Data management for informed decision-making:

analyzing Milan's rental property market with integrated news data

Matteo Altieri 897338
Enrico Mannarino 850859
Christian Persico 829558

OBIETTIVO

Realizzazione di un database integrato di annunci di affitti di appartamenti a Milano (Subito.it e Immobiliare.it), che risulti utile e informativo per gli utenti che si affacciano al mercato immobiliare della città.

Arricchimento con titoli e link di notizie recenti dal giornale MilanoToday (tramite Google News) per avere un quadro degli ultimi aggiornamenti in merito a cronaca e possibili eventi interessanti che citano la via dell'immobile.

KEYWORDS: Immobiliare.it, Subito.it, Openalfa.it, Google News - MilanoToday, NoSQL, MongoDB, Web scraping, API, PyMongo

INDICE

01

DATA ACQUISITION E CLEANING

FONTI

Subito.it

Tecniche di **web scraping** tramite BeautifulSoup per acquisire gli annunci di affitti di appartamenti.

Immobiliare.it

Tecniche di estrazione dei dati tramite **API** per gli annunci di affitti di appartamenti.

Openalfa.it - Stradario d'Italia

Tecniche di **web scraping** tramite BeautifulSoup per l'estrazione delle vie di Milano e dei rispettivi quartieri di appartenenza.

I dati sono aggiornati al 5 giugno 2023

Openalfa.it - Stradario d'Italia

Problema

Immobiliare.it e Subito.it hanno distinte suddivisioni in quartieri: se volessimo fare una ricerca per un appartamento in zona Arco della Pace, su Subito avremmo la voce "Vercelli, Fiera, Sempione" mentre su Immobiliare "Arco della Pace, Arena, Pagano".

Soluzione

Acquisizione dati dalla fonte esterna Openalfa, le cui informazioni sono ottenute e mantenute aggiornate dai dati liberamente disponibili su OpenStreetMap.

Limitazioni

Il codice può portare ad assegnazioni errate di strade al confine tra quartieri. Ad esempio, Via Padova, solitamente nel quartiere Padova, è stata associata al quartiere Parco Lambro - Cimiano.

Subito.it

WEB SCRAPING

1123 inserzioni

10 variabili:

1. Titolo dell'annuncio
2. Prezzo mensile (€)
3. Dimensione dell'appartamento (m^2)
4. Piano
5. Numero di locali
6. Numero di bagni
7. Descrizione dell'annuncio
8. URL
9. Proprietario dell'annuncio (privato o agenzia)
10. Indirizzo dell'agenzia immobiliare, se presente

Dopo l'applicazione dell'algoritmo e dell'arricchimento con i dati dello stradario

11. Indirizzo
12. Quartiere

PROBLEMA: presenza di annunci fake.

SOLUZIONE: come filtro vengono considerati a monte solo gli annunci con la thumbnail.

Immobiliare.it

API

2000 inserzioni

16 variabili

Tutte quelle di Subito in comune, eccetto la via dell'agenzia. In aggiunta:

- Numero di camere da letto
- Foto dell'appartamento
- Classe energetica
- Presenza dell'ascensore
- Presenza di optional (aria condizionata, ecc.)
- Condizioni dell'immobile
- Link alla pagina web dell'agenzia, se presente

Dopo l'applicazione dell'algoritmo e dell'arricchimento con i dati dello stradario

- Indirizzo
- Quartiere

PRIMA PULIZIA DEI DATI

DUPLICATI

Annunci ripetuti per avere più visibilità

Per poterli identificare:

1. Selezioniamo le osservazioni che abbiano in comune i valori per le variabili attinenti a: prezzo, dimensione, piano, quartiere, numero di bagni e locali, se è presente l'agenzia e il suo eventuale indirizzo (solo per Subito).

2. Di queste confrontiamo poi anche le descrizioni degli annunci, applicando un metodo basato sulla corrispondenza delle parole (non ripetute) in comune, scegliendo come soglia il 90%.

