Priscilla Vanny Amelia (0206021810012)

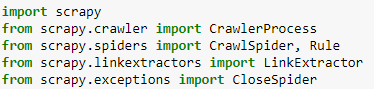
**PENJELASAN SOURCE CODE**

**Keterangan:**

1. Source Code akan terlebih dulu dijelaskan secara umum, baru kemudian perbedaan crawler untuk web besar dan web kecil.
2. Web Crawler ini menggunakan module “scrapy”
3. Semua crawler diletakkan pada 1 file yang sama, sehingga setiap mau berpindah crawler, harus me-reset kernel

**Penjelasan Crawler secara umum**:

1. Import semua modul dan package yang diperlukan



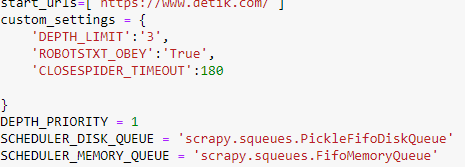
1. Buat kelas crawler yang merupakan turunan dari CrawlSpider:



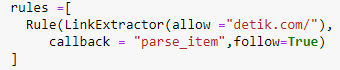
1. Deklarasi variable yang akan digunakan di parse:



1. Atur property settings untuk Crawler:
   1. DEPTH\_LIMIT: digunakan di web dengan jumlah halaman > 200, merupakan batas kedalaman crawler akan men-crawl web
   2. ROBOTSTXT\_OBEY: menentukan apakah crawler akan menaati robots.txt domain atau tidak
   3. CLOSESPIDER\_TIMEOUT: crawler akan tertutup secara otomatis ketika waktu sudah mencapai detik yang ditentukan



1. Menggunakan Package “Rule” dari “scrapy.spiders”, tentukan Batasan domain yang akan di crawl:



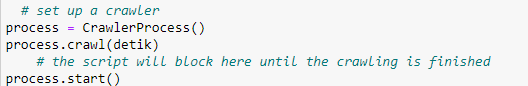
1. Buat function “parse\_item” yang merupakan callback function. Berisi algoritma pengolahan response crawler: (Penjelasan callback function akan dijelaskan di segmen selanjutnya)



1. Run Crawler:
   1. Instansiasi class Crawler Process
   2. Panggil crawl mana yang akan digunakan:

process.crawl(<kelas crawler turunan dari CrawlSpider yang dibuat di no. 2>

* 1. Mulai proses crawling



**Penjelasan “parse\_item” Callback Function:**

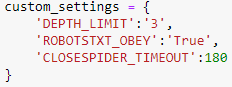


1. Untuk membuat file .txt penampung callback function:
   1. Deklarasi variable penampung index asal dan tujuan
   2. Menggunakan looping for, kita mencari indeks halaman referrer
   3. Apabila halaman referrer adalah baru/belum pernah tercatat di D, maka halaman tersebut akan dicatat
   4. Menggunakan looping for, kita mencari indeks halaman response
   5. Apabila halaman referer adalah baru/belum pernah tercatat di D, maka halaman tersebut akan ditambahkan di D
   6. Catat pasangan indeks halaman asal/referer dan tujuan/response di E
   7. Tuliskan D dan E di file .txt yang disediakan menggunakan .write()
2. Simpan setiap html ke dalalm folder yang sudah dibuat menggunakan open() dan .write():



Hal yang disimpan ke dalam file .html merupakan semua isi dalam tag <html> (didapatkan menggunakan response.css(“html”).get()

**Perbedaan Crawler Web Besar dan Web Kecil:** hanya terdapat di bagian ‘custom\_settings’. Web kecil tidak memakai parameter DEPTH\_LIMIT dan CLOSESPIDER\_TIMEOUT, karena untuk web kecil yg halamannya <200 akan di crawl semua

custom\_settings web kecil custom settings web besar

**PENJELASAN HASIL D DAN E TIAP WEB**

1. Hasil Web Indonesia Besar: <https://www.detik.com/>:

Dengan pembatasan depth = 3 dan time limit 180 detik, didapatkan 930 halaman html

1. Hasil Web Indonesia Kecil: <https://www.uc.ac.id/isb/>:

Didapatkan 75 halaman html

1. Hasil Web Bahasa Inggris Besar: <https://pcwonderland.com/>:

Dengan pembatasan depth = 3 dan time limit 180 detik, didapatkan 766 halaman html

1. Hasil Web Bahasa Inggris Kecil: <https://deepenglish.com/> :

Didapatkan 110 halaman html