Team Scrapy

CUSP-AlumniHackathon2017

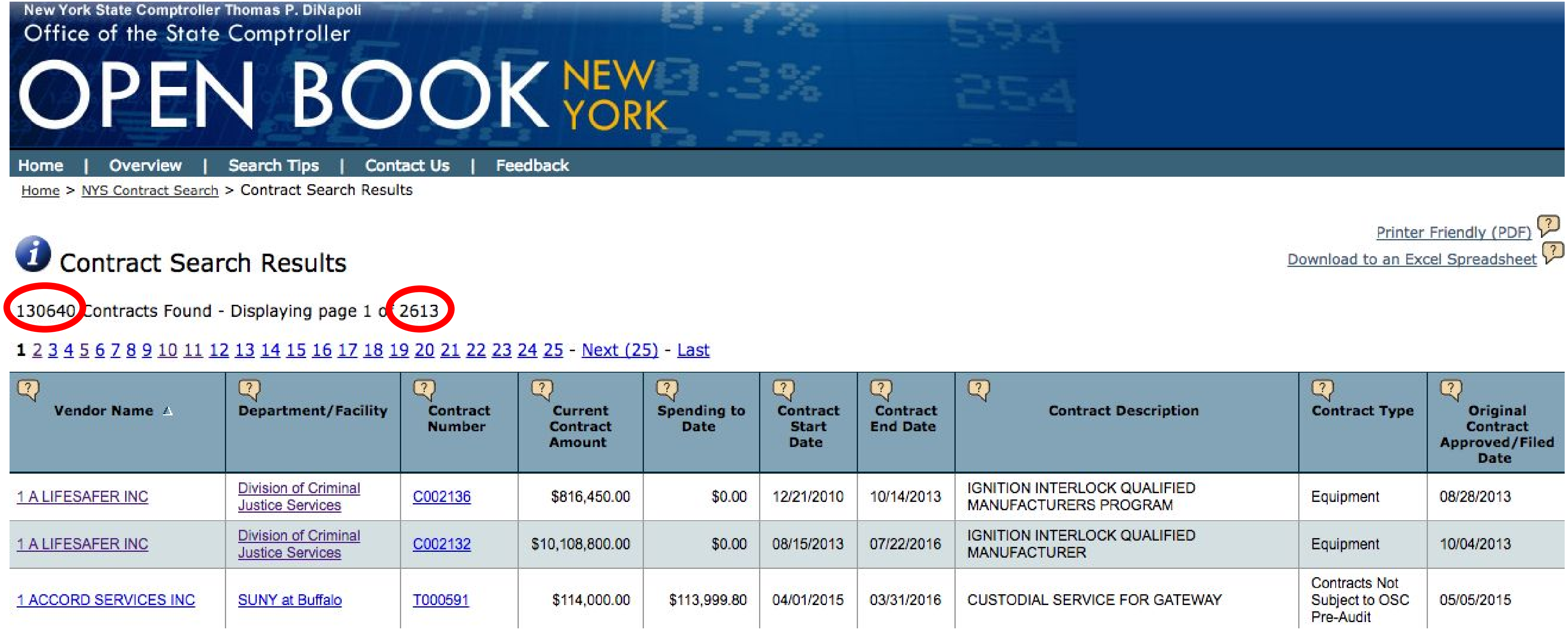
Connor Chen,
Dongjie Fan,
Christian Rosado,
Anastasia Shegay

Objective & Deliverables

- To automate the process of collecting and updating contract records from the Office of the State Comptroller website

- Deliverable is a web-scraping tool/script that can be run automatically

Data



New York State Comptroller Thomas P. DiNapoli
Office of the State Comptroller
OPEN BOOK NEW YORK

