Ekonomická fakulta
Faculty of Economics

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Jihočeská univerzita v Českých Budějovicích

Ekonomická fakulta

Katedra aplikované matematiky a informatiky

Bakalářská práce

Doporučovací algoritmy

Vypracovala: Veranika Kuliashova

Vedoucí práce: doc. Ing. Ladislav Beránek, CSc., MBA

České Budějovice 2023

JIHOČESKÁ UNIVERZITA V ČESKÝCH BUDĚJOVICÍCH

Ekonomická fakulta

Akademický rok: 2021/2022

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Veranika KULIASHOVA**
Osobní číslo:	**E20059**
Studijní program:	**B6209 Systémové inženýrství a informatika**
Studijní obor:	**Ekonomická informatika**
Téma práce:	**Doporučovací algoritmy**
Zadávající katedra:	**Katedra aplikované matematiky a informatiky**

Zásady pro vypracování

Cílem práce bude analýza známých doporučovacích algoritmů používaných u různých ecommerce aplikací. Hodnocení bude založeno na práci se zvolenou databází. Bude analýzována přesnost a efekt jejich jednotlivých parametrů. Práce bude obsahovat závěry a doporučení k aplikaci zvolených doporučovacích systémů.

Metodický postup:

1. Studium odborné literatury.
2. Úvod do problematiky doporučujících systémů.
3. Teoretický popis jednotlivých přístupů a jejich použití.
4. Doporučení pro aplikaci v různých případech.
5. Zhodnocení a možnosti dalšího vývoje.

Rozsah pracovní zprávy:	**40 – 50 stran**
Rozsah grafických prací:	**dle potřeby**
Forma zpracování bakalářské práce:	**tištěná**

Seznam doporučené literatury:

1. Aggarwal, Ch. C. (2016). *Recommender Systems.* Springer.
2. Balas, V. E., Chakrabarti, A., & Sharma, N. (2016). *Advances in Computing Applications.* Imprint: Springer.
3. Bokde, D., Girase, S., & Mukhopadhyay, D. (2015). *Matrix Factorization Model in Collaborative Filtering Algorithms: A Survey.* Procedia Computer Science.
   Dostupné z: <https://www.sciencedirect.com/science/article/pii/S1877050915007462#bibl0005>

Vedoucí bakalářské práce: **doc. Ing. Ladislav Beránek, CSc., MBA**
Katedra aplikované matematiky a informatiky

Datum zadání bakalářské práce: **11. ledna 2022**
Termín odevzdání bakalářské práce: **14. dubna 2023**

doc. Dr. Ing. Dagmar Škodová Parmová
děkanka

doc. RNDr. Tomáš Mrkvička, Ph.D.
vedoucí katedry

V Českých Budějovicích dne 28. února 2022

Prohlašuji, že svou bakalářskou práci jsem vypracovala samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury. Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své bakalářské práce, a to – v nezkrácené podobě vzniklé vypuštěním vyznačených částí archivovaných Ekonomickou fakultou – elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

Datum 21.04.2023 Podpis studenta

PODĚKOVÁNÍ

Za cenné konzultace a doporučení ráda bych poděkovala svému vedoucímu práci a školiteli, doc. Ing. Ladislavu Beránku, CSc., MBA, stejně jako za neustálou podporu, povzbuzení během celého mého výzkumného procesu a jeho trpělivost. Jeho zpětná vazba byla neocenitelná a bez jeho pomoci by tato práce nemohla vzniknout.

Ráda bych vyjádřila své nejhlubší poděkování také Mgr. Vladimiru Kadlecovi a BSc. Justinu Calvinu Schaeferovi za jejich cenné příspěvky k mému výzkumu.

Jsem také vděčná své přátelům, kteří mi byli stálým zdrojem podpory a povzbuzení, a rodině, zejména rodičům, kteří jsou pro mě velkou inspirací a největší oporou, a také důvodem, proč jsem dnes tam, kde jsem, když mám možnost studovat v zahraničí s úžasnými učiteli univerzity JČU.

Děkuji všem za podporu a povzbuzení.

Obsah

1. Úvod

Moderní svět se vyznačuje rychlými změnami a pokrokem. Je to dáno technologiemi a globalizací, které způsobily, že se svět pohybuje rychleji než kdykoli předtím. Počítač a další zařízení již nejsou jen nástrojem pro práci, studium nebo zábavu. Celý svět je zabudován do těchto malých zařízení, kde komunikují mezi sebou mnoha lidí, také prohlížejí si stránky, nakupují, sledují filmy a televizní seriály, poslouchají hudbu a čtou knihy. A díky internetu máme skoro všichni mají v současnosti neomezený přístup k informacím, zboží a službám, které v minulosti nebyly k dispozici. Od lidí se proto očekává, že se budou rychle rozhodovat, sledovat nejnovější trendy a udržovat si přehled o novinkách a událostech ve světě. To vytvořilo realitu, ve které je třeba být proaktivní a přizpůsobivý, aby člověk držel krok s měnící se dobou.

Relevantnost práce spočívá v tom, že s doporučovacími algoritmy se v posledních letech setkal téměř každý uživatel internetu. Tyto algoritmy se v průběhu let stávají stále sofistikovanějšími s inovací v této oblasti, s nástupem strojového učení a umělé inteligence. S vývojem efektivních doporučovacích algoritmů je však stále spojeno mnoho problémů, jako je řídkost dat, škálovatelnost a obavy o soukromí. Téma je relevantní nejen pro webové stránky elektronických obchodů, ale také pro platformy kontentu, jako jsou sociální média, zpravodajské webové stránky a služby pro streamování platformy, kde personalizovaná doporučení hrají zásadní roli při zapojení a udržení uživatelů. V důsledku toho, a právě kvůli mnohotvárnosti a přítomnosti v životech obyčejných, včetně mě, lidí, jsem se o toto téma začala zajímat a rozhodla jsem se ho zvolit pro svou bakalářskou práci.

Práce se bude skládat s 5 hlavních kapitol. První kapitola je úvodem k významu práce, ji relevantnosti a zabývá se jejími cíli. Následující druhá kapitola se zabývá problematikou doporučovacích algoritmů a systémů. Popisuje problémy, které tyto algoritmy řeší, a jejich dopad na společnost. Podrobně vysvětluje, jaké jsou různé typy algoritmů, popisuje jejich fungování a strukturu, charakteristické výhody, zabývá se využitím doporučovacích systémů v praxi a analyzuje využití algoritmů v elektronickém obchodě. Třetí kapitola popisuje metodiku praktické části. Kapitola čtyři popisuje vytvořený kolaborativní algoritmus pro filmovou databázi MovieLens. Závěrečná pátá kapitola obsahuje závěry a důsledky práce, možná zlepšení pro další výzkumné příležitosti. Následující kapitoly budou obsahovat seznamy všech použitých zdrojů,

obrázku, grafů a tabulek. Jedna z kapitol bude obsahovat také závěry a klíčová slova v angličtině.

1.1. Cíle práce

Cílem práce je získat a shrnout poznatky o doporučovacích algoritmech z odborné literatury, analyzovat co je doporučovací algoritmus a každý typ zvlášť, a jejich současné využiti na internetu v e-commerce, včetně e-shopů a streamovacích služeb. A na základě teoretických informací vytvořit vlastní kolaborativní doporučovací algoritmus pro filmovou databázi MovieLens se záměrem na přesnost.

2. Teorie doporučovacích algoritmu

Kapitola o teorii doporučovacích algoritmů je důležitou součástí této bakalářské práce. Poskytuje čtenářům veškeré informace o doporučovacích systémech, které nejsou běžným uživatelům internetu příliš známé. Jako autorka jsem v této kapitole shromáždila nejaktuálnější a nejtvořivější data pro pochopení základů doporučovacích algoritmů. Zde najdete definice a základní vysvětlení algoritmu, analýzu jeho typů, jako je kolaborativní filtrování a filtrování podle obsahu, a jejich použití v online službách. Doufám, že po přečtení této kapitoly budou moci i lidé, kteří o těchto algoritmech dosud neslyšeli, předem pochopit jejich základní funkce.

2.1. Základní představa o doporučovacích algoritmů

V této podkapitole prozkoumáme fascinující svět doporučovacích algoritmů, což je třída algoritmů strojového učení, které se používají k poskytování personalizovaných doporučení uživatelům na základě jejich preferencí, chování a chování podobných uživatelů. Ponoříme se do historie těchto algoritmů a podíváme se, odkud se vzaly. Používání doporučovacích algoritmů není bez etických důsledků. Proto se budeme zabývat také etickými aspekty spojenými s používáním těchto algoritmů. Probereme možná rizika a předsudky spojené s doporučovacími algoritmy, včetně otázek spravedlnosti a transparentnosti.

Na konci této kapitoly budete důkladně znát základní myšlenku doporučovacích algoritmů, jejich historii, fungování a etické aspekty, které jsou s jejich používáním spojeny. To poskytne solidní základ všem, kteří mají zájem porozumět světu doporučovacích algoritmů a roli, kterou hrají v našem životě.

2.1.1. Historie a původ doporučovacích algoritmů

Moderní doporučovací systémy a algoritmy se začaly objevovat na konci minulého století s rozvojem počítačových technologií a internetu, ale není tajemstvím, že od počátku věků a lidské komunikace lidé využívají doporučení od jiných lidí, např. přátel nebo známých. Zkušená a uznávaná osoba nebo několik lidí se stejným názorem mohou ovlivnit rozhodovací proces jednotlivce.

Při výběru, teoreticky řečeno, položky diskutovaly několik přátel o podobných věcech mezi sebou a se známými a sousedy, tj. shromažďovaly informace, a samy je

zpracovávaly s přihlédnutím k doporučením ostatních. Posloupnost činností v doporučovacích algoritmech je, jak vidíme, podobná.

Doporučovací algoritmy mají od 90. let 20. století bohatou historii. V roce 1992 S vývojem experimentálního poštovního systému Tapestry ve výzkumném centru Xerox Palo (Goldberg a kol., 1992) byla představena myšlenka "kolaborativního filtrování" (Collaborative Filtering), v němž bylo stejně jako účelem služby bylo umožnit uživatelům vytvářet pravidla filtrování pošty, která by mimo jiné mohla zohledňovat postoje a jednání ostatních lidí. Později, v roce 1994, byl představen systém filtrování zpráv GroupLens, jehož cílem bylo automatizovat proces kolaborativního filtrování založený na pravidlech systému Tapestry (Dong et al., 2022). Jako jeden z prvních systémů byl navržen systém GroupLens, který se spoléhal na explicitní hodnocení poskytovaná komunitou uživatelů a využíval strojové učení k předpovědi, zda se uživateli budou zdát určité neviděné zprávy atraktivní.

V následujících letech výzkumníci a společnosti pokračovali ve vývoji a zdokonalování doporučovacích algoritmů s cílem zvýšit jejich přesnost a účinnost. S rychlým rozvojem World Wide Webu v 90. letech 20. století se objevilo stále více nových aplikací pro doporučovací systémy. Úspěšné příklady využití doporučovacích systémů v elektronickém obchodě byly zaznamenány ještě před koncem tohoto desetiletí (Schafer, Konstan, & Riedl, 1999), přičemž jedním z prvních velkých uživatelů doporučovací technologie byla společnost Amazon.com (Linden, Smith & York, 2003).

V průběhu let výzkumníci a společnosti pokračovali ve vývoji a zdokonalování doporučovacích algoritmů s cílem zlepšit jejich přesnost a účinnost. Jedním z prvních a nejvlivnějších algoritmů bylo kolaborativní filtrování, které bylo vyvinuto v polovině 90. let 20. století výzkumníky na Minnesotské univerzitě (Herlocker a kol., 2004). Kolaborativní filtrování je založeno na myšlence, že lidé, kteří mají podobné preference nebo zájmy v jedné oblasti, budou mít pravděpodobně podobné preference nebo zájmy i v jiných oblastech.

Na počátku roku 2000 začal získávat popularitu nový přístup k doporučovacím algoritmům, který se nazývá filtrování na základě obsahu (Melville & Sindhwani, 2010). Filtrování založené na obsahu využívá algoritmy strojového učení k analýze charakteristik položek, jako jsou filmy, knihy nebo hudba, a poté doporučuje další položky, které jsou v určitém ohledu podobné.

V posledních letech jsou stále populárnější hybridní přístupy, které kombinují kolaborativní a obsahové filtrování (Adomavicius & Tuzhilin, 2005). Tyto algoritmy jsou

navrženy tak, aby využívaly silné stránky obou přístupů, což vede k přesnějším a efektivnějším doporučením.

Aktuálně moderní algoritmy jsou používány v e-commerce a na platformách streamovacích médií, v sociálních sítích. Vzhledem k tomu, že množství dat, která mají tyto algoritmy k dispozici, neustále roste, budou výzkumníci a společnosti pravděpodobně i nadále vyvíjet nové a sofistikovanější přístupy k doporučovacím algoritmům (Langville & Meyer, 2012).

2.1.2.Co je doporučovací algoritmus?

Abych se mohla začít zabývat tématem své práce, bylo nutné vysvětlit význam definice „Doporučovací algoritmus“. Podle Ottovu naučnému slovníku, algoritmus “znamená pravidlo nebo způsob početní”. Ve současnosti pojem se používán zejména při programování, algoritmus lze chápat jako je soubor pokynů, což vysvětluje, jak krok za krokem vyřešit určitý problém nebo splnit zadaný úkol. Zase doporučení je příznivým posudkem nebo návrhem nebo radou, která je někomu dána na základě jeho potřeb, preferencí nebo cílů. Obvykle je poskytována někým, kdo má odborné znalosti nebo zkušenosti v určité oblasti, a jejím cílem je pomoci dané osobě učinit informované rozhodnutí nebo přijmout konkrétní opatření.

Doporučovací algoritmus je typ algoritmu, který je určen k poskytování personalizovaných doporučení uživatelům na základě jejich dřívějšího chování, preferencí a zájmů (Ricci, Rokach, & Shapira, 2011). Tyto algoritmy běžně používají firmy k navrhovaní produktů nebo služeb zákazníkům a poskytovatelé obsahu k navrhování videí, hudby nebo jiných médií divákům.

Mezi doporučovacími algoritmy a doporučovacími systémy existuje úzký vztah. Doporučovací algoritmus je matematický model, který využívá analýzy dat a techniky strojového učení k vytváření personalizovaných doporučení pro uživatele. Je to je základním prvkem, klíčovou součástí doporučovacího systému, protože generuje doporučení na základě uživatelských dat a historických vzorců používání. Na druhé straně je doporučovací systém širší pojem, který označuje celý softwarový systém, který zahrnuje doporučovací algoritmus k poskytování personalizovaných doporučení uživatelům. Podle (Jannach et al., 2010) „„Doporučovací systém je podtřída systému filtrování informací, která se snaží předpovědět "hodnocení" nebo "preference", které by uživatel dal nějaké položce. Doporučovací systémy se využívají v různých oblastech,

přičemž běžně známé příklady mají podobu generátorů playlistů pro video a hudební služby, doporučovačů produktů pro internetové obchody nebo doporučovačů obsahu pro platformy sociálních médií a doporučovačů otevřeného webového obsahu.“

V následující podkapitole pronikneme hlouběji do fungování algoritmu. Takým způsobem je možné lépe pochopit, co se v základu doporučovací algoritmus dělá a jak dosahuje svých výsledků.

2.1.3. Jak funguje doporučovací algoritmus?

Hlavním cílem doporučovacích algoritmů je poskytovat personalizovaná doporučení pomocí funkce nebo matematického modelu.

Jednou ze základních součástí doporučovacího algoritmu jsou data použitá k jeho trénování a spuštění. Shromážděná data musí být kvalitní a ve správném formátu, aby byla zajištěna přesnost a účinnost algoritmu. Jedním z nejběžnějších způsobů, jak získat data pro algoritmus, je chování uživatelů. Webové stránky a aplikace mohou sledovat chování uživatelů shromažďováním údajů o vyhledávacích dotazech, kliknutích, zobrazeních stránek a historii nákupů. Kromě toho mohou být uživatelé požádáni, aby poskytli hodnocení a recenze produktů, služeb, které použili nebo konzumovali. Tyto údaje generované uživateli lze kombinovat s údaji z externích zdrojů, jako jsou demografické údaje nebo údaje ze sociálních sítí, a vytvořit tak komplexní profil uživatele. Po shromáždění dat je třeba je načíst do databáze nebo datového skladu a dále zpracovat. Data lze načíst do tradiční databáze či databáze jiného typu v závislosti na objemu a typu dat. Například, data mohou být uložena buď ve formátu NoSQL pro nestrukturovaná data nebo ve formátu SQL pro strukturovaná data. Po načtení dat je třeba je zformátovat do formátu použitelného pro doporučovací algoritmus. To může zahrnovat odstranění duplicit, převod kategoriálních proměnných na číselné proměnné a normalizaci dat pro zajištění konzistence proměnných. Sady uživatelských položek se často transformují do matice zvané matice hodnocení. K čištění a formátování dat pro algoritmy strojového učení mohou být použity nástroje, jako je knihovna Pandas v jazyce Python a hodně jiných.

