprender a descargar y trabajar con subtítulos de YouTube.	conexión con APIs. Descarga y manejo de subtítulos de YouTube como fuente textual.	subtítulos de YouTube usando herramientas específicas.			
5	2	Web scraping básico y scraping de redes sociales.	Enseñar los conceptos básicos y buenas prácticas del web scraping. Aprender a utilizar bibliotecas de R para recolectar datos textuales desde páginas web	Introducción al web scraping: conceptos, ética y limitaciones legales. Uso de la biblioteca rvest para scraping básico. Introducción a Selenium	archivo .qmd con teoría y práctica	Ejercitación práctica: Diseñar un scraping básico con rvest en un sitio web seleccionado. Guardar los datos obtenidos en un archivo CSV		

Clase N°	Unidad	Objetivos	Contenidos	Actividades a desarrollar	Material didáctico	Criterios de Evaluación	Recursos (tipo de soporte)	Bibliografía
			Librerías para recolección de datos de redes sociales (twitterscrapeR)					
6	3	Preprocesamiento básico: tokenización, lematización, limpieza y eliminación de ruido textual	Aprender a procesar textos en R mediante tokenización, eliminación de stopwords y lematización. Implementar técnicas para eliminar ruido textual (caracteres especiales, puntuación, normalización de texto).	Tokenización Stopwords y limpieza Lematización, raiz de palabra, stemming Eliminación de ruido textual	archivo .qmd con teoría y práctica	Ejercitación práctica.		

Clase N°	Unidad	Objetivos	Contenidos	Actividades a desarrollar	Material didáctico	Criterios de Evaluación	Recursos (tipo de soporte)	Bibliografía
7	3	Técnicas avanzadas: POS Tagging, WSD y NER con paquetes de R.	Implementar etiquetado de Partes del Discurso (POS Tagging) en textos en español. Introducir la Desambiguación del Significado de Palabras (WSD). Aplicar Reconocimiento de Entidades Nombradas (NER).	POS Tagging WSD NER Consulta previa a la primer entrega del trabajo final:Presentación de ideas iniciales, selección de corpus y metodología.	archivo .qmd con teoría y práctica	Ejercitación práctica.		
8	3	Análisis exploratorio de textos y visualización con nubes de palabras y	Aplicar técnicas de análisis exploratorio de texto en R. Introducir el análisis de	EDA en Text Mining: Medidas de diversidad léxica Medidas de similitud y	archivo .qmd con teoría y práctica	Ejercitación práctica. Entrega del proyecto de		

Clase N°	Unidad	Objetivos	Contenidos	Actividades a desarrollar	Material didáctico	Criterios de Evaluación	Recursos (tipo de soporte)	Bibliografía
		gráficos. Medidas de similitud, distancia y diversidad léxica en textos. Introducción al Análisis de Sentimiento	sentimiento Medir la diversidad léxica de un corpus textual. Visualizar la distribución de palabras y similitud entre documentos.	distancia entre documentos Visualización de textos Introducción al Análisis de Sentimiento		trabajo final de la materia: Documento con tema, objetivos, pregunta de investigación, metodología y fuentes de datos.		
9	4	Representación vectorial de textos (TF-IDF, word2vec, GloVe)	Comprender los métodos de representación vectorial de textos y sus aplicaciones. Implementar TF-IDF para modelar textos basados en frecuencia y	Introducción a la representación vectorial de textos Diferencias entre representaciones basadas en frecuencia y embeddings. TF-IDF (Term Frequency - Inverse Document	Archivo .qmd con teoría y práctica	Ejercitación práctica.		

Clase N°	Unidad	Objetivos	Contenidos	Actividades a desarrollar	Material didáctico	Criterios de Evaluación	Recursos (tipo de soporte)	Bibliografía
			relevancia. Explorar representaciones semánticas avanzadas como Word2Vec y GloVe.	Frequency) Embeddings (Word2Vec y GloVe)				
10	4	Aplicaciones del text mining						
11	4	Modelado de tópicos (LDA, STM) con paquetes de R	Introducir el modelado de tópicos como una técnica de agrupación de textos.	Introducción al modelado de tópicos. Interpretación de los resultados Espacio de consultas sobre avance del proyecto -	Archivo .qmd con teoría y práctica	Ejercitación práctica. Avance con ejemplos de preprocesamiento y primeras visualizaciones.		