Home | Overview | Search Tips | Contact Us | Feedback

Home > NYS Contract Search > Contract Search Results

Printer Friendly (PDF)
Download to an Excel Spreadsheet

Contract Search Results

130640 Contracts Found - Displaying page 1 of 2613

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 - Next (25) - Last

Vendor Name △	Department/Facility	Contract Number	Current Contract Amount	Spending to Date	Contract Start Date	Contract End Date	Contract Description	Contract Type	Original Contract Approved/Filed Date
1 A LIFESAFER INC	Division of Criminal Justice Services	C002136	$816,450.00	$0.00	12/21/2010	10/14/2013	IGNITION INTERLOCK QUALIFIED MANUFACTURERS PROGRAM	Equipment	08/28/2013
1 A LIFESAFER INC	Division of Criminal Justice Services	C002132	$10,108,800.00	$0.00	08/15/2013	07/22/2016	IGNITION INTERLOCK QUALIFIED MANUFACTURER	Equipment	10/04/2013
1 ACCORD SERVICES INC	SUNY at Buffalo	T000591	$114,000.00	$113,999.80	04/01/2015	03/31/2016	CUSTODIAL SERVICE FOR GATEWAY	Contracts Not Subject to OSC Pre-Audit	05/05/2015

Solution I - Web Scraping

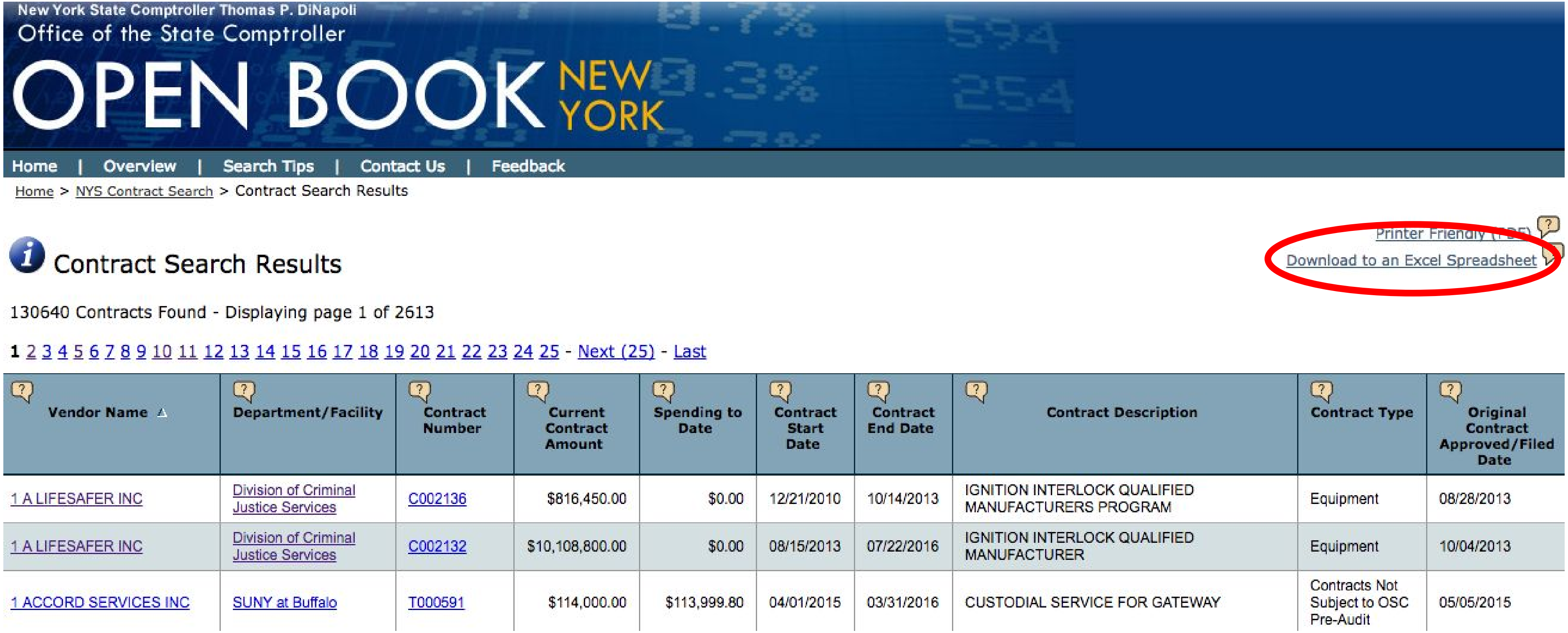
Advantages:

- Easily written and implementable code using Beautiful Soup Python package
- Allows downloading data in variety of formats: DataFrame, csv, text, etc.

