Università degli studi Modena e Reggio Emilia
Dipartimento di Ingegneria "Enzo Ferrari"
Corso di laurea in Ingegneria Informatica

Sviluppo di un crawler per l'estrazione di articoli e citazioni da Google Scholar

Relatore
Luca Gagliardelli

Candidata
Ferrari Chiara

Correlatore
Giovanni Simonini

Anno Accademico
2021/2022

Obbiettivo dell'elaborato

Sviluppare un crawler in grado di:

- estrarre da Google Scholar i dati relativi alle pubblicazioni di un determinato autore e tutti gli articoli che le citano
- memorizzare all'interno di file in formato .csv i dati estratti

PROGRAMMI UTILIZZATI

Python

Anaconda

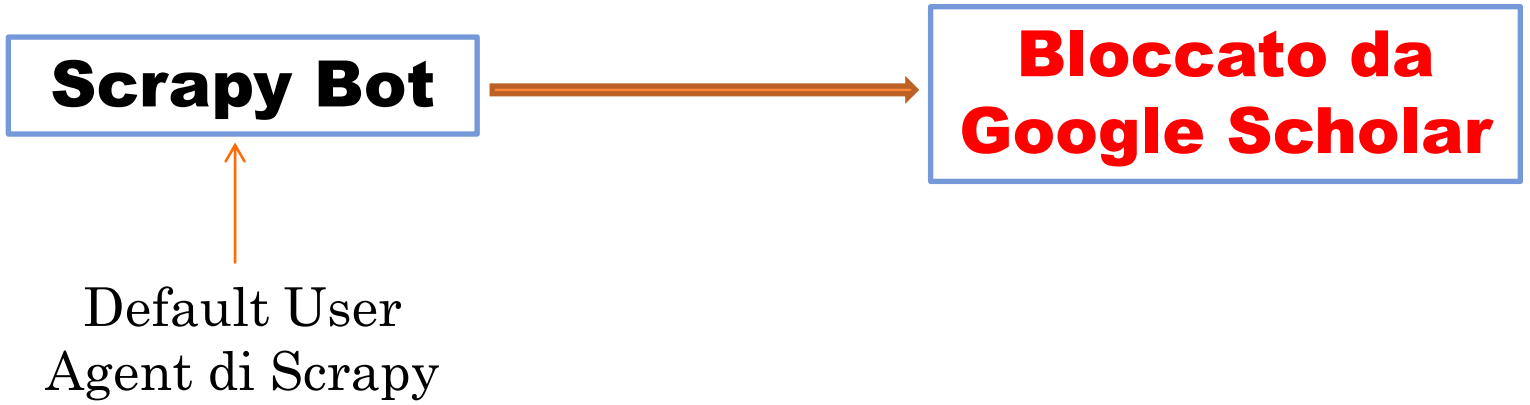
Scrapy

Selenium (in particolare Selenium Webdriver)

Chromedriver

User Agent

Stringa di testo utilizzata dal web server per identificare il web browser e il sistema operativo da cui proviene una richiesta.
Fa parte dell'header HTTP inviato al sito web.



User Agent: Soluzione

Sostituire l'User Agent di default con altri User Agent non bloccati da Google Scholar

-file "utils.py"

```
import random
user_agent_list = [
# Chrome
        'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML,
like    Gecko)  Chrome/60.0.3112.113 Safari/537.36',
]
def get_random_agent():
        return random.choice(user_agent_list)
```

- file "settings.py"

```
from tesina.utils import get_random_agent
USER_AGENT = get_random_agent()
```

Estrazione articoli e citazioni

Utilizzo selettori di Scrapy:

-XPath:

```
targets=driver.find_elements_by_xpath("//a[@class='gs_or_cit gs_or_btn gs_nph']")
```

-CSS:

```
for quote in sel.css('tr.gsc_a_tr'):
        titolo = quote.css('td.gsc_a_t a::text').get()
```

LINK

Follow() e Follow_all() : Funzioni di Python utilizzate per seguire i link

```
▼<tr class="gsc_a_tr">
  ▶<td class="gsc_a_t">…</td>
  ▼<td class="gsc_a_c">
    <a href="https://scholar.google.com/scholar?oi=bibs&hl=it&cites=15762210934181522530" class="gsc_a_ac gs_ibl">9</a> == $0
```

Link alla pagina successiva

```
for quote in sel.css('tr.gsc_a_tr'):
        next_page = quote.css('td.gsc_a_c a::attr(href)').get()
        if next_page is not None:
                yield response.follow(next_page,callback=self.parse_citazioni1)
```

Problema "Mostra altri"