Další fází je poskytování doporučení. Algoritmus hledá podobnosti mezi uživateli a položkami v databázi, v závislosti na typu algoritmu mezi žánry, podobnými rysy nebo hodnoceními. Systém po prozkoumání dat a identifikaci relevantních trendů a preferencí

poskytuje uživateli návrhy na míru. Tyto návrhy se mohou týkat produktů, částí obsahu nebo jiných relevantních údajů, které by spotřebitele zajímaly.

2.1.4. Pomocník nebo hrozba?

Doporučovací algoritmy se staly nedílnou součástí našeho digitálního života a pomáhají nám objevovat nové produkty, služby a informace, které by nás mohly zajímat. Někdy si ani neuvědomíte, jak skvělou práci odvádí, když všem uživatelům po celém světě poskytuje různá doporučení. Když se nad tím zamyslíte, je přirozené se ptát, zda je jejich použiti etické.

Z etického hlediska lze na používání doporučovacích systémů, stejně jako na všechny věci na tomto světě, nahlížet různými způsoby, ze dvou diametrálně odlišných úhlů. Může to být nástroj, který pomáhá spotřebitelům a zvyšuje prodej, něco, co usnadňuje nakupování a vyhledávání, nebo naopak něco, co může poškodit svědomí člověka.

Doporučovací algoritmy mohou být pro uživatele užitečnými nástroji v různých kontextech. Doporučení mohou uživatelům poskytnout orientaci a snížit kognitivní zátěž při rozhodování tím, že jim předloží vybrané možnosti, mohou také pomoci podnikům zvýšit příjmy a ziskovost a optimalizovat jejich provoz. To jsou jen některé ze skvělých věcí, které poskytuje.

Navzdory svým výhodám nejsou doporučovací algoritmy bez omezení a problémů. Jednou z možných výzev je otázka zaujatosti a diskriminace. Doporučovací algoritmy se při vytváření předpovědí a doporučení spoléhají na historická data a chování uživatelů, což může vést k posilování stávajících předsudků a stereotypů.

V následujících podkapitolách bych se chtěl blíže podívat na některé pozitivní a negativní jevy, se kterými se uživatelé mohou setkat při používání doporučovacích algoritmů.

2.1.4.1. Problém dlouhého chvostu

Koncept "dlouhého chvostu", v původním anglickém znění “The Long Tail’ představil Chris Anderson v roce 2004 v článku pro časopis Wired, který se zaměřuje na to, jak nové technologie ovlivňují kulturu, ekonomiku a politiku, a na základě tohoto článku byla napsána kniha (Anderson, 2006).

Obrázek 1. Graf dlouhého chvostu

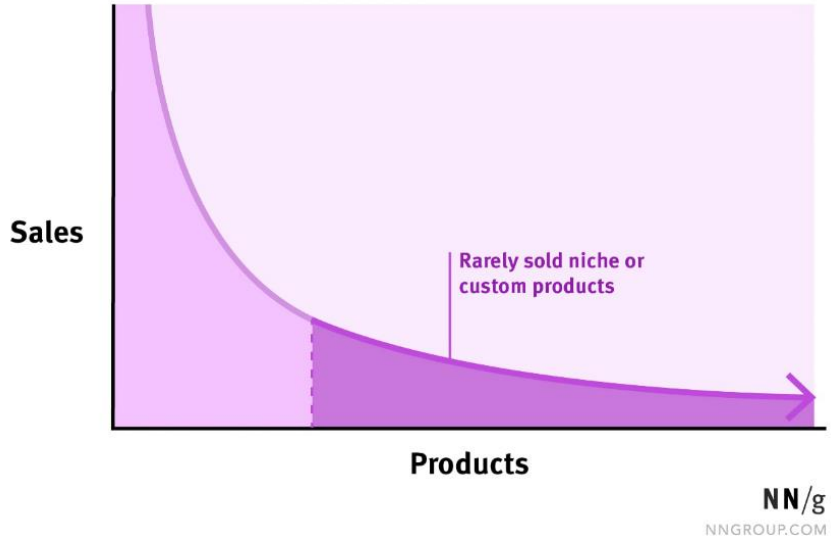
Zdroj: Recognize Strategic Opportunities with Long-Tail Data (Sunwall, 2021)

Problém dlouhého chvostu se týká jevu, kdy velký počet méně populárních nebo specializovaných produktů má dohromady podíl na trhu, který se vyrovná podílu několika nejprodávanějších produktů nebo jej převyšuje. Jinými slovy, většina poptávky po zboží nebo službách se nachází v "dlouhém chvostu" rozdělení, který představuje velký počet méně populárních položek, spíše než v "čele" rozdělení, které představuje malý počet nejprodávanějších položek.

Například než vznikl Internet, se veškeré nákupy odehrávaly v kamenných obchodech. Majitelé nakupovali převážně bestsellery obchodu, protože nebylo příliš výhodné držet na skladě zboží, které by si koupil jen malý počet lidí. Doporučující algoritmy s tímto problémem jsou velkým přínosem pro online prodejce, protože z takových specializovaných produktů získávají velký zisk navíc

Ve současnosti v e-komerci lze graf dlouhého chvostu použít k identifikaci nejoblíbenějších produktů, stejně jako k identifikaci produktů, které mají nižší poptávku, ale přesto v průběhu času generují významné příjmy. Může pomoci podnikům v oblasti elektronického obchodování optimalizovat nabídku produktů a cenové strategie s cílem maximalizovat zisky.

Obrázek 2. Graf dlouhého chvostu v e-komerci



Zdroj: Recognize Strategic Opportunities with Long-Tail Data (Sunwall, 2021)

V grafu představuje osa X různé produkty, které jsou k dispozici k prodeji, zatímco osa Y – počet prodejů nebo poptávku po jednotlivých produktech. Obvykle můžeme vidět dlouhý chvost na pravé straně osy X, což znamená, že existuje značný počet výrobků s nízkou poptávkou.

Problém dlouhého chvostu představuje pro podniky příležitosti i výzvy. Na jedné straně umožňuje podnikům proniknout na dříve nevyužité trhy a generovat příjmy z velkého množství úzce specializovaných produktů. Na druhou stranu vyžaduje, aby podniky spravovaly rozsáhlé a různorodé zásoby, což může být náročné a nákladné. Podniky navíc mohou mít problémy s efektivním marketingem a propagací méně populárních produktů, protože mohou být méně viditelné a oslovovat menší publikum.

2.1.4.2. Příhodnost

Doporučovací algoritmy, které automaticky navrhují materiály na základě předchozích akcí a zájmů uživatelů, spotřebitelům značně usnadňují práci. Tím, že uživatelé nemusí aktivně vyhledávat relevantní materiál, mohou ušetřit čas a úsilí a zároveň se dozvědět o nových a zajímavých věcech, které by jinak nenašli. Například online prodejci a služeb, jako je Amazon a Netflix, používají doporučovací algoritmy, aby svým uživatelům navrhovali produkty a filmy nebo televizní pořady na základě jejich nákupní a divácké historie a dalších faktorů, jako jsou hodnocení a recenze podobných uživatelů. To pomáhá uživatelům rychle najít produkty nebo obsah, které se jim

pravděpodobně budou líbit, aniž by museli trávit čas procházením nekonečných seznamů nebo kategorií. Podobně platformy sociálních médií, jako je Facebook a Instagram, používají výše uvedené algoritmy, které uživatelům navrhují obsah na základě jejich chování, například na základě toho, které příspěvky se jim líbily nebo které komentovali. To může uživatelům pomoci objevit nový obsah a spojit se s ostatními, kteří mají podobné zájmy, aniž by museli ručně vyhledávat nebo procházet celý svůj webový kanál.

Celkově lze říci, že doporučovací algoritmy poskytují uživatelům značné pohodlí, protože jim automaticky navrhují obsah na základě jejich předchozího chování a preferencí, což může ušetřit čas a úsilí, pomoci objevit nový obsah a snížit únavu z rozhodování. Je však důležité si uvědomit potenciální rizika a omezení těchto algoritmů.

2.1.4.3. Falešné hodnocení

Vzhledem k tomu, že algoritmy doporučují spíše produkty položek s poměrně dobrými recenzemi, získávaní skutečnou zpětnou vazby od spotřebitelů však není tak obtížný úkol, ale spíše, řekněme, dost problematický. Což může být pro některé prodejce záminkou k používání nekalých metod propagace a mezer v systému. Pak si podnikatelé nebo jejich zaměstnanci marketingu právě tyto hodnocení jednoduše píšou sami či vyhledávají takové služby u třetích stran. Nejhorší na této situaci je, že specifický jev je poměrně rozšířený, což znamená, že tato praxe podkopává důvěru spotřebitelů v posudky jako účinný nástroj propagace. Aby ukázat, že tento problém je opravdu skutečný jako příklad uvádím, že podle elektronických novin CNBC (Palmer, 2022), Amazon podal žaloby na několik společností za použiti falešných recenzí na své platformě. Podané žaloby obviňují společnosti AppSally a Rebatest z podpory falešných recenzí na online trhu Amazonu. Společnosti údajně propojily prodejce třetích stran se spotřebiteli, kteří výměnou za bezplatné produkty nebo platby zanechali pozitivní recenzi na jejich produkt. V prohlášení se dále uvádí, že AppSally a Rebatest uvádějí, že mají více než 900 000 uživatelů "ochotných psát falešné recenze".

Falešné recenze mohou poškodit pověst firmy a vést ke ztrátě příjmů, což může být obzvláště škodlivé pro malé podniky. Pokud se navíc spotřebitelé při rozhodování o nákupu spoléhají na falešné recenze, mohou si nakonec pořídit nekvalitní výrobek nebo službu, což může vést k frustraci a zklamání.

2.1.4.4. Filtrační bubliny

Filtrační bublina, nebo Filter Bubble, jsou fenoménem, kdy algoritmy selektivně poskytují jednotlivcům personalizovaný obsah na základě jejich předchozího chování, zájmů a preferencí, čímž vytvářejí situaci, kdy jsou vystaveni pouze informacím, které jsou v souladu s jejich stávajícími názory a hodnotami, a omezují jejich vystavení různým perspektivám a pohledům. To může vést k posilování předsudků a přesvědčení a k nedostatečnému vystavení alternativním pohledům a myšlenkám, což může omezit individuální růst a společenský diskurz (Pariser, 2011).

Koncept filtračních bublin si v posledních letech získal značnou pozornost, zejména v kontextu sociálních médií a politického diskurzu. Personalizovaná povaha algoritmů sociálních médií může vést k tomu, že uživatelé jsou vystaveni pouze obsahu, který potvrzuje jejich stávající přesvědčení, což vede k polarizaci politických a sociálních názorů a omezuje možnosti dialogu a porozumění (Sunstein, 2017). To může mít významné důsledky pro demokracii, protože jednotlivci mohou mít menší tendenci zapojovat se do občanské diskuse a častěji přijímat dezinformace nebo extremistické názory jako pravdu.

2.1.4.5. Manipulace

Pojem "manipulace" v kontextu doporučovacích algoritmů označuje záměrné nebo náhodné aktivity, které vedou k propagaci nebo potlačení určitých informací, často z ekonomických nebo politických důvodů. K tomu může docházet různými způsoby, například změnou váhy algoritmu nebo upřednostňováním určitých prvků nebo záměrným vyzdvihováním či znehodnocováním určitých typů informací v souladu s cíli nebo zásadami vývojářů nebo provozovatelů algoritmu.

Jedním z příkladů manipulace s doporučovacími algoritmy je tzv. astroturfing, kdy firmy nebo lidé vytvářejí falešné účty nebo materiály na podporu určitého zboží nebo názoru, často nepravdivým nebo zavádějícím způsobem. (Chen et al., 2021). Podobně mohou některé podniky nebo lidé používat techniky "clickbait", kdy jsou titulky nebo popisy materiálů vytvořeny tak, aby lákaly ke kliknutí nebo zobrazení, i když je samotný obsah nekvalitní nebo nesouvisí se zájmy uživatele.

Manipulací s doporučovacími systémy může výrazně utrpět soukromí a bezpečnost uživatelů. Například algoritmy, které sledují chování a preference uživatelů za účelem přizpůsobení obsahu, mohou také shromažďovat a ukládat citlivé údaje o uživatelích, což je vystavuje riziku bezpečnostních problémů nebo úniku dat (Huang, Liu & Tang, 2019).

2.2. Typy doporučovacích algoritmů

Existují různé typy doporučovacích algoritmů, z nichž každý má své silné a slabé stránky. Jedním z nejpoužívanějších typů doporučovacích algoritmů je kolaborativní filtrování, které při doporučování vychází z chování a preferencí skupin uživatelů. Algoritmy kolaborativního filtrování lze rozdělit do dvou hlavních kategorií: na základě uživatelů a na základě položek.

Dalším typem doporučovacího algoritmu je filtrování založené na obsahu, které doporučuje položky, které jsou svým obsahem nebo vlastnostmi podobné položkám, které si uživatel dříve oblíbil. Filtrování založené na obsahu se spoléhá na atributy, jako jsou klíčová slova, značky nebo žánry, aby identifikovalo položky, které budou uživatele pravděpodobně zajímat.

A konečně hybridní doporučovací algoritmy kombinují více doporučovacích technik, aby poskytovaly přesnější a rozmanitější doporučení. Hybridní algoritmus může například kombinovat kolaborativní filtrování s filtrováním založeným na obsahu a doporučovat položky, které jsou oblíbené mezi uživateli s podobnými preferencemi a zároveň mají podobný obsah jako položky, které si uživatel dříve oblíbil.

V následujících podkapitolách se budeme zabývat některými nejoblíbenějšími typy doporučovacích algoritmů a diskutovat o jejich výhodách a omezeních.

2.2.1.Kolaborativní filtrovaní

Kolaborativní filtrování nebo Collaborative filtering v angličtině je populární metoda doporučovacích algoritmů, která se hojně využívá v online platformách k poskytování personalizovaných doporučení uživatelům. Je oblíbené z několika důvodů. Zaprvé se jedná o jednoduchou a intuitivní techniku, kterou lze snadno implementovat a pochopí jak vývojáři, tak uživateli. Kolaborativní filtrování nevyžaduje složité modelování ani inženýrství prvků, takže je to algoritmus, který se v praxi implementuje poměrně snadno. Je efektivní při vytváření personalizovaných doporučení tím, že identifikuje podobnosti mezi uživateli na základě jejich chování a preferencí. Může dobře fungovat v situacích, kdy mají uživatelé různé zájmy a preference, protože dokáže identifikovat vzory v chování uživatelů, které nemusí být okamžitě zřejmé. Tento typ algoritmu lze použít s širokou škálou typů dat, včetně explicitních hodnocení, implicitní zpětné vazby a dalších typů chování uživatelů, jako jsou kliknutí nebo nákupy. Tato flexibilita z něj činí univerzální algoritmus, který lze použít v různých oblastech, od

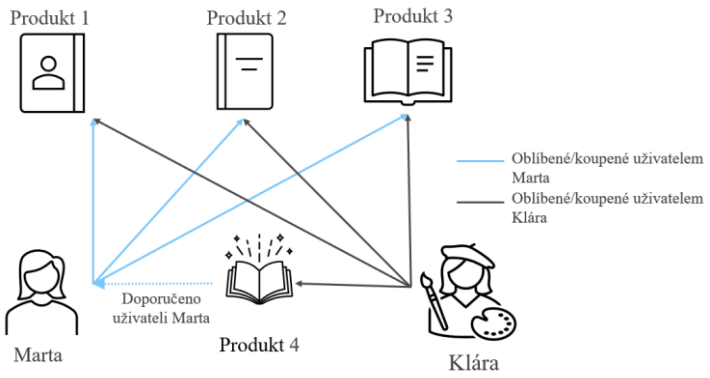
elektronického obchodu a zábavy až po sociální sítě a doporučování zpráv. V neposlední řadě lze kolaborativní filtrování kombinovat s jinými technikami doporučování a vytvářet tak hybridní algoritmy, které využívají silné stránky různých přístupů. Kolaborativní filtrování lze například kombinovat s filtrováním založeným na obsahu a poskytovat tak doporučení, která jsou personalizovaná a zároveň obsahově podobná položkám, které si uživatel dříve oblíbil.

