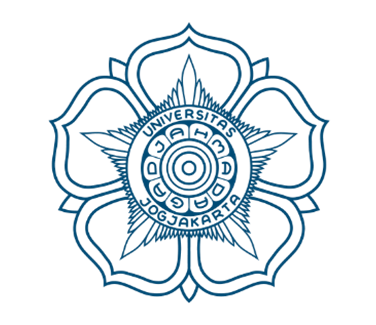
**SISTEM TEMU KEMBALI INFORMASI**

**TUGAS 3 METODE DAN HASIL CRAWLING - DATASET**

**DATA ANGGOTA ORGANISASI MAHASISWA (ORMAWA) ATAU UNIT KEGIATAN MAHASISWA (UKM)**

****

Disusun oleh :

Syuhra Putri Ganiswari 22/507327/NPA/19811

Ifta Mu’izana 22/507333/NPA/19813

**PROGRAM STUDI S1 ILMU KOMPUTER**

**DEPARTEMEN ILMU KOMPUTER DAN ELEKTRONIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS GADJAH MADA**

**YOGYAKARTA**

**2022**

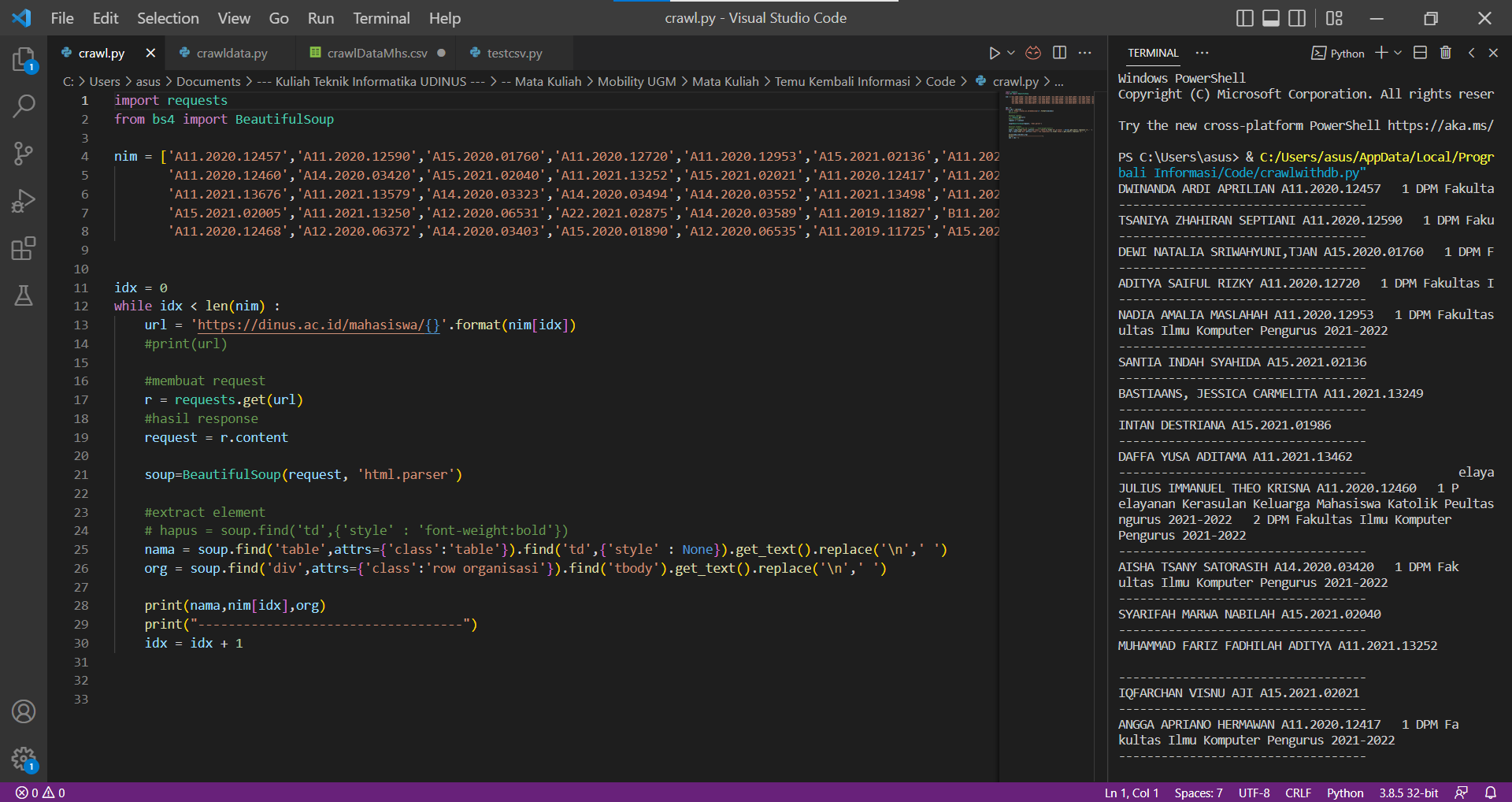
**Sistem Temu Kembali Informasi Data Anggota Organisasi Mahasiswa (Ormawa) atau Unit Kegiatan Mahasiswa (UKM)**

