ČVUT–FIT
petripat@fit.cvut.cz

Analýza dat z Netflixu s využitím data scraping

Patricie Petriľáková

2. ledna 2022

1 Úvod

V práci vytvářím webovou aplikaci s využitím knihovny Streamlit, kam uživatel nahraje svoje data z Netflixu. Tyto data jsou propojeny s daty získány z webu čsfd.cz, potom zanalyzovány a výsledky promítnuty na vytvořené stránce.

2 Vstupní data

O vstupní data jsem musela zažádat Netflix, který je do měsíce poskytnul. Z těchto dat k analýze využívám csv soubor ViewingActivity.csv, kde jsou informace o tom, co uživatel sleduje.

Využitelné pro analýzu jsou kolonky datum a čas sledování, jméno uživatele, délka sledování a země, ze které uživatel mohl sledoval Netflix. Na základě dat jsem pomocí názvu identifikovala, zda se jedná o seriál (rozděleno na název Jméno: Season n: Jméno episody (Episode n)) nebo film. Potom z ČSFD vytáhla rok/roky, mezi kterými byl film vydán nebo seriály vysílány, žánr, země původu, herce a hodnocení.

Data jsou velmi omezená svojí užitečností. Nelze poznat, zda uživatel viděl celý film nebo si jej jen pustil vícekrát a zda ho i dokoukal. Země sledování může být taky pozměněna při využívání VPN. Tady se nemusí taky jednat o skutečný ukazatel toho jaké země uživatel navštívil a kolik času v nich trávil sledováním. Názvy jsou občas v různých jazycích a nemusí se úplně shodovat s názvem v ČSFD. Také může nastat situace, kdy více filmů má stejný název a je vybrán špatný.

3 Metody/postupy/algoritmy

Po nahrání dat jsem je vyfiltrovala a vymazala jsem filmy/seriály s kratším sledováním než 2 minuty. Potom jsem vytvořila novou skupinu dat, kde jsem měla jen základní názvy, tedy například pro Brooklyn Nine-Nine: Season 2: The Chopper (Episode 22), jsem zkrátila název na Brooklyn Nine-Nine a odstranila

duplikáty. Pro tyto data jsem vytáhla informace z ČSFD a pak je znovu spojila s původními daty.

Na ČSFD není přesný seznam, přes který se lze dostat k filmům. Proto musím vyhledávat, projít stránku, naleznout požadovaný film, a hlavně jeho odkaz. Následně se připojit na odkaz a teprve tehdy mohu extrahovat data o filmu. Neustálému připojování způsobuje pomalejší načítání a tomu jsem se snažila zabránit lepší filtrací dat a snížením počtu hledáním, tedy nevyhledáváním každého sledovaného seriálu a všech jeho dílu. Také využívám caching, takže to při druhém kouknutí se na stejného uživatele urychluje načtení.

Výsledky také mohou být ovlivněny tím, že jediné, podle čeho mohu poznat film je název, takže může nastat situace, kdy je více filmů nebo seriálů se stejným názvem. V tomto případě beru ten první, protože na ČSFD jsou většinou seřazeny nalezené výsledky podle relevantnosti. Důležité upřesnění při hledání je, zda se jedná o film nebo seriál, podle toho dochází k vybrání identifikovaného filmu nebo seriálu. Také Netflix není systematický s názvy, mohou se vyskytnout názvy cizího filmu v angličtině nebo nemusí, ten se pak taky nemusí shodovat s názvem v ČSFD. Na ČSFD dochází ke kontrole obou poskytnutých názvu filmu, ale také to může ovlivnit správnost dat. K scrapování využívám knihovnu Beautiful Soup.

Pro analýzu používám Pandas a Plotly pro grafy. Výborným pomocníkem bylo z knihovny Plotly graph objects Figure, díky kterému jsem mohla přidat více vytvořených podgrafů do jednoho grafu. Jejich výhodou je, že jsou interaktivní, takže si uživatel může vyfiltrovat podgrafy podle potřeby. Největším problémem pro mě bylo strukturovat a správně pojmenovat jednotlivé sekce a podsekce, tak aby to bylo co nejkratší, ale zároveň srozumitelné. Pro představu, například sekce, která sleduje, jaké má uživatel návyky. Tedy, zda převážně sleduje Netflix např. večer, nebo v jaké dny a měsíce. Pro dny jsem tvořila hodinovou tabulku a pak rozpočítávala do jednotlivých hodin dne délku sledování. Tedy pokud začal sledovat v 19:58 a sledoval 2 hodinu, přičtou se 2 minuty v 19, pak je rozdělen zbytek do každé hodiny, dokud není vyčerpán celý čas na nulu. Obdobně v Figure graf postupné filtrování podle zemí všech dat, a následně zobrazení toho, jak uživatel sledoval v různých zemí, kolik času a kdy.

Pro znázornění webové stránky využívám Streamlit. Původně jsem začala s Django, ale přišlo mi to zbytečně mohutné a složité pro tak malou webovou aplikaci. Knihovna Streamlit je vcelku jednoduchá a výsledek vypadá jednoduše elegantní. Velké plus je, že mohu vytvářet stránku za chodu, protože se skript neustále obnovuje při každé změně, proto to bylo velmi uživatelsky přívětivé a snadno se ladily detaily.

4 Výsledky

V semestrální práci analyzuji a scrapuju data. Data jsou pak vykreslená a zobrazená na webovou stránku. V podstatě každý uživatel si může komplexně prohlédnout a analyzovat své data. Problém byl s hledáním seriálů/filmů na ČSFD, bylo to pomalejší, než jsem očekávala. Taky jsem nemohla ani příliš spěchat, protože bych byla označena za robota a mohlo by dojít k nepřesnosti v datech. Také by se hodily větší vědomosti k statistice, abych mohla vymyslet kreativnější analýzy.

5 Závěr

Vzhledem k tomu, že jsem ještě před tímto předmětem nepoužila Python, tak semestrální práce mě naučila hodně. Bylo to velice zajímavé, a mnohem časově náročnější, než jsem očekávala. Obzvlášť jsem ráda, že jsem si zkusila data scraping. Míst pro vylepšení je pravděpodobně hodně. Data poskytnuté Netflixem jsou sice velice omezená, ale z ČSFD by šlo určitě ještě extrahovat další data. Teoreticky na podobném principu by šlo zkusit extrahovat také jiné webové stránky s podobným zaměřením. S rozvojem streamingu by určitě komplexnější analýza dat, než nabízí streamingové společností mohla být pro jednotlivé osoby zajímavá, protože i členové rodiny byli překvapeni z prezentovaných dat.

6 Zdroje

Vzhledem k tomu, že to byla moje první zkušenost s Python, tak jsem se převážně učila pracovat s knihovnami a zdokonalit své dovednosti, tak nemám přímé citace k zpracování. Zdroje, které jsem však při práci využila byly převážně dokumentačního charakteru nebo články k tématu: docs.streamlit.io, discuss.streamlit.io, pandas.pydata.org, www.dataquest.io/blog/web-scraping-python-using-beautiful-soup/, naucse.python.cz/2020/pydata-ostrava-jaro/pydata/webscraping/, stackoverflow.com/, community.plotly.com/, plotly.com/python/, články na medium k práci s Pandas.