Existuje hodně služeb e-commerce, které využívají kolaborativní filtrování k personalizovaným doporučením pro své uživatele. Například, jde o e-shopu Amazon, který je jednou z největších společností v oblasti elektronického obchodování na světě, eBay, Zalando, Etsy, ASOS atd. Navíc, tento typ algoritmu využívají streamovací služby jako Netflix, Amazon Prime a Hulu, hudební streamovací služby jako Spotify a Last.fm, platformy sociálních médií jako Facebook a Twitter a hodně dalších.

Podle definice kolaborativní filtrování je technika pro doporučovací algoritmy a systémy, která předpovídá pravděpodobnost zájmu uživatele o položku na základě vzorců hodnocení nebo chování jiných uživatelů s podobnými zájmy" (Konstan & Riedl, 2012).

Pro lepší pochopení toho, jak algoritmus funguje, můžete se podívat na obrázek 3 na další. Představuji vám dvě uživatelky stejných webových stránek, Martu a Kláru. V tomto konkrétním případě mají rádi stejné knihy, takže tento algoritmus lze použít v jakémkoli internetovém obchodě, kde si můžete koupit knihy nebo kde na ně můžete zanechat recenzi, například na Goodreads. Kolaborativní filtrovaní je považuje za podobné, protože se jim líbí podobné, občas stejné objekty. A jak vidíme, Kláře se líbí i věci, které Marta nijak zatím neohodnotila. Takže díky podobnosti jejich zájmů a profilů bude Martě doporučeno to, co se líbí Kláře a čemu dala dobrou známku. Teoreticky to funguje vzájemně.

Obrázek 3. Doporučovací algoritmus založený na kolaborativním filtrování



Zdroj: vlastní tvorba

Na základě tohoto příkladu máme představu, že, hlavní myšlenka metody kolaborativního filtrování spočívá v tom, že tato nespecifikovaná hodnocení lze imputovat, protože pozorovaná hodnocení jsou často vysoce korelovaná mezi různými uživateli a položkami (Aggarwal, 2016). Uvažujme například dva uživatele s mimořádně podobným vkusem, Jana a Anetu. Základní algoritmus může určit jejich podobnost, pokud jsou hodnocení, která obě strany poskytly, zcela srovnatelná. V těchto situacích existuje velká pravděpodobnost, že hodnocení, kterým přidělil hodnotu pouze jeden z nich, budou rovněž srovnatelná. Tuto podobnost lze využít k vyvození závěrů o hodnotách, které nebyly plně uvedeny. Pro proces predikce využívá většina modelů kolaborativního filtrování buď korelace mezi položkami, nebo korelace mezi uživateli.

Existují dva hlavní typy kolaborativního filtrování: založené na uživatelích (User-Based) a na položkách (Item-Based). Například, algoritmus na obrázku Marty a Kláry je klasickým vzorcem User-Based kolaborativního algoritmu. Tyto dva typu algoritmů mají mnoho podobností. Například, oba přístupy například využívají k doporučením minulé chování uživatelů. Filtrování založené na uživateli identifikuje uživatele, kteří mají podobné preference a chování, zatímco filtrování založené na položkách identifikuje položky, které jsou si podobné na základě interakcí mezi uživateli. User-Based a Item-Based přístupy také vycházejí z předpokladu, že uživatelé, kteří měli podobné chování a preference v minulosti, budou mít pravděpodobně podobné chování a preference i v budoucnosti.

Jak filtrování založené na uživatelích, tak filtrování založené na položkách může trpět problémem studeného startu, kdy může být obtížné vytvořit doporučení pro nové uživatele nebo položky bez dostatečného množství dat. Také mohou být náchylné k takzvanému "zkreslení popularity", kdy jsou populární položky doporučovány častěji než méně populární položky.

Navzdory těmto podobnostem existují mezi kolaborativním filtrováním založeným na uživatelích a na položkách také některé důležité rozdíly, včetně jejich výpočetní složitosti, výkonu na řídkých souborech dat a citlivosti na změny v chování uživatelů v čase. Výběr nejvhodnějšího přístupu proto závisí na konkrétních vlastnostech datové sady a požadovaných kritériích výkonnosti.

Celkově je kolaborativní filtrování skvělou technikou pro personalizovaná doporučení uživatelům. Přestože má určitá omezení, zůstává oblíbenou volbou pro mnoho aplikací díky své účinnosti, škálovatelnosti a snadné implementaci. V následující podkapitole bych rád popsal jeho výhody a slabiny.

2.2.1.1. Výhody a nevýhody kolaborativního filtrování

Všechno na světě má své skvělé i méně skvělé stránky. Před výběrem vhodného algoritmu bychom se měli podívat na všechny výhody a nevýhody kolaborativního algoritmu. Ráda bych začal silnými stránkami:

- Popularita – kolaborativní filtrování je široce rozšířená technika, která se osvědčila v mnoha aplikacích, včetně elektronického obchodování, doporučování hudby a filmů. Podle Charu C. Aggarwal, "kolaborativní filtrování se ukázalo být jednou z nejúspěšnějších technik pro doporučovací systémy" (Aggarwal, 2016).
- Škálovatelnost – algoritmus na základě kolaborativního filtrování si poradí s velkými soubory dat a je škálovatelné. Je založen na jednoduchých a intuitivních konceptech, které se snadno implementují, což z něj činí oblíbenou volbu pro mnoho aplikací.
- Jednoduchost – tento typ filtrování je založen na intuitivních konceptech, které se snadno implementují. Technika vychází z předpokladu, že uživatelé, kteří měli podobné preference v minulosti, budou mít podobné preference i v budoucnosti a že položky, které jsou oblíbené mezi podobnými uživateli, budou oblíbené i u daného uživatele.

Pro efektivní využití kolaborativního filtrování je důležité pochopit jeho slabiny a uvědomit si možná omezení. Pochopením jeho omezení můžeme podniknout kroky k jejich odstranění a vyvinout přesnější doporučovací algoritmy a systémy. Toto jsou některé z těchto slabých stránek:

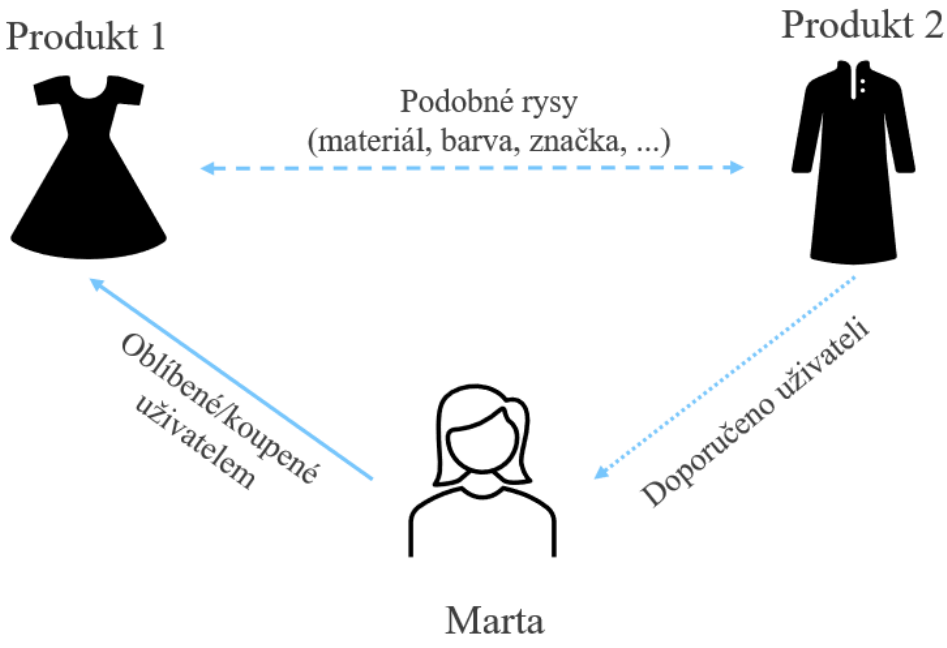
- “Cold Start” – občas kolaborativní filtrování může trpět problémem studeného startu, to znamená, že může být obtížné poskytovat doporučení pro nové uživatele nebo položky bez dostatečného množství dat. To může omezit užitečnost kolaborativního filtrování v aplikacích s novými uživateli nebo položkami.
- Citlivost na změny v chování uživatelů – jednou z nevýhod kolaborativního filtrování je také citlivé na změny v chování uživatelů v průběhu času, což může ovlivnit přesnost doporučení. Podle Aggarwala "kolaborativní filtrování není příliš účinné, pokud se preference uživatelů v průběhu času rychle mění" (Aggarwal, 2016).

2.2.2. Content-based filtrovaní

V Content-Based doporučovacích algoritmu se k doporučování používají popisné atributy položek. Tyto popisy se označují jako "obsah". V metodách, založených na obsahu či náplni, jsou informace o obsahu položky spojeny s uživatelskými hodnoceními a nákupními trendy. Uvažujme scénář, kdy Jan dal filmu „Sám doma“ vysoké hodnocení, ale nemáme přístup k názorům ostatních uživatelů jako při použiti kolaborativního filtrovaní. Tudíž popis této filmové položky má mnoho stejných žánrově příbuzných klíčových slov s jinými komediemi a rodinnými filmy, jako jsou “Noc v muzeu“ a „Matilda“. Tyto filmy mohou být Janu v takových situacích navrženy.

Obrázek 4 nám také pomůže podrobněji pochopit fungování algoritmu. Podstatou tohoto principu je, že pro každého uživatele je vytvořeno samostatné "virtuální portfolio", které zohledňuje charakteristiky položek (styl, rok, společnost, autor atd.). Portfolio se vytváří na základě preferencí uživatele nebo na základě přímého dotazu uživatele. Například, uživatel Marta kupuje sáty určitou značky v klasickém stylu, takže jí bude doporučeno další sáty a oblečení v podobném stylu a značek.

Obrázek 4. Doporučovací systém založený na metodě Content-Based filtrování



Zdroj: vlastní tvorba

Při použiti metod Content-Based filtrovaní se popisy položek, které jsou označeny hodnocením, používají jako trénovací data pro vytvoření klasifikačního nebo regresního modelovacího problému specifického pro uživatele. Pro každého uživatele odpovídají

trénovací dokumenty popisům položek, které si koupil nebo hodnotil. Třídní (nebo závislá) proměnná odpovídá zadaným hodnocením nebo nákupnímu chování. Tyto trénovací dokumenty se používají k vytvoření klasifikačního nebo regresního modelu, který je specifický pro daného uživatele (nebo aktivní uživatelku). Tento model specifický pro daného uživatele se používá k předpovědi, zda se příslušné osobě bude líbit položka, u níž není známo její hodnocení nebo nákupní chování (Aggarwal, 2016).

Metody založené na obsahu mají určité výhody při doporučování nových položek, pokud pro danou položku není k dispozici dostatek údajů o hodnocení. Je to proto, že aktivní uživatel mohl hodnotit jiné položky s podobnými atributy. Proto bude model pod dohledem schopen využít tato hodnocení ve spojení s atributy položky k vytvoření doporučení, i když pro danou položku není k dispozici historie hodnocení. Ačkoliv jako všechno má i toto své stinné stránky. Všechny výhody a nevýhody bych rád konkrétně popsal v následující podkapitole.

2.2.2.1. Výhody a nevýhody Content-Based filtrování

Výhody jakéhokoli algoritmu nebo technologie jsou důležitým faktorem při hodnocení jeho užitečnosti a účinnosti. Pokud jde o doporučovací systémy, porozumění přínosům nám může pomoci lépe ocenit hodnotu, kterou přinášejí různým odvětvím a aplikacím.

- Nezávislost od jiných uživatele – doporučující nástroje, založené na obsahu, využívají pouze hodnoceni, kteří byli poskytnutý uživatelem k vytvoření jeho vlastního profilu. Metody kolaborativního filtrování naopak potřebují hodnocení od jiných uživatelů, aby našly "nejbližší sousedy" aktivního uživatele, tj. uživatele, kteří mají podobný vkus, protože hodnotili stejné položky podobně. Poté se vyberou pouze položky, které se sousedům líbí nejvíce. aktivního uživatele, budou doporučeny.
- Transparentnost – vysvětlit, jak doporučovací systém funguje lze jednoduché díky poskytnuti velmi detailních vlastností obsahu nebo popisů. Tyto vlastnosti jsou indikátory, které je třeba konzultovat k rozhodnutí, zda doporučení důvěřovat. Naopak, kolaborativní systémy jsou černé skříňky, protože jediné vysvětlení doporučení položky je, že se daná položka líbila neznámým uživatelům s podobným vkusem.

- Používání nových položek – Content-Based algoritmy jsou schopny doporučovat položky, které dosud nebyly hodnoceny žádným uživatelem. V důsledku toho doporučeni netrpí problémem prvního hodnotitele.

Nicméně systémy založené na obsahu mají několik nedostatků:

- Přílišná specializace – jde o situaci, kdy se doporučovací algoritmus příliš soustředí na doporučování úzké množiny položek uživateli na základě jeho minulého chování nebo preferencí. K tomu může dojít, když se algoritmus příliš spoléhá na několik specifických vlastností nebo atributů uživatele nebo položek. Problémem přílišné specializace je, že může vést k nedostatečné diverzitě doporučených položek, což může omezit kontakt uživatele s novými nebo odlišnými produkty, službami nebo informacemi. To může mít za následek negativní uživatelskou zkušenost, kdy uživatel může mít pocit, že doporučení neodpovídají jeho potřebám nebo se příliš opakují.
- Nový uživatel – předtím, než Content-Based doporučující systém může skutečně pochopit preference uživatelů a poskytovat přesná doporučení, je třeba shromáždit mnoho hodnocení. Pokud je tedy k dispozici málo hodnocení, jako například u nového uživatele, systém nebude schopen poskytovat spolehlivá doporučení. (Lops, de Gemmis, & Semeraro, 2011)

2.2.3. Hybridní doporučovací algoritmus

Jedna se o typu algoritmu či systému, který kombinuje více doporučovacích algoritmu a technik poskytuje jich pro personalizovaná doporučení uživatelům. Tento přístup zlepšuje přesnost a pokrytí doporučení využitím silných stránek různých technik. V kontextu elektronického obchodování by hybridní doporučovací systém mohl kombinovat kolaborativní filtrování, Content-Based filtrování a další techniky, jako je filtrování na základě demografických údajů nebo kontextové filtrování, a poskytovat tak uživatelům přesnější a relevantnější doporučení.

V jedné studii (Ekstrand, Riedl, & Konstan, 2011) bylo zjištěno, že hybridní doporučovací systém, který kombinoval algoritmy kolaborativního a Content-Based filtrování výrazně překonaly jednotlivé přístupy na jejich souboru dat. Dospěli k závěru, že silné stránky jednotlivých technik lze kombinovat a poskytovat tak uživatelům přesnější a rozmanitější doporučení.

V další studii (Adomavicius a Tuzhilin, 2015) navrhli rámec hybridního doporučovacího systému, který kombinoval filtrování založené na obsahu (Content-Based), kolaborativní filtrování a filtrování založené na demografii. Systém nejprve použil filtrování na základě demografických údajů k identifikaci podmnožiny uživatelů s podobnými charakteristikami a preferencemi a poté použil filtrování na základě obsahu a kolaborativní filtrování k vytvoření doporučení pro každého uživatele v rámci této podmnožiny.

V příkladě, že doporučování algoritmus bude používán na webových stránkách o knihách by hybridní doporučovací systém mohl kombinovat techniky, jako je kolaborativní algoritmus na základě historie čtení, a algoritmus, založeny na Content-based filtrování podle autora, žánru nebo stylu a filtrování na základě demografických údajů podle věku, pohlaví nebo úrovně čtení. Proto uživateli, který dříve četl knihy Theodory Gossové a má rád žánr fantasy, mohou být doporučeny knihy podobných autorů, jako je Neil Gaiman, a také knihy, které jsou oblíbené mezi uživateli podobného věku a čtenářské úrovně.

Celkově lze říci, že hybridní doporučovací systém, který kombinuje více technik, může zlepšit přesnost a pokrytí doporučení a poskytnout uživatelům personalizovanější a uspokojivější zážitek.

2.3. Aplikace v moderní e-commerce

V dnešním světě se stálé používání internetu, nakupovaní v e-shopech a trávení večerů na streamovacích službách místo sledování televize nebo poslechu rádia stalo významnou součástí života. Teď skoro všichni si mohou zakoupit miliony produktů a konzumovat neuvěřitelné množství informací na jedno kliknutí. S tolika dostupnými možnostmi však může být pro zákazníky vyhledávání hledaných produktů zahlcující. V této situaci uživatelům pomáhají doporučovací algoritmy zabudované do online služeb.