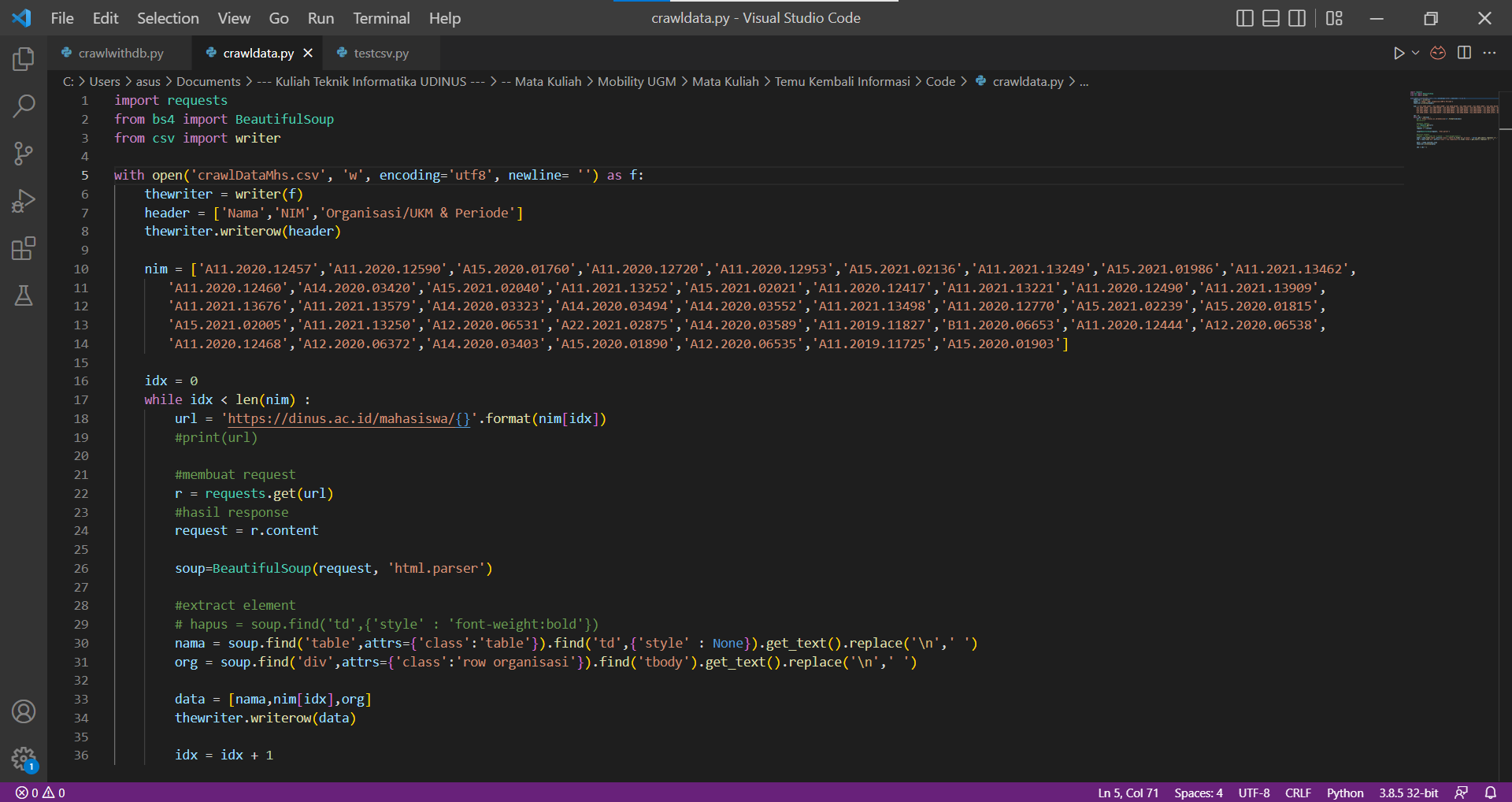
Anggota Kelompok :

