

Auto Scout24 scraping

Simone Zambetti

Background economico

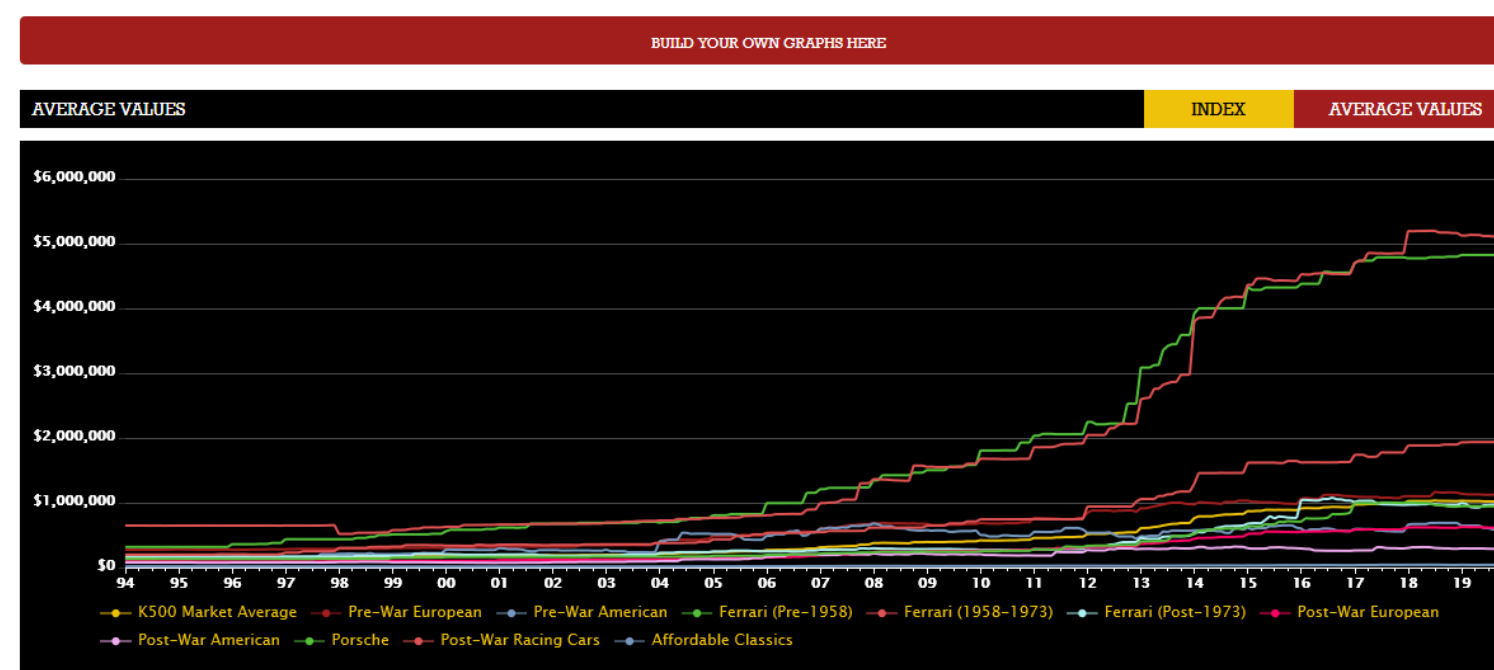
3/7/2020



[Github link](#)

Business target

- Target: scraping degli annunci di autoscout24, il portale leader in italia nel commercio di autovetture nuove o usate
- Focalizzandosi sul mercato emergente dei veicoli elettrici (EV), recuperare tutte le informazioni disponibili sulle pagine degli annunci elaborando anche i singoli annunci
- La base dati risultante può essere eventualmente usata per risolvere le seguenti domande di business, ad esempio:
 - E' vero che i deprezzamento delle auto elettriche inferiore rispetto alle endotermiche? Una domanda particolarmente interessante visto che, intuitivamente, il tasso di evoluzione tecnologico delle EV è molto più veloce oggi rispetto alle ICE (internal combustion engine), dunque si ci aspetterebbe il contrario.
 - Quale è lo share di mercato dei diversi marchi?
 - Quali variabili influenzano il valore di una vettura elettrica?
 - Come stanno influenzando gli ingenti incentivi statali e regionali il pricing delle vetture elettriche?
 - Costruire un indice di prezzo di mercato per vari modelli, come sottostante:



Indice di mercato per le vetture classiche k500, più info [qui](#)

