**INDEKS FILE (TEKS) MENGGUNAKAN SWISH-E**

*By: Maulana Hizbullah (1708107010055) – Penulusuran Informasi – 2020*

*Informatika – Universitas Syiah Kuala*

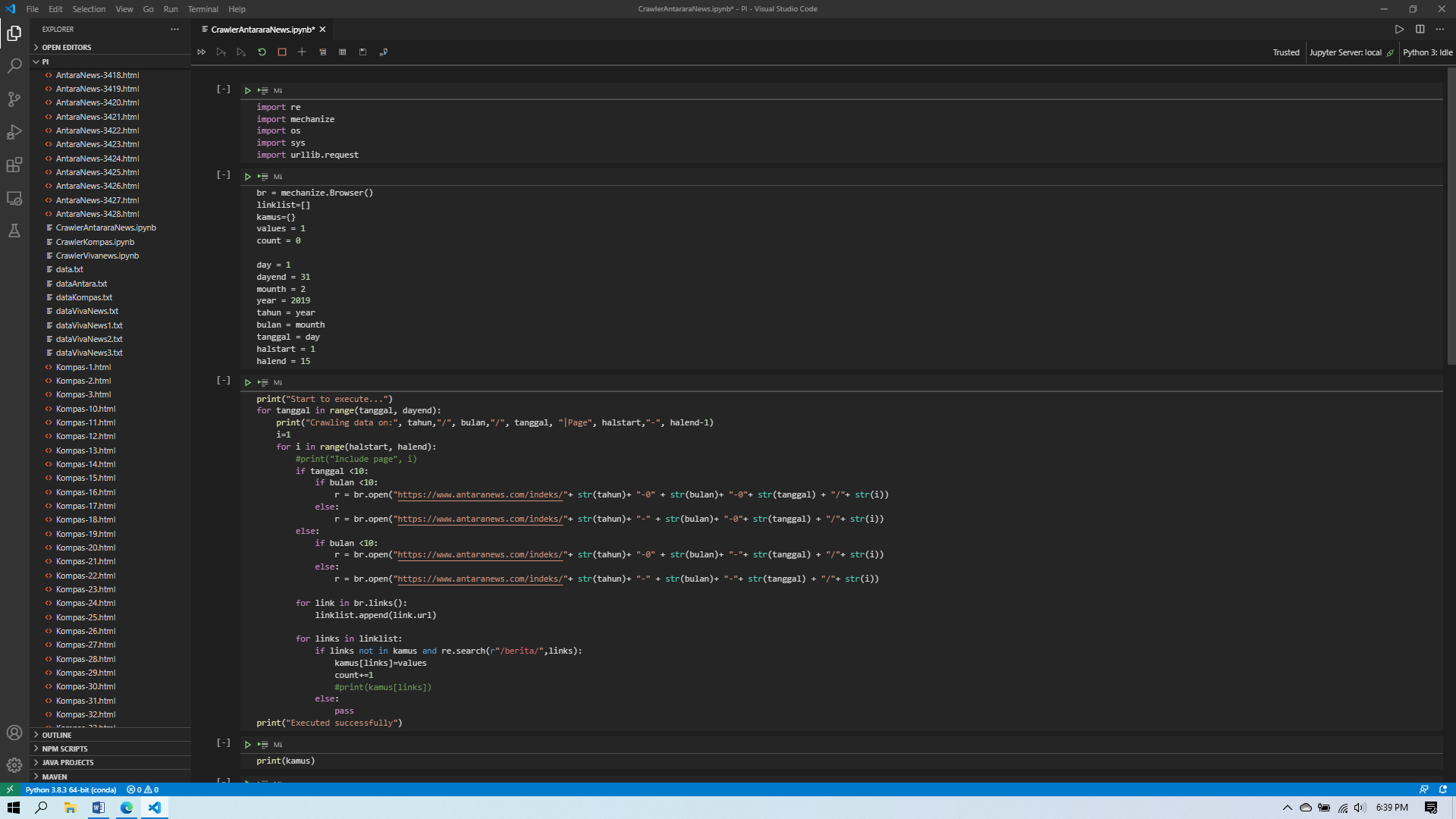
[**SWISH-e**](https://www.esa.org/tiee/search/html/index.html) adalah sebuah *tool* yang digunakan untuk meng-indeks teks dalam berbagai format, seperti PDF, html, txt, XML, PostScript, dll. Selain meng-indeks, *tool* ini juga dapat mencari dokumen berdasarkan *query* yang di-*input*-kan. *Tool* ini memiliki banyak fitur/kemampuan yang dimilikinya atau untuk lebih jelasnya dapat dilihat pada laman “https://www.esa.org/tiee/search/html/readme.html”