1. Syuhra Putri Ganiswari      (22/507327/NPA/19811)
2. Ifta Mu’izana                      (22/507333/NPA/19813)

Tugas 3 – Metode dan Hasil Crawling – Dataset

1. Metode Crawling

*Gambar 2. Source code crawling web Siadin*



*Gambar 2. Source code crawling web Siadin to CSV*

Pada aplikasi Sistem Pencarian Data Anggota Ormawa atau UKM yang kami buat, kami melakukan *crawling* atau pengambilan data dari web Siadin milik Universitas Dian Nuswantoro untuk mengambil data-data mahasiswa seperti Nama, NIM, dan Organisasi atau UKM yang diikuti oleh mahasiswa beserta periodenya. Kami melakukan *crawling* menggunakan Python yang mana hasil *crawling* web Siadin tadi akan diubah menjadi file .csv supaya dapat dijadikan dataset di excel.

Gambar 1 merupakan code crawl.py dimana data hasil *crawling* masih ditampilkan pada terminal. Kemudian pada gambar 2, code crawldata.py menggunakan tiga library python yaitu request, beautifulsoup, dan csvwriter. Library request digunakan untuk pengambilan data dari web Siadin dengan mengirimkan request ke server situs web tersebut. Library beautifulsoup digunakan untuk mengambil data HTML dan XML, serta sebagai parser guna memisahkan komponen-komponen HTML menjadi suatu rangkaian elemen yang mudah dibaca. Kemudian yang terakhir yaitu library csvwriter yang berfungsi untuk mengubah data hasil crawling dan menuliskannya menjadi file berbentuk .csv .

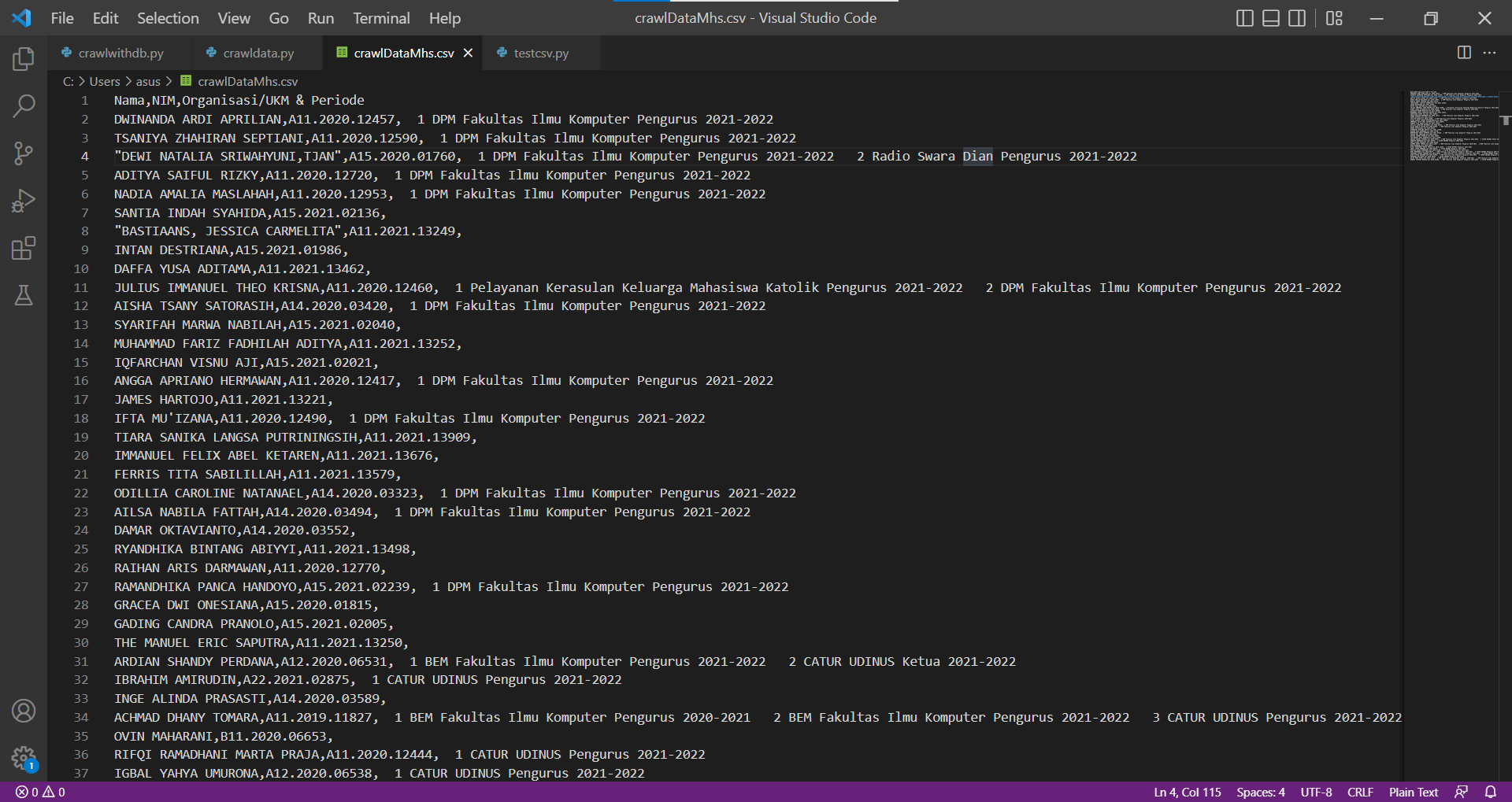
Untuk mendapatkan data setiap mahasiswa, harus dilakukan *crawling* pada banyak halaman web. Karena data tiap mahasiswa berada dalam satu halaman web dan hanya NIM yang membedakan link web tersebut, sehingga semua data NIM mahasiswa yang ingin di *crawling*, dimasukkan dalam array dengan nama variabel “nim”. Kemudian dilakukan *looping* dengan nilai awal variabel “idx” sama dengan 0, ketika nilai variabel “idx” kurang dari panjang array nim, maka akan dilakukan perulangan sebanyak jumlah NIM yang diinputkan, nilai “idx” akan ditambah 1 setiap melakukan perulangan.

