Prueba de evaluación 2

Entornos de data science con Python
Curso 2015/16.

Parte 1

En esta primera parte trabajaremos con el API REST de 4chan:

https://github.com/4chan/4chan-API

(es conveniente tratar de comprender los campos que devuelve el JSON antes de comenzar, siguiendo el ejemplo de clase)

Se pide lo siguiente:

- Obtener una muestra de dos boards de 4chan, sacando una instantánea concreta (todas las páginas de cada board). Solo es necesario obtener la información requerida para los puntos siguientes, no es necesario procesar todo.
- Contrastar diferencias en la media de respuestas por post en los dos boards elegidos.
- Obtener la distribución del tamaño de las imágenes en los posts de uno de los boards.

Parte 2

En esta segunda parte trabajaremos con información en XML. Concretamente, el gobierno británico nos ofrece datasets sobre eventos de tráfico:

https://data.gov.uk/dataset/live-traffic-information-from-the-highways-agency-road-network

Trabajaremos con los "Unplanned Events" que se proporcionan en un formato XML. Se puede descargar de esta URL:

http://hatrafficinfo.dft.gov.uk/feeds/datex/England/UnplannedEvent/content.xml

Concretamente, contiene un conjunto de situaciones ("`situation`") y dentro de la descripción de las situaciones, nos interesa el impacto ("`impact`").

Investiga el impacto de los eventos no planificados en las variables que aporten algo de información. Por ejemplo, la variable "`capacityRemaining`" indica la capacidad de la vía que la situación dejó sin afectar, y en ocasiones indica obstrucción total (valor 0.0) pero en otros casos sólo parcial. Estudia descriptivamente al menos 3 de las variables, sean numéricas o categóricas, dentro de las que describen el impacto solamente.

Parte 3

El sitio Web NOOA del gobierno de EEUU proporciona datasets de datos climáticos a través de esta página Web:

http://www.ncdc.noaa.gov/cdo-web/datasets

Entre ellos tenemos los datasets "**Quality Controlled Local Climatological Data (QCLCD)**" que se describen aquí:
http://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/quality-controlled-local-climatological-data-qclcd

Entre los datos que se encuentran en los datasets QCLCD están las precipitaciones por años y estaciones. Por ejemplo, podemos descargar los datasets de aquí:

http://www.ncdc.noaa.gov/orders/qclcd/

Y encontraremos ficheros con datos de precipitaciones como este:

```
Wban Number, YearMonthDay, Time, Hourly Precip
03013,19960701,0053,0
03013,19960701,0153,0
03013,19960701,0253,0
03013,19960701,0353,0
03013,19960701,0453,0
…
```

Se pide tomar datos de varios años (queda a la elección del estudiante) de este conjunto de datasets para las precipitaciones y obtener los siguientes resúmenes:

- Día en que ha habido más precipitaciones.
- Año en que ha habido más precipitaciones (obteniendo la media de cada año)

Se pide realizar el análisis en dos versiones:

- Una utilizando DataFrames y los ficheros de texto que se decargan directamente.
- Una segunda con un paso previo en el que se guardan los datos en un fichero HDF5 (que debe contener los metadatos descriptivos necesarios). Queda a la decisión del estudiante cómo organizar los datos en el fichero.

Y se pide comparar:

- El tamaño en disco que ocupan los datos en cada una de las versiones.
- El tiempo comparado de ejecución de los resúmenes anteriores.

Opcional: finalmente, se platea el almacenar en el fichero HDF5 los resúmenes mismos obtenidos y comparar el tiempo de recuperación de esos datos del fichero con el tiempo tardado en calcularlo.