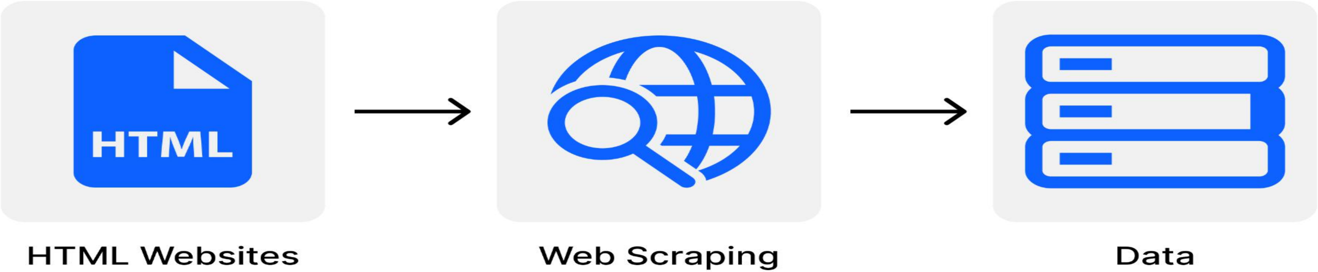
Real Estate Web Scraping

Team Tensorflow:
Kivanc, Quentin, Saina, Vincent

Process



- Data source: IMMOWEB
- Collecting URLs
- Scraping & Extracting information
- Creating dataset & Storing data

Task Distribution

The first day:
Each team member tried to implement small parts of the task

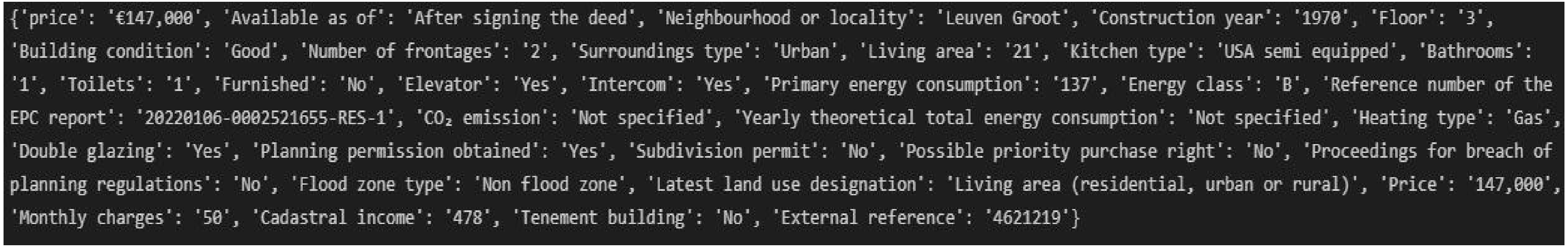
The other three days:

- Program for scraping pages: Quentin Saina Vincent
- Program for fetching web addresses: Quentin
- Program to create the csv file: Kivanc
- Finalizing codes: Kivanc Vincent
- Preparing slides: Quentin Saina

Challenges

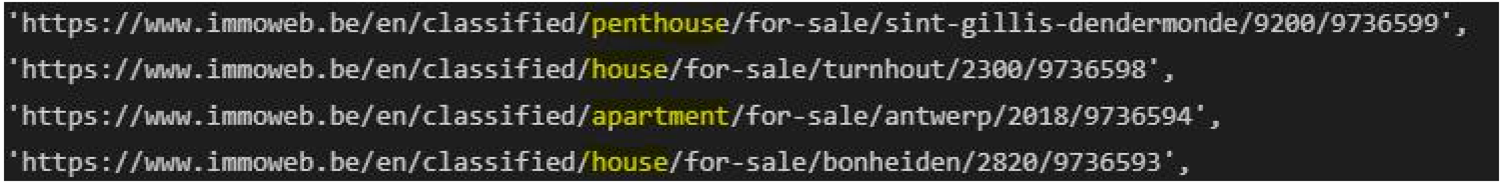
Web Scraping:

- Messy scraped data



```
{'price': '€147,000', 'Available as of': 'After signing the deed', 'Neighbourhood or locality': 'Leuven Groot', 'Construction year': '1970', 'Floor': '3',
'Building condition': 'Good', 'Number of frontages': '2', 'Surroundings type': 'Urban', 'Living area': '21', 'Kitchen type': 'USA semi equipped', 'Bathrooms':
'1', 'Toilets': '1', 'Furnished': 'No', 'Elevator': 'Yes', 'Intercom': 'Yes', 'Primary energy consumption': '137', 'Energy class': 'B', 'Reference number of the
EPC report': '20220106-0002521655-RES-1', 'CO₂ emission': 'Not specified', 'Yearly theoretical total energy consumption': 'Not specified', 'Heating type': 'Gas',
'Double glazing': 'Yes', 'Planning permission obtained': 'Yes', 'Subdivision permit': 'No', 'Possible priority purchase right': 'No', 'Proceedings for breach of
planning regulations': 'No', 'Flood zone type': 'Non flood zone', 'Latest land use designation': 'Living area (residential, urban or rural)', 'Price': '147,000',
'Monthly charges': '50', 'Cadastral income': '478', 'Tenement building': 'No', 'External reference': '4621219'}
```

- Missing information in webpage



```
'https://www.immoweb.be/en/classified/penthouse/for-sale/sint-gillis-dendermonde/9200/9736599',
'https://www.immoweb.be/en/classified/house/for-sale/turnhout/2300/9736598',
'https://www.immoweb.be/en/classified/apartment/for-sale/antwerp/2018/9736594',
'https://www.immoweb.be/en/classified/house/for-sale/bonheiden/2820/9736593',
```

Solution

Project repository link:

https://github.com/VincentPalau/challenge-collecting-data