Entity resolution and data fusion: An integrated approach 4 2019
D Beneventano, S Bergamaschi, L Gagliardelli, G Simonini
SEBD 2019: 27th Italian Symposium on Advanced Database Systems 2400

Articoli 1–20 MOSTRA ALTRI

Click utente necessario per caricare tutti gli articoli

Problema: Scrapy non supporta JavaScript

Soluzione: Selenium Webdriver

E' in grado di simulare interazioni da parte dell'utente su una pagina web:

```
driver = webdriver.Chrome()
driver.get(response.url)
button = driver.find_element_by_xpath("//button[@id='gsc_bpf_more']")
button.click()
```

IMPORTANTE: il selettore deve essere associato al driver:

```
sel = Selector(text=driver.page_source)
        for quote in sel.css('tr.gsc_a_tr'):
                titolo = quote.css('td.gsc_a_t a::text').get()
```

BIBTEX

Strumento utilizzato per la formattazione di testi bibliografici che utilizza un formato di file di tipo testuale,contenente un elenco di voci bibliografiche.

Citazione:

Blocking and filtering techniques for entity resolution: A survey
G Papadakis, D Skoutas, E Thanos... - ACM Computing Surveys ..., 2020 - dl.acm.org
Entity Resolution (ER), a core task of Data Integration, detects different entity profiles that correspond to the same real-world object. Due to its inherently quadratic complexity, a series ...
☆ Salva 99 Cita Citato da 98 Articoli correlati Tutte e 15 le versioni »

Formato Bibtex:

```
@article{papadakis2020blocking,
  title={Blocking and filtering techniques for entity resolution: A survey},
  author={Papadakis, George and Skoutas, Dimitrios and Thanos, Emmanouil and Palpanas,
Themis},
  journal={ACM Computing Surveys (CSUR)},
  volume={53},
  number={2},
  pages={1--42},
  year={2020},
  publisher={ACM New York, NY, USA}
}
```

MEMORIZZAZIONE DATI

1. Opzione modifica Settings.py
   PRO:
   - Facile da modificare
   CONTRO:
   - risultato poco chiaro
   -impossibilità di generare più file

```
# Desired file format
                FEED_FORMAT = "csv"
# Name of the file where data extracted is stored
                FEED_URI = "tesina.csv"
```

1	Authors,N_Citations,Title,Year
2	G Simonini, S Bergamaschi, HV Jagadish,89,BLAST: a loosely schema-aware meta-blocking approach for entity resolution,2016
3	F Benedetti, D Beneventano, S Bergamaschi, G Simonini,65,Computing inter-document similarity with context semantic analysis,2019
4	G Simonini, G Papadakis, T Palpanas, S Bergamaschi,53,Schema-agnostic Progressive Entity Resolution,2018

MEMORIZZAZIONE DATI

2. Opzione modifica script
   PRO :
   -risultato ordinato
   -possibilità di creare file separati
   CONTRO:
   -più complesso da modificare

```
header = ['titolo;' 'autori;' 'anno;' 'citazioni']
with open('scholar.csv', 'w', encoding='UTF8',newline='') as f:
        writer = csv.writer(f)
        writer.writerow(header)
        delimiter=';'
        data =[titolo2 + delimiter + autori2 + delimiter + anno2 + delimiter + citazioni2]
            writer.writerow(data)
```

1	titolo	autori	anno	citazioni
2	BLAST: a loosely schema-aware meta-blocking approach for entity resolution	G Simonini- S Bergamaschi- HV Jagadish	2016	91
3	Computing inter-document similarity with context semantic analysis	F Benedetti- D Beneventano- S Bergamaschi- G Simonini	2019	66
4	Schema-agnostic Progressive Entity Resolution	G Simonini- G Papadakis- T Palpanas- S Bergamaschi	2018	55

CONSIGLIATO

BLOCCO IP

CAUSA:
"Distributed Denial Of Service" (DDos): attacco informatico volto a sovraccaricare un sito Web di richieste al fine di renderlo inaccessibile.

EFFETTO:
Blocco da parte di un sito web di un IP a seguito di molteplici richieste in un breve lasso di tempo.

SOLUZIONE:PROXY

Nascondono il reale IP da cui proviene una richiesta impedendo che questa venga bloccata a causa delle troppe richieste provenienti da uno stesso IP

```
ROTATING_PROXY_LIST = [
                'https://114.55.111.137:22',
                'https://176.57.188.32:443',
                'https://157.230.34.219:3128'
]

DOWNLOADER_MIDDLEWARES = {
                'rotating_proxies.middlewares.RotatingProxyMiddleware': 610,
                'rotating_proxies.middlewares.BanDetectionMiddleware': 620
}
```

Grazie per l'attenzione