Scraping



- Partendo dalla pagina principale riferimento annunci, già filtrata per EV (fuel=E), massimo valori per pagina 20, paese italia, sorted by più recenti. Link di esempio: <https://www.autoscout24.it/lst?sort=age&desc=1&fuel=E&ustate=N%2CU&size=20&page=1&cy=I> (solo le prime 20 pagine e annunci correlati sono stati scaricati per ridurre cardinalità)
- Utilizzando **colab**, tutte le informazioni nei riquadri verdi sono «scraped» :
 - Dati non strutturati: in alcune annuncio abbiamo caratteristiche/descrizioni che possono essere completamente mancati in altri
 - ci sono anche immagini
 - Alcuni campi sono ripetuti, quindi non è necessario che vengano scaricati anche dalle singole pagine degli annunci

Pagina di ricerca annunci

Singolo annuncio



Scraping (pt. II)

- Output dello scraping:
- 2 file json

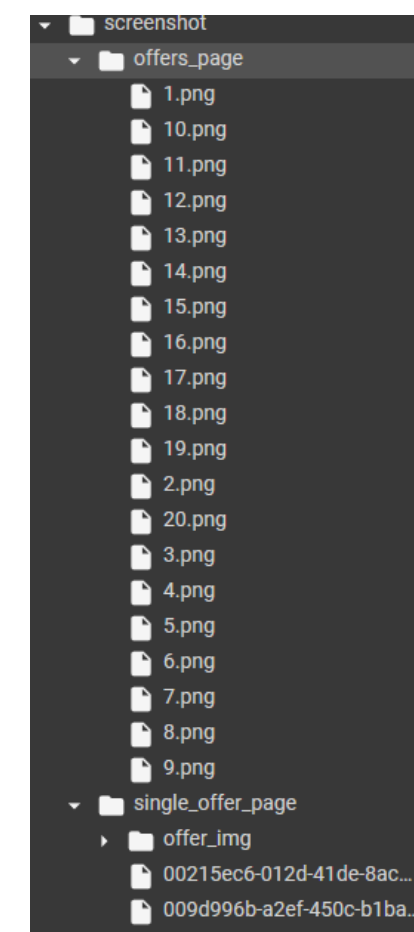
```
Git export_output_scraping.json
[
  {
    "id_annuncio": "084037c1-3fbe-48ce-b5b2-f1a278522e22",
    "link_annuncio": "https://www.autoscout24.it/annunci/smart-fortwo-electric-drive-coupe-elettrica-nero-084037c1-3fbe-48ce-b5b2-f1a278522e22?cldtidx=1&cldtsrc=listPage",
    "vehicle": "smart forTwo",
    "vehicle_user_desc": "electric drive coup\u00e9",
    "asking_price": "\u20ac 9.000,- 12",
    "vehicle_data": [
      "28.015 km",
      "02/2014",
      "35 kW (48 CV)",
      "Usato",
      "1 proprietario",
      "Automatico",
      "Elettrica",
      "-/- (l/100 km)",
      "0 g CO2/km (comb.) 4"
    ],
    "seller": "Unique Car Srl",
    "country": "IT",
    "address": "00166 Roma - Rm"
  },
  {
    "id_annuncio": "9dc886a9-480b-d914-e053-0100007f3eeb",
    "link_annuncio": "https://www.autoscout24.it/annunci/smart-fortwo-eq-edition-one-22kw-elettrica-grigio-9dc886a9-480b-d914-e053-0100007f3eeb?cldtidx=2&cldtsrc=listPage",

```

```
Git export_output_single_offers_scraping.json
[
  {
    "id_annuncio": "084037c1-3fbe-48ce-b5b2-f1a278522e22",
    "user_description": "vettura usata ottimo stato aziendale",
    "equipment": [
      "Alzacristalli elettrici",
      "Climatizzatore",
      "Autoradio",
      "Cerchi in lega",
      "ABS",
      "Airbag conducente",
      "Airbag laterali",
      "Airbag passeggero",
      "Chiusura centralizzata",
      "Controllo automatico trazione",
      "ESP",
      "Fendinebbia",
      "Immobilizzatore elettronico",
      "Servosterzo"
    ],
    "specs": {
      "Tipo di veicolo": "Usato",
      "Per neopatentati": "S\u00ec",
      "Proprietari": "1",
      "Usato Garantito": "12 mese",
      "Tagliandi certificati": "S\u00ec",
      "Veicolo per non fumatori": "S\u00ec",
      "Marca": "smart",
      "Modello": "forTwo",

```

- Immagini annunci
+ screenshot delle pagine:



Extraction



- I json sono caricati tramite shell su MongoDB, nel database autoscout_scraping, creando 2 collections con il prefix ds_*

```
#!/bin/bash
cd /media/sf_Condivisa #go to mounted share drive of master if running on virtual machine
sudo service mongod start
mongo --eval "use autoscout_scraping" #crea database

#import 1mo json in mongo
mongoimport --db autoscout_scraping --collection ds_main_offers_page --file export_output_scraping.json --jsonArray
mongo --eval "db.sc_main_offers_page.find({})" autoscout_scraping

#import 2ndo json in mongo
mongoimport --db autoscout_scraping --collection ds_single_offers_page --file export_output_single_offers_scraping.json --jsonArray
mongo --eval "db.sc_single_offers_page.find({})" autoscout_scraping
#the i use MongoDB compass to export to csv to process with OpenRefine as it is lengthy to specify all fields in bash
```

- MongoDB è stato montato su una VPS debian hostata su aruba.it
- Una volta caricate, le collections sono esportate tramite Mongo Compass in *.csv poichè possano essere letti correttamente da OpenRefine, visto che non riesce a leggere .json nel formato corretto

Transformation (pt. I)



- Trasformazione su OpenRefine dei csv esportati:

OpenRefine ds_main_offers_page csv Permalink

Facet / Filter Undo / Redo 29 / 29

330 rows

Show as: rows records Show: 5 10 25 50 rows

Extensions: Wikidata

« first < previous 1 - 10 next > last »

Filter:

12. Create new column year based on column country by filling 330 rows with
jython: return value[-4:].replace("nno"), "-")