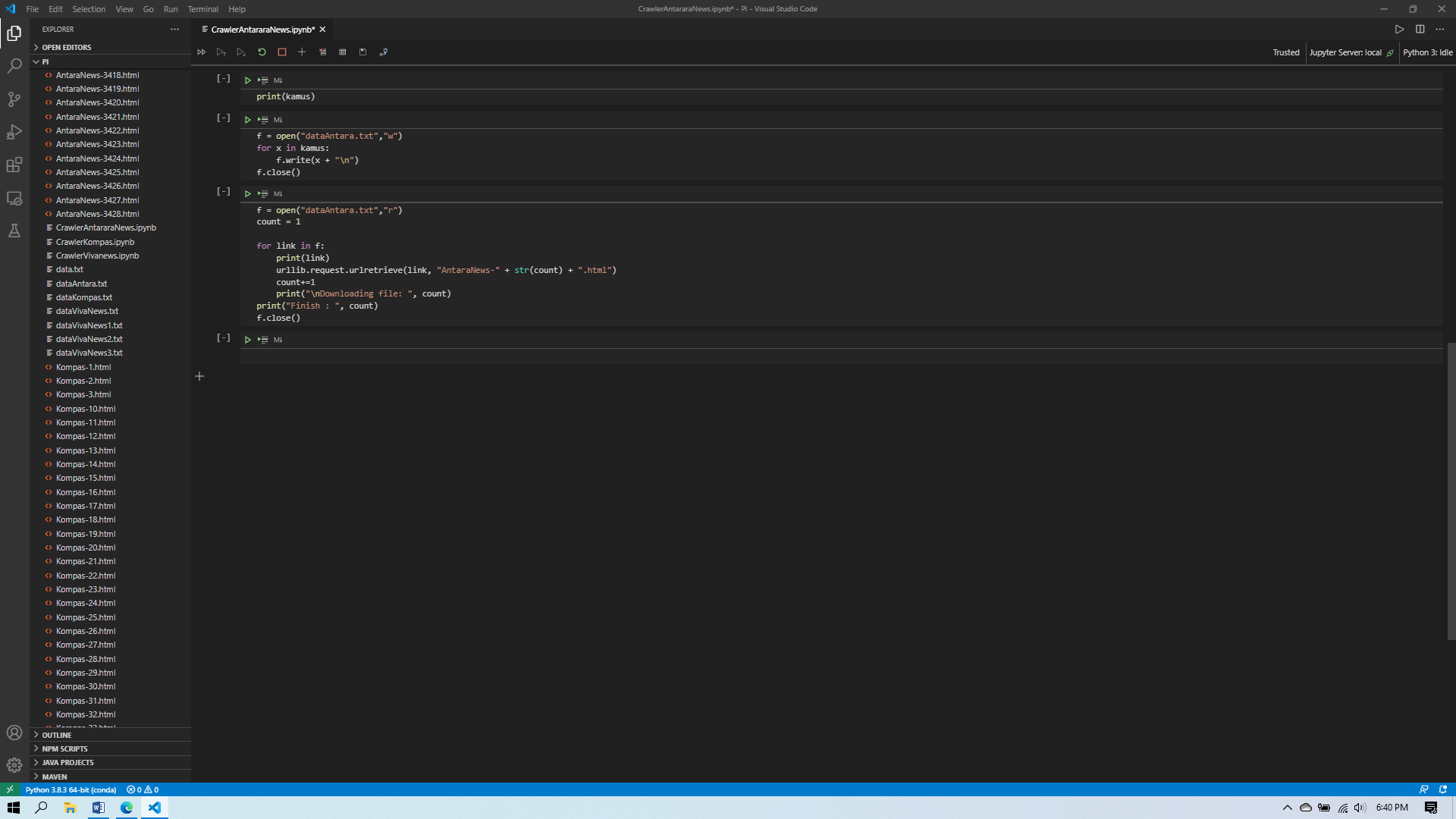
Penulis membuat skrip python untuk mengunduh (*crawling*) berita pada website:

* [AntaraNews](https://www.antaranews.com/)
* [KompasNews](https://www.kompas.com/)
* VivaNews

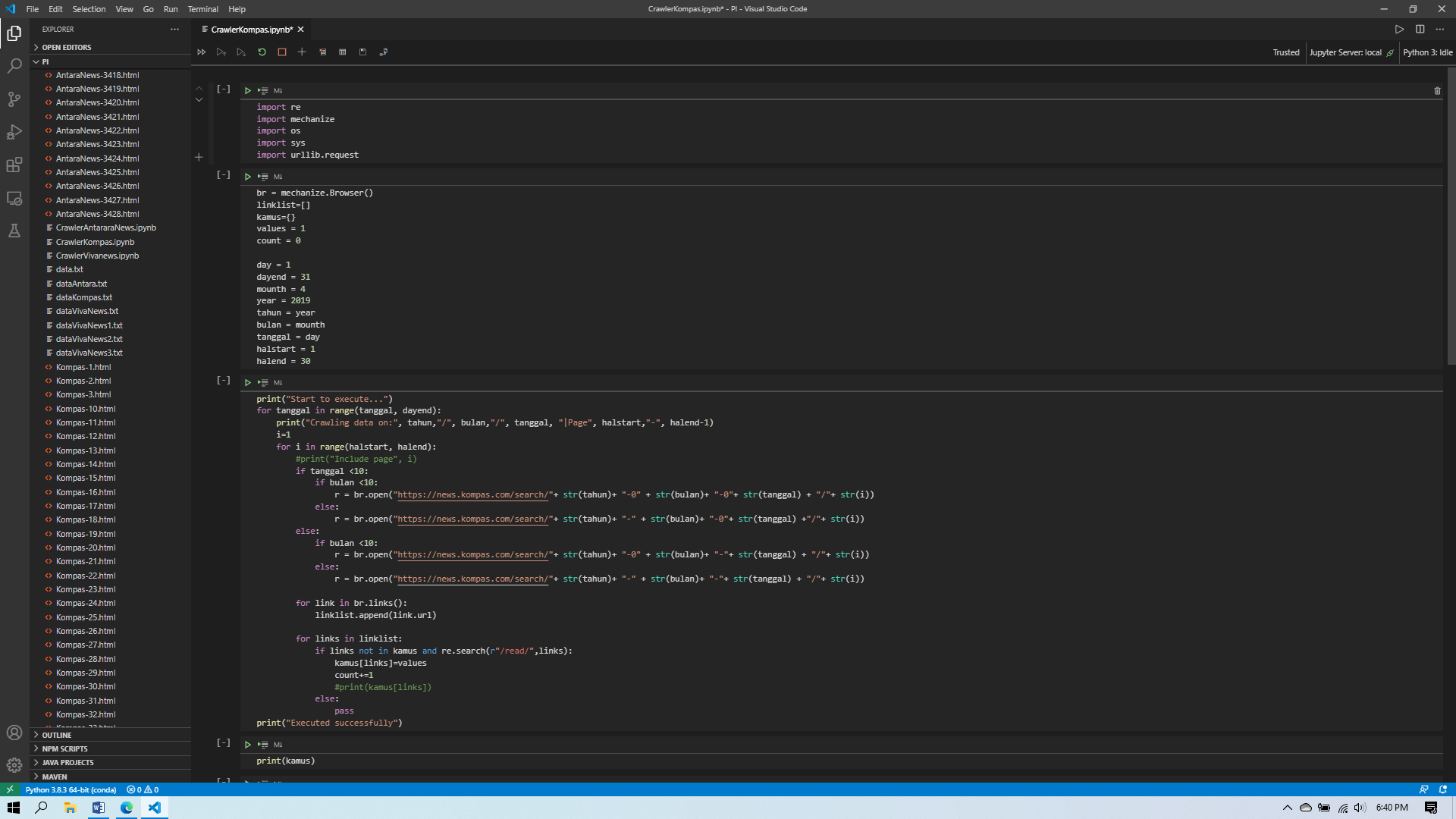
Berita yang terkumpul Dapat dilihat pada folder ./file. Secara keseluruhan berita yang diunduh sebanyaj 10K. Berita-berita ini nantinya dijadikan bahan *text* *indexing*. Berikut cuplikan skrip *crawling* berita:

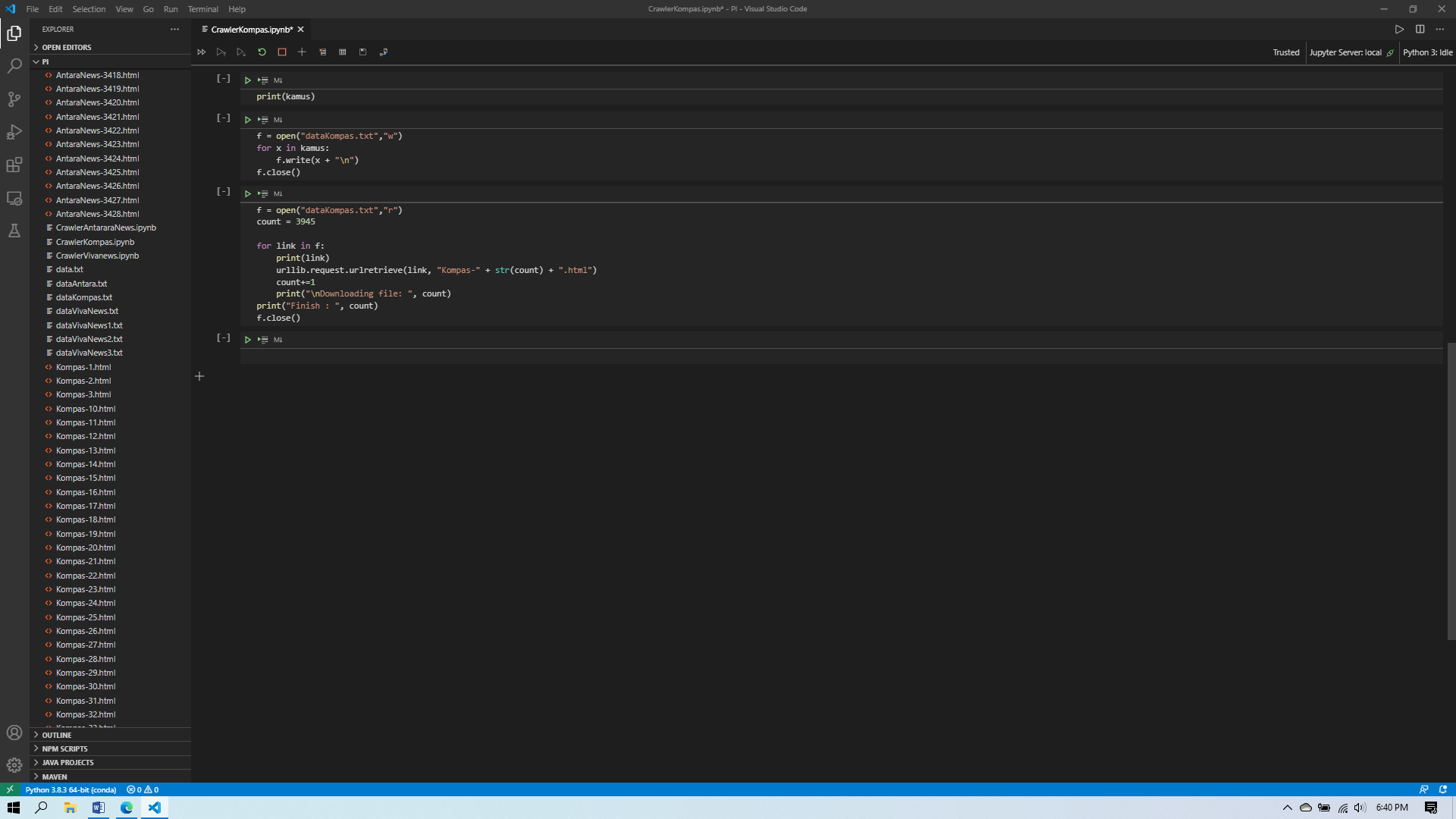
1. **AntaraNews**



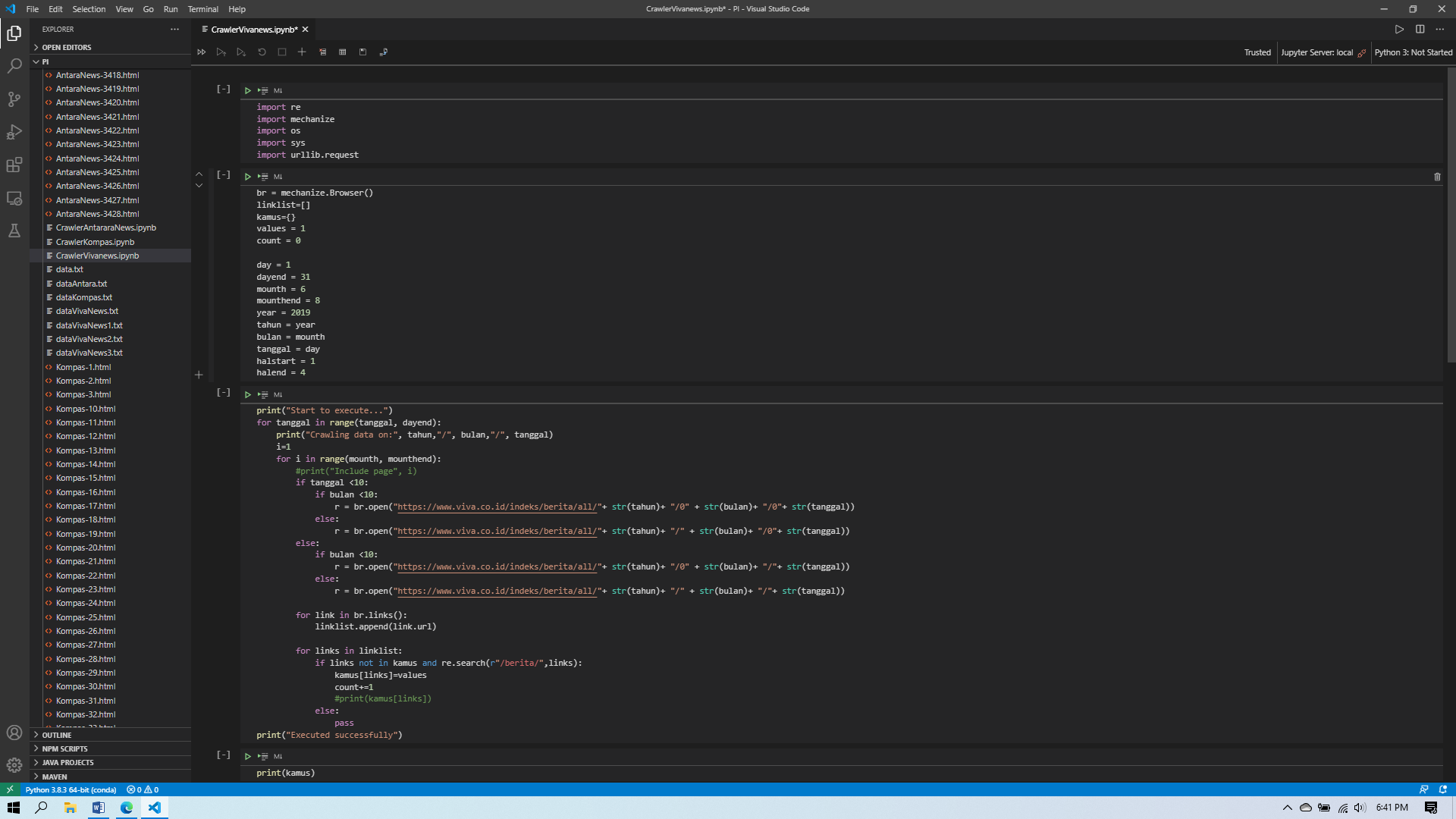


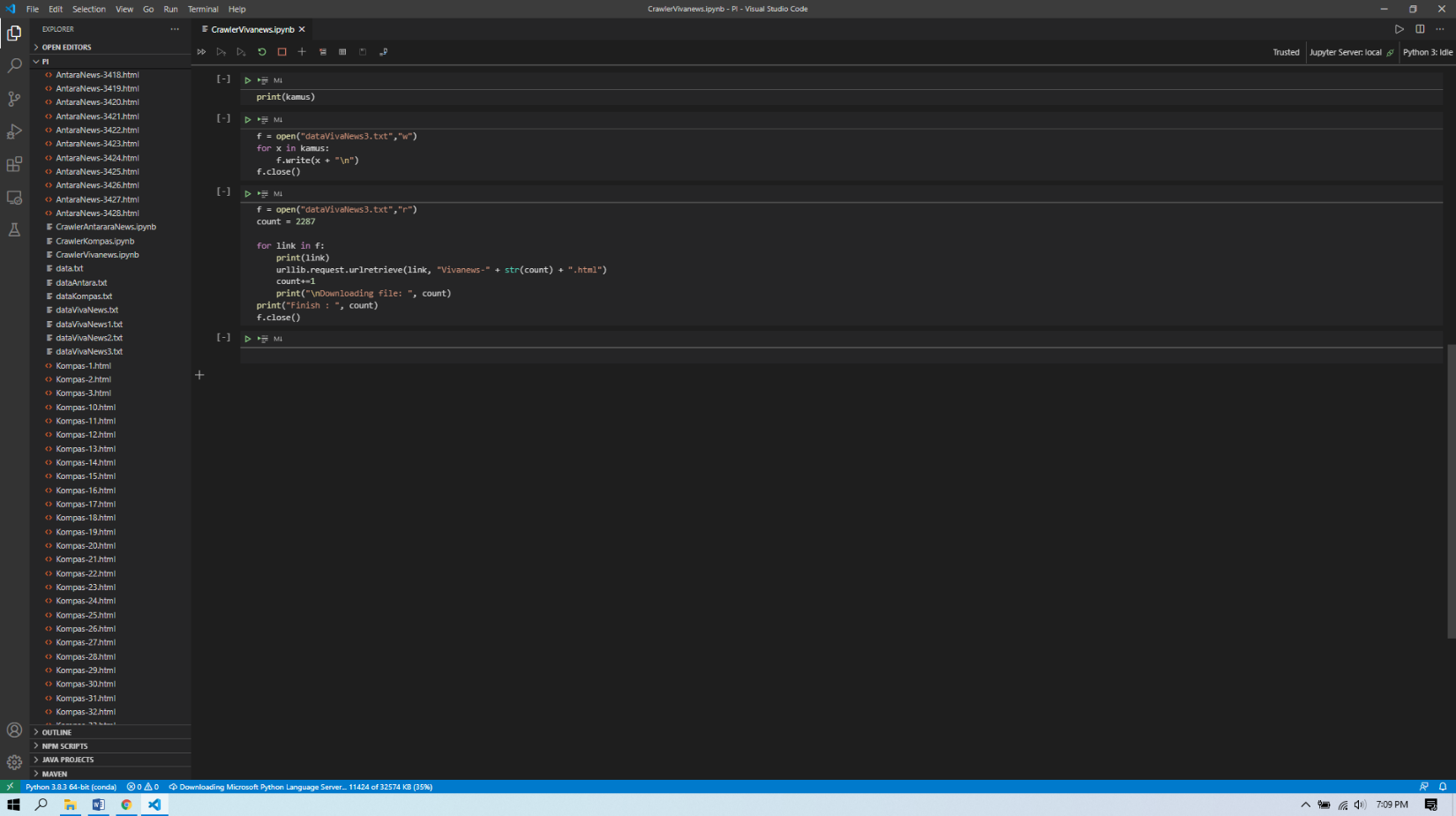
1. **KompasNews**