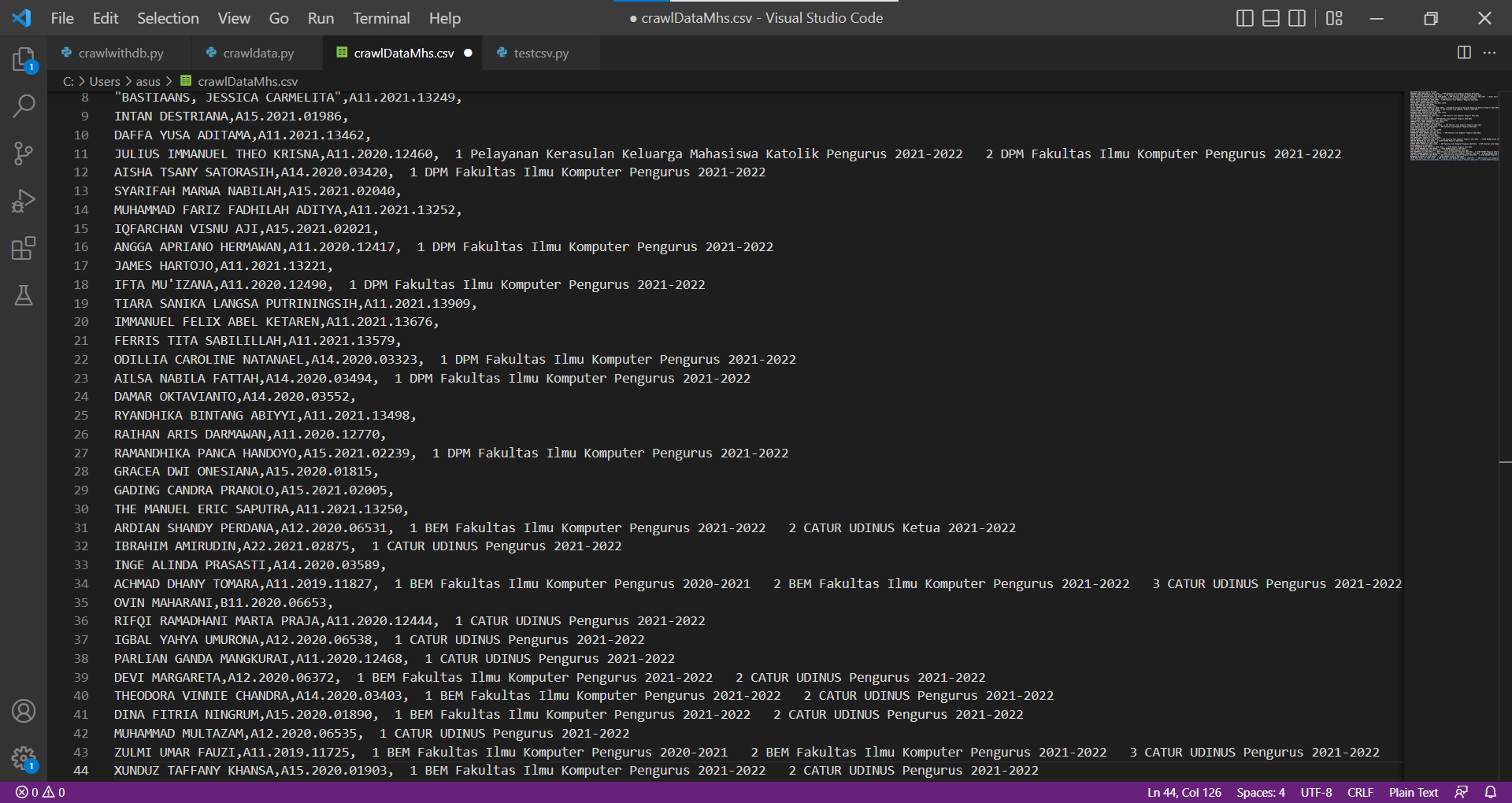
Link url akan selalu berubah-ubah sesuai nilai array di setiap indexnya, dilakukan request pengambilan data dari url, dan responnya akan ditampung di variabel “request”. Selanjutnya dilakukan parsing dengan *find method* untuk memisahkannya dan kemudian mendapatkan data nama mahasiswa yang ditampung dalam variabel “nama” serta data organisasi & periode yang ditampung di variabel “org”. Supaya output lebih rapi maka digunakan fungsi get\_text() untuk menghilangkan tag-tag HTML dan replace('\n',' ') supaya data berada dalam satu baris.

