TECHNICKÁ UNIVERZITA V KOŠICIACH

FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Manažment Znalostí

Odporúčací systém

2021 **Dmytro Lahunov, Dávid Lacko**

Ciele zadania

- **Pochopenie dát**
- **Predspracovanie dát**
- **Zhlukovanie používateľov**
- **Odporúčanie**
- **Vyhodnotenie systému**

Definícia odporúčacieho systému

Odporúčacie systémy dnes patria medzi najpopulárnejšie aplikácie dátovej vedy. Používajú sa na predpovedanie „hodnotenia“ alebo „preferencie“. Takmer každá veľká technologická spoločnosť ich v nejakej forme použila. Amazon ho používa na navrhovanie produktov zákazníkom, YouTube na rozhodovanie o tom, ktoré video prehrať ako ďalšie pri automatickom prehrávaní a Facebook ho používa na odporúčanie stránok, ktoré by sa nám mohli páčiť, a ľudí, ktorých by sme chceli sledovať. Cieľom odporovacieho systému je teda udržať používateľa na platforme, odporučiť produkt alebo službu na kúpu ale aj napríklad ako odporúčanie pre doktora pri liečbe pacienta.

Štruktúra riešenia

Recommendation_system

|

| _ Data

| | _ steam-200k.csv

| | _ steam_processed.csv

| _ data_processing.ipynb

| _ recommendation_system.py

Súbor “Data” zahrňuje csv súbory s základným datasetom a predspracovanými dátami. Hlavný repozitár zahrňuje skripty data_processing.ipynb(skript na pochopenie a predspracovanie dát) a recommendation_system.py(hlavná trieda, ktorá implementuje celú funkcionalitu systému).

Pochopenie dát

Dáta

Táto dataset je zoznamom správania používateľov na platforme steam, so stĺpcami: user-id, game-title, behavior-name, value. Zahrnuté spôsoby správania sú „nákup“ a „hranie“. Atribút “value” udáva, do akej miery bolo správanie vykonané – v prípade „nákup“ je hodnota vždy 1 a v prípade „hrania“ hodnota predstavuje počet hodín, počas ktorých používateľ hral hru.

Dataset ma rozmer 4 riadky a 200000 stĺpcov

Atribúty

user-id – Predstavuje ID jednotlivých steam užívateľov

game-title – názov hry ktorú hráč vlastni

behavior-name – obsahuje udaj o tom ci daný užívateľ hru hral alebo ju len vlastni

hours-played– je počet hodín strávených v jednotlivej hre užívateľom. Ak užívateľ hru vlastni ale nehral hodnota je nastavená na 1

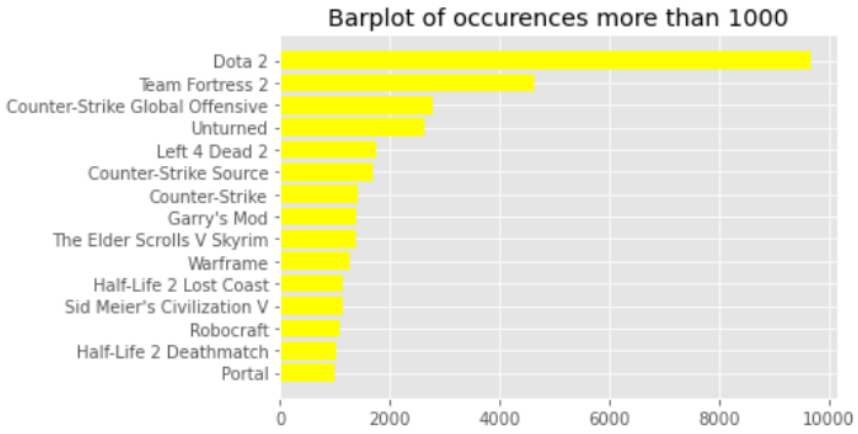
Column	Non-Null Count	Dtype
user-id	200000 non-null	int64
game-title	200000 non-null	object
behavior-name	200000 non-null	object
hours-played	200000 non-null	float64

Z vyzie uvedenej tabuľky sa vieme dozvedieť že dáta neobsahujú prázdne hodnoty.

