TECHNOLOGIES FOR INFORMATION SYSTEMS ROJECT

PROF. LETIZIA TANCA

WEB SCRAPING OF ANSA.IT

Alessandro Artoni - 899343

ACADEMIC YEAR 2017/2018

POLITECNICO
MILANO 1863

Introduction

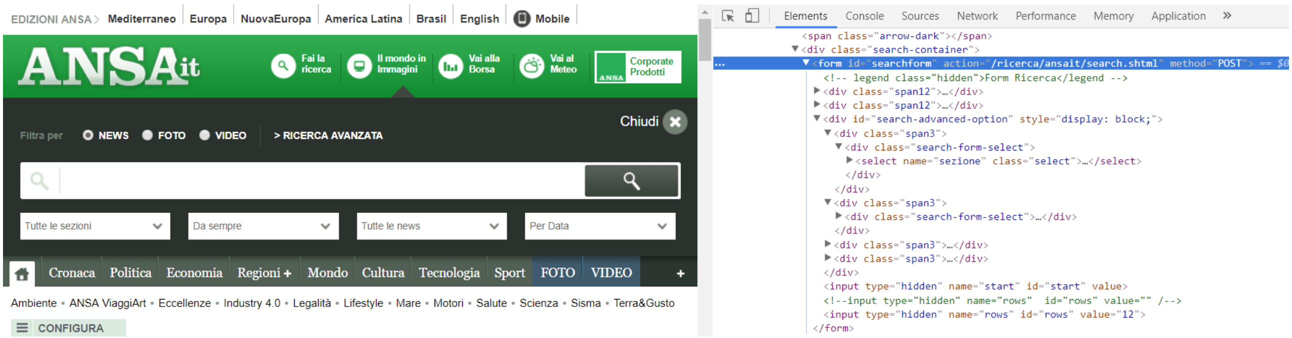
The purpose of this project is to collect economical articles from the website ansa.it using an automated process implemented with a script in python.
The script makes a request for a certain page, downloads it, takes the information that are needed and loops - making another request - until it reached the last article.

The Process

This section describes the steps followed to reach the objective of the project.

Landing page analysis of Ansa

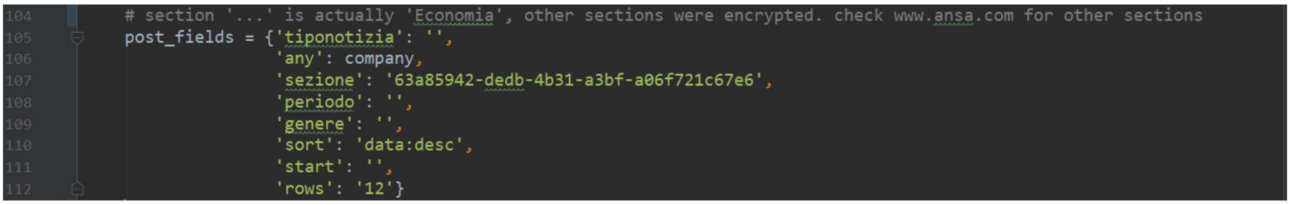
Firstly, the landing page of Ansa was analysed. We noticed that there is a form where an user can make a post to the database, and be redirect to a "result-page".
At this point, we needed to craft a proper post request, based on the attributes that the form needs.


EDIZIONI ANSA > Mediterraneo Europa NuovaEuropa America Latina Brasil English Mobile
ANSA.it
Fai la ricerca
Il mondo in Immagini
Vai alla Borsa
Vai al Meteo
Corporate Prodotti
Chiudi
Filtra per ● NEWS ● FOTO ● VIDEO > RICERCA AVANZATA
Tutte le sezioni
Da sempre
Tutte le news
Per Data
Cronaca Politica Economia Regioni + Mondo Cultura Tecnologia Sport FOTO VIDEO +
Ambiente • ANSA ViaggiArt • Eccellenze • Industry 4.0 • Legalità • Lifestyle • Mare • Motori • Salute • Scienza • Sisma • Terra&Gusto
CONFIGURA
Elements Console Sources Network Performance Memory Application »

```
      <span class="arrow-dark"></span>
    ▼<div class="search-container">
...   ▼<form id="searchform" action="/ricerca/ansait/search.shtml" method="POST"> == $0
        <!-- legend class="hidden">Form Ricerca</legend -->
       ▶<div class="span12">…</div>
       ▶<div class="span12">…</div>
       ▼<div id="search-advanced-option" style="display: block;">
         ▼<div class="span3">
           ▼<div class="search-form-select">
             ▶<select name="sezione" class="select">…</select>
            </div>
          </div>
         ▼<div class="span3">
           ▶<div class="search-form-select">…</div>
          </div>
         ▶<div class="span3">…</div>
         ▶<div class="span3">…</div>
        </div>
        <input type="hidden" name="start" id="start" value>
        <!--input type="hidden" name="rows"  id="rows" value="" /-->
        <input type="hidden" name="rows" id="rows" value="12">
      </form>
```