Disadvantages:

- Not scalable: took a couple of minutes to scrape dozen of pages, not scalable to 2500+ pages
- Not easy way to update records

Solution II - Download entire dataset



New York State Comptroller Thomas P. DiNapoli

Office of the State Comptroller

OPEN BOOK NEW YORK

Home > NYS Contract Search > Contract Search Results

Printer Friendly (PDF)

Download to an Excel Spreadsheet

Contract Search Results

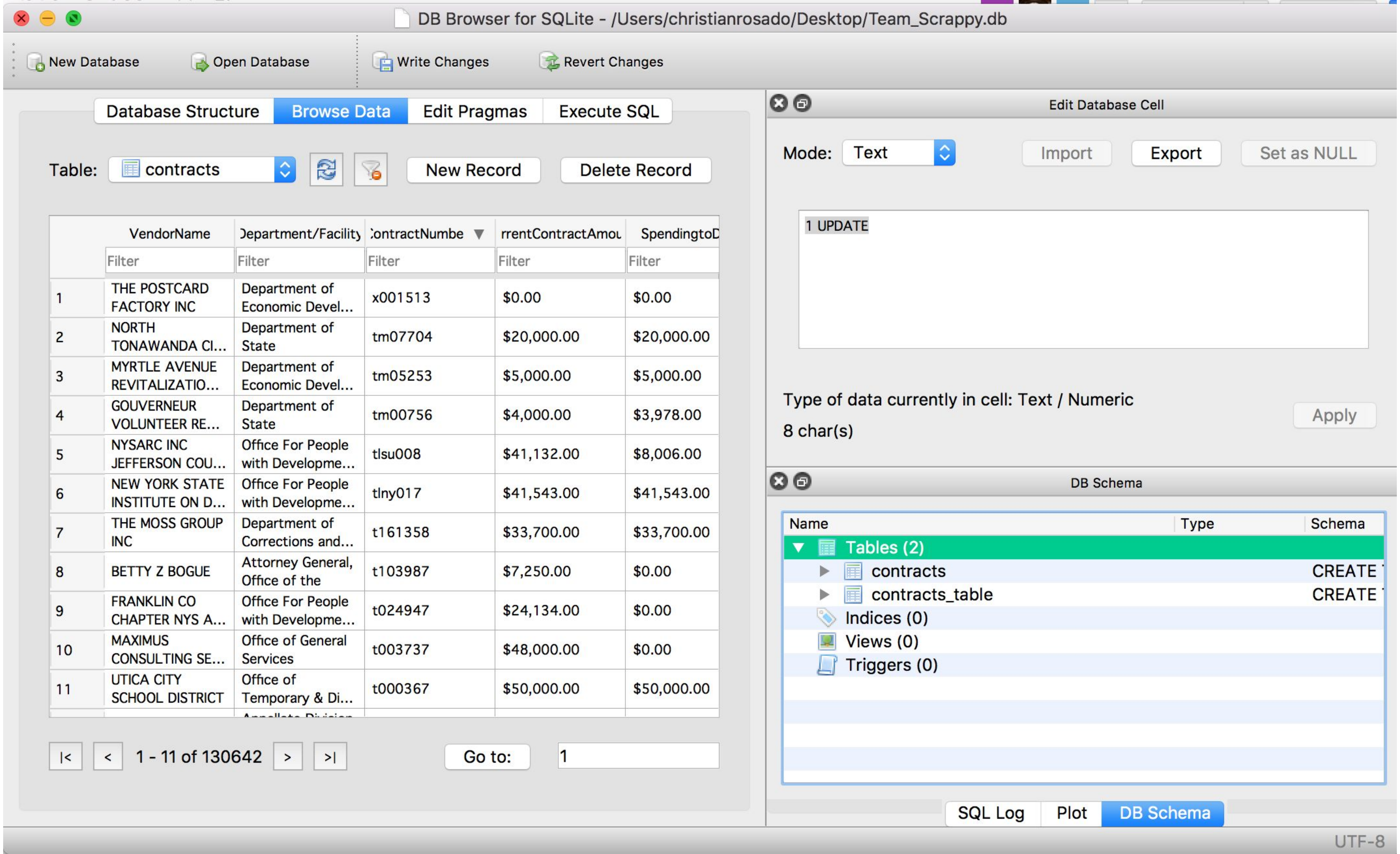
130640 Contracts Found - Displaying page 1 of 2613