PRIMA PULIZIA DEI DATI

OUTLIERS

- Annunci di affitti di breve periodo
- Errori di imputazione
- Inserimento dell'annuncio nella sezione sbagliata (vendita invece che affitto)

Prezzo mensile

- Limite minimo imposto: **€ 300**
- Limite massimo imposto: **€ 10.000**

Dimensione

- Limite minimo imposto: **28 m²** (*articolo 96 del Regolamento Edilizio del Comune di Milano*)

SUBITO.IT

DUPLICATI	OUTLIERS
52 (4.6%)	**78** (6.9%)

IMMOBILIARE.IT

DUPLICATI	OUTLIERS
121 (6%)	**119** (6%)

Algoritmo di estrazione delle vie: i problemi da affrontare

La specifica via in cui si trova un immobile in affitto non è fornita direttamente dai dati, ma è presente all'interno delle descrizioni o dei titoli. L'algoritmo confronta le strade acquisite dallo stradario con i testi degli annunci.

Casi problematici:

- Nelle descrizioni vengono spesso citate zone di interesse e strade famose per dare al lettore un'idea di dove si trovi la casa.
- Capita che la via dell'agenzia sia una delle vie presenti nel testo, a volte è anche l'unica.
- Le vie nelle descrizioni difficilmente vengono riportate per intero. Ad esempio 'Via Giovanni Pacini' spesso compare come 'Via Pacini'.

Stradario Testi Annunci

Via Giovanni Pacini ➡ Via Pacini

Via Pacini ➡ Via Giovanni Pacini

Dataset

STR. - STAND.

- Lowercase
- Punteggiatura
- Accenti
- Numeri
- StopWords ITA

BOW

- Adatta le vie a come sono scritte nel testo
- Quality Check: Stringhe con più di una parola con un “prefisso” all’interno

FUZZY WUZZY

- Assegna ad ogni via un indice di similarità
- Assegna la via migliore

RIPRESA VIA ORIGINALE

- Scarta la via dell’agenzia
- Viene presa la via con l’indice di similarità più alta
- Se ci sono più vie con indice massimo, sceglie quella con più parole

Algoritmo di estrazione delle vie: vantaggi e svantaggi

Vantaggi:

- Esclude le zone della città che non sono indirizzi come le diciture “Navigli”, “Duomo”, “Garibaldi” o “Porta Romana”.
- Adatta la via al linguaggio comune.

Svantaggi:

- Se il nome dell'agenzia non è presente nella variabile "*Agency_address*" ma è invece scritto nella descrizione, potrebbe essere inserito come indirizzo dell'immobile.
- Potrebbe non identificare la via corretta nel caso in cui più vie sono considerate “migliore via”.

02

DATA INTEGRATION ED ENRICHMENT

RECORD LINKAGE

Cross-posting

PROCEDURA DI RIMOZIONE

La procedura di identificazione e rimozione è la stessa della precedente:

1. Considerazione delle variabili in comune rilevanti.
2. Confronto delle descrizioni (soglia al 90% delle parole comuni non ripetute).

Numero di annunci **prima** della rimozione dei duplicati	Numero di annunci **dopo** la rimozione dei duplicati
2732	**2697**

ESPLORIAMO MILANO CON UN GIORNALE

QUALITÀ DELLA VITA

Il nostro obiettivo è comprendere meglio la situazione di una via ed il relativo quartiere, valutandone la vivibilità in modo indiretto attraverso le notizie di cronaca ed eventi ambientate nei suoi pressi. Ottenendo un'idea generale della sicurezza, dell'ambiente sociale e di altri fattori rilevanti che possono influenzare la qualità della vita di un nuovo residente della città.

ARRICCHIMENTO CON NOTIZIE

COME ABBIAMO FATTO?

Per raggiungere il nostro obiettivo, utilizziamo il web scraping, una tecnica automatizzata che ci consente di estrarre informazioni dai risultati di ricerca di **Google News**. In particolare, ci concentriamo sui titoli e sugli URL degli articoli della testata giornalistica **MilanoToday** che citano l'indirizzo di ogni annuncio del nostro dataset integrato. Questi dati vengono poi raccolti e organizzati in una nuova variabile chiamata "*News*".

