Scraping de datos de SEP por nombre ¶

Se escrapearon los daos de SEP con el nombre de cada funcionario, SEP da un score por cada nombre, se tomo el más probable y se almacena a la base de datos

In [3]:

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
import sys
reload(sys)
sys.setdefaultencoding('utf-8')
import json
import requests
import psycopg2
import jellyfish as jl


#conecta a una base de datos, regresa el cursor
def connect_database(dbname,u,p):
    #cleaning SEP files to integreate to Directory
    con=None
    try:
        con=psycopg2.connect(database=dbname, user=u, password=u)
        #cur = con.cursor()
        #cur.execute('SELECT nombre,primer_apellido,segundo_apellido from di
        return con
    except psycopg2.DatabaseError, e:
        print 'Error %s' % e
        sys.exit(1)
    return 0

#consulta la tabla de directorio y llama al scraping de SEP
def SEP_download(con):
    cursor=con.cursor()
    cursor.execute('SELECT id,nombre,primer_apellido,segundo_apellido from d
    rows=cursor.fetchall()
    c=0
    print len(rows)
    for row in rows:

        print row[0],row[1],row[2]
        escuela_json=scrap_name(str(row[1]),str(row[2]),str(row[3]))
        #r0=[row[1].upper().decode(encoding='UTF-8',errors='strict'),row[2].

        try:
            a= json.loads(escuela_json)
            c=c+1
            if 'docs' in  a and len(a["docs"])>0:
                for item in a["docs"]:
                    r=[]
                    r1=[item['nombre'].upper().decode(encoding='UTF-8',error
                    print len(item)
                    if 'numCedula' in item.keys() and 'titulo' in item.keys(
                        print item['numCedula'],item['titulo'],item['institu
                        query="INSERT INTO  SEP VALUES ('"+str(row[0])+"','"
                        print query#print valida_data(r0,r1)
                        #cursor.execute(query)
                        #unicode(row['PrimerApellido'],'utf8')
                        ##s=str(item[key])
                        #r.append(s.encode('utf8'))

                        #comparar los nombres, orden
                    #print r
```

```
            if c==1000:
                con.commit()
                c=0
        except ValueError:  # includes simplejson.decoder.JSONDecodeError
            print 'Decoding JSON has failed'
            pass
    con.commit()

#scraping SEP, devuelve un json
def valida_data(datos1,datos2):
    #convierte a UTF-8
    if jl.jaro_winkler(datos1[0],datos2[0])>0.60 and jl.jaro_winkler(datos1[
        return True
    else:
        return False
    #Comprar nombres y orden


def scrap_name(primerNombre,PrimerApellido, SegundoApellido):
    url = "http://search.sep.gob.mx/solr/cedulasCore/select?fl=%2A%2Cscore&q
    #print url
    try:
        r = requests.get(url)
        js = requests.get(url).json()

        if "response" not in js:
            jj="[]"
        else:
            jj=json.dumps(js["response"],encoding="utf-8")

        return jj
    except requests.exceptions.ConnectionError as e:
        print "no se pudo conectar he"
        pass


def start():
    database="dir"
    user="postgres"
    password="postgres"
    con=connect_database(database,user,password)
    SEP_download(con)

start()
```

```
dir=# select count(titulo) as C,titulo from SEP group by titulo order by C
desc;
   c   |                                                  titulo


-------+-----------------------------------------------------------------
----------------------
-----------------
 14868 | LICENCIATURA EN DERECHO
  4112 | LICENCIATURA COMO CONTADOR PÚBLICO
  3466 | PROFESOR EN EDUCACIÓN PRIMARIA
  3146 | LICENCIATURA EN ECONOMÍA
  2998 | LICENCIATURA EN ADMINISTRACIÓN
  2869 | LICENCIATURA EN CONTADURÍA
  2659 | LICENCIATURA COMO MÉDICO CIRUJANO
  2494 | LICENCIATURA EN INGENIERÍA CIVIL
  1872 | LICENCIATURA EN ADMINISTRACIÓN DE EMPRESAS
  1486 | LICENCIATURA EN PSICOLOGÍA
  1398 | LICENCIATURA EN ARQUITECTURA
  1195 | LICENCIATURA EN CONTADURÍA PÚBLICA
   992 | LICENCIATURA EN RELACIONES INTERNACIONALES
   963 | TÉCNICO EN ENFERMERÍA
   814 | LICENCIATURA EN INFORMÁTICA
   786 | LICENCIATURA EN INGENIERÍA QUÍMICA
   775 | LICENCIATURA COMO INGENIERO MECÁNICO ELECTRICISTA
   697 | LICENCIATURA EN BIOLOGÍA
   673 | LICENCIATURA EN INGENIERÍA INDUSTRIAL
   645 | LICENCIATURA COMO CIRUJANO DENTISTA
   627 | LICENCIATURA EN CIENCIAS POLÍTICAS Y ADMINISTRACIÓN PÚBLICA
   625 | LICENCIATURA EN CIENCIAS DE LA COMUNICACIÓN
   613 | LICENCIATURA COMO MÉDICO CIRUJANO Y PARTERO
   599 | LICENCIATURA COMO ABOGADO
   574 | LICENCIATURA EN PEDAGOGÍA
   537 | LICENCIATURA EN EDUCACIÓN PRIMARIA
   529 | LICENCIATURA EN INGENIERÍA EN SISTEMAS COMPUTACIONALES
   527 | LICENCIATURA COMO CONTADOR PÚBLICO Y AUDITOR
   501 | LICENCIATURA EN ACTUARÍA
   498 | LICENCIATURA EN INGENIERÍA MECÁNICA
   436 | PROFNAL. TEC. EN INFORMÁTICA
   432 | LICENCIATURA EN INGENIERÍA PETROLERA
   402 | LICENCIATURA COMO INGENIERO ELECTRICISTA
   396 | LICENCIATURA EN ENFERMERÍA
   353 | LICENCIATURA EN MERCADOTECNIA
   351 | LICENCIATURA EN INGENIERÍA EN COMUNICACIONES Y ELECTRÓNICA
   348 | LICENCIATURA COMO INGENIERO QUÍMICO INDUSTRIAL
   332 | PROFESOR DE EDUCACIÓN PREESCOLAR
   330 | LICENCIATURA EN MEDICINA VETERINARIA Y ZOOTECNIA
   318 | LICENCIATURA COMO INGENIERO ARQUITECTO
   313 | MAESTRÍA EN ADMINISTRACIÓN
   308 | LICENCIATURA EN SOCIOLOGÍA
   302 | LICENCIATURA EN RELACIONES COMERCIALES
   298 | LICENCIATURA EN ADMINISTRACIÓN INDUSTRIAL
   296 | LICENCIATURA COMO QUÍMICO FARMACÉUTICO BIÓLOGO
   287 | LICENCIATURA EN CIENCIAS DE LA INFORMÁTICA
   267 | LICENCIATURA COMO INGENIERO GEÓLOGO
   259 | LICENCIATURA EN EDUCACIÓN PREESCOLAR
   258 | LICENCIATURA COMO INGENIERO QUÍMICO
```

