Preproceso, Recolección y Visualización de Datos Requisitos del proyecto

Para la realización del proyecto es necesario escoger tres o cuatro fuentes de datos relacionadas y plantear un posible proyecto analítico que se pueda solventar con dichas fuentes. Una vez realizada la selección, nos centraremos en procesar, limpiar, unir y realizar finalmente un análisis descriptivo sobre los datos. También se puede obtener información derivada durante el procesado de datos.

El proyecto debe incluir un informe que responda, al menos, a las siguientes cuestiones:

1. **Objetivos del proyecto:** Toda selección de fuentes debe de responder a uno o más objetivos de análisis. Se deben exponer claramente qué objetivos se persiguen y qué requisitos deben de cumplir las fuentes para responder a estos objetivos de forma adecuada. A modo de ejemplo se plantean algunas preguntas: ¿Cómo de fiable debe de ser fiable la información a nivel individual? ¿Es necesario extraer información de forma continua o es suficiente con una sola extracción?
2. **Selección de fuentes:** A partir de los objetivos establecidos se debe determinar la naturaleza de las fuentes. ¿Son fuentes internas de una empresa o externas? ¿Son necesarios datos de redes sociales, deben obtenerse de un API o estamos buscando un dataset de Open Data? Tras tener clara su naturaleza se debe buscar la información en los repositorios correspondientes.
3. **Descripción de las fuentes seleccionadas:** Una vez seleccionadas las fuentes se debe presentar y documentar adecuadamente cada una de ellas. Tipo de fuente, formato, autoridad que las publica, enlace donde se encuentra, accesibilidad, disponibilidad, clasificación open data, fecha de última actualización de la publicación de las fuentes, etc.
4. **Metodología y análisis de fuentes individuales:** Herramientas utilizadas, proceso de búsqueda de valores anómalos, evaluación del cumplimiento de los requisitos establecidos en el paso 1. Resumen de problemas encontrados en los datos y cómo se han solucionado.
5. **Integración de datos:** Cómo se ha realizado el proceso de unión de las diferentes fuentes y qué restricciones, si existen, hay para la alineación de los datos de las distintas fuentes.
6. **Extracción de información:** Realización del análisis descriptivo a partir de los diferentes tipos de gráficas vistos en clase y extracción de conclusiones a partir de ellas
7. **Conclusiones del proyecto:** Dificultades generales del proyecto, mayor dificultad afrontada, mayores dificultades personales a la hora de abordar las diferentes tareas del proyecto.
8. **Conjunto de datos final obtenido:** Enlace a la descarga. Por ejemplo un archivo compartido en Google Drive.

Dado que la asignatura está centrada en el preprocesado, recolección y visualización de los datos, será esto mismo lo que tenga el mayor peso en la evaluación del proyecto. **NO SE PENALIZARÁ** si las conclusiones extraídas son más o menos robustas, o interesantes. La evaluación estará centrada en el proceso de recolección de los datos (APIs, repositorios), el preprocesado realizado para unir y limpiar los datos de las diferentes fuentes, y la visualización de los datos en diferentes gráficas.

Dado que la salida de este proyecto será un conjunto de datos preprocesado y limpio, esto permite

su reutilización como base para otros proyectos (minería, análisis, etc.), más centrados en el aspecto analítico de IA que puedan ser aplicables sobre el conjunto de datos obtenido.

Por ello, se recomienda que la selección de fuentes se haga pensando en el interés futuro del grupo. Por ejemplo, si hay mayor interés por el análisis de datos médicos, sería interesante buscar fuentes de datos relacionadas a algún campo específico (por ejemplo, COVID) seleccionando fuentes que contengan información que a priori permitan responder un determinado problema u objetivo analítico.

El proyecto se llevará a cabo en grupos de dos a cuatro personas (cinco de manera excepcional).

La **presentación** del proyecto se entregará a través de UA Cloud con fecha límite el 29 de Enero de 2026.