Pre porozumenie dát bol vybratý jeden účet no ktorom môžeme vidieť že datataset obsahuje redundantne dáta. Ak hráč hral danú hru je 100% iste že danú hru aj vlastni.

	user-id	game-title	behavior-name	hours-played
0	151603712	The Elder Scrolls V Skyrim	purchase	1.0
1	151603712	The Elder Scrolls V Skyrim	play	273.0
2	151603712	Fallout 4	purchase	1.0
3	151603712	Fallout 4	play	87.0
4	151603712	Spore	purchase	1.0
5	151603712	Spore	play	14.9
6	151603712	Fallout New Vegas	purchase	1.0
7	151603712	Fallout New Vegas	play	12.1
8	151603712	Left 4 Dead 2	purchase	1.0
9	151603712	Left 4 Dead 2	play	8.9

Graf vykresľuje 15 hier ktoré vlastni najviac hráčov



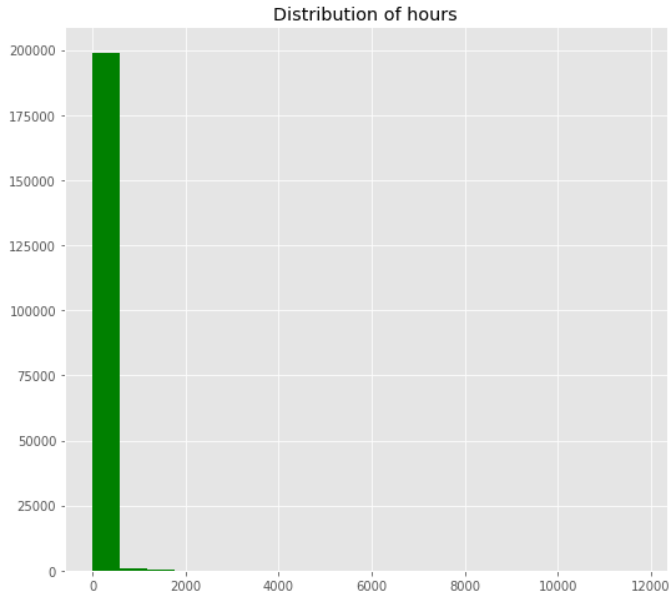
Ďalším cieľom bolo zistiť koľko hodín užívatelia strávia hraním jednej hry. Nasledujúca tabuľka predstavuje niektoré štatistické miery:

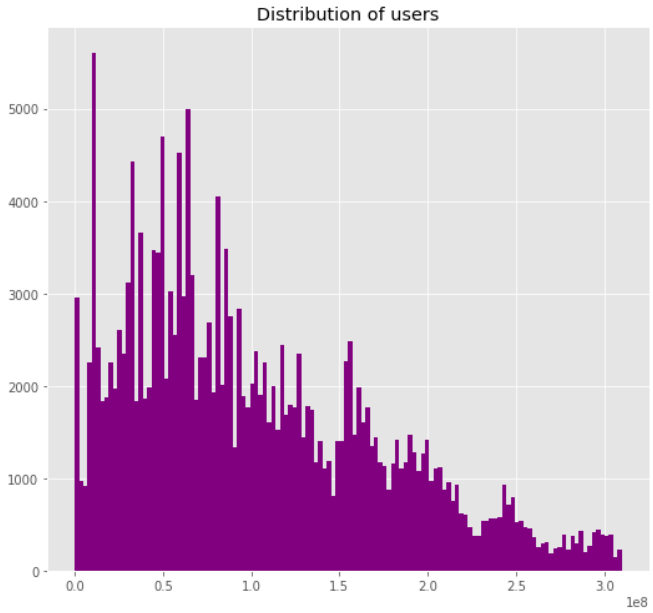
	hours-played
count	200000.000000
mean	17.874384
std	138.056952
min	0.100000
25%	1.000000
50%	1.000000
75%	1.300000
max	11754.000000

