Interaktivní aplikace vizualizující scrapovaná data z realitní stránky 'reality.idnes.cz'

Markéta Minářová
ČVUT–FIT
minarma5@fit.cvut.cz

19. prosince 2021

1 Úvod

Tento report byl vytvořen za účelem informovat o semestrální práci v rámci předmětu BI-PYT, která byla vytvořena v zimním semestru B211.

Semestrální práce se zabývá web scrapingem, což je technika, pomocí které můžeme strojově číst obsah webových stránek na internetu. To se hodí zejména, pokud jsou data dostupná pouze v HTML dokumentech a nikde jinde (nebo alespoň ne veřejně k dispozici).

Cílem této práce je vytvořit program, který bude extrahovat data o nemovitostech z webu a následně je ukládat do .csv souboru. Poté by měla aplikace extrahovaná data vizualizovat a porovnávat a měla by být interaktivní.

2 Web scraping

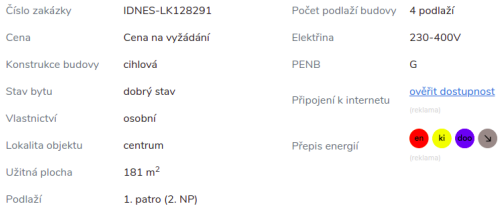
K extrahování dat z webu jsem použila Scrapy, což je open-source framework pro web-crawling napsaný v jazyce Python.

Architektura Scrapy projektu je založená na tzv. pavoucích (spiders), kde každý pavouk představuje jednoho automatizovaného crawlera/bota, který vykonává instrukce dané třídy a stahuje data z webu.

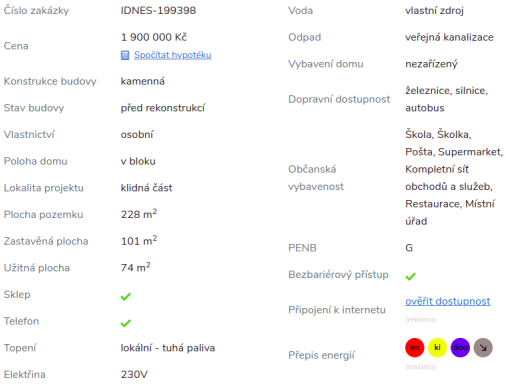
Narazila jsem na problém s ukládáním dat do více souborů. Původní plán byl mít pro každý kraj jeden soubor a podle nich při vizualizaci rozlišovat kraje. Scrapy sice umožňuje nějakou oklikou ukládat data od jednoho spidera do více souborů, ale nepovedlo se mi to, takže jsem nakonec zvolila variantu ukládat vše do jednoho souboru.

Nicméně přišel další problém s odlišením jednotlivých krajů (v tuto chvíli byla všechna data v jednom souboru). Nakonec jsem to vyřešila tak, že response (což je objekt, který dostávají všechny funkce) obdrží meta informace ohledně příslušnosti do kraje.

Další problém byl, že extrahovaná data se často v HTML dokumentu nacházela v tabulce, ve které nebyla pevně daná struktura. Například při extrakci položky 'Užitná plocha' jsem občas získala záznam o 'Lokalita projektu' (viz obrázek 1 a 2).



Obrázek 1: Údaje o nemovitosti *1* - položka 'Užitná plocha' je na 7.řádku tabulky



Obrázek 2: Údaje o nemovitosti *2* - položka 'Užitná plocha' je na 10.řádku tabulky

Tento problém byl vyřešen pouze 'hádáním' místa příslušné položky v tabulce podle jejích nejčastějších výskytů.

3 Dash

Aplikace pro vizualizaci dat byla vytvořena pomocí frameworku Dash, který je opět open-source a slouží ke snadnému vytváření analytických webových aplikací.

Dash aplikace se skládá ze dvou částí - layout a callbacks. Layout je část aplikace, která udává vzhled aplikace a callbacks její interaktivitu, logickou část.

