Obteniendo y tratando datos con NumPy y Pandas: Prueba de evaluación

ASIGNATURA: Entornos de Data Science en Python
Última actualización: 30 de enero de 2015

Objetivos

- Adquirir práctica en el manejo de la herramienta IPython Notebook.
- Saber importar datos como DataFrames de pandas y manipularlos.

Criterio de evaluación

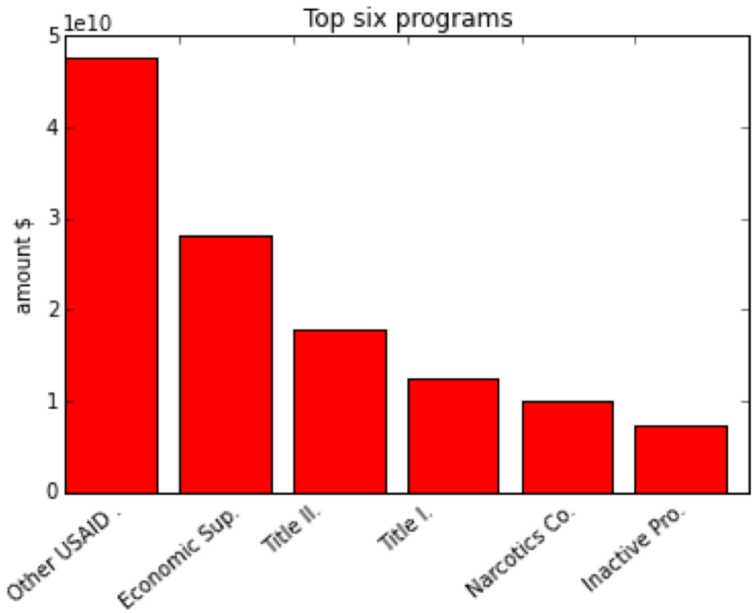
Esta es una práctica básica de obtención de datos externos y manejo de los mismos con NumPy y Pandas. Se evaluará simplemente el ser capaces de realizar las tareas solicitadas dentro de un IPython Notebook.

Esta prueba tiene un peso de un 40% de la evaluación del módulo

1. Manipulación de XML (30%)

Utiliza el dataset en XML “U.S. Overseas Loans and Grants (Greenbook)” en un Notebook para lo siguiente:

- Cargar los datos en un pandas DataFrame
- Encontrar los diferentes programas (“ProgramName”) en el data set. Necesitarás una function para obtener los **valores únicos.**
- Obten los seis programas que han gastado más dinero (recuerda que las cifras monetarias no son homogéneas).
- Dibujas las cantidades con una apariencia como la siguiente:



Recuerda que el dataset se puede descargar aquí:
http://catalog.data.gov/dataset/us-overseas-loans-and-grants-greenbook-usaid-1554

2. Uso de APIs Web (35%)

Volvemos al AngelList API. Estudia la documentación de esta llamada:
https://api.angel.co/1/startups?filter=raising

(la doc está aquí:
https://angel.co/api/spec/startups#GET_startups%3Ffilter%3Draising)

Encontrarás que proporciona información de las startups que están en fase de "public fundraising". Los resultados están paginados.

Después de estudiar la información y la estructura del JSON que devuelve, haz lo siguiente:

- Obtén la primera página.
- Carga los datos del JSON a un DataFrame con por lo menos los siguientes campos: URL de la compañía en AngelList, el campo "quality" y la "location" (solo la primera si tiene varias), fecha de aperture de la ronda, cantidad que se quiere obtener (raising amount), pre-money valuation y cantidad obtenida (raised amount).
- Después obtén la siguiente información:
    - Porcentaje de la cantidad obtenida respect a la esperada (para cada startup), ordenada de mayor porcentaje a menor.
    - Investiga si el número "quality" correlaciona con la cantidad solicitada o el pre-money valuation.

- Finalmente, dibuja las cinco startups que tienen un ratio mayor de raising amount/pre-money valuation, dibujando ambos valores en un stacked bar plot.

Recuerda que hay que estar seguros de que los datos están limpios.

3. Uso de HDF5 (35%)

Copiar los contenidos del dataset churn.txt utlizado en clase, de la siguiente forma:

- En la raíz, guardar el fichero tal cual, con todos los datos.
- Crear dos subcarpetas en la raíz, C1 y C2, y guardar una tabla en cada subcarpeta, que corresponda a los clientes que abandonaron y a los que no, respectivamente (hay que dividir los datos por el valor de la columna “Churn?”).

Comprobar que se pueden leer los datos y contrastar el tiempo de cómputo para obtener la media de uno de los atributos numéricos del dataset si utilizamos solo el fichero de texto original o si utilizamos el fichero HDF5, leyendo el dataset entero.

Después, toma tres de las columnas numéricas del dataset y utilizando regresión lineal, describe si esas variables que has elegido son buenas para tener un modelo del abandono.