V současné době doporučovací systémy využívají především velké společnosti s obrovskými rezervami zdrojů, stejně jako společnosti, které jsou ochotny experimentovat s různými algoritmy, umělou inteligencí a strojovým učením. Mezi ti, kteří aktivně využívají a inovují v této oblasti, patři téměř všechny populární sociální sítě, například, Facebook, Instagram, Twitter, LinkedIn, stejně jako streamovací služby – Apple Music, Spotify, Netflix, Disney+, a co je nejdůležitější e-shopy – Amazon, Alza, Aukro atd. Doporučovací systémy, jak bylo uvedeno, se používají v tolika oblastech, že je každý

alespoň jednou použil. V následující tabulce bych například rád představila některé služby, kde je můžete najít a jak jsou prezentovány. Tyto příklady samozřejmě nejsou všechny, ale i kdybych napsal všechny věci, které mě napadnou, mohl by tento seznam pokračovat velmi dlouho.

Tabulka 1. Jak jsou doporučovací algoritmy prezentovány online

Webová stránka	***Doporučovaní***
YouTube	Videa
Netflix, HBO Max, Disney+, Hulu atd.	Filmy a série
AppStore, Google Play Market	Aplikace a mobilní hry
Jobs.cz, LinkedIn	Inzeráty, Práce
Facebook	Návrhy přátel, Příspěvky, Reklama
Amazon	Knihy a další zboží
Aukro	Zboží
Goodreads	Knihy
Google News	Noviny
Twitter	Tweety, Uživateli

Zdroj: vlastní tvorba

V této podkapitole kapitole bych ráda analyzovala, jak doporučovací algoritmy těchto populárních služeb fungují ze strany uživatele, a pokud možno se podíval, jak fungují v praxi.

2.3.1.Amazon

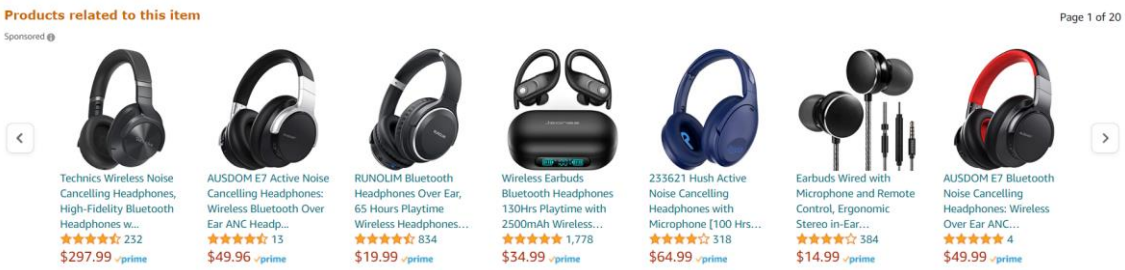
Amazon je přední platforma elektronického obchodování, která způsobila revoluci v nakupování produktů online. Jedním z klíčových prvků, kterými se Amazon odlišuje od ostatních internetových obchodů, je využívání doporučovacích algoritmů k poskytování personalizovaných doporučení produktů svým zákazníkům. Tyto doporučovací algoritmy jsou navrženy tak, aby analyzovaly chování zákazníků, jejich preference a historii nákupů a navrhovaly jim produkty, které budou pro každého jednotlivého zákazníka s největší pravděpodobností zajímavé.

Jednou z hlavních výhod doporučovacích algoritmů společnosti Amazon je, že zákazníkům umožňují plynulejší a příjemnější nakupování. Díky personalizovaným doporučením je Amazon schopen pomoci zákazníkům objevit nové produkty, které by jinak možná nenašli. To může zvýšit loajalitu zákazníků a podpořit opakované nákupy. Kromě toho je společnost Amazon díky navrhování relevantních produktů zákazníkům schopna zvýšit prodeje a příjmy.

Používání algoritmů má však i potenciální nevýhody. Někteří zákazníci se mohou cítit nepříjemně z úrovně údajů, které Amazon shromažďuje o jejich chování a preferencích. Navíc existuje riziko, že doporučovací algoritmy mohou posílit stávající předsudky.

Doporučovací algoritmy e-shopu Amazon jsou založeny na různých technikách, včetně kolaborativního filtrování, Content-Based filtrovaní, User-Based kolaborativního filtrování a personalizovaného řazení (Linden, Smith, & York, 2003). Abych vyzkoušet, jak výše uvedené algoritmy fungují, šla jsem na webové stránky Amazon a rozhodla se vyhledat produkt, o který jsem měla zájem. Tímto předmětem pro mě byla bezdrátová sluchátka od společnosti Sony. Po přečtení a prostudování vlastností a charakteristik produktu jsem jako běžný uživatel webu a potenciální kupující posunula jsem se dolů na konec stránky, abych si přečetla recenze. Tehdy jsem narazila na sekci doporučených produktů.

Obrázek 5. Sekce doporučování „Produkty související s touto položkou“

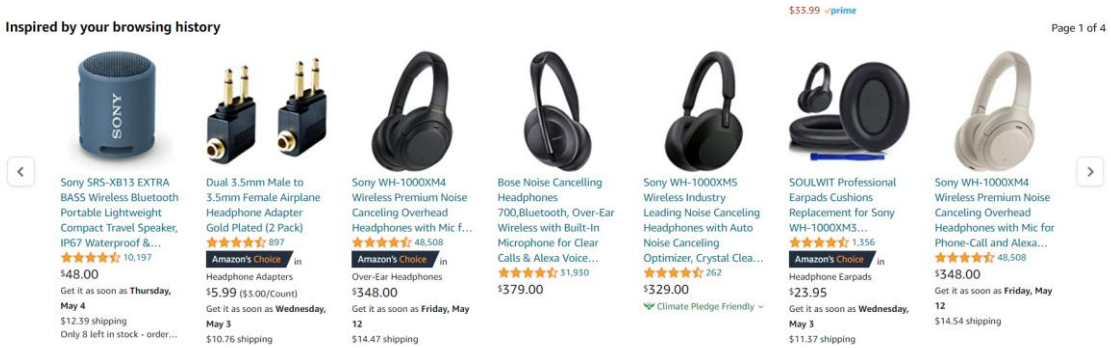


Zdroj: vlastní tvorba

Na obrázku můžete vidět jednu ze sekci pro doporučovaní produktů s názvem „Produkty související s touto položkou“. Je to sekce, která se zobrazuje pod popisem a podrobnostmi o produktu, obvykle na pravé straně stránky. V této části se obvykle můžete vidět seznam dalších produktů, které souvisejí s položkou, kterou si zákazník právě prohlíží. Jsou obvykle podobné produktu, který si zákazník prohlíží, pokud jde o kategorii, funkce nebo zamýšlené použití. Mohou to být také produkty, které si jiní zákazníci zakoupili v souvislosti s položkou, kterou si zákazník právě prohlíží.

Sekce "Produkty související s touto položkou" má zákazníkům pomoci objevit další produkty, o jejichž koupi by mohli mít zájem, a povzbudit je k prozkoumání dalších možností na Amazon. Zobrazením souvisejících produktů se Amazon snaží zvýšit angažovanost zákazníků a prodej a také poskytnout individuálnější nákupní zážitek pro každého jednotlivého zákazníka.

Obrázek 6. Sekce doporučováni "Inspirováno vaší historií prohlížení"



Zdroj: vlastní tvorba

Na obrázku výše můžete vidět Sekce "Inspirováno vaší historií prohlížení" na Amazonu. Je to personalizovaná sekce, která se zobrazuje na domovské stránce nebo na stránkách kategorií, hned pod hlavním navigačním panelem. V této sekci se zobrazuje výběr produktů, které souvisejí se zbožím, které si zákazník nedávno prohlížel nebo hledal na Amazon. Produkty zobrazené v této sekci jsou určeny doporučovacími algoritmy společnosti Amazon, které analyzují historii prohlížení a nákupů zákazníka a navrhují mu produkty, které ho s největší pravděpodobností zajímají. Algoritmy při vytváření těchto doporučení berou v úvahu faktory, jako je historie předchozích nákupů zákazníka, vyhledávací dotazy a hodnocení.

Sekce "Inspirováno historií prohlížení" je navržena tak, aby zákazníkům pomohla objevit nové produkty, o jejichž nákup by mohli mít zájem. Poskytováním personalizovaných doporučení na základě historie prohlížení stránek zákazníka se společnost Amazon snaží vytvořit zajímavější a příjemnější nákupní zážitek pro každého jednotlivého zákazníka. Je důležité si uvědomit, že produkty zobrazené v této sekci nejsou omezeny na položky, které zákazník již dříve zakoupil. Místo toho se v této sekci mohou zobrazovat i položky, které si zákazník prohlédl nebo přidal do košíku, ale ještě je nezakoupil. Kromě toho se produkty zobrazené v této sekci mohou v průběhu času měnit podle toho, jak se vyvíjí historie prohlížení a nákupů zákazníka.

Abych to shrnula, společnost Amazon používá doporučovací algoritmy k poskytování personalizovaných doporučení produktů svým zákazníkům. Tato doporučení

se mohou zobrazovat v různých sekcích na webových stránkách, například "Produkty související s touto položkou" a "Inspirováno vaší historií prohlížení". Algoritmy také analyzují chování zákazníků, jejich preference a historii nákupů, aby jim navrhly produkty, které budou pro každého jednotlivého zákazníka s největší pravděpodobností zajímavé. Poskytováním personalizovaných doporučení je e-shop Amazon schopen pomoci zákazníkům objevovat nové produkty, zvyšovat jejich loajalitu a podporovat opakované nákupy.

2.3.2. Netflix

Pokud jde o doporučovací algoritmy a systémy, nejčastěji se uživatelům internetu vybaví Netflix. což není nic divného, protože Netflix je králem všech – mezi spotřebiteli je s velkým náskokem nejoblíbenější (Floorwalker, 2021). Podle obrovského množství internetových portálů a kritiků je doporučovací systém této streamovací služby nejlepší, pokud jde o kvalitu algoritmů, velikost knihovny pořadů a filmů, uživatelský komfort.

Netflix je online streamovací platforma, která svým předplatitelům nabízí rozsáhlou knihovnu filmů a televizních pořadů. Při tak rozsáhlé sbírce obsahu může být pro uživatele nepřehledné rozhodnout se, co si pustit příště. K vyřešení tohoto problému vyvinula společnost Netflix doporučovací systém, který svým uživatelům navrhuje personalizovaný obsah na základě jejich historie sledování, hodnocení a dalších faktorů.

Společnost Netflix byla založena v roce 1998 jako služba půjčování DVD poštou (Floorwalker, 2021). V této fázi je doporučovací systém společnosti založen na historii půjčování a vyhledávacích dotazech zákazníka a nyní je nejrozšířenější známou streamovací službou, která umožňuje uživatelům sledovat filmy a televizní pořady online.

2.3.2.1. Algoritmus

Doporučovací systém Netflixu se v průběhu let výrazně vyvinul a zahrnuje nové technologie a chování uživatelů, aby poskytoval přesnější a personalizovanější doporučení. Podle článku v časopise Wired je doporučovací systém společnosti Netflix založen na kombinaci algoritmů strojového učení, analýzy dat a chování uživatelů (Plummer, 2017). Systém funguje na základě analýzy historie sledování a hodnocení milionů uživatelů a následně na základě těchto dat identifikuje vzorce a podobnosti mezi uživateli a obsahem. Doporučovací systém využívá k navrhování obsahu techniky kolaborativního filtrování a filtrování na základě obsahu. Kolaborativní filtrování je

založeno na předpokladu, že lidé, kteří mají podobné zvyky při sledování, budou mít podobné preference. Filtrování založené na obsahu naproti tomu analyzuje samotný obsah a hledá podobnosti v tématech, žánrech a dalších atributech.

Jedním z klíčových aspektů doporučovacího systému společnosti Netflix je využití algoritmů hlubokého učení. V článku na serveru Medium společnost Netflix vysvětluje, že tyto algoritmy se používají k analýze obrazových, zvukových a textových dat za účelem navrhování obsahu (Maity, 2020). Může například využívat rozpoznávání obrazu k analýze vizuálních prvků filmu nebo televizního pořadu a na základě těchto prvků navrhovat podobný obsah. Tyto algoritmy se také používají k identifikaci vzorců chování uživatelů, například které pořady jsou sledovány v režimu “binged-watched“ a které jsou sledovány delší dobu.

Dalším důležitým aspektem doporučovacího systému společnosti Netflix je důraz na personalizaci. Jak je uvedeno v článku v článku Zacha Bulyga, doporučovací systém společnosti Netflix je navržen tak, aby poskytoval personalizované návrhy každému jednotlivému uživateli (Bulyga, n.d.). Toho je dosaženo pomocí sofistikovaného algoritmu, který zohledňuje historii sledování, hodnocení a další faktory uživatele. Algoritmus se neustále učí a přizpůsobuje, takže doporučení jsou postupem času stále přesnější. Tato personalizace je klíčovým faktorem úspěchu Netflixu, protože pomáhá udržet uživatele v zájmu a spokojenosti s obsahem platformy.

2.3.2.2. Doporučovaní v hlediska používatele

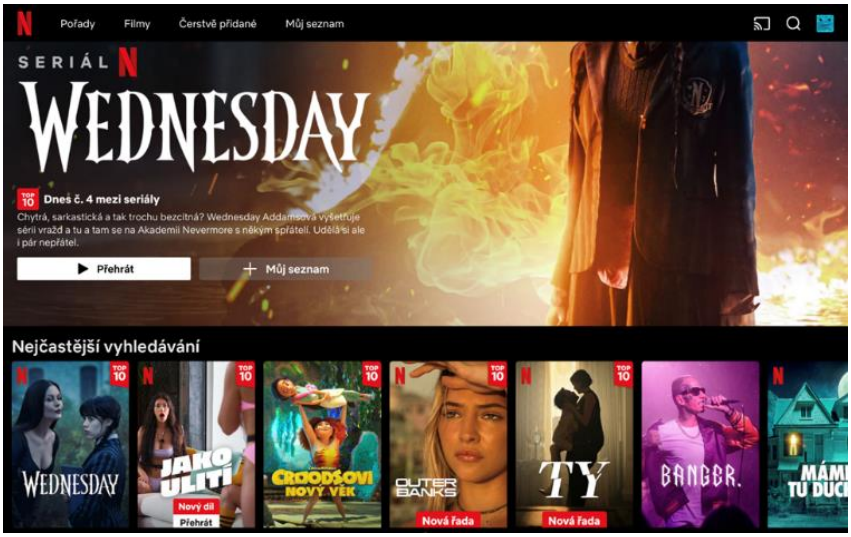
Začněme se seznamovat se systémem doporučování služeb Netflix z hlediska spotřebitele.

Při vytváření nového účtu nebo nového profilu ve stávajícím účtu nejsou jasné zájmy uživatele, proto doporučovací systém platformy začíná s čistým štítem a nemá žádné informace o historii sledování nebo preferencích uživatele. Z tohoto důvodu se nejprve zobrazí populární obsah – Netflix může na základě chování ostatních uživatelů doporučovat obsah, který je na platformě aktuálně populární. Jak vidíme na obrázku 7, na domovské stránce aplikaci právě vytvořeného před chvíli profilu byl doporučen pořad "Wednesday". Jednalo se o jeden z nejoblíbenějších seriálů roku 2022, takže dává smysl, aby se poprvé zaměřil na něco, co má většina ráda. Je důležité zmínit, že tento seriál je také produktem studia Netflix, což také znamená, že pořady vytvořené platformou budou doporučovány lidem bez předchozí historie. Avšak, Systém však rychle začne shromažďovat údaje na základě chování uživatele na platformě. Takže se nám zobrazí

standardní složky v závislosti na vaší poloze, např. oblíbené pořady a filmy v České republice a Top-10 dne, spolu s obecně oblíbenými projekty na webu.

Čím více službu používáte, tím více informací o vás shromažďuje o tom, co sledujete, na jakých zařízeních (počítač, tablet, telefon, televize atd.), v jakém čase a jak dlouho. K výběru doporučení se používají informace o tom, jak se službou pracujete, jaké jsou vaše zvyky při sledování a jak ji hodnotíte.

Obrázek 7. Hlavní stránka servisu Netflix pro nového uživatele



Zdroj: vlastní tvorba

Podívejme se také na doporučení na mém účtu, kde o mně služba již několik let shromažďuje údaje. Po přihlášení do aplikace a výběru účtu se na úvodní stránce může zobrazit jedna z věcí, které vám služba nabízí ke sledování. To může záviset na tom, zda je navrhovaný titul nový, nebo zda se jedná o něco, co vychází pouze z vašich zájmů. Poté se vám mohou zobrazit kategorie vybrané speciálně pro vás. Při každém vstupu do aplikace mohou měnit své umístění, které je rovněž založeno na doporučení algoritmu. Například, v tuto chvíli je první pro mě řadou doporučená kategorie Americké TV seriály, protože jsem právě dokoukal seriál tohoto žánru, konkrétně The Office či Kancl v překladě na český.