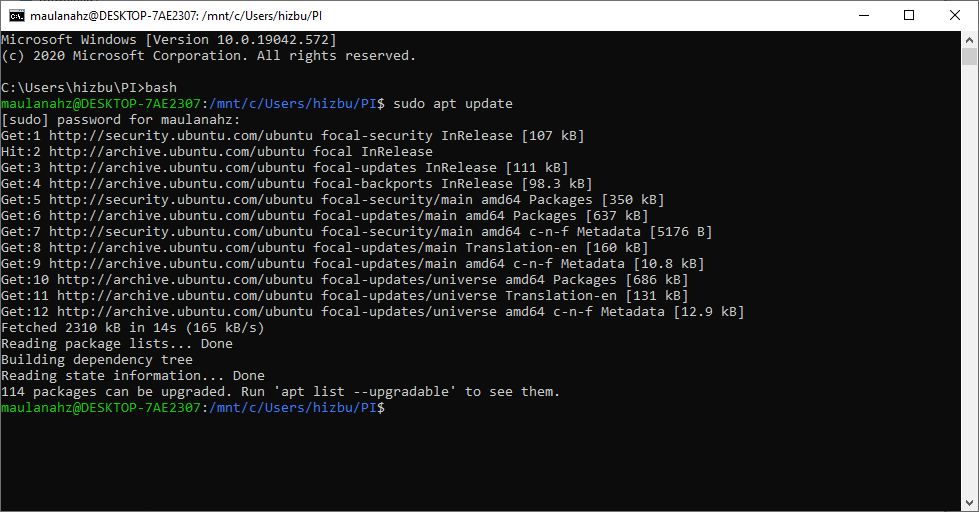
1. **VivaNews**





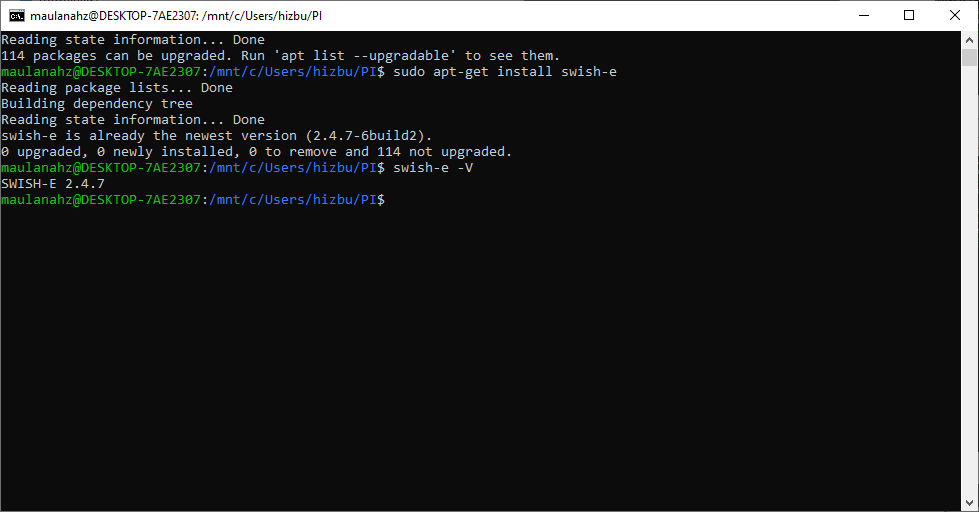
Setelah *crawling*, dilanjutkan dengan proses *indexing*:

* 1. Update terminal: **$ sudo apt update**

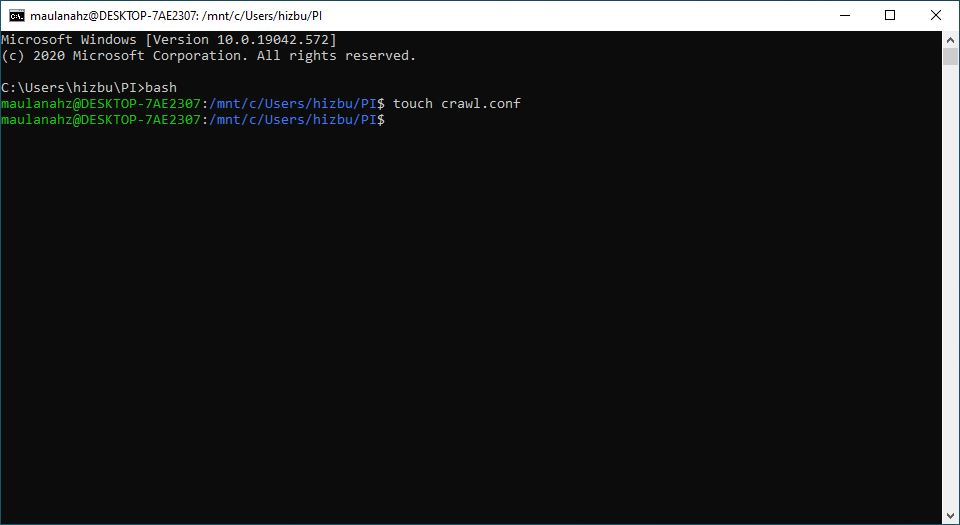


* 1. Install SWISH-e: **$ sudo apt-get install swish-e**

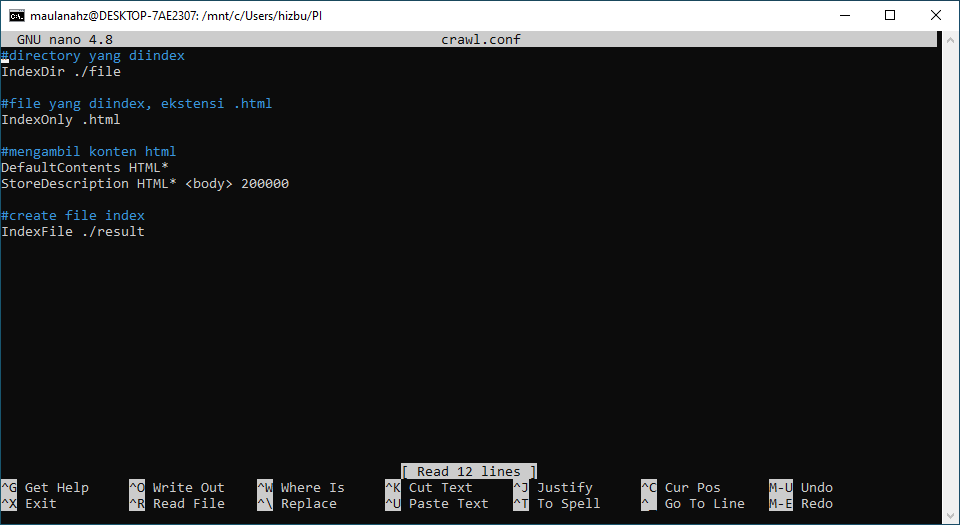
Cek Versi SWISH-e: **$ swish-e -V**



* 1. Membuat file konfigurasi SWISH-e untuk *indexing*:**$ touch crawl.conf**

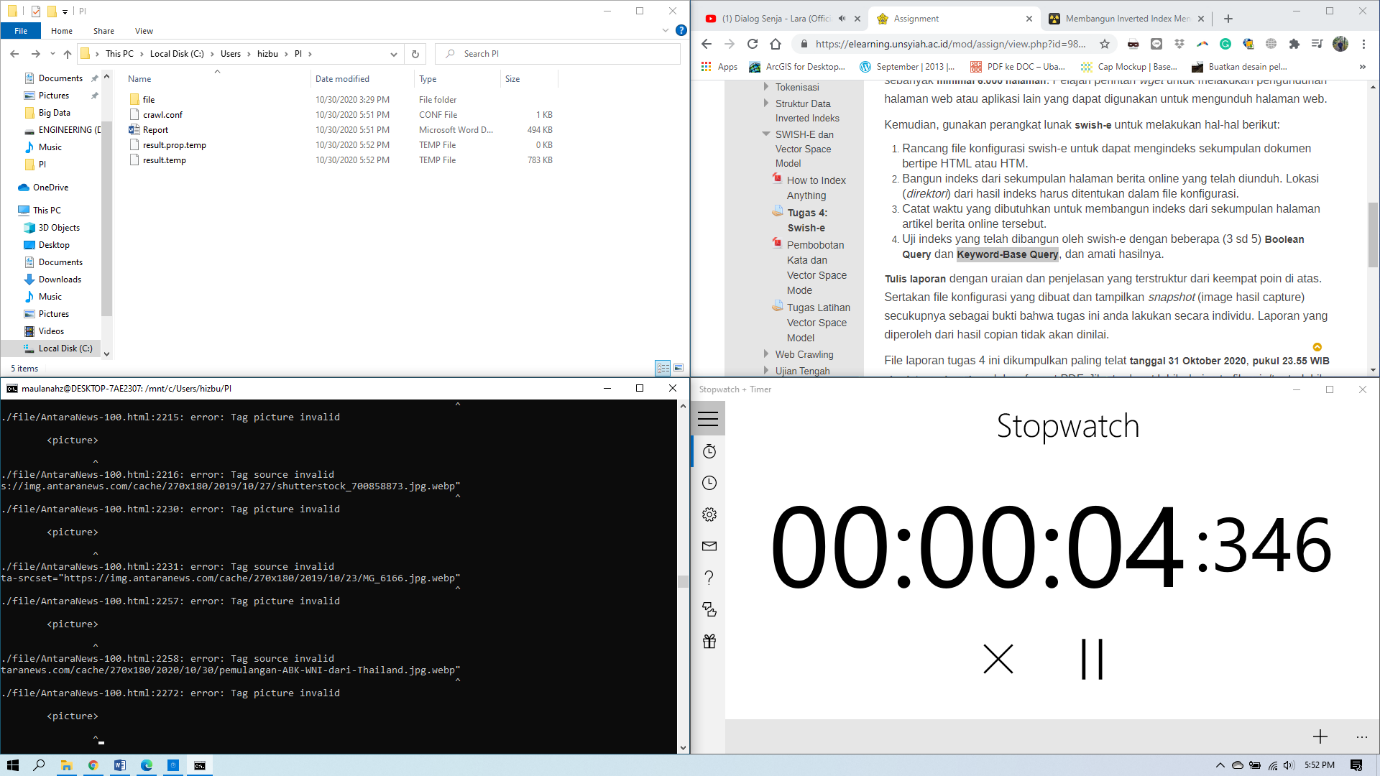


* 1. Menulis file konfigurasi yang telah dibuat sebelumnya: **$ pico crawl.conf**

