REIN Laboratori 3

Daniel Beltrán Drago y Toni Cifré Vicens

Novembre 2020

Contents

Per el nostre treball hem realitzat dos *scrappers* complementaris, un que es dedica a realitzar la busca de pel·lícules d'una pagina web i un altre per trobar aquestes pel·lícules a una pagina de descarregues "poc legals", retornant aixi l'arxiu per la descarrega

Per el primer rastrejador l'hem anomenat Sensacine - Scrappy, donat que actua a la pagina web anomenada, mentres que el segon, l'hem anomenat Mejor-Torrent - Scrappy per el mateix motiu

1 Sensacine

1.1 Objectiu

El nostre objectiu a l'hora de plantejar el rastrejador es obtenir les pel·lícules de la pagina web `www.sensacine.com` per poder així realitzar consultes per poder obtenir el titul d'aquestes, l'autor, els actors principals, les valoracions, etc..

1.2 Modificaicons realitzades

En primer lloc, per a poder optimitzar la velocitat en la qual es guarden els ítems recol·lectats pel nostre scraper, hem modificat la manera en la qual es guarden dintre de `l'Elasticsearch`, de tal forma que hem evitat que és faig una petició cada cop que s'obté un document, sinó que hem creat una llista d'ítems, la qual es bolca a la nostra base de dades un cop te `n` documents dintre, evitat que fer una petició cada vegada que recollim un ítem.

```
class SensaCineElasticPipeline(object):
    collection_name = 'scrapy-sensacine'

    def __init__(self):
        self.client = Elasticsearch()
        self.elastic_uri = 'http://localhost:9200/'
        self.elastic_db = 'scrapy-sensacine'

        self.l_docs = []
        self.n_docs = 500
        self.index_dic = {'_op_type': 'index', '_index': self.elastic_db, '_ty

    def open_spider(self, spider):
        ...

    def close_spider(self, spider):
        if len(self.l_docs) > 0:
            print('[INFO] Indexing ...', end='')

            bulk(self.client, self.l_docs)
            self.l_docs.clear()
```

```
            print('    [OK]')
        self.client.close()

    def process_item(self, item, spider):
        self.l_docs.append({**self.index_dic, **item})

        if len(self.l_docs) > self.n_docs:
            print('[INFO] Indexing ...', end='')

            bulk(self.client, self.l_docs)
            self.l_docs.clear()

            print('    [OK]')

        return item
```

D'altra banda, per a poder aplicar aquest nou pipeline creat s'ha de modificar l'arxiu settings.py indicant quina de les classes dintre de l'arxiu s'ha d'utilitzar, aquesta indicació es fa amb el següent paràmetre:

```
ITEM_PIPELINES = {
    'reinscrapy.pipelines.MejorTorrentElasticPipeline': 300,
    # 'reinscrapy.pipelines.ReinscrapyPipeline': 300,
    # 'reinscrapy.pipelines.ReinscrapyElasticPipeline': 300,
}
```

Dintre d'aquest paràmetre, l'enter indica la prioritat d'execució, en el cas que es vulgui executar més d'un pipeline es pot indicar afegint la prioritat d'execució.

Per últim hem afegit dos paràmetres més de configuració, el primer és per indicar que la informació que recol·lecta'm i guardem es processi en 'UTF-8' per evitar que en recuperar-la es perdi o modifiqui, ja que en aquesta pàgina s'utilitza aquesta codificació. L'últim paràmetre indica que només mostrarà per consola la informació rellevant, per tal de poder seguir l'execució de forma correcta i no rebre massa informació irrellevant.

```
FEED_EXPORT_ENCODING = 'utf-8'
LOG_LEVEL = 'INFO'
```

Per a poder observar el resultat, hem modificat l'arxiu `SearchIndex.py`, afegint dues funcions, una per cercar i mostrar de forma adient la informació dels índexs de `Sensacine` i `MejorTorrent` depenent dels arguments introduïts per l'usuari, de tal forma que la funció per `Sensacine` ens queda de la següent manera.

```
def search_sensacine():
    try:
```

```
        print('=== SensaCine ===\n')

        s = Search(using=client, index='scrapy-sensacine')

        q = Q('query_string', query=query)
        s = s.query(q)
        response = s[0:10].execute()

        for r in response:
            print(f'\n----- {r["Titulo"]} ----------')
            print(f'Genero: {r["Genero"]}')
            print(f'Director: {r["Director"]}')
            print(f'Reparto: {r["Reparto"]}')
            print(f'Puntuacion: {r["SensaCine"]} / {r["Medios"]} '
                  f'/ {r["Usuarios"]}')
            print(f'Sinopsis:\n{r["Sinopsis"]}')
            print('==========')

        print('%d Documents' % response.hits.total.value)
    except NotFoundError:
        print('Index %s does not exist' % index)
```

1.3 Problemes a l'implementació

A l'hora de realitzar l'implementació d'aquest rastrejador no hem tingut una gran quantitat de problemes, més enllà de entendre el codi de la pròpia pagina web.