Figure 1 - Analysis of ANSA landing page

From Fig. 1, we can see the fields needed, and from Fig. 2, the corresponding post request.
The first span12, is a container for the type of the news (news, foto, video). If it is left empty, it selects news.
The second span12, is a container for the search request. In the program there's "company" which is a variable that contains the name of the company from wich we wish to retrieve information.
"Sezione" means section. Since we were interested in "Economia", we selected '63a85942-dedb-4b31-a3bf-a06f721c67e6' which should be encrypted but it's actually written clear-text on the source html. Other fields display how do we want to sort dates (in our case from the most recent to the least recent), and the number of rows we wanted to display.

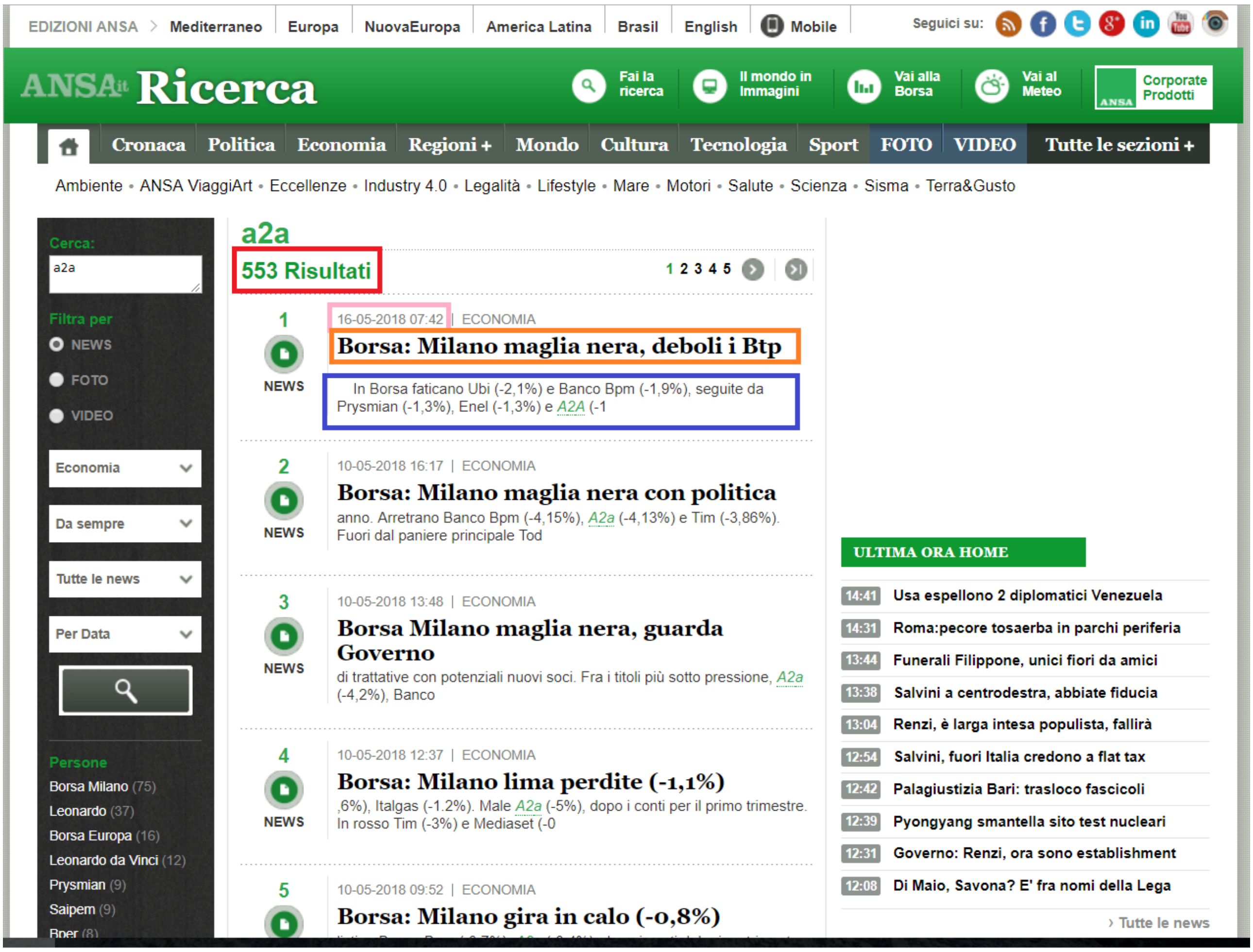


```
    # section '...' is actually 'Economia', other sections were encrypted. check www.ansa.com for other sections
    post_fields = {'tiponotizia': '',
                   'any': company,
                   'sezione': '63a85942-dedb-4b31-a3bf-a06f721c67e6',
                   'periodo': '',
                   'genere': '',
                   'sort': 'data:desc',
                   'start': '',
                   'rows': '12'}
```