WEB SCRAPING

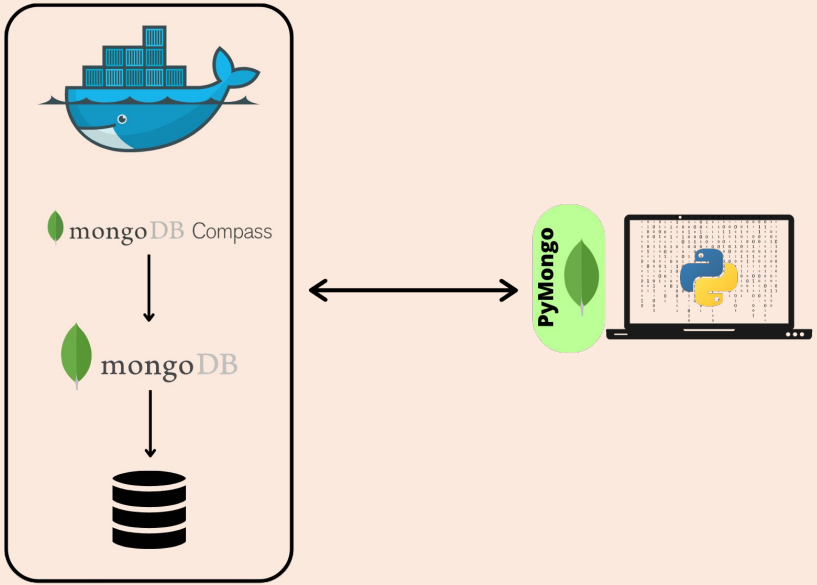
03

DATA STORAGE

DATA STORAGE

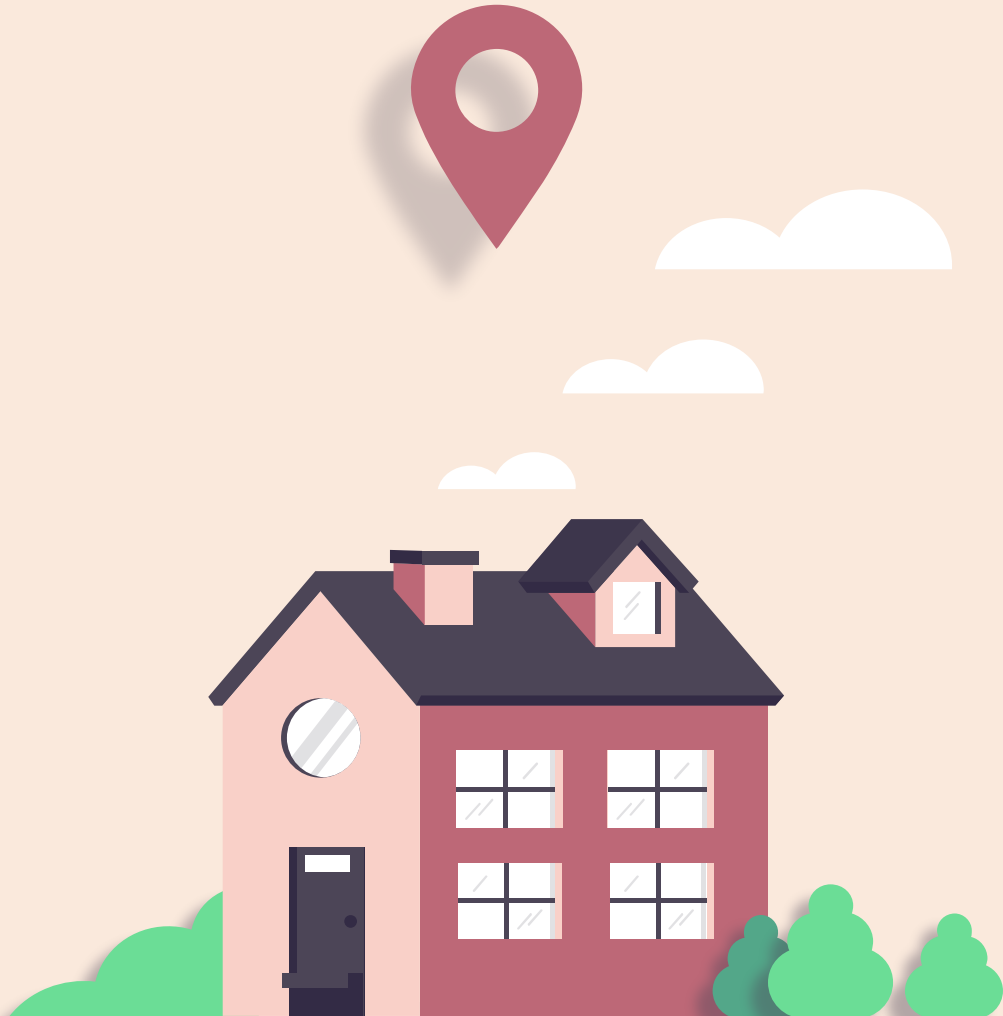
PROCEDIMENTO

Il database integrato viene immagazzinato in un server tramite **MongoDB**, un DBMS non relazionale orientato ai documenti. Il server è creato all'interno di un container in locale tramite il software Docker. L'importazione è effettuata tramite il tool MongoDB Compass. Di fatto, si ha a disposizione una collezione con un totale di 2697 documenti diversi, uno per ogni annuncio, disponibili per analisi future.



04

QUERIES

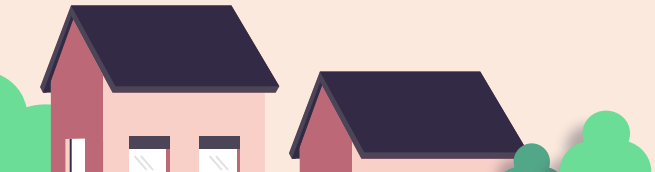


QUERIES

1. Pipeline per il calcolo del prezzo medio mensile d'affitto su metro quadrato per ogni quartiere.

2. Pipeline per il calcolo del numero totale di annunci di appartamenti in affitto per quartiere.

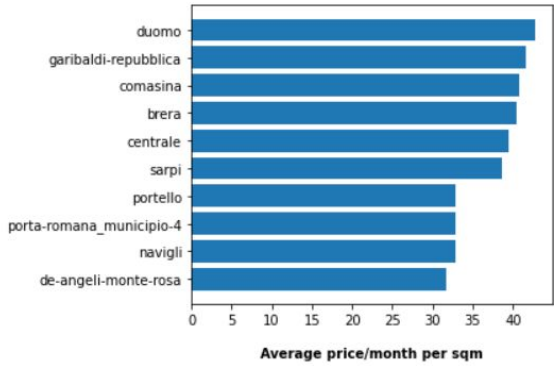
3. Pipeline per ricavare, per ogni quartiere, il totale di articoli scritti nell'ultimo anno da MilanoToday che citino le rispettive vie presenti nel database.