Donat que l'estructura estava realment clarificada, el temps de realitzar l'anàlisi per saber com indicar els paràmetres que ens interessen guardar ha set pràcticament minúscul

1.4 Sortida del rastrejador

Després de l'execució del nostre scraper, hem obtingut un total de 53106 documents (pel·lícules) amb un tamany de 47.5mb amb una velocitat mitjana de recol·lecció de 500 documents per minut.

Per a veure un exemple sé sortida, podem buscar la pel·lícula Vengadores on aparegui l'actor Robert Deniro

```
$ python SearchIndex.py --index scrapy-sensacine --query Titulo:Vengadores
AND Titulo:Endgame AND Reparto:Robert


=== SensaCine ===

----- Vengadores: Endgame ------------------------
Género: Acción Fantasía Aventura
```

```
Director: Joe Russo Anthony Russo
Reparto: Robert Downey Jr. Chris Evans
Puntuación: 4.5 / 4.1 / 4.5
Sinopsis:
Después de los devastadores eventos ocurridos en Vengadores: Infinity War ...
================================================
1 Documents
```

También es pot obtenir les pel·lícules on els medis les han classificades amb un 5, és a dir, la nota més alta.

```
$ python SearchIndex.py --index scrapy-sensacine --query Medios:5

...
51 Documents
```

2 MejorTorrent

2.1 Objectiu

La principal utilitat d'aquesta eina es la de recuperar links de descarrega per les pel·lícules de la pròpia pagina web. A més, al combinar els dos rastrejadors podem obtenir links de descarrega de pel·lícules que ens recomana el *Sensacine* sense la necessitat d'entrar directament a la pagina, minimitzant aixi el risc donat que es una pagina poc segura.

2.2 Modificacions

Per a poder executar el scraper MejorTorrent hem hagut d'ajustar el fitxer settings.py, sumades a les anteriorment fetes pel scraper SensaCine, per a no ser detectes per la pàgina com un scraper maliciós, perquè, al ser detectats com a tal disposa d'un mecanisme per evitar l'extracció d'informació a través d'un rastrejador.

Per evitar ser detectat i bloquejats pel firewall, hem afegit un temps d'espera entre cada petició de dos segons en els setting afegint l'opció `DOWNLOAD_DELAY = 2`.

Per altra banda, la pàgina, també disposa del document robot.txt per tal de fer la petició de no ser rastrejat al robot. En el nostre cas, a l'utilitzar la informació només per un estudi de classe, hem decidit no fer cas a la seva petició afegint l'opció `ROBOTSTXT_OBEY = False` dels settings.

Per últim hem creat una altra classe dintre del fitxer `pipelines.py` de la mateixa forma que amb el scraper SensaCine però amb les modificacions pertinents de l'índex amb el qual és guarda't dintre de l'Elasticsearch.

2.3 Problemes a l'implementació

Per aquesta implementació hem tingut que ajustar paràmetres del propi SearchIndex, degut a que esta pagina contava amb un firewall, per augmentar el temps entre peticions per aixi no ser detectats.

També remarcar que aquesta implementació ens a set més costos pel fet del mal estructurada que es trobava aquesta pagina, fent aixi una mol mala tant lectura com a implementació del nostre rastrejador

Per ultim comentar, que degut al augment de temps entre peticions (forçat per la pròpia pagina web), tardem molt al realitzar l'execució, ralentitzant-nos aixi el flux de treball.

2.4 Sortida del rastrejador

En aquest cas farem la mateixa consulta que amb SensaCine, però per obtenir els torrents disponibles de la pel·lícula dels Vengadores

```
$ python SearchIndex.py --index scrapy --query Titulo:Vengadores AND
Titulo:Endgame AND Actores:Robert
```

```
=== Mejor Torrent ===

----- Vengadores: Endgame -------------------------
Género: Ciencia Ficción - Fantasía - Acción.
Actores: Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, Scarlett Johanss
ana, Evangeline Lilly.
Formato: HDRip
Tamaño: 2,85 GB
Descripción:
 Después de los devastadores eventos ocurridos en Vengadores: Infinity War, el universo
Link Descarga = http://www.mejortorrentt.net/tor/peliculas/Vengadores_Endgame.torrent
================================================

----- Vengadores: Endgame (Open Matte Imax Proper) -------------------------
Género: Ciencia Ficción - Fantasía - Acción.
Actores: Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, Scarlett Johanss
ana, Evangeline Lilly.
Formato: BLRey
Tamaño: 13,65 GB
Descripción:
 Después de los devastadores eventos ocurridos en Vengadores: Infinity War, el universo
Link Descarga = http://www.mejortorrentt.net/tor/peliculas/Vengadores_Endgame_OPEN_MATTE
================================================
2 Documents
```

Com es pot observar obtenim dos documents, ja que disposem de dos torrents diferents de qualitats diferents