Proyecto programado 1: Arañador web

Aurelio Sanabria
Recuperación de la información textual
I semestre, 2022

Motivación

Descargar información de sitios web es esencial para mantener actualizadas colecciones de documentos con información fresca y actualizada. Es por esto que para este proyecto diseñaremos un arañador web que se ajuste a los intereses de las personas estudiantes del curso. Tal como se discutió en clase, la temática y funcionalidad de este arañador estará sujeto a las decisiones que tome cada grupo, siempre y cuando estas se justifiquen debidamente.

Objetivos formativos

5. Desarrollar un conocimiento práctico para resolver problemas relacionados con el manejo de texto.
6. Aprender diferentes herramientas y lenguajes usados para el procesamiento de información textual

Especificación del proyecto

Para este proyecto cada grupo debe recorrer toda la información construida durante el curso, ampliarla con investigación propia y decidir lo siguiente:

1. Tema central de la información a descargar
2. Detalles de implementación del arañador
3. Políticas a implementar

Cada decisión debe ser documentada y justificada de forma que se evidencie un criterio basado en hechos y fuentes bibliográficas confiables.

Metodología

Para completar este proyecto se les sugiere a lxs estudiantes de cada grupo pasen por las siguientes etapas.

1. Repasen la materia vista en clase.
2. Investiguen para complementar lo visto en clase.
3. Discutan resultados y los ajusten con respecto al tema y tipo de arañador a construir.
4. Diseñen y documenten sus decisiones sobre la implementación
5. Implementen en `Python 3` un arañador.

Rúbrica

1. Documento

- 15pts **Motivación**: Que se va a buscar (tema) y que impulsó a elegirlo
- 15pts **Justificación de la implementación**: Describir y justificar las políticas elegidas para el modelo. Uso del calendarizador, multihilo y partes del preprocesamiento.
- 10pts **Análisis de resultados**: Analizar los resultados obtenidos de la elaboración del proyecto de forma instrospectiva.
- 10pts **Memes (4 x integrante)**: ¿Cómo se sintieron programando el proyecto?

2. Colección

- 15pts **Captura correcta del documento**: Documentos almacenados de forma correcta y legible.
- 15pts **Relevancia de los documentos**: Que los documentos sean relevantes a la búsqueda que se realizó

3. Arañador

- 15pts Funcionamiento del código
- 15pts Calidad del código
- 5pts Documentación interna

Estimación de tiempo

1. (2 hora) Repasen la materia vista en clase.
2. (2 hora) Investiguen para complementar lo visto en clase.
3. (1 hora) Discutan resultados y los ajusten con respecto al tema y tipo de arañador a construir.
4. (4 horas) Diseñen y documenten sus decisiones sobre la implementación
5. (7 horas) Implementen en `Python 3` un arañador.

Fecha de entrega

- 25 de abril, 2022. 10:00 pm (GMT -6).

Aspectos Generales

1. Los trabajos de entrega tardía no se calificarán
2. Se aclararan dudas sobre la progra vía Telegram (en horario diurno, en días hábiles)
3. Las clases sincrónicas tienen disponibilidad para destinar tiempo a comentarios o consultas sobre la progra.
4. Es posible y esperable comentar, discutir, compartir y construir en conjunto a otros grupos el proyecto. Sin embargo, es necesario que las implementaciones sean realizadas de forma independiente para que el proyecto cumpla con los objetivos específicos de este proyecto.