Práctica 1: Web scraping

Análisis de Redes Sociales 2024/25 (1C)

Fecha	Descripción de la versión
2024-09-20	Enunciado de la práctica 1

Objetivo

El objetivo de esta práctica es desarrollar un programa en Python que realice web scraping del subreddit r/spain para crear un conjunto de datos que contenga información sobre esta red social. Se recomienda usar el dominio "antiguo" (https://**old**.reddit.com/r/spain/).

Requisitos

- Utilizar python.
- Utilizar obligatoriamente las siguientes librerías:
  - `requests`
  - `beautiful soup`
- Puede ser útil también usar:
  - `ratelimit`
  - `click`
- El resultado debe ser almacenado en el sistema de ficheros en un formato estándar (por ejemplo, CSV, JSON, XML). Esto nos ayudará a guardar resultados parciales y no repetir peticiones http, lo que nos facilitará el trabajo y además será más respetuoso con la plataforma, pero tendremos que pensar en cómo almacenar la información de forma efectiva.
- El programa debe ser ejecutable desde la terminal, utilizando un entorno virtual `python`. Se recomienda el uso de la herramienta `uv`.
- La configuración del programa puede ser realizada a través de parámetros de línea de comando o mediante un archivo de configuración (por ejemplo usando `toml`.
- El proyecto debe ser un repositorio git bien estructurado, con su README (o README.md) y su historia de desarrollo.

Información a capturar

1. Los nodos de nuestra red van a ser: posts, comentarios y usuarios. Las aristas van a ser las relaciones entre estos nodos.

2. Para cada post, hay que guardar:

    - título
    - fecha
    - descripción, si la hay
    - autor (arista hacia usuario)

3. Para cada comentario:

    - texto
    - fecha
    - post al que responde (arista hacia post)
    - autor (arista hacia usuario)

4. Para cada usuario:

    - nombre
    - karma
    - posts creados (arista hacia post)
    - comentarios hechos (arista hacia comentario)

Lógicamente, las aristas aparecen dos veces en la descripción anterior, pero eso no quiere decir que haya que guardarlas por duplicado. No debemos incluir posts ni comentarios que no pertenezcan al reddit `r/spain`, ni ficheros binarios (imágenes, etc), ni el código de la interfaz de usuario (javascript, css...).

Entrega

1. El conjunto de datos obtenido se entregará como archivo `.zip` en el campus virtual.

2. El código se entregará a través de un repositorio git, que se usará para todas las prácticas del curso. La url del repositorio se subirá en el campus virtual. Se deberá crear una etiqueta `entrega_p1` en el repositorio, que será la versión del código que el profesor corregirá. El commit deberá estar subido al repositorio remoto a tiempo para la fecha límite de entrega.

3. El repositorio debe incluir todas las instrucciones necesarias para ejecutar el programa en el archivo README, así como la documentación adicional que se considere oportuna. El repositorio debe incluir también cualquier archivo de configuración que sea necesario.

4. La evaluación de la tarea se realizará en el campus virtual. Se ha incluido una rúbrica donde se puede ver los criterios de corrección que se usarán.

5. La fecha de entrega límite es: **2024-10-06 23:55**.