13. Rename column Column 13 to transmission

14. Rename column Column 14 to fuel

15. Text transform on 320 cells in column previous_owners: value.toNumber()

16. Rename column Column 15 to fuel_consumption

17. Text transform on 88 cells in column Column 16: value.replace(" g CO2/km (comb.) 4\"")

18. Rename column Column 16 to emissions_gCO2/km

19. Remove column Column 21

20. Remove column Column 22

21. Remove column Column 23

22. Remove column Column 24

23. Remove column Column 29

24. Remove column Column 25

25. Remove column Column 26

26. Remove column Column 27

27. Remove column Column 28

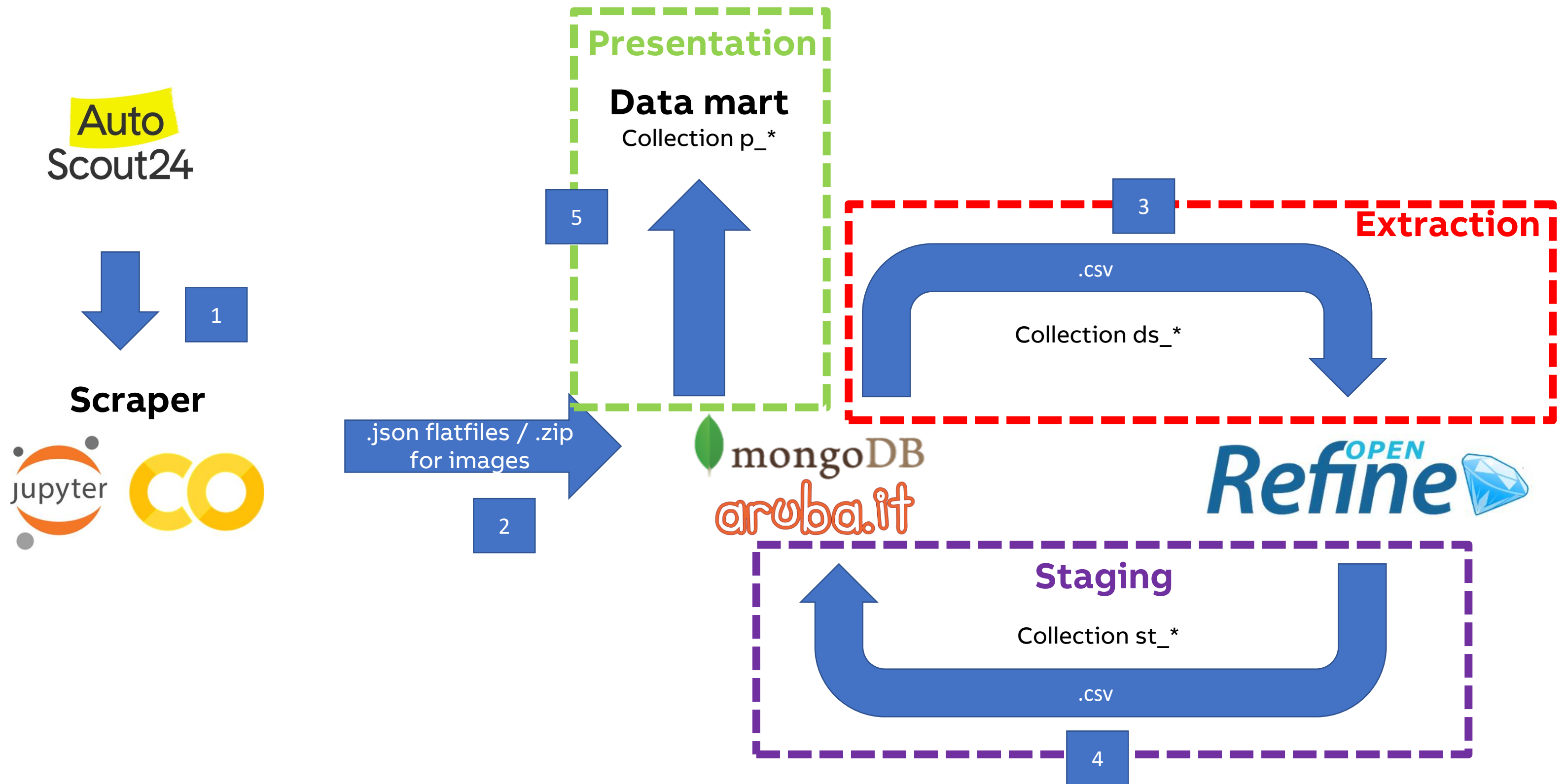
28. Rename column Column 17 to seller

29. Rename column country to registration

		<input type="checkbox"/> All	<input type="checkbox"/> _id	<input type="checkbox"/> id_annuncio	<input type="checkbox"/> link_annuncio	<input type="checkbox"/> vehicle	<input type="checkbox"/> vehicle_user_de	<input type="checkbox"/> asking_price	<input type="checkbox"/> km	<input type="checkbox"/> registration	<input type="checkbox"/> year	<input type="checkbox"/> address	<input type="checkbox"/> status	<input type="checkbox"/> previous_owner	<input type="checkbox"/> transmission	
☆	1.			Objectld("5eff3b6c156a97c639b6f7b3")	084037c1-3f8e-48ce-b5b2-f1a278522e22	https://www.autoscout24.it/annunci/smart-fortwo-electric-drive-coupe-elettrica-nero-084037c1-3f8e-48ce-b5b2-f1a278522e22?cldtid=1&cldtsrc=listPage	smart forTwo	electric drive coupé	€ 9.000	28015	02/2014	2014	48	Usato	1	Automatico
☆	2.			Objectld("5eff3b6c156a97c639b6f7b4")	9dc886a9-480b-d914-e053-0100007f3eeb	https://www.autoscout24.it/annunci/smart-fortwo-eq-edition-one-22kw-elettrica-grigio-9dc886a9-480b-d914-e053-0100007f3eeb?cldtid=2&cldtsrc=listPage	smart forTwo	EQ Edition One (22kW)	€ 31.900	10	07/2020	2020	56	Usato	0	-/- (Tipo di cambio)
☆	3.			Objectld("5eff3b6c156a97c639b6f7b5")	36179b32-ce8a-490a-9ac0-4135dc143a5c	https://www.autoscout24.it/annunci/volkswagen-up-elettrica-grigio-36179b32-ce8a-490a-9ac0-4135dc143a5c?cldtid=3&cldtsrc=listPage	Volkswagen up!		€ 12.600	35000	07/2015	2015	82	Usato	0	Automatico
☆	4.			Objectld("5eff3b6c156a97c639b6f7b6")	7b2985a9-bc13-3b49-e053-0100007fcf0b	https://www.autoscout24.it/annunci/reault-twizy-80-technic-elettrica-bianco-7b2985a9-bc13-3b49-e053-0100007fcf0b?cldtid=5&cldtsrc=listPage	Renault Twizy	80 Technic	€ 5.500	27389	06/2012	2012	11	Usato	0	-/- (Tipo di cambio)
☆	5.			Objectld("5eff3b6c156a97c639b6f7b7")	2587ee6d-6d0e-48a2-86f7-ca85503b369c	https://www.autoscout24.it/annunci/smart-fortwo-electric-drive-passion-elettrica-argento-2587ee6d-6d0e-48a2-86f7-ca85503b369c?cldtid=4&cldtsrc=listPage	smart forTwo	electric drive Passion	€ 14.890	22340	01/2018	2018	56	Aziendale	0	Automatico
☆	6.			Objectld("5eff3b6c156a97c639b6f7b8")	ce0193ba-0632-4c4b-b369-d40ebdd8f48a	https://www.autoscout24.it/annunci/peugeot-ion-active-elettrica-bianco-ce0193ba-0632-4c4b-b369-d40ebdd8f48a?cldtid=7&cldtsrc=listPage	Peugeot iOn	Active	€ 13.500	9550	10/2017	2017	48	Usato	0	Automatico
☆	7.			Objectld("5eff3b6c156a97c639b6f7b9")	9074bb95-8f77-48a6-875f-e362562b71a7	https://www.autoscout24.it/annunci/smart-forfour-eq-electric-drive-prime-elettrica-bianco-9074bb95-8f77-48a6-875f-e362562b71a7?cldtid=8&cldtsrc=listPage	smart forFour	EQ Electric Drive Prime	€ 16.900	16985	12/2018	2018	56	Aziendale	0	Automatico
☆	8.			Objectld("5eff3b6c156a97c639b6f7ba")	04ad05b0-97eb-4732-a7d5-3ee56a2da323	https://www.autoscout24.it/annunci/volkswagen-up-e-5p-my20-elettrica-grigio-04ad05b0-97eb-4732-a7d5-3ee56a2da323?cldtid=9&cldtsrc=listPage	Volkswagen up!	e- 5p my20	€ 17.800	1	01/2020	2020	82	Usato	1	Automatico
☆	9.			Objectld("5eff3b6c156a97c639b6f7bb")	be59766e-2dac-40d0-9072-59782d6952ab	https://www.autoscout24.it/annunci/audi-e-tron-spb-55-quattro-s-line-edition-elettrica-be59766e-2dac-40d0-9072-59782d6952ab?cldtid=10&cldtsrc=listPage	Audi e-tron	SPB 55 quattro S line edition	€ 118.860	0	-/- (Anno)	-	41	Nuovo	0	Automatico
☆	10.			Objectld("5eff3b6c156a97c639b6f7bc")	c7861535-b64b-45a2-8f74-6d157046e451	https://www.autoscout24.it/annunci/nissan-leaf-i-leaf-tekna-30kw-109cv-my17-elettrica-bianco-c7861535-b64b-45a2-8f74-6d157046e451?cldtid=11&cldtsrc=listPage	Nissan Leaf	I leaf Tekna 30kW 109cv my17	€ 17.900	12000	06/2017	2017	0	Usato	0	Automatico