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 - Next (25) - Last

Vendor Name ▵	Department/Facility	Contract Number	Current Contract Amount	Spending to Date	Contract Start Date	Contract End Date	Contract Description	Contract Type	Original Contract Approved/Filed Date
1 A LIFESAFER INC	Division of Criminal Justice Services	C002136	$816,450.00	$0.00	12/21/2010	10/14/2013	IGNITION INTERLOCK QUALIFIED MANUFACTURERS PROGRAM	Equipment	08/28/2013
1 A LIFESAFER INC	Division of Criminal Justice Services	C002132	$10,108,800.00	$0.00	08/15/2013	07/22/2016	IGNITION INTERLOCK QUALIFIED MANUFACTURER	Equipment	10/04/2013
1 ACCORD SERVICES INC	SUNY at Buffalo	T000591	$114,000.00	$113,999.80	04/01/2015	03/31/2016	CUSTODIAL SERVICE FOR GATEWAY	Contracts Not Subject to OSC Pre-Audit	05/05/2015

Solution III - DB Browser for SQLite

- DB Browser for SQLite is a high quality, visual, open source tool to create, design, and edit database files compatible with SQLite.
- It is for users and developers wanting to create databases, search, and edit data. It uses a familiar spreadsheet-like interface, and you don't need to learn complicated SQL commands.
    - Create and compact database files
    - Create, define, modify and delete tables
    - Update existing tables/records
    - Query via SQL



DB Browser for SQLite - /Users/christianrosado/Desktop/Team_Scrappy.db

Table: contracts

	VendorName	Department/Facility	ContractNumbe ▼	rrentContractAmou	SpendingtoD
	Filter	Filter	Filter	Filter	Filter
1	THE POSTCARD FACTORY INC	Department of Economic Devel...	x001513	$0.00	$0.00
2	NORTH TONAWANDA CI...	Department of State	tm07704	$20,000.00	$20,000.00
3	MYRTLE AVENUE REVITALIZATIO...	Department of Economic Devel...	tm05253	$5,000.00	$5,000.00
4	GOUVERNEUR VOLUNTEER RE...	Department of State	tm00756	$4,000.00	$3,978.00
5	NYSARC INC JEFFERSON COU...	Office For People with Developme...	tlsu008	$41,132.00	$8,006.00
6	NEW YORK STATE INSTITUTE ON D...	Office For People with Developme...	tlny017	$41,543.00	$41,543.00
7	THE MOSS GROUP INC	Department of Corrections and...	t161358	$33,700.00	$33,700.00
8	BETTY Z BOGUE	Attorney General, Office of the	t103987	$7,250.00	$0.00
9	FRANKLIN CO CHAPTER NYS A...	Office For People with Developme...	t024947	$24,134.00	$0.00
10	MAXIMUS CONSULTING SE...	Office of General Services	t003737	$48,000.00	$0.00
11	UTICA CITY SCHOOL DISTRICT	Office of Temporary & Di...	t000367	$50,000.00	$50,000.00
		Appellate Division			