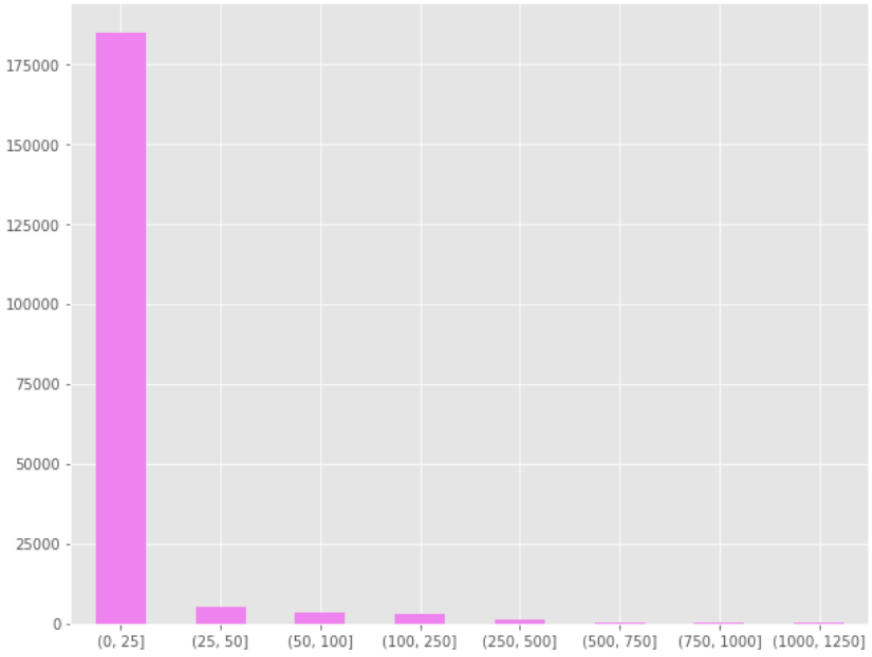
Distribúcia hodín - väčšina užívateľov hrala menej ako 1000 hodín, do 2000 hodín hralo 99.9% užívateľov.

Najviac hodín strávených na jednej hre užívateľom je 11756.

Distribúcia užívateľkou - reprezentuje počet hier používateľov.







Ako je možné vidieť na obrázkoch, frekvencia priemerných hodín strávených pri videohre sa riadi mocninovým rozdelením. To naznačuje, že existuje len veľmi málo videohier s množstvom hodín hrania, zatiaľ čo väčšina videohier sa hrá oveľa menej. To isté si môžeme všimnúť podľa hodnôt kvartilov v štatistickej tabuľke. To môže viesť k skresleným odporúčaniam uprednostňujúcim obľúbené hry. Metóda, ktorú sme zvolili, rieši tento problém vyhodnotením ratingu videohry na základe počtu hodín strávených používateľmi.

Predspracovanie dát

Fáza pochopenia dát ukázala, že dáta majú veľa duplikácií, teda atribút “behavior” pred tým ako nadobudnúť hodnotu “play” nadobúda hodnotu “purchase”, čo je zbytočnou informáciou, lebo keď hráč hral hru, to znamená, že si ju kúpil. Tieto duplikácií sme odstránili a potom aj atribút “behavior”.

Následné sme zvolili metódu výpočtu ratingu pomocou frekvencií výskytu hier podľa článku [1].

$$f_{i,u} = \frac{h_{i,u}}{\sum_{v \in U} h_{i,v}}.$$

Vyššie je uvedený vzorec, ktorý použijeme na nájdenie frekvencie pre hru ***i*** a používateľa ***u***, kde h_i, ***u*** predstavuje počet odohraných hodín. Teda sa zoberú hodiny, ktoré používateľ strávi v hre, a vydelí sa súčtom všetkých hodín, ktoré iní používatelia strávili pri hre.

Nasledujúci vzorec reprezentuje výpočet ratingu, ktorý dáva používateľ podľa frekvencie výskytu hier, ktorá sa počíta podľa vyššie uvedeného vzorca.

$$r_{i,u} = 4 \left(1 - \sum_{k=1}^{k-1} f_{i,k} \right) + 1.$$

Tu ***k*** je ***k***-tý používateľ hry ***i*** pri zoradení podľa frekvencie v zostupnom poradí a f_i, ***k*** je frekvencia vypočítaná v predchádzajúcom vzorci pre používateľa ***k*** hry ***i***.

Rating užívateľov sa dáva v rozmedzí 1-5.

Po výpočtu frekvencie a ratingu dáta sú pripravené na vytvorenie odporúčacieho systému.

Zhlukovanie používateľov

KNN

Algoritmus K Nearest Neighbor spadá do kategórie Supervised Learning a používa sa na klasifikáciu a regresiu. Je to všestranný algoritmus, ktorý sa používa aj na imputovanie chýbajúcich hodnôt a prevzorkovanie súborov údajov. Algoritmus KNN predpokladá, že podobné veci existujú v tesnej blízkosti. Algoritmus závisí od toho, že tento predpoklad je dostatočne pravdivý na to, aby bol algoritmus užitočný. [2]

Pre zhlukovanie užívateľov sme použili algoritmus NearestNeighbors

Hľadanie najbližšieho suseda, je problém optimalizácie nájdenia bodu v danej množine, ktorý je najviac podobný danému bodu. Blízkosť vyjadrujeme pomocou funkcie odlišnosti, čím menej podobné sú objekty, tým väčšie sú hodnoty funkcie.