Obrázek 8. Kategorie „Americké TV seriály“, doporučená servisem Netflix

Zdroj: vlastní tvorba

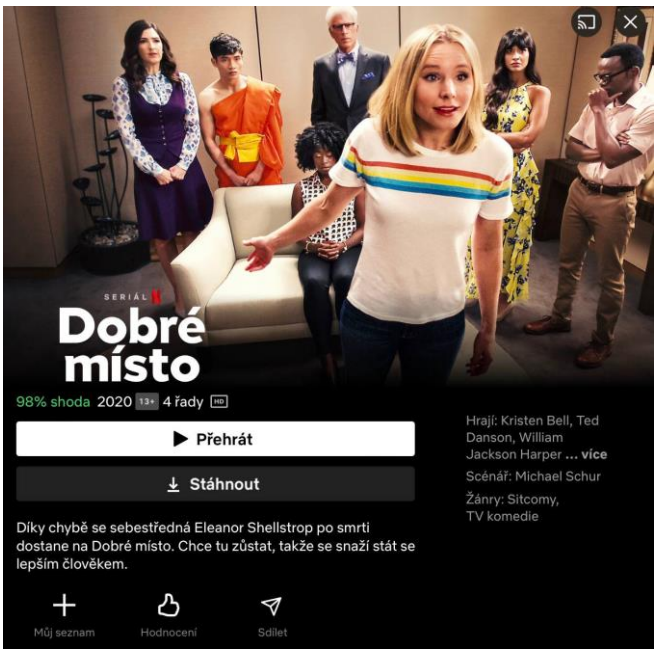
Jak vidíme, na prvním místě je seriál z produkce Netflixu, co nám říká, že Netflix se vždy snaží propagovat pořady vlastní produkce. Totéž lze říci o jednotlivých kategoriích, kteří pomocí nabízených kategorií propaguji série a filmy výroby Netflix, případně Uvádí jedině Netflix, nebo Netflix Originals v původním jazyce. To je pro tým důležité, protože Netflix vynakládá spoustu peněz na výrobu vlastního obsahu a v tomto případě také šetří peníze, protože služba nemusí platit peníze vlastníkům obsahu (jiným studiím apod.) (Falk, 2019).

Obrázek 9. Kategorie „Uvádí jedině Netflix“, doporučená algoritmem Netflix

Zdroj: vlastní tvorba

Navíc pokud se rozhodnete najet myší na jeden z poskytnutých plakátů na počítačové zařízení nebo kliknout na obrázek na mobilním zařízení, zobrazí se Vám procento podobnosti s vaším vkusem (přes obrázek 10. – 98 % shoda), které bylo vypočítáno jedním z algoritmů. Takým způsobem Netflix prohledal žebříček obsahu a vypočítal procentuální pravděpodobnost, že se film nebo seriál shoduje s vaším vkusem.

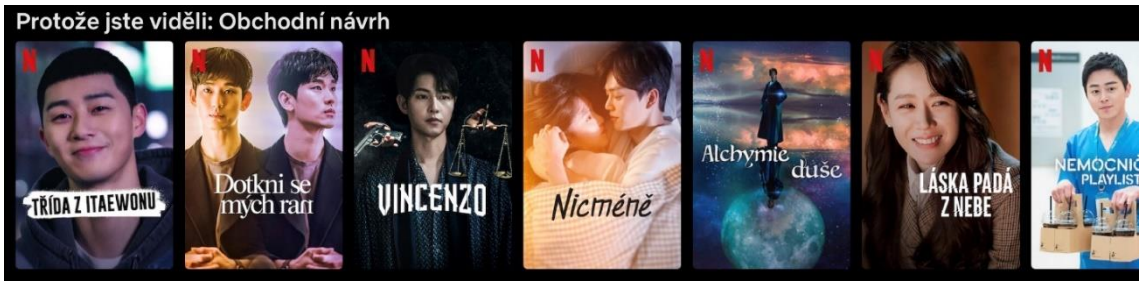
Obrázek 10. Přehled konkrétního doporučení



Zdroj: vlastní tvorba

Dalším příkladem doporučení je, že služba může také vytvořit kategorii, která vám doporučí něco po zhlédnutí určitého pořadu. Například, nějakou dobu poté, co jsem sledovala seriál „Obchodní návrh". Má to žánr K-drama, a mi služba doporučila kategorii „Protože jste viděli: „název seriálu"", v mem případě „Protože jste viděli: Obchodní návrh". Tento seznam doporučení většinou obsahoval seriály se stejnými žánry – K-Drama a Romantické TV Drama. Z toho vyplývá, že algoritmus doporučuji také podobné k-dramatické seriály, což není překvapivé, protože Netflix v poslední době na tyto seriály vynakládá hodně prostředků.

Obrázek 11. Přehled doporučené kategorie „Protože se viděli"



Zdroj: vlastní tvorba

Jednou z nepovinných informací, které můžete existuje na platformě, je hodnocení každého kusu media, pokud jste libí anebo ne.

Tlačítka "To se mi líbí" a "To se mi nelíbí" jsou na Netflixu klíčovou součástí doporučovacího systému platformy. Když uživatel ohodnotí film nebo televizní pořad palcem nahoru nebo dolů, algoritmus Netflixu tuto informaci zohlední při vytváření budoucích doporučení.

Obrázek 12. Nastroj hodnocení na servisu Netflix

Zdroj: vlastní tvorba

Algoritmus používá k vytváření doporučení kombinaci kolaborativního filtrování a filtrování na základě obsahu. Kolaborativní filtrování zkoumá historii sledování uživatele a porovnává ji s ostatními uživateli s podobnými zvyklostmi, zatímco filtrování podle obsahu zkoumá charakteristiky samotného obsahu, jako je žánr, herci a děj.

Když uživatel ohodnotí film nebo televizní pořad palcem nahoru, algoritmus Netflixu si všímá charakteristik obsahu, které se uživateli líbily, a zapracuje je do budoucích doporučení. Pokud, například, uživatel dá palec nahoru romantické komedii s určitým hercem, může mu algoritmus doporučit další romantické komedie se stejným hercem.

Naopak, pokud uživatel ohodnotí film nebo televizní pořad palcem dolů, algoritmus vezme na vědomí vlastnosti obsahu, které se uživateli nelíbily, a v budoucnu se vyhne doporučování podobného obsahu.

2.3.3. Spotify

Spotify je služba digitálního streamování hudby, která způsobila revoluci v poslechu hudby. Byla založena v roce 2008 a od té doby se stala jednou z nejoblíbenějších platforem pro streamování hudby na světě. Je jedna ze společností zodpovědných za globalizaci hudby na internetu: tato společnost vznikla jako start-up, což byl nápad lidí, kteří se rozhodli nahrávat hudbu do cloudového úložiště, aby k ní měli lepší přístup. Úspěch služby Spotify lze přičíst jejím inovativním funkcím, které uživatelům usnadňují objevování nové hudby a vytváření personalizovaných playlistů.

Díky úspěchu služby je v současnosti k dispozici více než 100 milionů skladeb, jejichž počet se každým dnem zvyšuje, a ani ten nejvášnivější milovník hudby nemůže poslechnout všechny skladby ve streamovací službě, protože by mu to trvalo déle než celý život. Právě k tomuto účelu slouží doporučovací algoritmy v tomto systému.

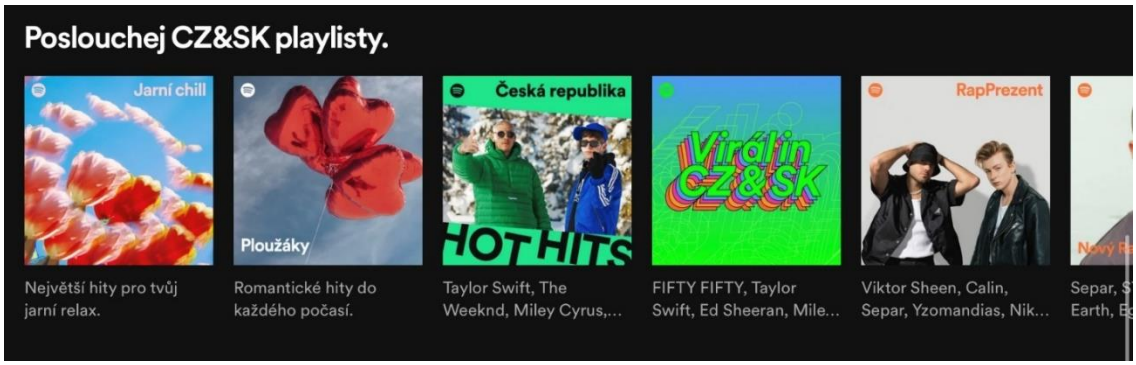
Jednou z nejvýznamnějších inovací Spotify je personalizace. Společnost Spotify využívá analýzu dat a algoritmy strojového učení k vytváření personalizovaných playlistů pro své uživatele.

2.3.4.1. Doporučovaní v hlediska používatele

Když se poprvé zaregistrujete na této platformě a dáte základní údaje o svém účtu, stejně jako v předchozí podkapitole, podobně Netflixu, algoritmus nebude mít žádné údaje o uživateli. Ale aplikace stále začne s doporučeními, jako například je seznam se

skladbami oblíbenými ve vaší zemi, a to v našem případě Česká republika a Slovenská republika. Mohou tam být také doporučeny playlisty podle ročního období – pokud si účet zaregistrujete v zimě, tak uvidíte například vánoční seznamy skladeb. Tento seznam doporučených playlistů však není osobně přizpůsoben vašim zájmům, ale stále pomůže shromáždit informace o vás pro další nejlepší doporučení.

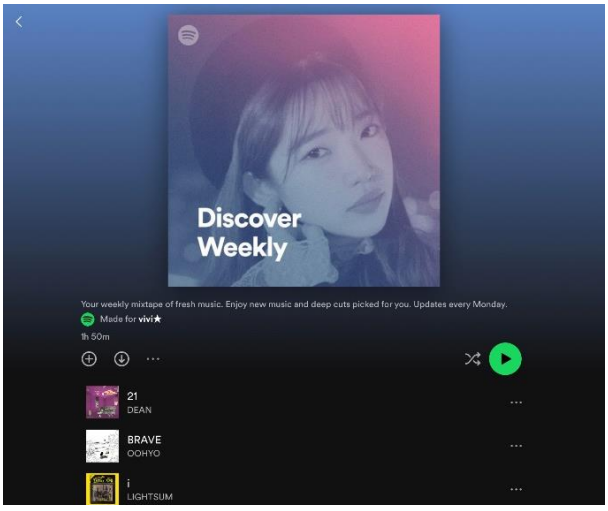
Obrázek 13. Kategorie doporučení pro země Česko a Slovensko servisu Spotify



Zdroj: vlastní tvorba

Většina uživatelů služby si na personalizovaných doporučeních pochvaluje playlist "Discover Weekly". Podle popisu služby z obrázku 14, která je "Tvůj týdenní mixtape s čerstvou hudbou. Novinky a lahůdky z archivu vybrané speciálně pro tebe." Playlist je aktualizován vždy na začátku týdne – v pondělí, a jeho hlavním účelem je seznámit uživatele s novou hudbou a umělci. Může obsahovat skladby umělců, které právě vyšly a jsou ve vašich oblíbených, a také skladby umělců, jejichž žánry jsou stejné jako ty, které posloucháte. Hlavní výjimkou oproti všem ostatním playlistům Spotify je, že dané skladby jste nikdy neposlouchali.

Obrázek 14. Přehled nástroje doporučeni „Discover Weekly“



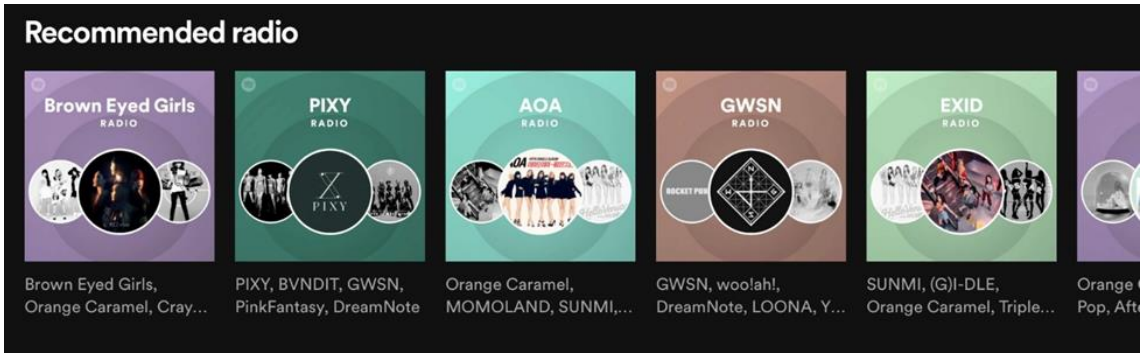
Zdroj: vlastní tvorba

Doporučení jsou pro tento seznam skladem generována, jako i pro vše doporučení v aplikace, pomocí kombinace kolaborativního filtrování a technik zpracování přirozeného jazyka. Kolaborativní filtrování se používá k identifikaci ostatních uživatelů s podobnými poslechovými zvyklostmi a k doporučování skladeb, které se těmto uživatelům líbily. Zpracování přirozeného jazyka se používá k analýze textů a hudebních charakteristik písní, aby bylo možné vytvářet diferencovanější a personalizovanější doporučení.

Dalším nástrojem v rámci platformy pro streamování hudby Spotify, který uživatelům poskytuje personalizovaná hudební doporučení na základě jejich historie poslechu a preferencí, je Rádio. Dle vlastního výběru existuji rádia umělce, skladby a alba. Použití této funkce uživatelé zahájí výběrem některé z dostupných možností, většinou skladby, kolem které chtějí vytvořit rádio. Po výběru skladby použije služba Spotify svůj algoritmus k analýze zvukových vlastností dané skladby, jako je tempo, rytmus a instrumentace. Na základě těchto zvukových vlastností vytvoří služba Spotify vlastní rádio playlist, který přehrává podobné písničky. Při poslechu rádio máte možnost poskytnout zpětnou vazbu k přehrávaným skladbám tím, že uvedete, zda se vám jednotlivé skladby líbí, nebo nelíbí. Na základě vaší zpětné vazby rádio Spotify upraví výběr přehrávaných skladeb tak, aby lépe odpovídal vašim preferencím.

Rádio si můžete vytvořit podle vlastního uvážení a také vám může být nabídnuto na panelu na domovské obrazovce aplikace, jak vidíte na obrázku 15. V této situaci mi algoritmus zodpovědný za to, jak vypadá domovská obrazovka, nabídl několik rádií podle interpretů a kapel, které na platformě poslouchám poměrně často.

Obrázek 15. Doporučení radia umělců servisu Spotify

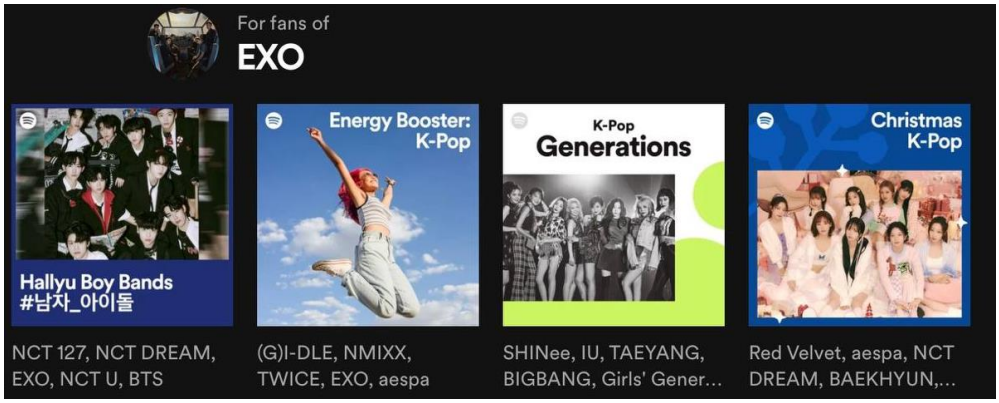


Zdroj: vlastní tvorba

Další funkce doporučení "pro fanoušky" ve službě Spotify má uživatelům pomoci objevit novou hudbu, která je podobná jejich oblíbeným interpretům nebo kapelám. Vytváří se seznam doporučených skladeb a interpretů na základě historie poslechu

uživatele, včetně jeho oblíbených skladeb, alb a interpretů. Když uživatel přejde na stránku umělce ve službě Spotify, zobrazí se mu část označená "Fanouškům se také líbí", která obsahuje seznam podobných umělců, playlistů, alb atd. Umožňuje uživatelům objevovat novou hudbu podobnou jejich oblíbeným interpretům a žánrům. Po kliknutí na profil umělce se zobrazí část označená jako to může být skvělý způsob, jak najít novou hudbu, která by se vám mohla líbit na základě vašich stávajících preferencí.