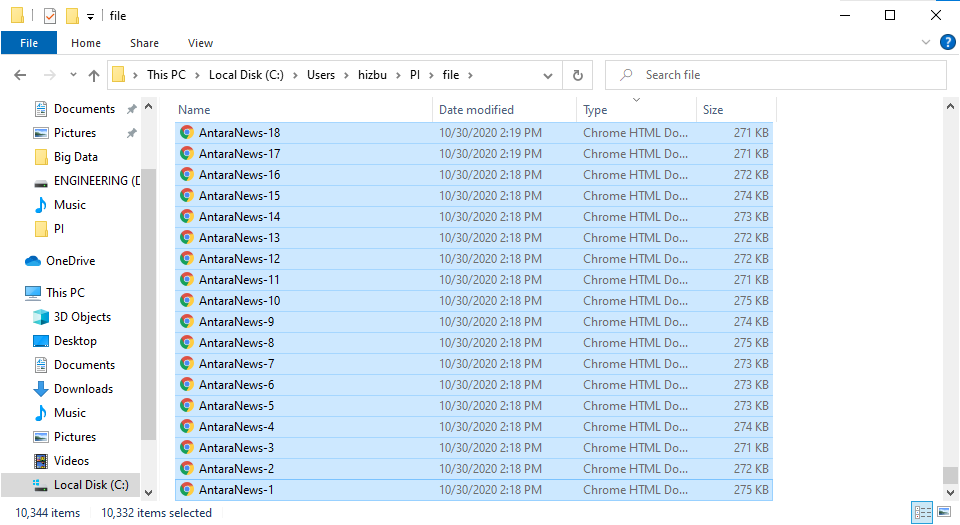


* 1. Menjalankan SWISH-e: **$ swish-e –c crawl.conf**

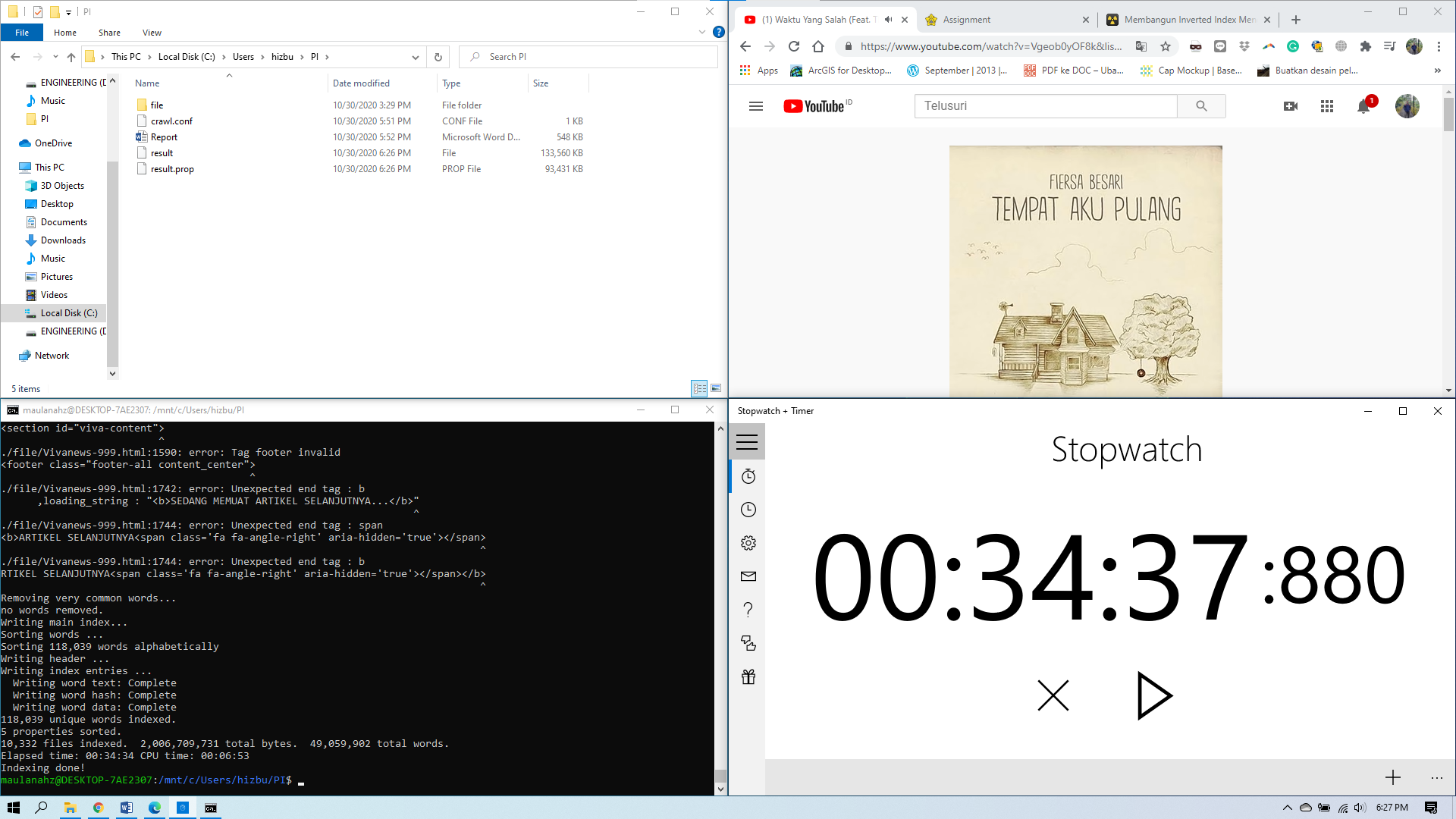
Proses *indexing* di SWISH-e dihitung menggunakan *stopwatch*.



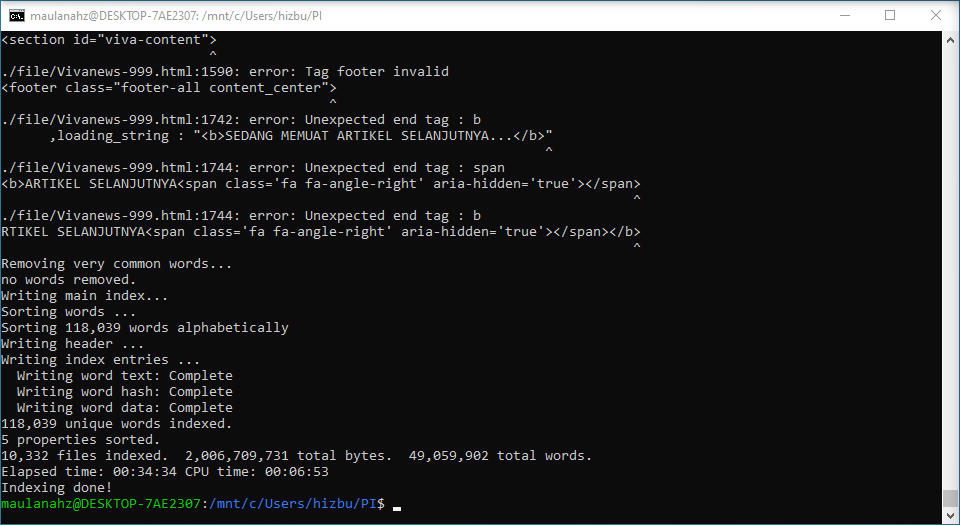