Clase N°	Unidad	Objetivos	Contenidos	Actividades a desarrollar	Material didáctico	Criterios de Evaluación	Recursos (tipo de soporte)	Bibliografía
				preprocesamiento y exploración de datos				
12	5	Clasificación de documentos y métricas de evaluación.	Introducir modelos supervisados de clasificación de texto. Implementar Naive Bayes, Regresión Logística y SVM en R. Evaluar el rendimiento de los modelos con métricas de clasificación.	Introducción a la clasificación de textos Modelos de clasificación supervisados Métricas de evaluación	Archivo .qmd con teoría y práctica	Ejercitación práctica.		

Clase N°	Unidad	Objetivos	Contenidos	Actividades a desarrollar	Material didáctico	Criterios de Evaluación	Recursos (tipo de soporte)	Bibliografía
13	5	Algoritmos de clustering: k-means, DBSCAN y ejemplos aplicados	Introducir el concepto de clustering y su aplicación en textos. Implementar y comparar los algoritmos k-means y DBSCAN en datos textuales. Evaluar los resultados de los modelos en función de la estructura de los datos.	Introducción al clustering en minería de texto k-means: Algoritmo de clustering basado en particiones DBSCAN: Algoritmo basado en densidad	Archivo .qmd con teoría y práctica	Ejercitación práctica.		
14	5	Introducción a Network Analysis: creación de	Introducir el análisis de redes aplicado a textos. Construir grafos	Introducción al análisis de redes Construcción de grafos a partir de	Archivo .qmd con teoría y práctica	Ejercitación práctica.		

Clase N°	Unidad	Objetivos	Contenidos	Actividades a desarrollar	Material didáctico	Criterios de Evaluación	Recursos (tipo de soporte)	Bibliografía
		grafos y métricas clave	de co-ocurrencias de palabras y redes semánticas. Calcular métricas clave de centralidad y modularidad en redes textuales.	textos Métricas clave en análisis de redes Visualización				
15	5	Clase de consulta				Participación en clase		
16	5	Presentación del Proyecto Final y Evaluación						

IV. EVALUACIÓN Y PROMOCIÓN

La materia contempla la posibilidad de promoción directa. Para acceder a esta instancia, se requiere un mínimo del 75% de asistencia a las clases o participación en actividades asincrónicas y la aprobación del trabajo final con una calificación mínima de 7 (siete) puntos.

Para aquellos que no cumplan con los requisitos de promoción, pero deseen rendir el examen final, se exigirá la entrega del trabajo de investigación con al menos una calificación de 4 (cuatro) puntos.

La evaluación de la materia estará basada en la realización de trabajos prácticos y un proyecto de investigación. Los trabajos prácticos se desarrollarán a lo largo de la cursada y estarán orientados a la aplicación de técnicas específicas vistas en clase. El Trabajo de Investigación Final se desarrollará a lo largo del curso y tendrá dos entregas obligatorias: (1) presentación del proyecto de investigación y, (2) informe final con resultados. Se espera que se construya a partir de la aplicación de las técnicas trabajadas en la materia y análisis de hallazgos.

La materia prevé una instancia de recuperación del trabajo final en caso de aplazo. La regularidad del curso tendrá una vigencia de 2 años.

V. RECURSOS

Para las actividades sincrónicas se requiere una sala de reunión virtual que permita grabar videos y utilizar herramientas de chat y participación. Para las actividades asincrónicas se requerirá un aula virtual que permita el armado de foros, consignas, alojar materiales y recursos.

VI. BIBLIOGRAFIA OBLIGATORIA y BIBLIOGRAFÍA COMPLEMENTARIA

Becerra, G., & López Alurralde, J. P. (2021). Topic modeling y los desafíos de la investigación cualitativa. *XIV Jornadas de Sociología*. http://jornadasdesociologia2021.sociales.uba.ar/wp-content/uploads/ponencias2021/2092_282.pdf

Jurafsky, D., & Martin, J. H. (2006). *Speech and Language Processing: An introduction to natural language processing*. Prentice Hall.

Bechmann, A., & Bowker, G. C. (2019). Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data and Society*, 6(1), 1–11. https://doi.org/10.1177/2053951718819569

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297. https://doi.org/10.1093/pan/mps028

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of Predictive Text Mining*. Springer London. https://doi.org/10.1007/978-1-84996-226-1

Silge, J., & Robinson, D. (2017). *Text Mining with R: A tidy approach* (1st ed.). O'Reilly Media. ISBN: 978-1491981658. Disponible en **https://www.tidytextmining.com/**

Friedl, J. E. F. (2006). *Mastering regular expressions: Understand your data and be more productive* (3.ª ed.). O'Reilly Media.