Obrázek 16. Přehled kategorii „For fans of " servisu Spotify



Zdroj: vlastní tvorba

Uvedl jsem téměř všechny způsoby, jakými Spotify doporučuje skladby a plazlisty, ale v poslední době platforma propaguje nejen svou hudbu, ale také některé podcasty. Doporučovací algoritmus Spotify pro podcasty je podobný jako u hudby, analyzuje chování a preference uživatelů a vytváří personalizovaná doporučení. Jak uživatelé poslouchají více podcastů a poskytují zpětnou vazbu k doporučením, algoritmus se zpřesňuje a přizpůsobuje jejich preferencím.

2.4. Závěr teoretické kapitoly

V této teoretickou kapitole bakalářské práce jsem probrala základy doporučovacích algoritmů, jejich typy a použití v elektronickém obchodě. Zjistila jsem, co to vlastně jsou doporučovací algoritmy, a jak souvisí s doporučovacími systémy, jak v základních krocích funguje. Důležitým krokem základního pochopení je z mého pohledu pochopení etických hodnot doporučovacích algoritmů. Tyto podkapitoly se týkaly všech běžných známých strohých stránek a hrozeb těchto algoritmů. Poté byly probrány různé typy doporučovacích algoritmů, včetně kolaborativního filtrování, filtrování na základě obsahu (Content-Based filtrovaní) a hybridních přístupů, které kombinují více technik. U každého typu doporučovacího algoritmu jsme uvedli stručný přehled jeho fungování a jeho silné a slabé stránky. Dále jsme upozornili na aplikace doporučovacích algoritmů v

elektronickém obchodě. Zmínila jsem, jak se algoritmy hojně využívají v online tržištích, na streamovacích platformách atd. Probrala jsem, jak mohou doporučovací algoritmy zlepšit uživatelskou zkušenost, zvýšit zapojení zákazníků a podpořit prodej v prostředí elektronického obchodu.

Překvapením pro mě bylo, jak běžné jsou dnes doporučovací algoritmy a systémy. Obklopují vás téměř každou minutu vašeho online sezení. Staly se všudypřítomnými v našich online zkušenostech a často se s nimi setkáváme, aniž bychom si to uvědomovali. Od personalizovaných doporučení produktů na stránkách e-shopů po navrhovaná videa na YouTube a Netflixu – doporučovací algoritmy neustále pracují v zákulisí, analyzují naše data a poskytují nám relevantní a zajímavý obsah.

3. Metodika

Účelem této kapitoly je poskytnout podrobný popis metodiky použité při vývoji algoritmu. To čtenáři umožní pochopit proces a kroky, které byly podniknuty při návrhu a vývoji algoritmu. Cílem kapitoly je také poskytnout vhled do problematik, které se byli vyskytnuty během procesu vývoje. Kapitola o metodice začíná popisem použitého kompilátoru kódu, programovacího jazyka a nejpoužívanější knihovny s názvem Surprise. Popsány jsou také použité k hodnocení přesnosti algoritmů hodnoty chyb. To umožní pochopit proces a provedené kroky.

3.1. Google Colab

Google Colaboratory, známá také jako Google Colab, je bezplatná online platforma pro vývoj a spouštění algoritmů, zejména strojového učení, a dalších projektů datové vědy. Jedná se o cloudovou platformu, což znamená, že uživatelé nemusí instalovat žádný software ani hardware do svých počítačů. Místo toho uživatele mohou ke službě Colab přistupovat prostřednictvím webového prohlížeče na jakémkoli zařízení s připojením k internetu. Systém umožňuje uživatelům psát a spouštět kód v jazyce Python, ukládat a sdílet zápisníky kódu a spolupracovat s ostatními uživateli v reálném čase. Lze do Colab nahrávat svá vlastní data, a to buď přímo do platformy, nebo připojením služby ke svému účtu na Disku Google. To umožňuje snadné ukládání, vyhledávání a sdílení dat.

Tento kompilátor kodu je dodáván s několika předinstalovanými knihovnami, navíc uživatelé mohou také nainstalovat další knihovny pomocí příkazů *pip*. Colab poskytuje uživatelům GPU (Graphics Processing Unit či Grafický Procesor), CPU (Central Processing Unit či Centrální Procesorová Jednotka) nebo TPU (Tensor Processing Unit či Jednotka pro Zpracování Tenzorů od společnosti Google) pro trénování modelů hlubokého učení, což je užitečné zejména pro výpočetně náročné úlohy. Google Colab je vynikající nástroj pro datové vědce, výzkumníky a studenty, kteří nemají přístup ke špičkovému hardwaru a potřebují provádět rozsáhlé experimenty. Umožňuje také bezproblémovou spolupráci a sdílení kódu, takže je ideální platformou pro týmové projekty nebo příspěvky do open source.

Platforma Google Colaboratory je výkonná a všestranná, díky snadnému používání, funkcím pro spolupráci a integraci s dalšími službami Google je ideálním nástrojem pro každého, kdo potřebuje pracovat na projektech datové vědy nebo strojového učení.

Použila jsem tento kompilátor kódu kvůli snadnému procesu instalace knihoven pro její používání si získal místo v mém výzkumu. Navíc, z důvodu že můj počítač není v současné době nejmodernější ani nejvýkonnější, takže možnost používat cloudovou platformu pro mě byla ideální možnost.

3.2. Python

Python je vysokoúrovňový programovací jazyk s otevřeným zdrojovým kódem, který vytvořil Guido van Rossum a který byl poprvé vydán v roce 1991. Python je známý svou jasnou syntaxí a čitelností, což z něj činí ideální jazyk pro začátečníky (Pilgrim, 2014). Jedná se o objektově orientovaný programovací jazyk, který podporuje více programovacích paradigmat, včetně funkcionálního a imperativního programování.

Jednou z hlavních výhod Pythonu je jeho rozsáhlá knihovna vestavěných funkcí a modulů, která vývojářům umožňuje snadno implementovat složité algoritmy a datové struktury Python má také rozsáhlou sbírku knihoven a frameworků třetích stran, včetně NumPy pro vědecké výpočty a mnoha dalších. Python je multiplatformní jazyk, což znamená, že jej lze používat na více operačních systémech, například Windows, MacOS a Linux. Je to také interpretovaný jazyk, což znamená, že kód jazyka Python se provádí řádek po řádku bez nutnosti kompilace (Pilgrim, 2014).

Dostupnost velkého množství knihoven je pro jazyk Python velkou výhodou. Jejich používání může vývojářům ušetřit značné množství času a úsilí při psaní kódu od nuly, protože poskytují předem připravená a otestovaná řešení pro běžné programovací úlohy. Umožňují také vývojářům vytvářet složité aplikace s menším počtem řádků kódu, což urychluje a zefektivňuje proces vývoje.

Tento programovací jazyk byl pro bakalářskou práci vybrán díky svému bohatému ekosystému knihoven a nástrojů, velké komunitě vývojářů, která je ochotna programátorům pomáhat na mnoha sociálních sítích, a také díky své flexibilitě a všestrannosti.

Při psaní praktické části bakalářské práce budeme k dosažení našich cílů využívat řadu knihoven programovacího jazyku Python, ale největší vliv bude mít knihovna Surprise, kterou chci podrobněji popsat v následující podkapitole.

3.2.1. Knihovna Surprise

Surprise je knihovna pro Python, která se používá k vytváření a analýze doporučovacích systémů. Vytvořil ji Nicolas Hug s hlavní předností na poskytovaní

jednoduché a snadno použitelné rozhraní API pro vývoj a testování algoritmů kolaborativního filtrování.

Knihovna poskytuje několik předpřipravených algoritmů pro kolaborativní filtrování, včetně: algoritmy založené na faktorizaci matic, jako je singulární rozklad (Singular Value Decomposition nebo SVD) a faktorizace nezáporných matic (Non-negative Matrix Factorization nebo NMF), algoritmy založené na sousedství, jako jsou K-Nearest Neighbours (KNN) a BaselineOnly, SlopeOne – jednoduchý a rychlý algoritmus, který se často používá jako základní pro srovnání (Hug, 2020).

Knihovna rovněž poskytuje nástroje pro vyhodnocování a porovnávání výkonnosti těchto algoritmů, včetně různých metrik, jako je střední kvadratická chyba (RMSE), střední absolutní chyba (MAE) a další metriky.

Knihovnu Surprise lze použít pro různé aplikace, například pro doporučování filmů nebo hudby, personalizované zpravodajské kanály a doporučování produktů v elektronickém obchodě. Dokáže také efektivně zpracovávat řídké a rozsáhlé soubory dat, takže je vhodný pro scénáře z reálného světa.

Celkově je Surprise výkonná a flexibilní knihovna pro vytváření a vyhodnocování doporučovacích systémů a může být cenným nástrojem pro datové vědce, inženýry strojového učení a výzkumníky, kteří pracují na doporučovacích systémech.

3.3. MovieLens 10M

Datová sada MovieLens je oblíbenou srovnávací datovou sadou pro výzkum doporučovacích systémů. Jedná se o sbírku hodnocení filmů a metadat o filmech, která byla v průběhu let shromážděna z webových stránek MovieLens. Tato datová sada byla široce využívána při výzkumu doporučovacích algoritmů a sloužila jako měřítko pro hodnocení výkonnosti doporučovacích modelů. Datová sada MovieLens byla poprvé zveřejněna v roce 1997 výzkumnou skupinou GroupLens Research z Minnesotské univerzity (Harper & Konstan, 2015). Datová sada byla v průběhu let několikrát aktualizována, přičemž poslední verzí je MovieLens 25M, vydaná v roce 2019. Data set obsahuje různé velikosti, včetně původního data setu se 100 000 hodnoceními a dalších s 1M, 10M, 20M a 25M hodnoceními.

Pro svou práci jsem použila databázi 10 milionů, abych získala přesnější hodnoty. K mé velké lítosti jsem po pokusu o práci s databází 20 milionů hodnocení zjistil, že tak velký objem nebudu schopen zpracovat co nejpřesněji kvůli omezením mého počítače.

Soubor dat MovieLens 10M obsahuje přibližně 10 milionů hodnocení filmů, která poskytlo zhruba 70 000 uživatelů k víc než 10 000 filmům. Hodnocení jsou na stupnici od 1 do 5, přičemž 5 je nejvyšší hodnocení, s půlhvězdičkovými přírůstky (0,5 hvězdičky - 5,0 hvězdičky). Soubor dat obsahuje další informace o filmech, například název filmu, rok vydání a žánry. Data byla shromážděna v období od ledna 1995 do září 2015, a v databázi jsou použity pouze údaje o uživatelích a filmech, které mají 20 nebo více hodnocení.

3.4. Směrodatná odchylka chyb

RMSE (Root-Mean-Square Deviation či směrodatná odchylka chyb) je standardní způsob měření chyby modelu při předpovídání kvantitativních dat. Formálně je definována jako následující vzorec:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}i - yi)^2}{n}}$$

Zde jsou:

- Zde $\hat{y}1, \hat{y}2, ..., \hat{y}n$ jsou předpokládané hodnoty.
- $y1, y2, ..., yn$ jsou pozorované hodnoty.
- n je počet pozorování.

RMSE měří průměrný rozdíl mezi předpovídanými hodnoceními a skutečnými hodnoceními, která uživatelé udělili souboru položek. Čím nižší je hodnota RMSE, tím algoritmus lépe předpovídá hodnocení položek uživateli, a proto je přesnější.

3.5. Průměrná absolutní chyba

Průměrná absolutní hodnota nebo Mean Average Error (MAE) je jednou z mnoha metrik pro shrnutí a hodnocení kvality modelu strojového učení. Chyba se zde vztahuje k odečtení předpovídané hodnoty od skutečné hodnoty, jak je vidět na vzorci níže.

Chyba Předpovědi = Skutečná Hodnota – Předpovězená Hodnota

Chyba predikce se bere pro každý záznam a poté se všechny chyby převedou na kladné. Toho dosáhneme tak, že pro každou chybu vezmeme absolutní hodnotu, jak je uvedeno níže:

Absolutní chyba → | *Chyba Předpovědi* |

Nakonec vypočítáme průměr pro všechny zaznamenané absolutní chyby (průměrný součet všech absolutních chyb).

$$MAE = \frac{\sum_{i=1}^{n} |\mathrm{yi} - \mathrm{xi}|}{n}$$

Zde jsou:

- Zde y_i jsou předpokládaná hodnota.
- x_i jsou pozorovaná hodnota.
- n je počet pozorování.

Průměrná absolutní hodnota (MAE) je důležitou metrikou pro hodnocení přesnosti doporučovacích algoritmů a může poskytnout užitečné informace o silných a slabých stránkách různých přístupů. Je oblíbená a široce používaná díky své jednoduchosti, protože je jednoduchá a intuitivní metrika, kterou lze snadno pochopit a interpretovat. Je citlivá také na velké i malé chyby, což znamená, že dokáže zachytit celkovou přesnost algoritmu. MAE je méně citlivá na odlehlé hodnoty než jiné metriky, například RMSE (Root Mean Squared Error), což může být důležité v přítomnosti řídkých a zašuměných dat. Střední průměrná chyba navíc poskytuje standardizovaný způsob porovnávání výkonnosti různých doporučovacích algoritmů, protože měří stejný typ chyby u všech algoritmů.

3.6. Křížová validace

Křížová validace, či cross validation, je technika používaná k hodnocení výkonnosti doporučovacích algoritmů, včetně SVD (Singulární rozklad), NMF (Faktorizace nezáporných matic) a PMF (Pravděpodobnostní faktorizace matic). Běžně se používá také ve strojovém učení k testování zobecňováních schopností modelu na neznámých datech.

V kontextu doporučovacích algoritmů zahrnuje křížová validace rozdělení dat na více složek nebo podmnožin. Jedna podmnožina se použije k testování a ostatní podmnožiny se použijí k trénování modelu. Tento proces se několikrát opakuje, přičemž každá podmnožina se použije k testování jednou, a výsledky se zprůměrují, aby se získala celková míra výkonnosti algoritmu. umožňuje testovat algoritmus na různých částech dat, což poskytuje robustnější posouzení jeho výkonnosti.

4. Vytvoření doporučujícího algoritmu

Praktická část mého projektu spočívajícího ve vytvoření doporučovacího algoritmu pro doporučování filmů, nejprve to zahrnuje analýzu dat a testování doporučovacích různých algoritmů které pomocí křížového ověřování předpovídají, které filmy se budou uživateli s největší pravděpodobností líbit na základě jeho předchozí historie sledování, s cílem najít ten nejlepší pro daný problém. Díky studiu teorie doporučovacích algoritmů jsem si pro testování a sestavení svého algoritmu vybral algoritmus založený na kolaborativním filtrování.

V následujících podkapitolách bych se chtěl věnovat jednotlivým částem studie, zejména analýze data setu MovieLens, porovnaní doporučovacích algoritmů a následujícímu doporučení filmů pro konkrétního uživatele.

4.1. Začátek kódu

Aby mohli začít pracovat s kódem, měli bychom co nejprve rozhodnout, které knihovny a funkce z nich budeme používat, aby importovat vše, co potřebujeme. Začínáme pomocí *pip,* co je instalátorem balíčků Python, abychom do projektu nainstalovat klíčové knihovny, které nejsou předinstalovaný v jazyce Python. V našem případě jsou to knihovny Surprise a Pandas.

Knihovna Pandas je populární open-source knihovna pro manipulaci s daty v Pythonu. Je široce používána pro práci se strukturovanými daty, jako jsou tabulky nebo tabulkové procesory, a poskytuje řadu nástrojů pro analýzu dat a datových struktur.

Jak víme z kapitoly o metodologii, Surprise je knihovna Pythonu pro vytváření a analýzu doporučovacích systémů. Poskytuje jednoduché rozhraní pro práci s běžně používanými algoritmy kolaborativního filtrování, jako je faktorizace matic a metody založené na sousedství. Kromě výše uvedených funkcí je určena pro práci se soubory dat o hodnocení, kde uživatelé vyjádřili své preference pro položky, jako jsou filmy nebo produkty, prostřednictvím hodnocení nebo zpětné vazby. Cílem doporučovacího systému vytvořeného pomocí knihovny Surprise je předpovědět hodnocení, které by uživatel udělil položce, kterou ještě nehodnotil, na základě předchozích hodnocení uživatele a hodnocení podobných uživatelů nebo položek. Knihovna Surprise poskytuje řadu předpřipravených algoritmů pro vytváření doporučení, včetně rozkladu singulární hodnoty (SVD) a faktorizace nezáporných matic (NMF) atd. Nabízí také nástroje pro vyhodnocování výkonnosti doporučovacího algoritmu, například křížovou validaci a

různé metriky pro měření přesnosti předpovědí. Kvůli těmto vlastnostem knihovnu bude použito v mém výzkumu.