1 - 11 of 130642

Go to: 1

Mode: Text

1 UPDATE

Type of data currently in cell: Text / Numeric

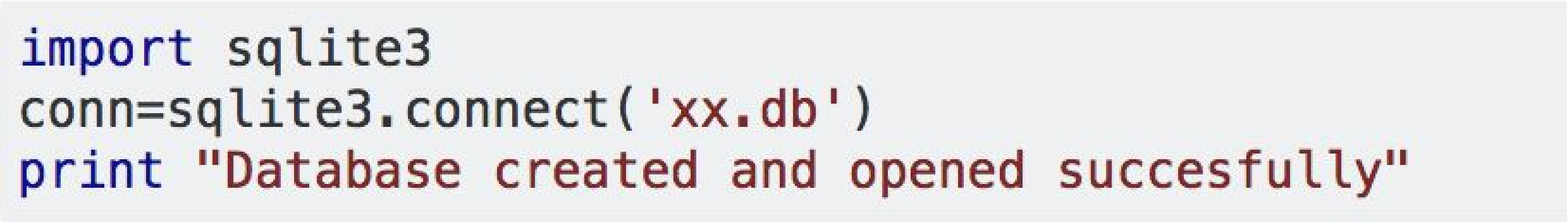
8 char(s)

DB Schema

Name	Type	Schema
▼ Tables (2)		
▶ contracts		CREATE
▶ contracts_table		CREATE
Indices (0)		
Views (0)		
Triggers (0)		

