TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

PRA 1 – Web Scraping

Alumno: Luis Manuel Molina Coca (luimoco)

Web scraping: Consideraciones del user-agent según el libro de texto

Tema 1

- *Analizar el archivo robots.txt*: Para observar la política de restricciones que tenga la página para los user-agents
- *Examinar el mapa del sitio web*
  (sitemap: <urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">) y código fuente.
- Tamaño de la página web
- *Tecnología*
- Propietario

Tema4

- *Formato del dataset* ¿CSV, Json? Dependerá de la información a introducir en él. Un dataset de features sería propicio el csv y un dataset de texto o información binaria mejor Json
- *Licencia* para compartir el dataset en el repositorio

Tema 5: Prevención del web scraping

- Bloqueo de ips
- Robots.txt
- Control de exceso de tráfico
  - *Rutinas para "frenar" al user-agent* durante su rastreo según los tiempos de carga o factor para simular comportamiento humano.
- Utilización de Captchas
- CSS Sprites
- Variaciones de datos con HTMLCSS

Tema 6: Resolución de obstáculos

- Modificar la *cabecera de la petición* para simular ser un navegador (si no permiten user-agents en la web)
- Gestión de logins y cookies de sesión.
- Respetar el robots.txt
- *Espaciado de peticiones* http
- Bloqueo de ips: HAbría que usar un proxy, espero que no haga falta
- *Configurar timeouts y excepciones*: El user-agent debe tener control de excepciones de las peticiones web realizadas.
- *Trampas de araña*: Para webs con generación dinámica

Tema 8: Prácticas y consejos

- *No parsear el html manualmente*: Utilización de la librería Beautifulsoup.
- *No saturar de peticiones* el servidor web: Aplicar rutina para frenar el user-agent
- Modificar el user-agent: Si no se permite el web-scraping para simular ser navegador
- Asumir que el web-scraper dejará de funcionar: Incorporar las *rutinas de excepciones* para errores 404 500, etc.
- Calidad y robustez de los datos
- Recordar *aspectos legales*.