Přejdeme k importu knihoven, které jsme instalujeme nebo které již máme k dispozici v našem prostředí. Pomocí klíčového slova "import" importujeme knihovnu a přiřadíme jí název, obvykle zkrácenou verzi názvu knihovny. Tento název nám umožní odkazovat na knihovnu v našem kódu, aniž bychom museli pokaždé psát celý název knihovny.

Obrázek 17. Import všech důležitých knihoven a funkcí

```
!pip install scikit-surprise
!pip install pandas

import surprise
from surprise import Dataset
from surprise import Reader
from surprise import SVD
from surprise import NMF
from surprise.model_selection import cross_validate
import pandas as pnd
import numpy as np
import matplotlib.pyplot as matplt
import matplotlib
import math
import statistics
from statistics import mode

from google.colab import drive
drive.mount('/content/drive')
```

Zdroj: vlastní tvorba

Část kódu, kterou vidíme na obrázku 17, instaluje knihovny Surprise a Pandas a poté importuje několik modulů z těchto knihoven, stejně jako další běžné knihovny Pythonu používané při analýze dat, jako jsou Numpy a Matplotlib.

Důležite poznamenat, že kód importuje několik modulů z knihovny Surprise, včetně modulů Dataset, Reader, SVD a NMF, které budou používaný později. Tyto moduly poskytují funkce pro načítání a zpracování dat a také pro sestavování a trénování různých algoritmů doporučovacích systémů.

Kód také importuje funkci *cross_validate* z modulu *surprise.model_selection*, která slouží k vyhodnocení výkonnosti doporučovacího systému pomocí křížové validace.

Kromě knihovny Surprise kód importuje také několik dalších běžných knihoven jazyka Python používaných při analýze dat, včetně výše uvedenou knihovny Pandas: knihovny NumPy pro numerické výpočty a knihovny Matplotlib pro vizualizaci dat.

Nakonec kód připojí úložiště Google Drive, aby z něj bylo možné přistupovat k datovým souborům a načítat je. Jedná se o běžný krok v sešitech Google Colab, které umožňují uživatelům spouštět kód Pythonu v cloudu a přistupovat k různým službám Google.

Další části kódu budou věnovány analýze vybraného souboru dat MovieLens. Zde zjistíme množství uživatelů, filmů a jejich hodnocení, vytvoříme databáze pro další práci s algoritmy a vizualizujeme data pomocí grafů.

4.2. Analýza databáze MovieLens 10M

Jak je uvedeno v kapitole o metodice, pro můj výzkum byla použita významná databáze filmů MovieLens, verze MovieLens 10M Dataset. Daný data set obsahuje přibližně 10 milionů hodnocení provedených uživateli filmů poskytnutých servisem DataLens (Harper, Konstan, 2015). Všechny ID uživatelů je kompletní anonymně, z toho důvodu, že uživatelé MovieLens byli pro zařazení vybráni náhodně a jejich identifikační údaje byly anonymizovány. Proto pro průzkum nebyly použity žádné demografické údaje ani žádné detaily mimo ID.

Tato určitá verze datového souboru se skládá ze tří souborů formátu .dat, movies.dat, ratings.dat a tags.dat.

- movies.dat:

Tento soubor obsahuje informace o filmech v souboru dat, včetně ID filmu, názvu, roku vydání a žánrů. Každý řádek v souboru představuje jeden unikalni film a je formátován takým způsobem: ID filmu::Název::Žánry. Jako příklad uvadim řádky ze souboru, kteří mohou vypadat například takto:

1::Toy Story (1995)::Adventure|Animation|Children|Comedy|Fantasy
2::Jumanji (1995)::Adventure|Children|Fantasy
3::Grumpier Old Men (1995)::Comedy|Romance

Můžeme vidět ID, názvy a to, co mě překvapilo – většinou má film více než 1 žánr, možná 2 nebo 3.

- ratings.dat:

Tento soubor obsahuje hodnocení filmů poskytnutá uživateli v datovém souboru. Každý řádek v souboru představuje jedno hodnocení ve formátu userID::movieID::rating::timestamp. Pro názornost opět uvedu první tři řádky souboru jako přiklad:

1::122::5::838985046
1::185::5::838983525
1::231::5::838983392

Tyto řádky představuji udělené uživatelem 1 filmu 122 hodnocení 5 v časovém razítku 838985046, hodnocení 5 filmu s movieID 185 v časovém razítku 838983525 a hodnocení 5 filmu, movieID, kterého je 231, v časovém razítku 838983392.

- tags.dat:

Soubor obsahuje tagy, které uživatelé v datovém souboru použili u filmů. Není pro můj výzkum relevantní, protože neobsahuje údaje o hodnocení filmů uživateli – pouze jejich osobní poznámky, názory nebo něco, co si o filmu zapamatovali. Přesto si myslím, že by se soubor měl brát v úvahu, protože je součástí databáze. Každý řádek souboru představuje jednu značku a je formátován takto: userID::movieID::tag::timestamp. První tři řádky v souboru vypadají například takto:

15::4973::excellent!::1215184630
20::1747::politics::1188263867
20::1747::satire::1188263867

Jeden ze řádku představuje značku "excellent", kterou použil uživatel 15 na film 4973 v časovém razítku 1215184630.

Aby bylo možné pracovat s databázemi, bylo nutné je převést do formátu vhodného pro zpracování informací v jazyce Python, konkrétně pro knihovnu Surprise. K tomu jsem použila knihovnu Pandas, která se používá pro práci s databázemi a jejich formátování. Pomocí funkci *read_csv()* bylo provedeno čtení dat ze souboru formátu .dat (hodnoty oddělené dvěma dvojtečkami) a vytvořen *DataFrame*. *DataFrame* je dvourozměrná označená datová struktura se sloupci potenciálně různých typů.

Na následujících obrázcích vidíte fragmenty kódu použitého k vytvoření vhodných datových sad pro budoucí analýzu. Kromě toho jsou zobrazeny ukázky datových sad.

Ačkoli, pro algoritmy budou použity pouze datové sady s názvy film a hodnoceni, databáze tag, která neobsahuje pro nás užitečné údaje, byl vytvořen pro zobrazení údajů, které se v něm nacházejí, v rámci analýzy data setu MovieLens.

Obrázek 18. Vytvoření datovou sady "film"

```
film = pnd.read_csv('/content/drive/MyDrive/Dataset10M/movies.dat', header=None, engine = 'python',sep='::', encoding="latin-1")
film.columns = ['movieID', 'title', 'genres']
```

Zdroj: vlastní tvorba

Obrázek 19. Ukázka datovou sady "film"

```
film.head(5)
```

	movieID	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

Zdroj: vlastní tvorba

Obrázek 20. Vytvoření datovou sady "hodnoceni"

```
hodnoceni = pnd.read_csv('/content/drive/MyDrive/Dataset10M/ratings.dat', header=None, engine = 'python', sep='::', encoding="latin-1")
hodnoceni.columns = ['userID', 'movieID', 'rating','timestamp']
```

Zdroj: vlastní tvorba

Obrázek 21. Ukázka datovou sady "hodnoceni"

```
hodnoceni.head(5)
```

	userID	movieID	rating	timestamp
0	1	122	5.0	838985046
1	1	185	5.0	838983525
2	1	231	5.0	838983392
3	1	292	5.0	838983421
4	1	316	5.0	838983392

Zdroj: vlastní tvorba

Obrázek 22. Vytvoření datovou sady "tag"

```
tag = pnd.read_csv('/content/drive/MyDrive/Dataset10M/tags.dat', header=None, engine = 'python',sep='::', encoding="latin-1")
tag.columns = ['userId','movieId','tag','timestamp']
```

Zdroj: vlastní tvorba

Obrázek 23. Ukázka datovou sady "tag"

```
tag.head(5)
```

	userId	movieId	tag	timestamp
0	15	4973	excellent!	1215184630
1	20	1747	politics	1188263867
2	20	1747	satire	1188263867
3	20	2424	chick flick 212	1188263835
4	20	2424	hanks	1188263835

Zdroj: vlastní tvorba

Pomocí údajů z vytvořených databází jsem mohla shromáždit přesné informace o počtu hodnocení, aktivních uživatelů webu a filmů v data setu MovieLens. Jak můžete vidět na následující ukázce kódu, počet unikátních uživatelů je 69878, počet unikátních filmů je 10681, a počet hodnocení je 10000054 celkem.

Obrázek 24. Ukázka kódu pro analýzu počtu uživatelů, filmů a hodnocení

```
n_uziv = len(hodnoceni.userID.unique())
n_film = len(film.movieID.unique())
n_hodnoceni = len(hodnoceni)
print('Počet unikátních uživatelů:', n_uziv)
print('Počet unikátních filmů:', n_film)
print('Počet hodnocení celkem:', n_hodnoceni)
```

```
Počet unikátních uživatelů: 69878
Počet unikátních filmů: 10681
Počet hodnocení celkem: 10000054
```

Zdroj: vlastní tvorba

Chtěla bych také zjistit nejoblíbenější žánr, který se v hodnocených filmech vyskytuje nejčastěji. Za tímto účelem jsem žánry seskupil do sloupce a použil statistickou funkci *mode()*. Jak to můžete vidět na obrázku 25, tímto žánrem je drama.

Obrázek 25. Ukázka kódu pro zjištěni nejběžnějšího žánrů

```
vse_zanry = film['genres'].to_numpy()

def most_common(vse_zanry):
    return(mode(vse_zanry))

print(most_common(vse_zanry))
```

Zdroj: vlastní tvorba

Nyní můžeme přistoupit k další fázi analýzy data setu MovieLens, a to znamená k vytvoření grafů pro vizuální pochopení. Vše grafy budou vytvořeny díky knihovně Matplotlib, která se používá pro vizualizaci a vykreslování dat.

V prvním grafu bych rád zjistil vztah mezi všemi typy hodnocení a počtem hodnocení. Můžete se podívat na kód použitý k vytvoření grafu 1 na obrázku 26.

Obrázek 26. Ukázka kódu pro vytvoření grafu 1

```
group_by_rating = hodnoceni.groupby('rating')['movieID'].count()
y = group_by_rating.to_numpy()
```

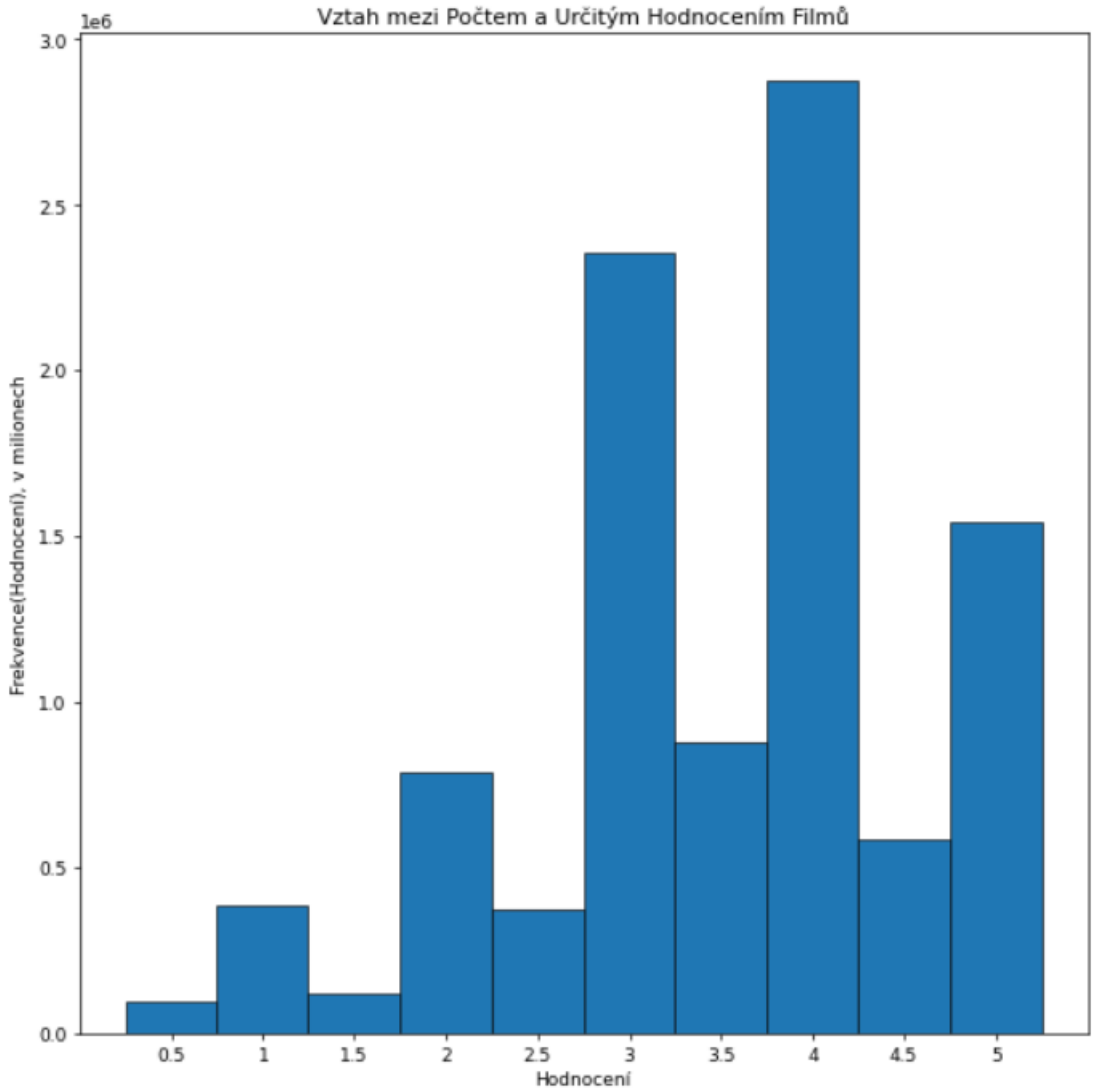
```
matplotlib.rcParams['figure.dpi'] = 65
fig = matplt.figure(figsize =(10, 10))
y = group_by_rating.to_numpy()
x = [0.5,1,1.5,2,2.5,3,3.5,4,4.5,5]
tick_label = ['0.5', '1', '1.5', '2', '2.5', '3', '3.5', '4', '4.5', '5']

matplt.bar(x, y, tick_label = tick_label, width=0.5, edgecolor="black", linewidth=0.7)

matplt.xlabel('Hodnocení')
matplt.ylabel('Frekvence(Hodnocení), v milionech')
matplt.title('Vztah mezi Počtem a Určitým Hodnocením Filmů')
```

Zdroj: vlastní tvorba

Tento kód využívá knihovny Pandas a výše uvedenou Matplotlib v jazyce Python k vytvoření sloupcového grafu, který vizualizuje vztah mezi existujícími hodnocení a jejich počtem. Poprvé jsem seskupovala datovou sadu hodnoceni podle sloupce „hodnocení“ a funkce spočetla počet výskytů každé recenzi. Výsledek se přiřadí proměnné *group_by_rating*. Pak jsem převedla objekt *group_by_rating* na array Numpy a přiřadila tomu proměnné y. Pak jsem vytvořila seznam souřadnic x pro sloupce s hodnotami od 0,5 do 5 v krocích po 0,5 a přiřadí jej proměnné x. Funkce *matplt.bar* vytvoří sloupcový graf pomocí funkce *bar()* z knihovny Matplotlib, přičemž vstupy jsou

x a y a jsou v něm různé možnosti formátování, například popisky zaškrtnutí, šířka sloupce, barva okraje a šířka čáry. V posledních radkách jsem nastavila popisky osy x a y, nastavila název grafu na "Vztah mezi Počtem a Určitým Hodnocením filmů".

Graf 1. Vztah mezi počtem filmů a hodnocením



Zdroj: vlastní tvorba

Graf 1 ukazuje vztah mezi počtem a hodnocením filmů. Osa x představuje vše existujici možnosti hodnocení – od 0,5 do 5 hvězdiček, zatímco osa y představuje počet určitym hodnocení filmů. Z grafu vyplývá, že většina filmů má hodnocení 4 a 3 hvězdičky, menší počet filmů má hodnocení 5 nebo 3,5 hvězdiček. Kromě toho je méně hodnocení 0,5 až 1,5 mají nejměnší počet hodnocení.

Díky tomu, že je v databázi tolik různých recenzí, s větším počtem pozitivních, si můžeme být jisti, že každá z těchto hodnocení bude hrát důležitou roli při jejich doporučování filmů uživatelům.