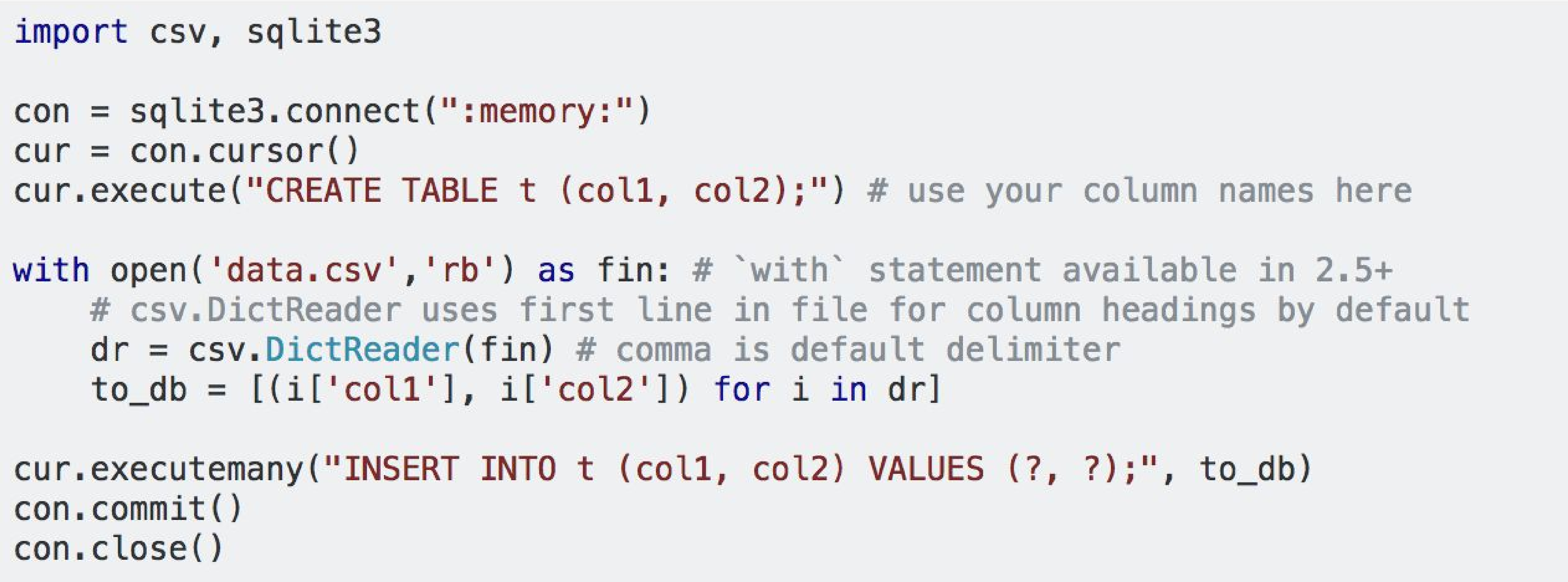
UTF-8

Connect to Database Using Python



```
import sqlite3
conn=sqlite3.connect('xx.db')
print "Database created and opened succesfully"
```

Import CSV files as Tables into Database



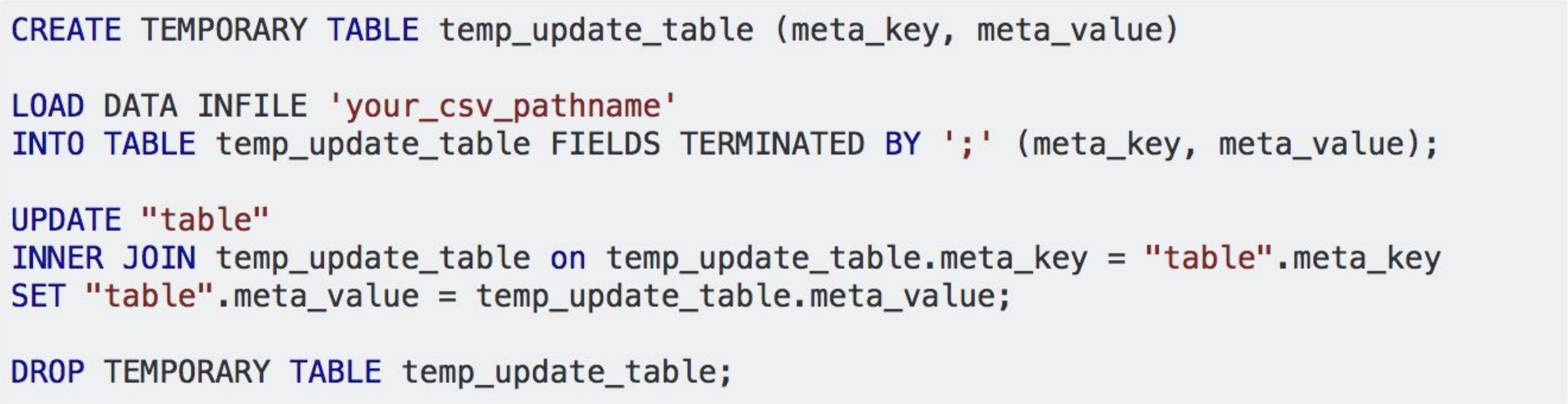
```
import csv, sqlite3

con = sqlite3.connect(":memory:")
cur = con.cursor()
cur.execute("CREATE TABLE t (col1, col2);") # use your column names here

with open('data.csv','rb') as fin: # `with` statement available in 2.5+
    # csv.DictReader uses first line in file for column headings by default
    dr = csv.DictReader(fin) # comma is default delimiter
    to_db = [(i['col1'], i['col2']) for i in dr]

cur.executemany("INSERT INTO t (col1, col2) VALUES (?, ?);", to_db)
con.commit()
con.close()
```

Update Existing Records/Append New



```
CREATE TEMPORARY TABLE temp_update_table (meta_key, meta_value)

LOAD DATA INFILE 'your_csv_pathname'
INTO TABLE temp_update_table FIELDS TERMINATED BY ';' (meta_key, meta_value);

UPDATE "table"
INNER JOIN temp_update_table on temp_update_table.meta_key = "table".meta_key
SET "table".meta_value = temp_update_table.meta_value;

DROP TEMPORARY TABLE temp_update_table;
```