```
254 | LICENCIATURA EN TURISMO
253 | LICENCIATURA EN INFORMÁTICA ADMINISTRATIVA
250 | LICENCIATURA COMO INGENIERO AGRÓNOMO
249 | LICENCIATURA EN COMUNICACIÓN
246 | LICENCIATURA EN TRABAJO SOCIAL
246 | LICENCIATURA EN COMERCIO INTERNACIONAL
241 | LICENCIATURA EN DISEÑO GRÁFICO
240 | LICENCIATURA EN SISTEMAS DE COMPUTACIÓN ADMINISTRATIVA
232 | LICENCIATURA COMO INGENIERO EN COMPUTACIÓN
232 | LICENCIATURA EN INGENIERÍA EN ELECTRÓNICA Y COMUNICACIONES
224 | LICENCIATURA EN INGENIERÍA EN COMPUTACIÓN
214 | ESPECIALIDAD EN ANESTESIOLOGÍA
189 | LICENCIATURA COMO MÉDICO VETERINARIO ZOOTECNISTA
181 | LICENCIATURA COMO INGENIERO AGRÓNOMO FITOTECNISTA
172 | LICENCIATURA EN INGENIERÍA INDUSTRIAL Y DE SISTEMAS
170 | LICENCIATURA COMO MÉDICO CIRUJANO DENTISTA
168 | LICENCIATURA EN EDUCACIÓN FÍSICA
163 | LICENCIATURA COMO INGENIERO EN SISTEMAS COMPUTACIONALES
163 | LICENCIATURA COMO INGENIERO EN COMUNICACIONES Y ELECTRÓNICA
160 | TÉCNICO EN ENFERMERÍA GENERAL
158 | MAESTRÍA EN FINANZAS
157 | LICENCIATURA EN INGENIERÍA ELECTRÓNICA
152 | LICENCIATURA EN NEGOCIOS INTERNACIONALES
149 | LICENCIATURA EN NUTRICIÓN
146 | LICENCIATURA EN INGENIERÍA INDUSTRIAL EN ELÉCTRICA
142 | LICENCIATURA EN ADMINISTRACIÓN PÚBLICA
138 | ESPECIALIDAD EN PEDIATRÍA
132 | TÉCNICO EN TRABAJO SOCIAL
131 | ESPECIALIDAD EN CIRUGÍA GENERAL
129 | LICENCIATURA COMO BIÓLOGO
129 | LICENCIATURA COMO INGENIERO MECÁNICO
126 | LICENCIATURA EN HISTORIA
122 | LICENCIATURA COMO INGENIERO BIOQUÍMICO
119 | MAESTRÍA EN DERECHO
119 | LICENCIATURA EN DERECHO Y CIENCIAS SOCIALES
118 | LICENCIATURA COMO INGENIERO QUÍMICO PETROLERO
117 | LICENCIATURA COMO INGENIERO INDUSTRIAL MECÁNICO
116 | LICENCIATURA EN PERIODISMO
115 | LICENCIATURA EN CIENCIAS JURÍDICAS
114 | LICENCIATURA EN SISTEMAS COMPUTACIONALES
114 | LICENCIATURA COMO INGENIERO AGRÓNOMO ZOOTECNISTA
113 | MAESTRÍA EN EDUCACIÓN
113 | LICENCIATURA EN CIENCIA POLÍTICA
110 | MAESTRÍA EN ADMINISTRACIÓN PÚBLICA
105 | LICENCIATURA COMO INGENIERO  AGRÓNOMO ESP. EN FITOTECNIA
105 | LICENCIATURA COMO INGENIERO EN AERONÁUTICA
105 | LICENCIATURA EN ADMINISTRACIÓN MILITAR
104 | LICENCIATURA COMO QUÍMICO
101 | ESPECIALIDAD EN MEDICINA FAMILIAR
```