Dapat dilihat *file* hasil *crawling* yang diindeks berjumlah 10.322 *files*.



Dari hasil pengukuran waktu dengan *stopwatch*, dihasilkan lama proses *indexing* yaitu 34 menit dengan jumlah 10K *files*.

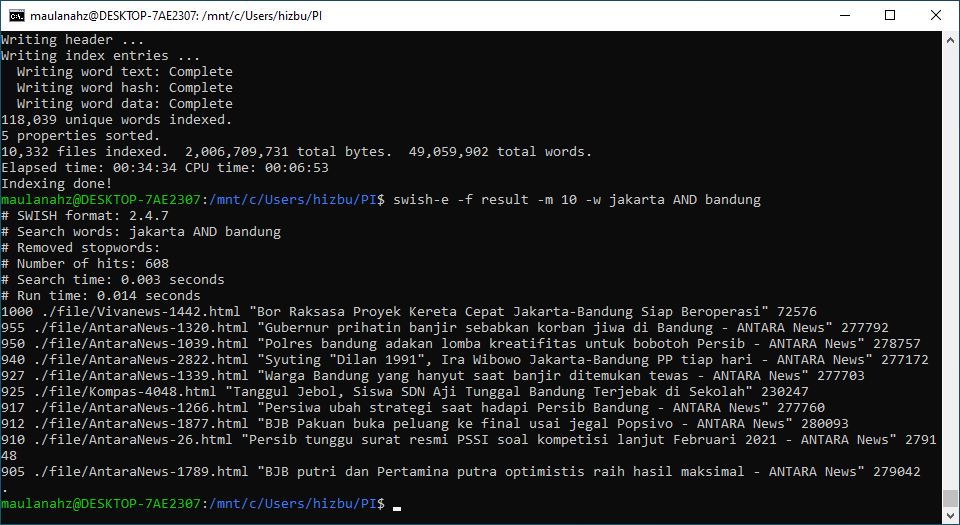


Hasil *indexing* dengan SWISH-e.