Ve vytvořené aplikaci může uživatel vybírat dva kraje, které chce porovnávat, a dále zda chce nemovitosti na prodej či k pronájmu. První graf zobrazuje jednoduché rozložení cen nemovitostí, aby měl uži-

vatel představu, kde se zhruba pohybují. Druhý graf udává minimální, maximální a střední hodnotu cen nemovitostí za metr čtvereční. Třetí graf vykresluje závislosti ceny na počtu místností, kdy (v rámci implementace) například 3 + 1 představuje 4 místnosti a 3 + kk představuje 3,5 místností. Čtvrtý a pátý graf jsou grafy koláčového typu a ukazují rozložení vlastnictví a typů nemovitostí.

Nejtěžším úkolem této části bylo vymyslet grafy, které budou dostatečně a srozumitelně interpretovat extrahovaná data.

4 Získaná data

Díky nastavení v konfiguračním souboru settings.py Scrapy automaticky ukládá extrahovaná data do výsledného .csv souboru. Tyto data bylo nutné před vizualizací očistit a vyfiltrovat, případně vytvořit nové sloupce pro snadnější práci s nimi. K tomu jsem použila knihovny Pandas a Numpy, díky kterým jsem například mohla snadno ze sloupce 'price' odstranit koncové ' Kč/měsíc' či vytvořit sloupec 'payment' na základě jiných.

5 Výsledky

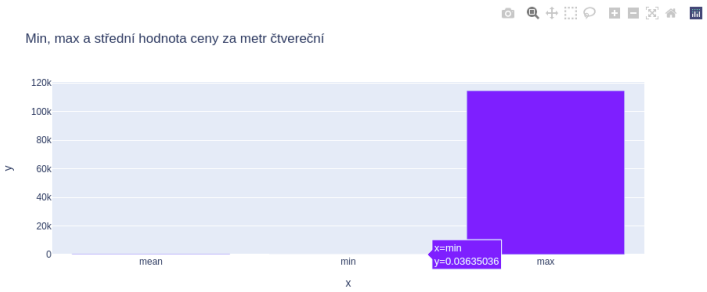
V rámci této práce byly extrahovány informace o téměř 50 000 nemovitostech, což je důvod, proč program běží tak dlouho (zhruba hodinu).

Dále byla vytvořena jednoduchá aplikace, která umožňuje porovnávání dat mezi jednotlivými kraji České republiky.

Zpočátku jsem myslela, že budu mít více nápadů co se smysluplné vizualizace týče. Určitě by se vizualizace dala mnohem více rozšířit, jelikož Plotly nabízí ohromné množství grafů. Nicméně u grafů, které jsem zařadila do své aplikace to někdy vypadá, že na nich nic není (viz Obrázek 3).

Obrázek 3: Graf 1

Je to kvůli velkému rozdílu mezi min a max bodem na ose x. Napadlo mě data normalizovat nebo standardizovat, ale ten rozdíl by tam pořád byl, čili si uživatel prostě musí najet na graf, aby viděl reálnou hodnotu (viz Obrázek 4).



Obrázek 4: Graf 2

6 Závěr

Tento program se hodí pro rychlou analýzu trhu nemovitostí podle krajů. Například pokud se člověk bude chtít stěhovat, tak toto může využít pro rychlé porovnání cen.

Dal by se vylepšit vzhled aplikace tak, aby byl více uživatelsky přívětivý a přehledný. Dále by se mohly přidat další vizualizace sloupců, porovnávání vícero krajů najednou atd.

Webscraper by se mohl poupravit, aby stahoval data z vícero stránek a následně promazával duplikované záznamy, protože jedna realitní stránka nemusí obsahovat veškeré nemovitosti. Na druhou stranu řada nejznámějších realitních stránek má zablokované scrapování dat, takže je otázka, zda by se toho našlo reálně víc.

Nakonec by šlo samozřejmě program rozšířit tak, aby doloval data nikoliv pouze z ČR, ale z dalších zemí.

7 Zdroje

https://dash.plotly.com/

https://scrapy.org/

https://docs.python.org/3/