Hodnota p je nastavená na p = 2 to znamená že používame euklidovskú vzdialenosť.

Použitý algoritmus „auto“ sa pokúsi rozhodnúť o najvhodnejšom algoritme na základe hodnôt odovzdaných metóde prispôsobenia.

Ako metriku vzdialenosti používame minkowski, je to metrika určená pre skutočné vektorové priestory. Minkowského vzdialenosť môžeme vypočítať iba v normovanom vektorovom priestore, čo znamená v priestore, kde vzdialenosti môžu byť reprezentované ako vektor, ktorý má dĺžku a dĺžky nemôžu byť záporné.

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

V našom prípade hľadáme 5 najbližších susedou. (tento počet susedou nám dáva dostatočne veľký počet odporúčaných hier)

Susednosť získavame podlá hodnotenia hry.

Výstupom algoritmu sú hodnoty distance a indices (Vráti indexy vzdialenosti od susedov každého bodu)

Na základe týchto hodnôt pridávame každému užívateľovi jeho susedov.

Výstupom nasej funkcie je list susedou pre každého užívateľa s jeho vzdialenosťou.

Odporúčanie

Mechanizmus odporúčania pre používateľa funguje na základe ratingu, ktoré používateľa s podobnými záujmami dávajú danej hre. Na začiatku sa zoberú hry podobných používateľov(teda najbližších susedov), ktoré nemá používateľ, pre ktorého sa robí odporúčanie. Prejde sa cez všetky hry a na odporúčanie sa zoberú len tie hry, ktoré používateľ už nevlastni a ktorých rating od susedov je väčší ako 3.8, čo tzv. prahom pre odporúčanie.

Content-based recommenders: navrhuje podobné položky na základe konkrétnej položky. Tento systém používa metadáta o položkách, ako je žáner, režisér, popis, herci atď. pre filmy na vytvorenie týchto odporúčaní. Všeobecnou myšlienkou týchto odporúčacích systémov je, že ak sa niekomu páči konkrétny predmet, bude sa mu páčiť aj predmet, ktorý je mu podobný. A na odporúčanie použije minulé metadáta položky používateľa.

Collaborative filtering: tieto systémy sú široko používané a snažia sa predpovedať hodnotenie alebo preferencie, ktoré by používateľ dal položke, na základe predchádzajúcich hodnotení a preferencií iných používateľov. Kolaboratívne filtre nevyžadujú metadáta.

Pre našu úlohu sme použili Collaborative filtering

Na konci v hlavnej funkcie sa vyberie 3 náhodných používateľa a následné sa pre nich vytvorí odporúčanie na základe bližších susedov.

Vyhodnotenie

Colaborative-filtering je užitočnou a ľahkou metódou odporúčacích systémov, ktorá používa históriu užívateľov na budúce odporúčania. V našom prípade odporúčací systém navrhuje vhodne hry pre používateľa, pre používateľov u ktorých susedia vlastnia tie iste hry/ pre nových užívateľov/ pre užívateľov bez susedov by sa odporúčali hry podlá ich celkovej popularity.

Úspešnosť výberu susedou sme vyhodnocovali podlá hodnotenia vlastnených hier pre každého používateľa.

Úspešnosť predikcie sa pohybovala priemerne nad 98% . Čo znamená že predikcia je úspešná.

Vzorové odporúčanie:

User:	10809782	308238255	133217805
Games:	Half-Life Deathmatch Source', 'Portal 2'	The elder scrolls - Skyrim', 'Fallout 4'	Age of Empires II HD The Forgotten', 'Alan Wake', 'Arma 3', 'Arma 3 Helicopters'
Accuracy:	98,66%	97,54%	99,47%

Mean_Accuracy:	98,55%

Zdroje

[1] Pérez-Marcos, Javier & Sánchez-Moreno, Diego & Batista, Vivian & Muñoz, María. (2019). Estimated Rating Based on Hours Played for Video Game Recommendation. 10.1007/978-3-319-99608-0_34.

[2] *Sklearn.neighbors.kneighborsclassifier*. scikit. (n.d.). Retrieved December 14, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html