SISTEMAS INTELIGENTES PARA CIENCIA DE DATOS

Maestría en Ciencia de Datos

Universidad de la Ciudad de Aguascalientes

Anuncios

- Retroalimentación tarea 1
- Dudas de la sesión pasada
- Retroalimentación del curos
- Concluir API
- Próxima semana serán 2 sesiones

Web scraping

- Extraer información automáticamente
  - Texto
  - Imágenes
  - Archivos
- Hace peticiones http
  - Analiza la respuesta
  - Almacena datos relevantes

Beautiful Soup

- Extrae información de archivos HTML XML
- Ve el documento como un árbol

- Permite
  - Navegar
  - Analizar
  - Buscar
  - Modificar

```
pip install beautifulsoup4
```

Ejercicio

S03_webscraping.ipynb

Mejores Prácticas

- Las consultas son caras. Evitar retrabajo
- Las paginas cambian constantemente. Nuestro código también debería
- No sobrecargar un solo servidor
- Revisar Robots.txt
- Revisar términos y condiciones

Otras herramientas

- Scrapy https://scrapy.org
- Selenium https://selenium-python.readthedocs.io