Pipeline per il calcolo del prezzo medio mensile d'affitto su metro quadrato per ogni quartiere.

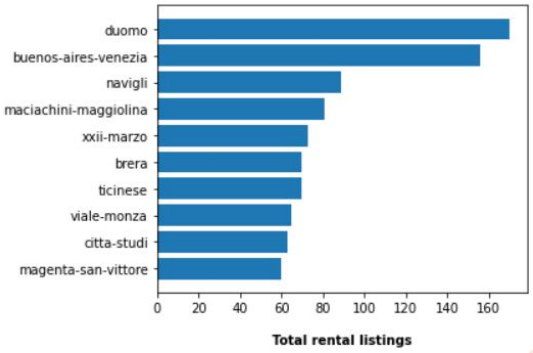
```
pipeline1 = [
    {'$match': {
        "Price/month": {"$exists": True, "$ne": np.nan},
        "Size (m^2)": {"$exists": True, "$ne": np.nan}}},
    {'$group': {
        '_id': '$District',
        'Avg_price_per_sqm': {'$avg': {'$divide': ['$Price/month',
                                                   '$Size (m^2)']}}}},
    {"$project": {
        "_id": 0,
        "District": "$_id",
        "Avg_price_per_sqm": 1}},
    {"$sort": {"Avg_price_per_sqm": -1}}
]
```

	Avg_price_per_sqm	District
0	42.7	duomo
1	41.6	garibaldi-repubblica
2	40.7	comasina
3	40.5	brera
4	39.5	centrale
5	38.7	sarpi
6	32.9	portello
7	32.9	porta-romana_municipio-4
8	32.8	navigli
9	31.7	de-angeli-monte-rosa



Pipeline per il calcolo del numero totale di annunci di appartamenti in affitto per quartiere.