Figure 2 - Post request fields

Research page

Then, we analysed the next research page, and from that we started to retrieve some information.



Cerca: a2a

Filtra per ● NEWS ○ FOTO ○ VIDEO

Persone
Borsa Milano (75)
Leonardo (37)
Borsa Europa (16)
Leonardo da Vinci (12)
Prysmian (9)
Saipem (9)
Bper (8)

a2a

553 Risultati

1 NEWS
16-05-2018 07:42 | ECONOMIA
Borsa: Milano maglia nera, deboli i Btp
In Borsa faticano Ubi (-2,1%) e Banco Bpm (-1,9%), seguite da Prysmian (-1,3%), Enel (-1,3%) e A2A (-1

2 NEWS
10-05-2018 16:17 | ECONOMIA
Borsa: Milano maglia nera con politica
anno. Arretrano Banco Bpm (-4,15%), A2a (-4,13%) e Tim (-3,86%). Fuori dal paniere principale Tod

3 NEWS
10-05-2018 13:48 | ECONOMIA
Borsa Milano maglia nera, guarda Governo
di trattative con potenziali nuovi soci. Fra i titoli più sotto pressione, A2a (-4,2%), Banco

4 NEWS
10-05-2018 12:37 | ECONOMIA
Borsa: Milano lima perdite (-1,1%)
,6%), Italgas (-1.2%). Male A2a (-5%), dopo i conti per il primo trimestre. In rosso Tim (-3%) e Mediaset (-0

5
10-05-2018 09:52 | ECONOMIA
Borsa: Milano gira in calo (-0,8%)

ULTIMA ORA HOME

14:41	Usa espellono 2 diplomatici Venezuela
14:31	Roma:pecore tosaerba in parchi periferia
13:44	Funerali Filippone, unici fiori da amici
13:38	Salvini a centrodestra, abbiate fiducia
13:04	Renzi, è larga intesa populista, fallirà
12:54	Salvini, fuori Italia credono a flat tax
12:42	Palagiustizia Bari: trasloco fascicoli
12:39	Pyongyang smantella sito test nucleari
12:31	Governo: Renzi, ora sono establishment
12:08	Di Maio, Savona? E' fra nomi della Lega

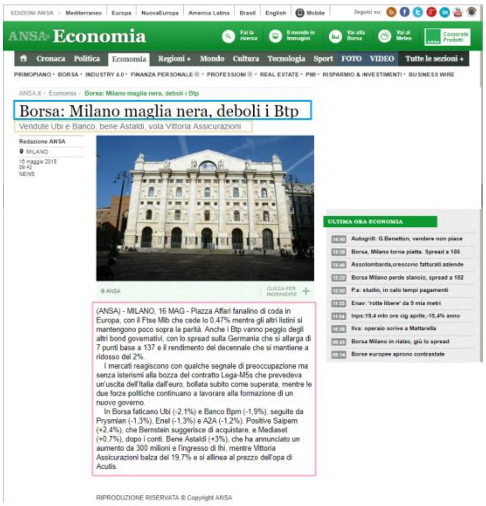
› Tutte le news

Figure 3 – Research Page

In *red* we see the number of results, in *pink* there's the date and the time, in *orange* there's the title and in *blue* there's the subtitle.

From this page on, we retrieve the links of each article in the page. Then we make a simple get request to the article and we reach the article page.

