Data Wrangling - Session 1

Web Scraping and Pandas

Content

1. Introduction
2. Web Scraping
3. Pandas
4. Hands-On:

Introduction

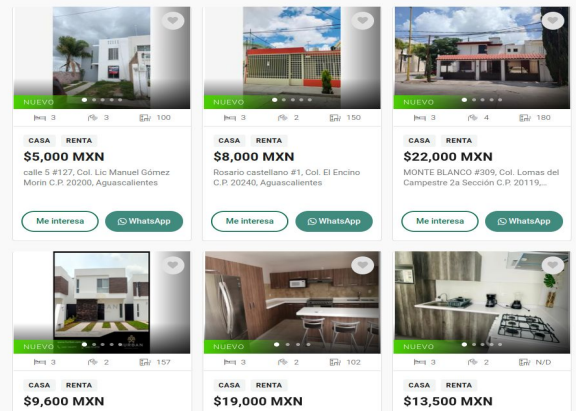
Data wrangling means the involved procedures and operations to create a clean dataset in order to perform analysis or model training. For example:

- Extract information from original (‘wild’) sources such as: text, webpages, images, recordings, etc.
- Extract data from databases (relational or not relacional)
- Join different data sources
- Clean, a data source: remove or replace missing values, remove heterogeneity in data types, etc.

Web Scraping

- Data obtention from web pages can be done via web scraping
- Web scraping uses HTML and DOM (Document Object Model) concepts to extract precise information

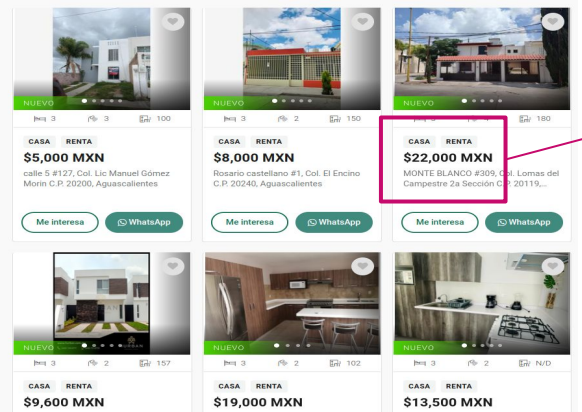
For example, a bot to extract prices from departments:



Web Scraping

- Data obtention from web pages can be done via web scraping
- Web scraping uses HTML and DOM (Document Object Model) concepts to extract precise information

For example, a bot to extract prices from departments:



The bot needs to find clickable elements

Web Scraping

Typical libraries, depending on the complexity of web sites, include:

- Requests
- Urllib
- Beautiful Soup (bs4)
- Selenium

Pandas

- High-level data structures and functions designed to make working with structured or tabular data intuitive and flexible.
- Main objects:
  - DataFrame, a tabular, column-oriented data structure with both row and column labels
  - Series, a one-dimensional labeled array object
- Extra capabilities:
  - Includes relational operations as in SQL
  - Time-series manipulation
  - Flexible handling of missing data

Hands-On