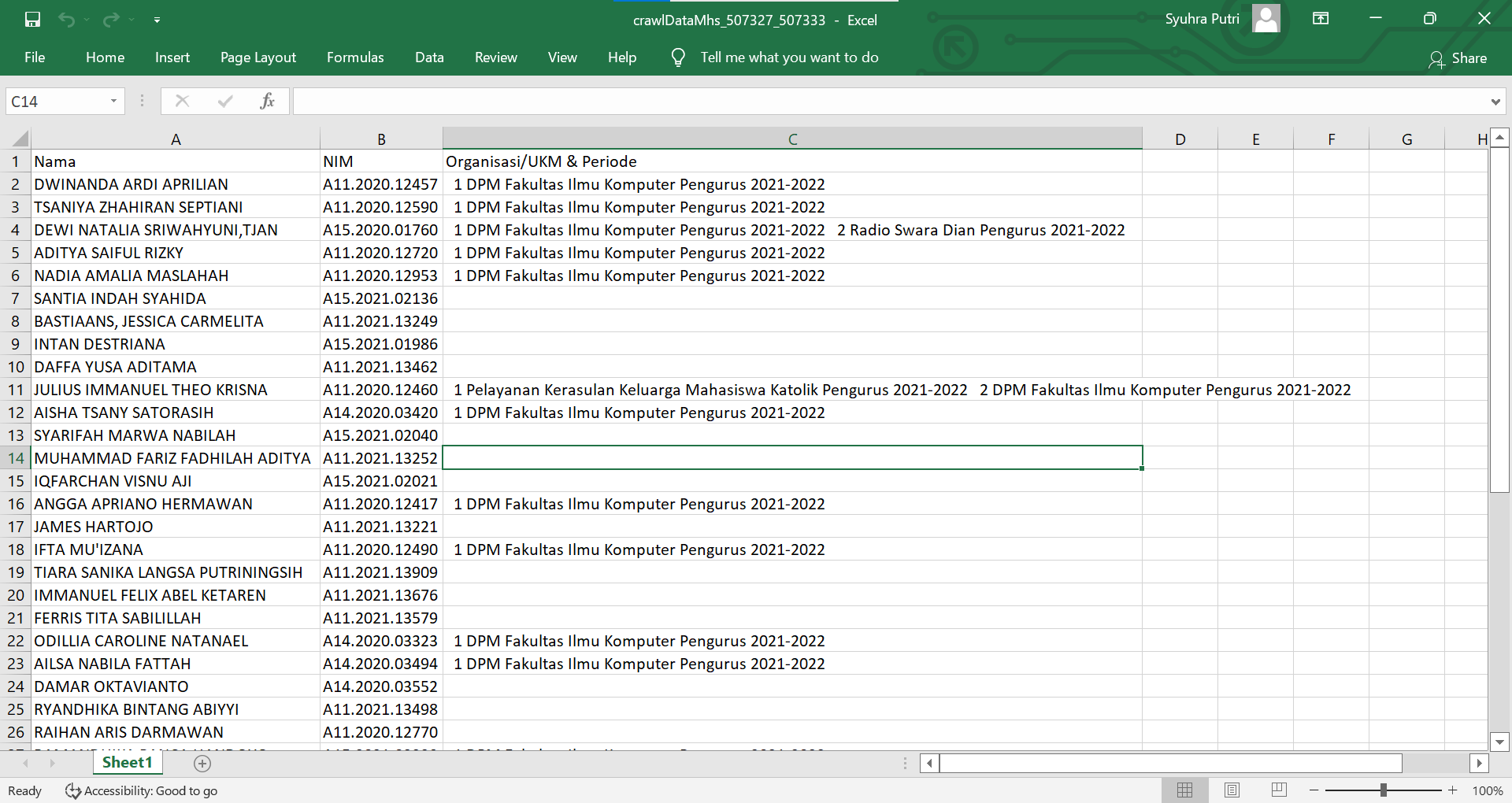
Langkah yang selanjutnya yaitu menampung variabel nama, nim[idx], dan org ke dalam variabel data. Supaya data-data hasil *crawling* tersebut dapat dituliskan menjadi file .csv, digunakan code thewriter.writerow(data). Hasil dari crawldata.py akan menjadi crawlDataMhs.csv setelah di run, pada baris pertama akan menampilkan header Nama, NIM, Organisasi/UKM & Periode, dan baris selanjutnya akan menampilkan data-data nama mahasiswa, NIM mahasiswa serta organisasi yang diikuti beserta periodenya.

