

Test Assignment

The goal is to extract about 5000 property objects from a segment of the site quoka.de:

<http://www.quoka.de/immobilien/bueros-gewerbeflaechen/>

The Result of the work should be:

1. Github repository with all files
2. MySQL dump file with collected data, placed in Github or Dropbox for download
3. Comments on major challenges and decisions

Critical requirements:

- Usage of Python Scrapy framework is obligatory, no Selenium, no BeautifulSoup
- Usage of Linux, MySQL, Git (Github) is obligatory
- Total number of objects in this site segment need to be collected
- All fields marked red below need to be identified, extracted and saved correctly
- The script must be able to collect all data fully automated

You should involve search filters on the site. We need to collect only property offers (Angebote), no search postings (Gesuche). There are two type of offerers: private (Private) and commercial (Gewerbliche). We need to collect both types and mark the property entries accordingly (set field Gewerblich=0 or =1). You should additionally use the City filters, in order to build smaller chunks to paginate.


Use our example files if you like:

<https://drive.google.com/file/d/0BzGpdWXz2Bn-Ti16VHd4VzJ0eEE/view?usp=sharing>

For questions ask Andreas: as@lot-internet.de.

Good luck!

Scrape all property objects from <http://www.quoka.de/immobilien/bueros-gewerbeflaechen/>



Eine Marke von **RUSSMEDIA**

KLEINANZEIGEN

KOSTENLOS INSERIEREN

MEIN QUOKA

Was suchst du?

Büros, Gewert ▾

PLZ oder Ort

+ 25 km ▾

Finden

Kleinanzeigen > Immobilien > Büros, Gewerbeflächen

Meine Suchen | ★ Merkliste (0)

Meine Suche

Büros, Gewerbeflächen

Suchagenten anlegen

Rubriken

Alle Rubriken

Immobilien

Büros, Gewerbeflächen 5.431

Preis eingrenzen

von

-

bis

>

Anbieter

Set both filter: private / Gewerbliche

nur Private 4.929 Gewerblich=0

nur Gewerbliche 502 Gewerblich=1

Angebotstyp

Set only „Angebote“ filter

nur Angebote 5.287

~~nur Gesuche 144~~

Ort

Büros, Gewerbeflächen - 5.431 Anzeigen

neueste Anzeigen ▾

Google Anzeigen

trigo28.de

Büro, Hallen, Freiflächen

Unschlagbar günstig Viel Platz für Ihren Erfolg!

Impressum

Kontakt

Wir Bieten

Übersicht

Anfahrt

News

immonet.de

(3.9)

Mietwohnung

1 bis 6-Zimmerwohnungen in Ihrer Nähe. Sofort online suchen!

Über 1,495 Mio Objekte · Aktuelle Neubauprojekte · Immobilienbewertung

1.786 Personen folgen Immonet auf Google+

Wohnung mieten


Haus kaufen

Wohnung kaufen

Haus mieten

Immobilien mieten

Immobilien kaufen



Heller und ruhiger Kurs- /Seminarraum ...


Wir bieten Ihnen helle, freundliche und ruhige Räume, in denen sich Ihre Teilnehmer wohlfühlen. ...

50,-

D-53111
Bonn Zentrum

TOP

☆



Gepflegte Stahlhalle B20.5xL23xT5.5m ...

Stahlbauhalle inkl. Demontage u. Verladung zu verkaufen. Zwei

VHS

D-14120

TOP

Open every detail page and scrape all relevant data there

- nur Angebote 5.287
- nur Gesuche 144

Ort

Alle Städte

- Berlin 221
- München 76
- Nürnberg 49
- Frankfurt 44
- Hamburg 39
- Köln 39

[weitere Städte](#)

Crawl all cities

 Suchagenten anlegen

 Jetzt kostenlos inserieren!



Gepflegte Stahlhalle B20.5xL23xT5.5m ...

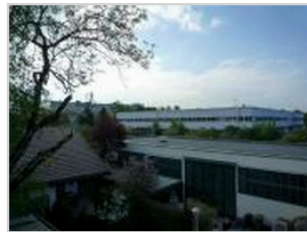
Stahlbauhalle inkl. Demontage u. Verladung zu verkaufen. Zwei freitragende Hauptrahmen ...

VHS

D-14129
Berlin



TOP



Gewerbehalle

Gewerbehalle, 330 qm, ebenerdig einfahrbar, heizbar, für ein Jahr (November 2015 bis Oktober 2016) ...

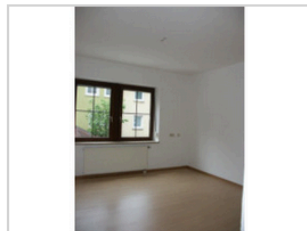
1.500,-

D-80992
München Moosach



TOP

[alle Top-Anzeigen](#)



Erst mal. .. in die Lange Straße- IHR Bü...

Crailsheim, Lange Straße, B, Bj. 1996, KT 2 MM, frei ab: 1.10.2015

600,- EUR bei [Immobilienscout24](#)

600,-

Partner-Anzeige

If you find entries like this, do not open the target page.
Save „Immobilienscout“ as „Anbieter_ID“ and only fields visible here



Exklusives Gewerbeobjekt in Thüringen!! ...

Sie suchen das Besondere, einen gewerblichen Standort in Thüringen? Aus gesundh. Gründen und damit ...

