Angel Pablo Hinojosa

Scraping y Obtención de Datos para Big (y no tan Big) Data (Parte I)

Big Data

Lo que significa…

Obtención, gestión y manipulación de grandes volúmenes de datos.

Little Big Data

(Lo mismo, pero menos)

¿De dónde saco la información?

Datos primarios:

Datos de nuestras propias fuentes: usuarios, tráfico, logs, mails, redes, aplicaciones biométricas...

¿De dónde saco la información?

Datos de terceros:

Fuentes abiertas, proveedores de datasets…

(Por ejemplo `http://opendata.ugr.es/`)

...y scraping

Formatos

No todos los formatos son iguales:

- Formatos manipulables y no manipulables
- Formatos abiertos y cerrados

El infierno de los formatos

...y el PDF como ejemplo de todo lo malo

`http://sl.ugr.es/bigdataPDF`

El infierno de los formatos

Ejemplos:

- PRIMER_TRIMESTRE_2015.pdf
-
- presupuesto_ayuntamiento_2015.pdf
-
- presupuesto_parque_ciencias_2015.pdf
-
- tabla.pdf

El infierno de los formatos

...y el PDF como ejemplo de todo lo malo

`http://sl.ugr.es/bigdataPDF`

El infierno de los formatos

Son tus amigos:

- JSON
- CSV

Herramientas online:

Multiconversor:

http://www.cometdocs.com/

(con Limitaciones si no pagas)

Herramientas online:

Extraer tablas de PDF:

`https://pdftables.com/`

(Sólo lee las tablas)

Herramientas online:

PDF a Excell (y otros):

`https://www.pdftoexcelonline.com/`

Lo envía a tu correo

Herramientas online:

OCR de PDF

`http://www.onlineocr.net/`

(una sola pagina)

Herramientas online:

OCR de PDF

`http://free-online-ocr.com/`

(hasta 10 páginas)

NOTA: Ningún OCR es bueno

Lento, complejo, propenso a errores

Requiere supervisión y revisión posterior

Huye del OCR como de la peste

En local

Tabula

`http://tabula.technology/`

(fácil y simple, pero son OCR)

Conversor de formatos

Pandoc

`pandoc.org/`

(no convierte DE PDF)

Para programadores (Python)

pdfminer

`https://euske.github.io/pdfminer/`

pypdfocr

`https://pypi.python.org/pypi/pypdfocr`

Gracias

(Ruegos y preguntas)