Article page



Borsa: Milano maglia nera, deboli i Btp

Vendute Ubi e Banco, bene Astaldi, vola Vittoria Assicurazioni

Redazione ANSA
MILANO
16 maggio 2018
09:42
NEWS

(ANSA) - MILANO, 16 MAG - Piazza Affari fanalino di coda in Europa, con il Ftse Mib che cede lo 0,47% mentre gli altri listini si mantengono poco sopra la parità. Anche i Btp vanno peggio degli altri bond governativi, con lo spread sulla Germania che si allarga di 7 punti base a 137 e il rendimento del decennale che si mantiene a ridosso del 2%.

I mercati reagiscono con qualche segnale di preoccupazione ma senza isterismi alla bozza del contratto Lega-M5s che prevedeva un'uscita dell'Italia dall'euro, bollata subito come superata, mentre le due forze politiche continuano a lavorare alla formazione di un nuovo governo.

In Borsa faticano Ubi (-2,1%) e Banco Bpm (-1,9%), seguite da Prysmian (-1,3%), Enel (-1,3%) e A2A (-1,2%). Positive Saipem (+2,4%), che Bernstein suggerisce di acquistare, e Mediaset (+0,7%), dopo i conti. Bene Astaldi (+3%), che ha annunciato un aumento da 300 milioni e l'ingresso di Ihi, mentre Vittoria Assicurazioni balza del 19,7% e si allinea al prezzo dell'opa di Acutis.

RIPRODUZIONE RISERVATA © Copyright ANSA

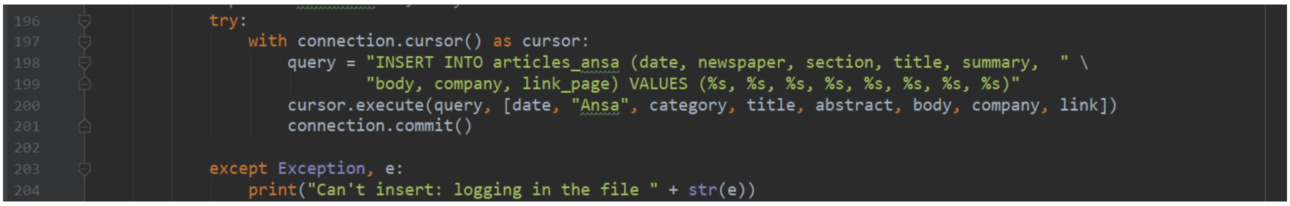
ULTIMA ORA ECONOMIA

14:30 Autogrill: G.Benetton, vendere non piace
13:38 Borsa, Milano torna piatta. Spread a 186
12:46 Assolombarda,crescono fatturati aziende
12:22 Borsa Milano perde slancio, spread a 182
12:00 P.a: studio, in calo tempi pagamenti
11:33 Enav: 'rotte libere' da 9 mia metri
11:04 Inps:19,4 mln ore cig aprile,-15,4% anno
10:38 Ilva: operaio scrive a Mattarella
09:43 Borsa Milano in rialzo, giù lo spread
09:14 Borse europee aprono contrastate

In the article page, we can double check that there is still the same title, a new subtitle and the body of the article itself.
Finally, it is important to note that the "number of results" is important because from that we know how many times we need to iterate in order to retrieve all the articles talking about a particular company.

Query

Once we retrieved all these information, we clean all data we collected from weird html characters, in order to make them using several times the function 'replace(html char, ascii -char)'.
After the clean, we can finally insert them into our database.



```
        try:
            with connection.cursor() as cursor:
                query = "INSERT INTO articles_ansa (date, newspaper, section, title, summary,  " \
                        "body, company, link_page) VALUES (%s, %s, %s, %s, %s, %s, %s, %s)"
                cursor.execute(query, [date, "Ansa", category, title, abstract, body, company, link])
                connection.commit()

        except Exception, e:
            print("Can't insert: logging in the file " + str(e))
```

Figure 4 – Final query

Final remarks on the project

- There are no explicit authors on ANSA due to the fact that it is an agency. From an article page you will always see "*Riproduzione riservata © Copyright ANSA*"
- Some articles, marked "professional" are unavailable. You need to get your credential from ANSA as a journalist in order to read them.
- At the end of the project, we noticed that we saved a "string" date instead of a datetime type. Since it could be useful make some queries on the most recent news, we provided python script to fix the problem. The python script though, needs a user to add a column on the database. The process could be fully automate but we preferred to avoid the script to drop/create columns inside tables.

Conclusions

Thanks to this script, we provided an automated way to collect and store articles from ansa.it. At the end, more then 38 800 articles were registered, with all their corresponding information.