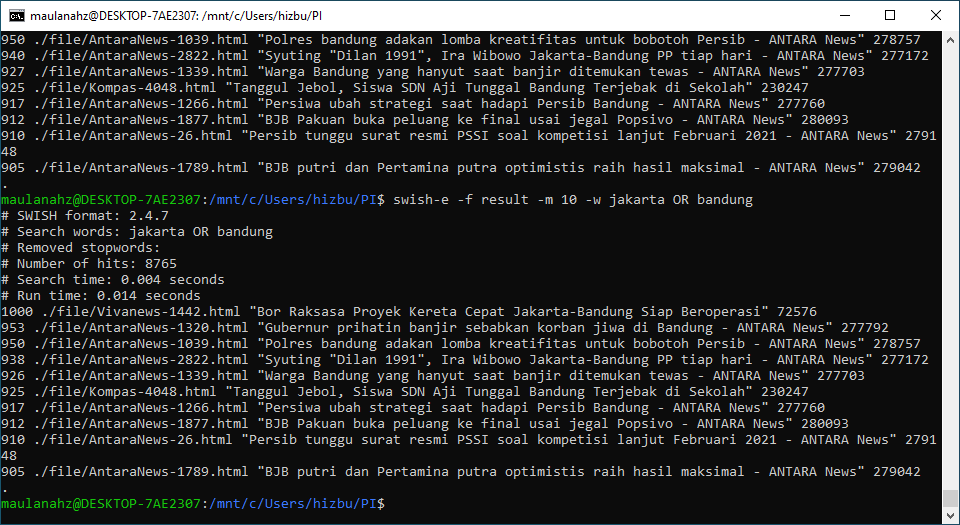


* 1. Melakukan proses *Search Word*
* search word **Boolean Query**

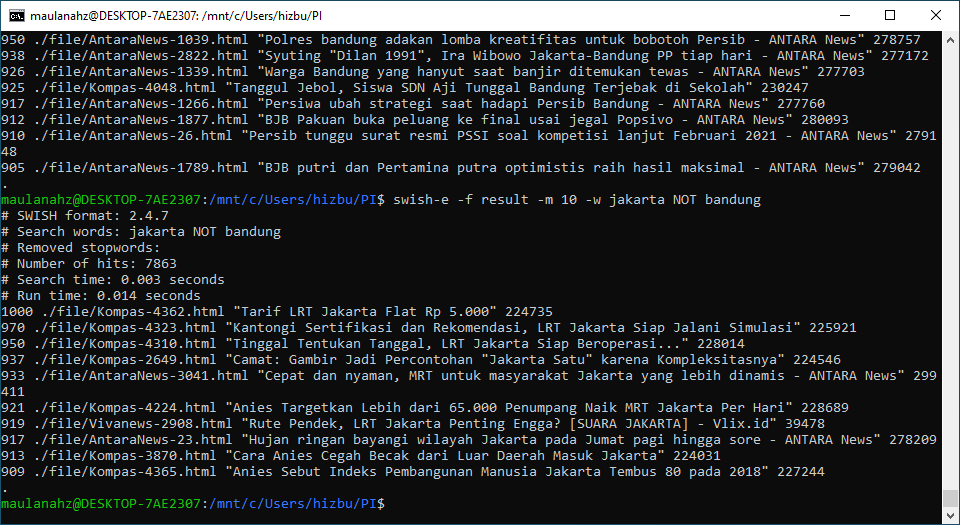
**$ swish-e –f result -m 10 –w Jakarta AND bandung** : menghasilkan 608 hit.



**$ swish-e –f result -m 10 –w Jakarta OR bandung** : menghasilkan 8765 hit.

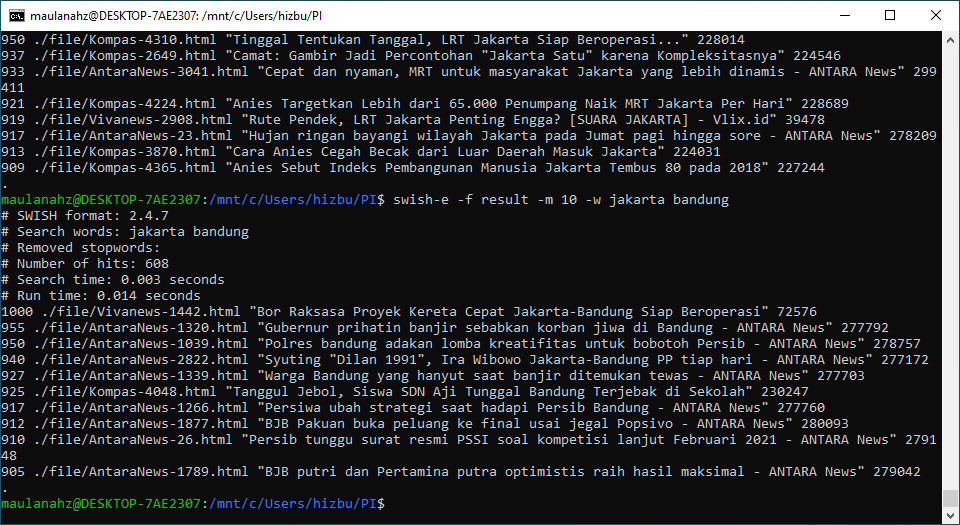


**$ swish-e –f result -m 10 –w Jakarta NOT bandung** : menghasilkan 7863 hit.



* search word **Keyword-Base Query**

**$ swish-e –f result -m 10 –w Jakarta bandung** : menghasilkan 608 hit.



* 1. Hasil **Boolean Query** dan **Keyword-Base Query**

Dari kata “jakarta bandung” menghasilkan **Keyword-Base Query** yang sama dengan hasil **Boolean Query** “AND”, karena menghasilkan hit yang sama yaitu 608. Hal ini menunjukkan proses *default* *keyword-base query* merupakan proses *Boolean query “AND”.*