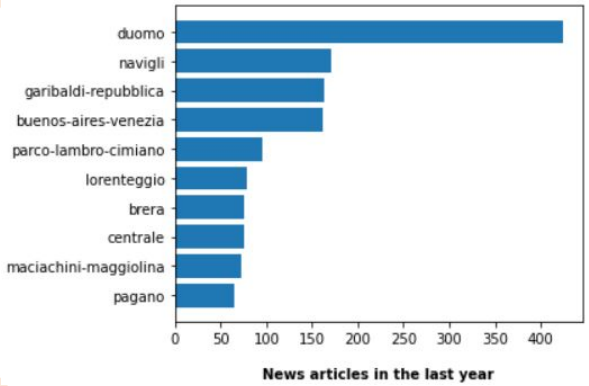
```
pipeline2 = [
    {'$match': {
        "District": {"$exists": True, "$ne": ""}
        }},
    {'$group': {
        '_id': '$District',
        "Listings": {"$sum": 1}
        }},
    {"$sort": {"Listings": -1}},
    {"$project": {
        "_id": 0,
        "District": "$_id",
        "Listings": 1
        }}
]
```



Stradario di Duomo, Municipio 1

	Listings	District
0	170	duomo
1	156	buenos-aires-venezia
2	89	navigli
3	81	maciachini-maggiolina
4	73	xxii-marzo
5	70	brera
6	70	ticinese
7	65	viale-monza
8	63	citta-studi
9	60	magenta-san-vittore

Pipeline per ricavare, per ogni quartiere, il totale di articoli scritti nell'ultimo anno da MilanoToday che citino le rispettive vie presenti nel database.



	Total_News	District
0	424	duomo
1	171	navigli
2	164	garibaldi-repubblica
3	162	buenos-aires-venezia
4	96	parco-lambro-cimiano
5	79	lorenteggio
6	76	brera
7	75	centrale
8	73	maciachini-maggiolina
9	65	pagano

```
pipeline3 = [
    {'$match': {
        "District": {"$exists": True, "$ne": np.nan}}},
    {'$group': {
            '_id': '$District','news': {'$push': '$News'}}},
    {'$match': {
            'news': {
                '$not': {
                    '$elemMatch': {
                        'title': '',
                        'URL': ''}}}}},
    {'$project': {
            'news': {'$reduce': {
                        'input': '$news',
                        'initialValue': [],
                        'in': {
                            '$setUnion': ['$$value', '$$this']}}}}},
    {'$addFields': {'Total_News': {'$size': '$news'}}},
    {"$sort": {"Total_News": -1}},
    {"$project": {"_id": 0,"District": "$_id","Total_News": 1}}
]
```

05

DATA QUALITY

Data Quality: Openalfa

Consistenza e Coerenza:

E' necessario analizzare la copertura delle strade della città e dei quartieri. Per i quartieri è stato effettuato un matching di similarità con i nomi degli 88 NIL ufficiali. Per le strade è stato effettuato un confronto della numerosità tra il dataset e la lista di tutte le vie di Milano.

I quartieri mancanti sono 'Ronchetto delle Rane' e 'Giambellino'

Numero di quartieri con nome simile ad un NIL e totale quartieri	Numero di vie nello stradario contro numero di vie nel viario
86/88 96	4527/4383

Abbiamo più vie, ma ciò è dovuto al fatto che è uno stradario e non un viario

Data Quality: estrazione delle vie

Accuratezza:

In questo caso l'accuratezza semantica delle vie estratte è stata svolta controllando un campione casuale di 100 annunci di affitti.

