Almacenes y Minería de Datos

PostgreSQL+Python, Agrupamiento

Joaquín Ángel Triñanes Fernández
Instituto de Investigacións Tecnolóxicas
Joaquin.Trinanes@usc.es Ext: 16001 Externo: 881816001

Python/Postgresql

Python

There are several Python drivers for PostgreSQL. This is the incomplete feature matrix for them; please help complete it as you see fit.

In general, Python users want to use psycopg2 unless they have a strong reason to try another driver, most of which are no longer maintained. Since DBAPI allows drivers to have different semantics, porting applications from one driver to another is non-trivial.

Software	License	Platforms	Python versions	DB API 2.0	Native (uses libpq)	Last Release	Notes
Psycopg2	LGPL	Unix, Win32	2.6-3.6	yes	yes	2019	Most popular python driver, required for most Python+Postgres frameworks
pg8000	BSD	any (pure Python)	3.3+	yes	no	2019	Used by Web2Py. current updated official site
py-postgresql	BSD	any (pure Python)	3.0+	yes	no	2018	Pure Python with optional C accelerator modules,<br\>extensive custom API. Python 3 only.
PyGreSQL	BSD	Unix, Win32	2.6 thru 3.6	yes	yes	2017	The first PostgreSQL adapter for Python. Still actively maintained.
ocpgdb	BSD	Unix	2.3-2.6	yes	yes	2010	PG8.1+
bpgsql	LGPL	any (pure Python)	2.3-2.6	yes	no	2009	labeled alpha
aiopg	BSD	any	3.52+	(ish)	Native	2019	

Psycopg2

Conforme al estándar DB-API 2.0 (promueve el uso de técnicas comunes por parte de módulos que acceden a bases de datos desde Python).

Pequeño, rápido y estable. Driver más popular.

Instalar: pip install psycopg2 (última versión 2.9). Psycopg3 ya está disponible!

Python/Postgresql

Diseñado para aplicaciones multihilo que hacen un uso intensivo de la BD

```
import psycopg2
conn_str = "host='localhost' dbname=‘midb’ port=5432 user=’usuario' password=’pwd’”
print “Conectando a la BD\n->%s" % (conn_str)
# Conectamos y si no se puede realizar, lanzamos una excepción
conn = psycopg2.connect(conn_str)
# conn.cursor devuelve un objeto cursor que usaremos para realizar las consultas
cursor = conn.cursor()
print "Connectados!\n"
```

Python/Postgresql

```
cursor.execute("SELECT * FROM mitabla")
# bajar los registros
registros= cursor.fetchall()
# Mostrarlos en pantalla
pprint.pprint(registros)
##Si la tabla fuera muy grande y no quisieramos bajarla toda de una vez
contador_fila= 0
for fila in cursor:
        contador_fila+= 1
        print “fila: %s    %s\n" % (contador_fila, fila)
##Cerrar el cursor y la conexión
cursor.close()
conn.close()
```

Python/Postgresql

###Cómo hacemos con los datos?
import pandas as pd

##Queremos insertar los datos de la consulta en un data frame de Pandas
##No necesitamos crear un cursor o hacer fetchall

df_pandas=pd.read_sql(consulta,con=conn)

Python Agrupamiento/Clasificación

Tips: (ejemplo DBSCAN)

```
import numpy as np

from sklearn.cluster import DBSCAN

# Importamos datos. Necesitamos escalar?

c=DBSCAN(eps=0.1,min_samples=10)

c.fit(DATOS)
```

- Todos los estimadores implementan: fit(), predict(), fit_predict()
  - Ejemplo: from sklearn import svm c=svm.SVC(gamma=0.01)

    c.fit(X_train,Y_train) y_pred=c..predict(X_test)

Ejercicio#7

El código de está práctica será implementado en Python.

Agrupamiento- Vamos a partir del siguiente conjunto de datos (podéis seleccionar un subconjunto en caso de problemas de rendimiento):

https://www.dropbox.com/s/k2vs1tvqoe6u70y/AMD_clustering_0.csv.gz?dl=0

Aplicar 1 metodología para la detección de valores atípicos en el conjunto de datos de trabajo. Describir los procesos y explicar los resultados.

a) Aplicar los siguientes algoritmos de agrupamiento sobre los datos:

a.1) K-means a.2) DBSCAN a.3) SOM

b) World Bank provee Los World Development Indicators (WDI). Importar en PostgreSQL los datos del fichero:

https://www.dropbox.com/s/71ybnugxdbelcn6/Databank_World_Development_Indicators.csv?dl=0

Metadatos:

https://www.dropbox.com/s/ynobjtvfuvg5s13/Databank_World_Development_Indicators_Metadata.csv?dl=0

Analiza los agrupamientos dentro de este conjunto de datos, accediendo a la BBDD en PostgreSQL.

Se valorará la claridad de la descripción y código, así como el análisis de resultados.

Esta práctica será realizada en equipos de 3 componentes y consistirán en un documento PDF con el código y la descripción de la práctica.