```
# Name: Trien Bang Huynh
# Lab 3: Web scraping and data storage with requests, beautifulsoup, sqlite3,
review tkinter.
# lab3back.py

import requests
from bs4 import BeautifulSoup
import re
import json
import sqlite3


def webScrape(ROOT_URL):
    '''
    A function which takes in a root url, fetches and returns list of restaurant
info
    '''
    mainPage = requests.get(ROOT_URL)
    mainSoup = BeautifulSoup(mainPage.content, "lxml")
    # # get links for all next pages
    pagesURL = set()
    for link in mainSoup.select('.btn.btn-outline-secondary.btn-sm'):
        pagesURL.add("https://guide.michelin.com/" + link['href'])

    pagesURL = sorted(pagesURL)
    pageInfoDic = []
    for pageURL in pagesURL:
        page = requests.get(pageURL)
        soup = BeautifulSoup(page.content, "lxml")

        index = 0
        for link in soup.select('.card__menu-content--title.pl-text.pl-big a') :
            tempDic = {}

            # URL of a restaurant
            resURL = "https://guide.michelin.com/" + link['href']
            tempDic['URL'] = resURL

            # Name of a restaurant
            name = link.text.strip()
            tempDic['Name']  = name

            # City of a restaurant
            city = soup.select('.card__menu-footer--location.flex-fill.pl-text')
[index].text.strip().split(',')[0]
            tempDic['City']  = city

            # Cost + Cuisine of a restaurant
            detailText = soup.select('.card__menu-footer--price.pl-text')
[index].text.strip()
            detail = re.split(r'[,\s.]+', detailText)

            cost = detail[0]
            tempDic['Cost'] = cost

            cuisine = detail[2]
            tempDic['Cuisine'] = cuisine
```

```
            # # Go to single restaurant page to get address
            pageSub = requests.get(resURL)
            soupSub = BeautifulSoup(pageSub.content, "lxml")

            # Address of a restaurant
            address = soupSub.select('.restaurant-details__heading--address')
            tempDic['Address'] = address[0].text


            pageInfoDic.append(tempDic)
            index += 1
    return pageInfoDic
def writeToJsonFile(jsonFile, pageInfoDic):
    '''
    A function which write a list of restaurant info into the JSON file for the
current city
    '''

    with open(jsonFile, 'w') as fh:
        json.dump(pageInfoDic, fh, indent=3)


def createDatabase(jsonFile, dataFile):
    '''
    A function which read data from the JSON file to a db file and create tables
database
    '''
    # fetch data from the JSON file
    with open(jsonFile, 'r') as fh:
        d = json.load(fh)

    # Connect to database
    conn = sqlite3.connect(dataFile)
    cur = conn.cursor()

    ####  Create tables #####

    # Locations table

    cur.execute("DROP TABLE IF EXISTS Locations")
    cur.execute('''CREATE TABLE Locations(
                   id INTEGER NOT NULL PRIMARY KEY,
                   city TEXT UNIQUE ON CONFLICT IGNORE)''')
    # Costs table
    cur.execute("DROP TABLE IF EXISTS Costs")
    cur.execute('''CREATE TABLE Costs(
                   id INTEGER NOT NULL PRIMARY KEY ,
                   cost TEXT UNIQUE ON CONFLICT IGNORE)''')
    # Cuisine table
    cur.execute("DROP TABLE IF EXISTS Cuisine")
    cur.execute('''CREATE TABLE Cuisine(
                   id INTEGER NOT NULL PRIMARY KEY ,
                   cuisine TEXT UNIQUE ON CONFLICT IGNORE)''')

    # Main table
    cur.execute("DROP TABLE IF EXISTS Main")
    cur.execute('''CREATE TABLE Main(
                   id INTEGER NOT NULL PRIMARY KEY UNIQUE,
                   name TEXT,
```

```
                    url TEXT,
                    loc INTEGER,
                    cost INTEGER,
                    kind INTEGER,
                    addr TEXT)''')

    # Insert data to tables
    for restaurant in d:
        cur.execute('INSERT INTO Locations (city) VALUES (?)',
(restaurant['City'],))
        cur.execute('SELECT id FROM Locations WHERE city = ? ',
(restaurant['City'], ))
        city_id = cur.fetchone()[0]

        cur.execute('INSERT INTO Costs (cost) VALUES (?)', (restaurant['Cost'],))
        cur.execute('SELECT id FROM Costs WHERE cost = ? ', (restaurant['Cost'], ))
        cost_id = cur.fetchone()[0]

        cur.execute('INSERT INTO Cuisine (Cuisine) VALUES (?)',
(restaurant['Cuisine'],))
        cur.execute('SELECT id FROM Cuisine WHERE cuisine = ? ',
(restaurant['Cuisine'], ))
        cuisine_id = cur.fetchone()[0]

        cur.execute('''INSERT INTO Main (name, url, loc, cost, kind, addr) VALUES (
?, ?, ?, ?, ?, ? )''',
        (restaurant['Name'], restaurant['URL'], city_id, cost_id, cuisine_id,
restaurant['Address']) )


    conn.commit()
    conn.close()



def main():
    ROOT_URL = 'https://guide.michelin.com/us/en/california/san-jose/restaurants'
    pagesInfo = webScrape(ROOT_URL)
    writeToJsonFile('SJdata.json',pagesInfo)
    createDatabase('SJdata.json','data.db')


if __name__ == "__main__":
    main()
```