290.000,-

Heute, 11:19 Uhr

D-98693
Ilmenau



Top Gewerbeanwesen in Thüringen! ...

Sie suchen das Besondere, einen gewerblichen Standort in Thüringen? Aus gesundh. Gründen und damit ...

290.000,-

Heute, 11:19 Uhr

D-07318
Saalfeld



Mittendrin statt nur dabei: Top- BÜRO- L...

354.288,-

Partner-Anzeige

Berlin, Wolliner Straße, B, Bj. 1897

354.288,- EUR bei Immobilienscout24

Google Anzeigen



immowelt.de

★★★★★ (4.1)

Kauf Gewerbeimmobilien

Aktuelle Gewerbeimmobilien. Läden, Büros, Lager und mehr!

Kostenfreier Suchauftrag · Immobilien anbieten · ab 24,90 € inkl. MwSt.

12.720 Personen folgen immowelt auf Google+

[Immobilienbewertung](#)

[Immobilien anbieten](#)

[Baufinanzierung](#)

[Suchanzeige aufgeben](#)

[Haus kaufen](#)

[Bausparen](#)



deutsche-wohne...

Wohnangebote in Berlin

Provisionsfreie Wohnungen in vielen Stadtteilen Berlins!

Fritz-Reuter-Allee 46, Berlin



europa-center.de

Gewerbeimmobilien mieten

provisionsfrei - zentral - flexibel Jetzt informieren!

Rudower Chaussee 11, Berlin

Crawl all pages with selected filters (city, gewerblich...)

Seite 1 von 13



1

2

3

4

5

6

7

8

9



Datenschutzinfo

[► Haus Berlin](#)

[► Büroräume](#)

[► Büros](#)

[► Inserieren](#)



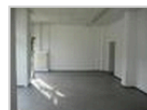
← zurück zur Anzeigenliste **Überschrift**

Anzeigen blättern < >

**Büro, Laden in Sterndamm 50, 12487 Berlin
Schönenweide**

€ **997,30**

Kaufpreis Festpreis



[Nachricht senden](#)

0163/ 3929523

030... [anzeigen](#) **Telefon**

Gewerblicher Inserent

Aktiv seit: 22.03.12

[Alle Anzeigen des Inserenten](#)

☆ In Merkliste speichern

📊 Anzeigenstatistik

🖨️ Anzeige drucken

📍 Auf Karte anzeigen

⚠️ Anzeige melden

🛡️ Sicherheitshinweise

Standort: D-12487 Berlin Johannisthal

Anzeige: 156183596

Datum: 18.08.2015

Klicks: 13

Stadt

Erstellungsdatum

Beschreibung

Das Büro Ladengeschäft befindet sich im Ortsteil Johannisthal gegen über Kino Astra Filmpalast in einem sehr gepflegten Wochbauensemble, das 1930erbaucht wurde. Wir haben letzte komplette Modernisierung Mai 2012 vorgenommen Fußboden ist mit ein Laminat verlegt. Büro ist ab 01.10.2015 frei. Objekt hat

MONSTER

Personalreferent (m/w)
Alcoa Fastening Systems Fairch

Personalreferent m/w
Postbank Systems AG

Personalreferent/-in
Knauf Insulation Operation Gmb

MySQL table format

Field name	Type	Format	Length	Description
1 id	Numerisch	Primärschlüssel	8	increment number
2 Boersen_ID	Numerisch		8	website number (fix =21)
3 OBID	Numerisch		8	Objekt ID of the offer on the website, extract from detail page URL or field „Anzeige:“
4 erzeugt_am	Numerisch	Datum	8	crawling date
5 Anbieter_ID	Numerisch		8	empty or "Immobilienscout24"
6 Anbieter_ObjektID	Alphanumerisch		100	empty
7 Immobilientyp	Alphanumerisch		50	"Büros, Gewerbeflächen"
8 Immobilientyp_detail	Alphanumerisch		200	empty
9 Vermarktungstyp	Alphanumerisch		50	"kaufen"
10 Land	Alphanumerisch		30	"Deutschland"
11 Bundesland	Alphanumerisch		50	empty
12 Bezirk	Alphanumerisch		150	empty
13 Stadt	Alphanumerisch		150	City
14 PLZ	Alphanumerisch		10	ZIP
15 Strasse	Alphanumerisch		100	empty
16 Hausnummer	Alphanumerisch		40	empty
17 Überschrift	Alphanumerisch		500	Title of the offer
18 Beschreibung	Alphanumerisch		15000	Offer description. Please remove all HTML and line breaks, tabs and other control characters like \n etc.
19 Etage	Numerisch		30	empty
20 Kaufpreis	Numerisch	ohne €-Zeichen	8	Preis
21 Kaltmiete	Numerisch	ohne €-Zeichen	8	Empty in this category
22 Warmmiete	Numerisch	ohne €-Zeichen	8	empty
23 Nebenkosten	Numerisch	ohne €-Zeichen	8	empty
24 Zimmeranzahl	Numerisch		8	empty
25 Wohnflaeche	Numerisch	ohne ca. und m²	8	empty
41 Monat	Numerisch		8	current month
42 url	Alphanumerisch		1000	detail page URL
43 Telefon	Numerisch		100	Phone number
44 Erstellungsdatum	Numerisch		8	Creating date on the site
45 Gewerblich	Numerisch		8	Which filter set: Gewerblich = 1, Privat = 0