Datos no estructurados y semiestructurados

Especialización en Ciencia de Datos (Decon)

Consigna de Trabajo Final

2024

Características generales

Trabajo individual para aprobación del curso *Datos no estructurados y semiestructurados* - Especialización en Ciencia de Datos (Decon/Facultad de Ciencias Sociales - UdelaR).

Plazo de entrega

18/6/2024 a las 23:59 horas

Formato

Subir a la plataforma EVA del curso y por correo electrónico (elina.gomez@cienciassociales.edu.uy | nschmidt@cienciassociales.edu.uy un archivo .Rmd, .pdf y/o .html correspondiente a su compilación, así como archivos y/o base de datos si son de libre acceso (o una muestra de las mismas) que permita la reproducción del documento.

Consigna

- *Recuperación:* Elegir una fuente de datos de las vistas en el curso (archivos de texto brutos, OCR, web scraping, scrapeo parlamentario, prensa digital, búsquedas de google, transcripción de audio, subtítulos de YouTube, APIs de Google, entre otras) e incluir el proceso de extracción o reconstrucción de la información.
- *Pre-procesamiento:* Crear un corpus de datos textuales tabulados, realizar el pre-procesamiento utilizando alguna/s de las herramientas vistas en el curso (manipulación de strings, limpieza o pre/codificación manual)
- *Análisis:* Incluir al menos tres técnicas de minería de texto vistas en el curso (frecuencia de ocurrencia, asociación de palabras, contexto de aparición de palabras o frases, diccionarios, análisis de sentimiento, modelado de temas, entre otros posibles)
- *Visualización:* Ilustrar el documento con al menos dos visualizaciones que surjan del procesamiento del texto precedente y den cuenta de los resultados del mismo.

Se valorará

- Propuesta de fuentes de datos novedosa
- Originalidad en las técnicas de procesamiento
- Resultados e interpretaciones relevantes

Consultas

elina.gomez@cienciassociales.edu.uy | nschmidt@cienciassociales.edu.uy