1. Hasil Crawling

Crawling adalah proses dimana mesin pencarian dapat mencari dan memindai konten yang berada di situs web berisikan data, gambar, daftar, artikel dll. Berikut adalah hasil *crawling* yang diambil dari web SIADIN milik Universitas Dian Nuswantoro untuk mendapatkan data mahasiswa aktif yang mengikuti ORMAWA dan/atau UKM. Data yang diambil dari *website* SIADIN antara lain adalah : Nama, NIM, Ormawa / UKM dan periode. Mulanya, output dari proses *crawling* adalah dengan format file .csv kemudian dari file .csv tersebut di import ke file excel. Tujuannya adalah agar seluruh data yang sudah terkumpul melalui proses *crawling* dapat dilihat lebih rapi di file excel dengan tampilan berbentuk tabel sehingga menjadi dataset.

*Gambar 2 Output berupa file .csv (1)*

*Gambar 3 Output berupa file .csv (2)*

*Gambar 4 Dataset (Hasil import dari file .csv)*

1. Kesimpulan

Pada tugas kali ini dilakukan proses pembuatan program *web crawling* yang akan digunakan untuk mengambil data mahasiswa yang mengikuti ORMAWA dan/atau UKM secara otomatis dari *website* SIADIN milik UDINUS. Proses *web crawling* data dari tugas ini dapat diimplementasikan menggunakan bahasa pemrograman *Python*. Proses *web crawling* pada tugas ini sudah berhasil dilakukan dan data yang di dapat pun sudah sesuai. Dari tugas ini dapat disimpulkan bahwa *web crawling* merupakan salah satu metode yang bisa digunakan untuk mengumpulkan dan mengambil informasi dari suatu *website*. *Website* yang dapat dijadikan objek untuk melakukan proses *web crawling* juga beragam, salah satunya *website* SIADIN milik UDINUS yang merupakan *website* berisi data mahasiswa yang mengikuti ORMAWA dan/atau UKM yang dijadikan *website* acuan.

1. Referensi
2. <https://youtu.be/RvCBzhhydNk>
3. <https://youtu.be/Uu9Ipxipi9k>
4. <https://www.pythontutorial.net/python-basics/python-write-csv-file/>