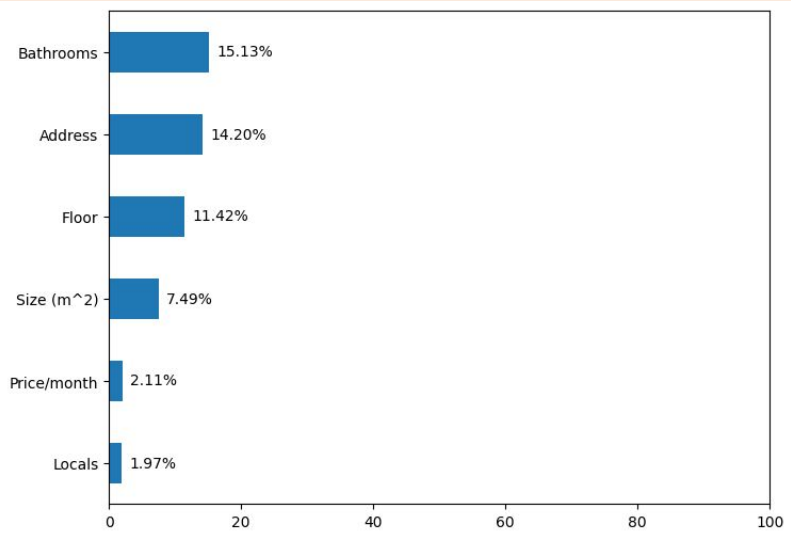
Percentuale di vie assegnate correttamente nel campione casuale di 100 annunci:
87%

- Chi ha postato l'annuncio ha scritto male l'indirizzo (es. Viale Padova)
- Non è specificato l'indirizzo dell'agenzia
- Ci sono più "vie migliori"

Data Quality: valori mancanti

Completezza:

Sono state selezionate le variabili riportate nel grafico perché sappiamo già che tutte le osservazioni considerate sono provviste di titolo, descrizione, url e via dicendo. Inoltre, alcune variabili non sono in comune tra le due fonti, pertanto non avrebbe senso considerarle.



06

CONCLUSIONI E SVILUPPI FUTURI

Conclusioni

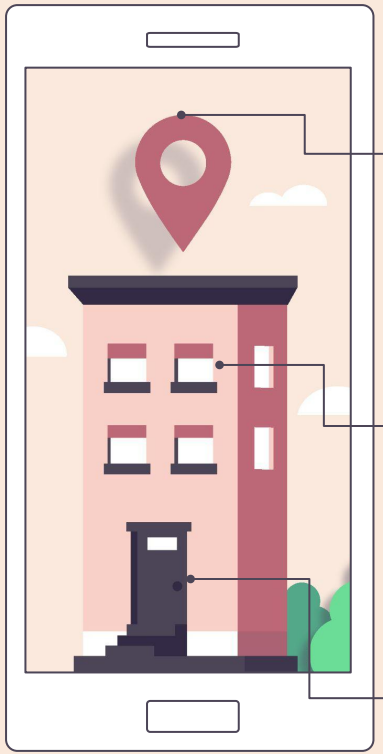
Obiettivo raggiunto:

- Realizzare un database unico e informativo per la raccolta di annunci di affitti da fonti diverse.

Punti di forza:

- Considerare sia la parte di offerta proveniente da agenzie immobiliari che da privati, permettendo una visione più ampia e completa del mercato di affitti di appartamenti nella città di Milano.
- Un utente ha a disposizione le notizie in merito a cronaca ed eventi recenti inerenti alla posizione di ogni singolo immobile.

Sviluppi futuri



Possibilità di miglioramento:

1. Implementazione di un processo di aggiornamento automatico e periodico.

2. Implementazione di metodi avanzati di Natural Language Processing per consentire l'esclusione a priori di annunci che domandano una casa in affitto invece di offrirla.

3. Tali metodologie risolverebbero gli errori e renderebbero impeccabile anche il processo di estrazione delle vie da descrizioni e titoli degli annunci, laddove presenti.

GRAZIE

Matteo Altieri
Enrico Mannarino
Christian Persico