- Ad esempio, i seguenti problem sono stati risolti:
 - Annunci con id duplicate con facet (sono tutti singoli)
 - Replacement heading e cleanup dei valori in alcune colonne, ad esempio classe emissioni, campo proprietary conversion a numerico
 - Alcuni annunci non sono auto elettriche (errori di input da parte degli utenti), per risolvere questo problema le seguenti logiche sono stati applicate:
 - Filtrati solo gli annunci di veicoli immatricolati dopo il 2005
 - Esclusi annunci che indicavano emissioni > 0
 - Escluse alcune marche che al momento non possiedono veicoli elettrici puri (e.g. fiat)
- Una volta ultimati questi cambiamenti, i 2 file .csv sono stati esportati su MongoDB nuovamente, su collections con prefisso st_*

Overall architecture



Migliorie e criticità

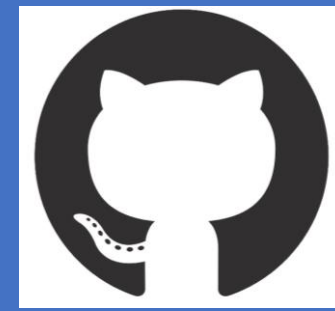
- Possibili migliorie:

- Automatizzare il lancio dello scraper ogni giorno, possibilmente passandolo alla VPS di aruba da Colab
- Automatizzare in python il passaggio di json -> csv e il processing di OpenRefine
- Usare le restful API di autoscout, chiedendo account developer per avere accesso a tutte le pagine (per il momento lo scraper scarica solo le prime 20 pagine, un workaround sarebbe scaricare solo gli annunci pubblicati in giornata, per cui autoscout ha un filtro)
- Utilizzare tutte le immagini delle annunci scaricati

- Criticità:

- La struttura dell'output dello scraper andrebbe cambiato in modo tale da non convertire in csv i file in json
- Anche se gli id annunci sono diversi, al momento non vi è un modo per capire 2 o più annunci siano effettivamente duplicati (ad esempio: carico la stessa offerta in molti annunci per avere migliore visibilità)

Github



- Tutto il materiale è disponibile sulla repo:

[https://github.com/szambetti/Data-science-Master-UNIMIB/tree/master/Module%204%20-%20Big%20Data%20%26%20Analytics/Home work](https://github.com/szambetti/Data-science-Master-UNIMIB/tree/master/Module%204%20-%20Big%20Data%20%26%20Analytics/Home%20work)