1 lawschool_csv_crawler.py

This Python script fetches data from lawschoolnumbers.com in .csv format.

```python
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import NoSuchElementException
from bs4 import BeautifulSoup
import csv
import time


def fetch_data(year):
    # Format the year to the specified format
    if (int(year) < 2009):
        formatted_year = f"0{year % 100}0{(year + 1) % 100}"
    elif (int(year) == 2009):
        formatted_year = "0910"
    else:
        formatted_year = f"{year % 100}{(year + 1) % 100}"
    url = f"https://michigan.lawschoolnumbers.com/applicants/{formatted_year}"
    driver = webdriver.Chrome()
    driver.get(url)
    wait = WebDriverWait(driver, 10)

    try:
        # Wait until the table is loaded
        wait.until(EC.presence_of_element_located((By.CLASS_NAME, ’table-application’)))
        return driver
    except:
        print(f"Failed to load the page for {formatted_year}")
        driver.quit()
        return None

def parse_page(driver):
    data = []
    soup = BeautifulSoup(driver.page_source, ’html.parser’)
    table = soup.find(’table’, class_=’table-application’)
    if table:
        rows = table.find_all(’tr’)
        for row in rows[1:]:  # skip header row
            cols = row.find_all(’td’)
            if len(cols) >= 8:
                username_link = cols[0].find(’a’)
                username = username_link.text if username_link else ’N/A’
                signifiers = ’’.join(signifier.text for signifier in cols[0].find_all(’span’, class_=’signifier’))
                urm = 1 if ’U’ in signifiers else 0
                inter = 1 if ’I’ in signifiers else 0
                full_username = f"{username} {signifiers}".strip()
                status = cols[1].text.strip()
                if ("Waitlisted" in status or "Pending" in status):
                    continue
                elif ("Accepted" in status):
                    status = "Accepted"
                elif ("Rejected" in status):
                    status = "Rejected"
                lsat = cols[2].text.strip().split(’:’)[-1].strip()
                gpa = cols[3].text.strip().split(’:’)[-1].strip()

                data.append([full_username, status, lsat, gpa, urm, inter])
    return data

def save_to_csv(data, formatted_year):
    with open(f"michigan_law_{formatted_year}.csv", ’w’, newline=’’, encoding=’utf-8’) as file:
        writer = csv.writer(file)
        writer.writerow(["Username with Signifiers", "Status", "LSAT", "GPA", "URM", "Intl"])
        writer.writerows(data)

def main():
    years = range(2003, 2024)
    for year in years:
        if (int(year) < 2009):
            formatted_year = f"0{int(year) % 100}0{(int(year) + 1) % 100}"
        elif (int(year) == 2009):
            formatted_year = "0910"
        else:
            formatted_year = f"{int(year) % 100}{(int(year) + 1) % 100}"
        print(f"Processing year: {formatted_year}")
        driver = fetch_data(year)
        if driver:
            all_data = []
            while True:
                # Parse the current page
                page_data = parse_page(driver)
                all_data.extend(page_data)
                # Check if there's a next button and click it
                try:
                    next_button = driver.find_element(By.CSS_SELECTOR, '.pagination-holder .pagination .next_page')
                    next_button.click()
                    time.sleep(3)  # Wait for the next page to load
                except NoSuchElementException:
                    break  # No next button found, exit loop
                except Exception as e:
                    print(f"Error occurred while clicking next button: {e}")
                    break  # Exit loop on any error
            # Save all collected data from all pages
            save_to_csv(all_data, formatted_year)
            print(f"Data for {formatted_year} saved successfully.")
            driver.quit()
        else:
            print(f"No data available for {formatted_year}")

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

2 combine_csv_files.py

This Python script combines the .csv files of each application cycle's data into one big .csv file.

```python
import os
import glob
import pandas as pd

def combine_csv_files(directory, output_file):
    # Change this to the directory where your CSVs are saved
    os.chdir(directory)
    all_files = glob.glob('michigan_law_*.csv')
    all_data = []

    for filename in all_files:
        df = pd.read_csv(filename, index_col=None, header=0)
        # Extract the year from the filename
        year = filename.split('_')[-1].split('.')[0]
        df['Year'] = year
        all_data.append(df)

    combined_csv = pd.concat(all_data, axis=0, ignore_index=True)
    combined_csv.to_csv(output_file, index=False)
    print(f"Combined CSV has been saved as {output_file}")

combine_csv_files('/Users/joonchoi/Desktop/STATS451/finalproj/data', 'combined_michigan_law.csv')
```