Závěrem lze říci, že většina filmů v tomto souboru dat má hodnocení 4 hvězdičky, což znamená, že oni byly diváky obecně dobře přijaty.

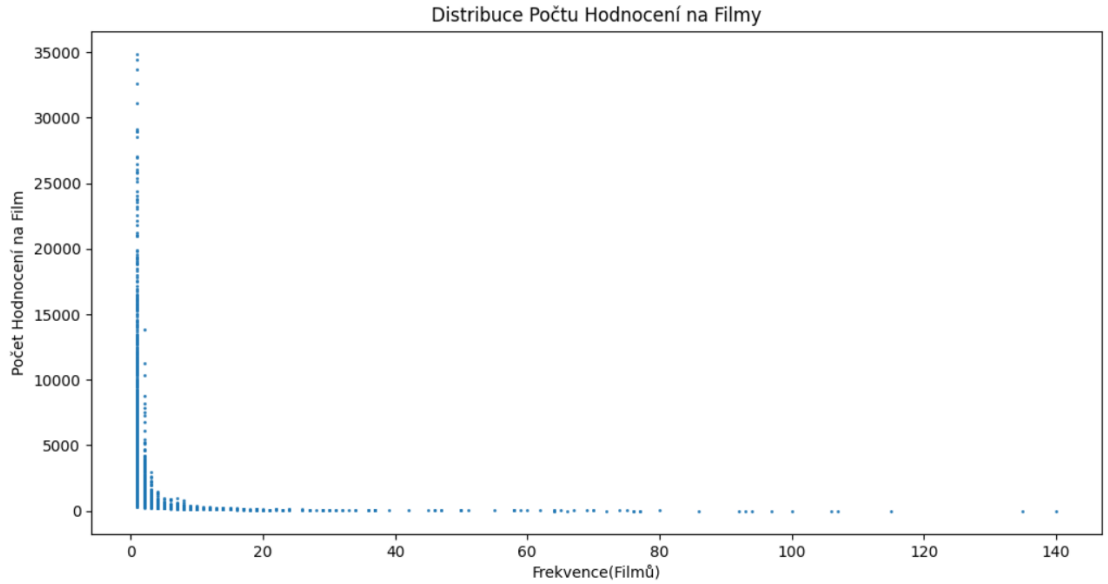
Dalším grafem bych chtěla ukázat distribuce počtu hodnocení na filmy. Pokud se podíváte na obrázek 27 s kódem pro vytvoření tohoto grafu, uvidíte, že funguje podobným principem jako ten předchozí. Rozdíl je v tom, že výstup funkce *np.unique()* je pak přiřazen dvěma proměnným, unique a counts. Proměnná unique obsahuje pole unikátních prvků ve film_hod a counts obsahuje pole četností jednotlivých unikátních prvků. pomocí těchto proměnných byl vytvořen graf rozptylu, který je v práce znázorněn jako graf 2.

Obrázek 27. Ukázka kódu pro vytvoření grafu 2

```
fig = matplt.figure(figsize =(12, 6))
film_pocet_hod = hodnoceni.groupby('movieID')['rating'].count()
film_hod = film_pocet_hod.to_numpy()
unique, counts = np.unique(film_hod, return_counts=True)
x = counts
y = unique
matplt.scatter(x, y, s=1)
matplt.ylabel('Počet Hodnocení na Film')
matplt.xlabel('Počet Filmů')
matplt.title('Distribuce Počtu Hodnocení na Filmy')
```

Zdroj: vlastní tvorba

Graf 2. Distribuce Počtu Hodnocení na Filmy



Zdroj: vlastní tvorba

Tento graf nám skvěle ukazuje, že ve skutečnosti často existuje problém dlouhého chvostu, protože vidíme, že maximální počet recenzí má malý počet filmů. Dále se nejčastěji vyskytují počty recenzí na jeden film v hodnotách mezi 50 a přibližně 4000.

Poté máme opět velké množství shluků, kde některé filmy mají ve skutečnosti málo recenzí.

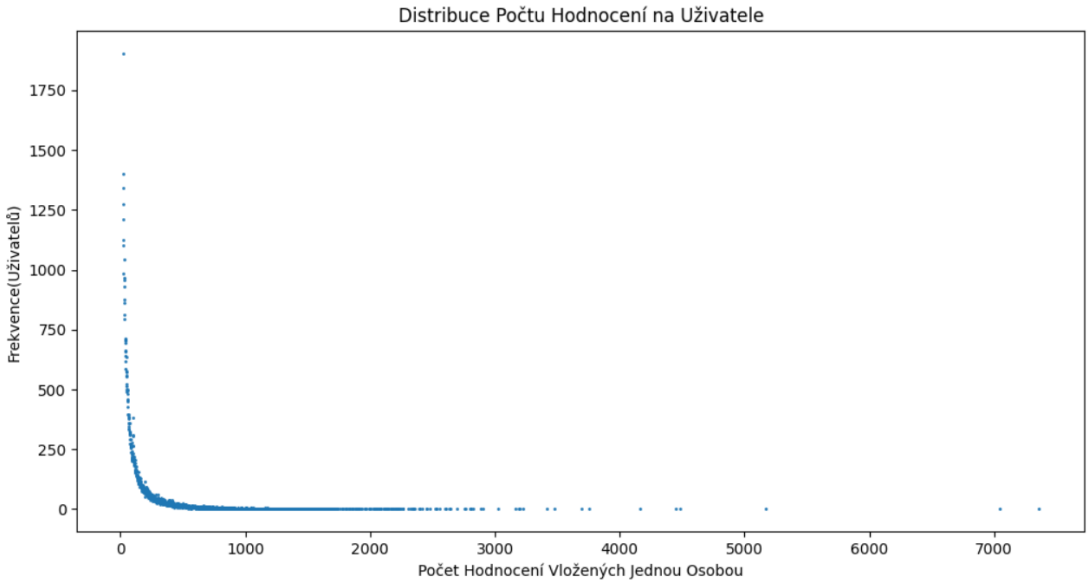
Pomocí grafu 3 bych chtěla ukázat distribuce počtu hodnocení na uživatele. Vytvoření tohoto grafu je velmi podobné grafu 2, pokud to vidíte z obrázku 28. Rozdíl je v tom, že jsme na začátku seskupili hodnocení a ID uživatelů místo ID filmů jako ve předchozím příklade.

Obrázek 28. Ukázka kódu pro vytvoření grafu 3

```
fig = matplt.figure(figsize =(12, 6))
user_pocet_hod = hodnoceni.groupby('userID')['rating'].count()
user_hod = user_pocet_hod.to_numpy()
unique, counts = np.unique(user_hod, return_counts=True)
x = unique
y = counts
matplt.scatter(x,y, s=1)
matplt.ylabel('Počet Uživatelů')
matplt.xlabel('Počet Hodnocení Vložených Jednou Osobou')
matplt.title('Distribuce Počtu Hodnocení na Uživatele')
```

Zdroj: vlastní tvorba

Graf 3. Distribuce Počtu Hodnocení na Uživatele



Zdroj: vlastní tvorba

Díky 3. grafu můžeme opět vidět graf dlouhého chvostu, což znamená, že uživatelé v průběhu let udělili velké množství hodnocení populárním filmům, pak průměrné populární filmy máji většinu hodnocení (od cca 100 do cca 500) a většina filmů má opět nízké hodnocení.

Po analýze dat z databází bych chtěla začít vybírat nejpřesnější algoritmus pro doporučování filmů. V následujících podkapitolách budou analyzovány algoritmy SVD, NMF a PMF, aby byl nalezen ten nejpřesnější.

4.3. SVD Algoritmus

SVD je zkratka pro singulární rozklad nebo Singular Value Decomposition v angličtině. Jedná se o techniku faktorizace matice, která rozkládá matici na další matice, což lze využít pro řadu aplikací v lineární algebře, statistice a analýze dat.

V kontextu doporučovacích algoritmů se SVD často používá k faktorizaci matice interakce mezi uživatelem a položkou na dvě méně rozměrné matice: jednu matici obsahující latentní faktory, které představují preference uživatele, a druhou matici obsahující latentní faktory, které představují atributy položky. Tyto faktory pak lze použít k předpovědím interakcí mezi uživatelem a položkou a ke generování personalizovaných doporučení pro uživatele. V doporučovacích systémech rozklad na singulární hodnoty hojně používá, protože je to výpočetně efektivní metoda pro zpracování velkých, řídkých matic, které jsou v doporučovacích systémech běžné. Algoritmus SVD lze implementovat pomocí různých optimalizačních technik, a knihovna Surprise, kterou jsem importovala dříve, nám s tím pomůže. Má samostatnou funkci pro rozklad na singulární hodnoty – *SVD()*, takže nemusíme vytvářet složité funkce pro vytvoření tohoto algoritmu.

Obrázek 29. SVD Algoritmus

```
reader=Reader()
data = Dataset.load_from_df(hodnoceni[["userID", "movieID", "rating"]], reader)

algoritmus = SVD()

result = cross_validate(algoritmus, data, measures=['MAE','RMSE'], cv=5, verbose=True)
print(result)
```

Zdroj: vlastní tvorba

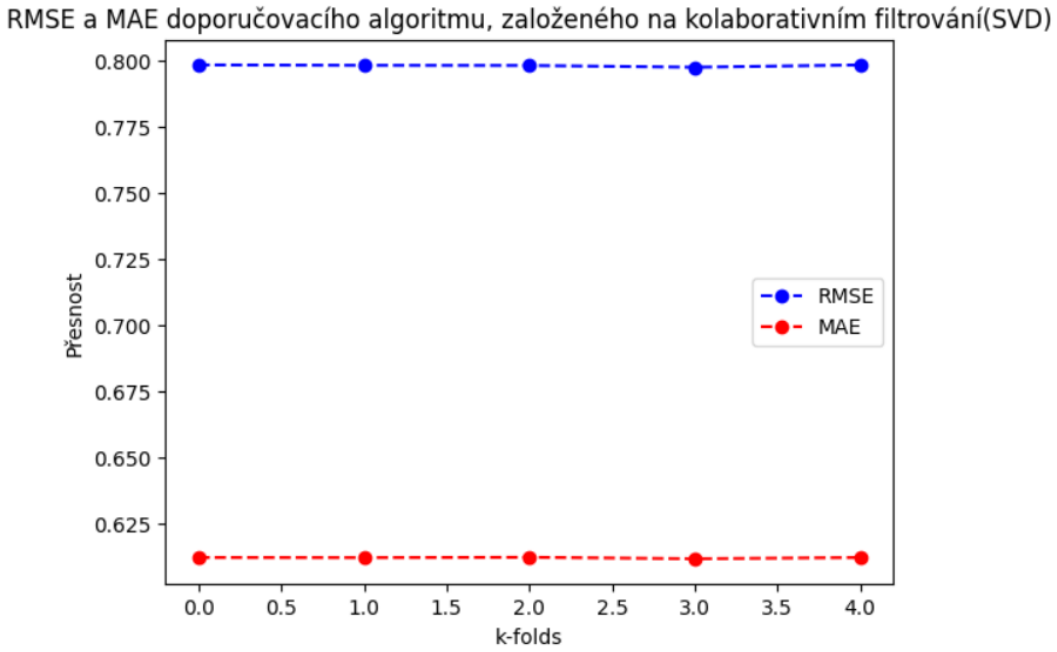
Po spuštění algoritmu, který je zobrazen na obrázku 29, získáte hodnoty MAE a RMSE, které byly zadány do tabulky 2. Provedla jsem pětinásobnou křížovou validaci dat hodnocení filmů pomocí algoritmu SVD. Argument “measures“ určuje, které metriky hodnocení se mají vypočítat (v tomto případě střední absolutní chyba (MAE) a střední kvadratická chyba (RMSE)). Argument „verbose“ říká funkci *cross_validate*, aby vypisovala aktualizace průběhu. Na grafu 4 vidíme, že všechny hodnoty chyb mají přibližně stejné hodnoty, což je dobré znamení, že algoritmus pracuje správně. Aby to vypadalo vizuálně přehledně, byl z této tabulky 1 je vytvořen graf 5.

Obrázek 30. Ukázka kódu pro vytvoření grafu 4

```
matplt.plot(result['test_rmse'], 'bo--', label='RMSE')
matplt.plot(result['test_mae'], 'ro--', label='MAE')
matplt.xlabel('k-folds')
matplt.ylabel('Přesnost')
matplt.legend()
matplt.title('RMSE a MAE doporučovacího algoritmu, založeného na kolaborativním filtrování(SVD)')
fig = matplt.gcf()
```

Zdroj: vlastní tvorba

Graf 4. RMSE a MAE doporučovacího algoritmu (SVD)



Zdroj: vlastní tvorba

Ukázka kódu na obrázku 30 vytváří graf 4, který zobrazuje hodnoty RMSE a MAE získané z doporučovacího algoritmu založeného na kolaborativním filtrování (SVD) pro hodnoty pěti složení křížovou validace. Graf má legendu, nadpis, osy x a y jsou vhodně označeny.

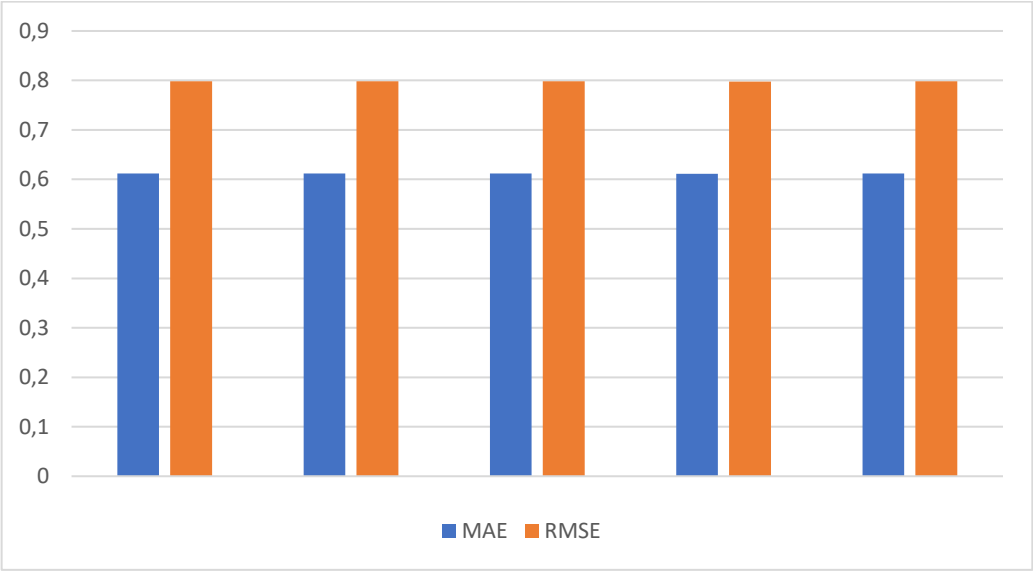
Tabulka 2. Hodnoty MAE a RMSE SVD algoritmu

	Složení 1	Složení 2	Složení 3	Složení 4	Složení 5	Průměr	STD
MAE	0.6120	0.6120	0.6121	0.6115	0.6120	0.6119	0.0002
RMSE	0.7983	0.7981	0.7980	0.7974	0.7982	0.7980	0.0003

Zdroj: vlastní vypočet

Tabulka 2 obsahuje hodnoty MAE a RMSE testovaného algoritmu singularniho rozkladu – SVD. Vidíme zde hodnoty každého z pěti maticových členů, na kterých byl algoritmus ověřován. Také jsou v tabulce průměrné hodnoty, které pro nás budou důležité v podkapitole porovnávající 3 algoritmy, abychom zjistili, který z nich je nejpřesnější.

Graf 5. RMSE a MAE doporučovacího algoritmu, založeného na SVD



Zdroj: vlastní tvorba

Jak vidíme, hodnoty MAE a RMSE je malé, což znamená, že pravděpodobnost chyby doporučení v tomto algoritmu je mala. Tyto údaje jsme získali díky velikosti použité databáze. Z důvodu, že obsahuje 10 milionů hodnocení přesnosti, je přesnější.

4.4. NMF Algoritmus

NMF (Non-negative Matrix Factorization) je technika faktorizace matic, kterou lze použít k rozkladu nezáporné matice na dvě nezáporné matice. Cílem NMF je najít dvě nezáporné matice s nízkým řádem, které se po vynásobení přibližují původní nezáporné matici.

V kontextu doporučovacích algoritmů lze NMF použít k faktorizaci matice hodnocení uživatelů a položek na dvě nezáporné matice s nižší hodnotou, z nichž jedna představuje vložené hodnoty uživatelů a druhá vložené hodnoty položek. Výsledná vložení lze použít k doporučením tak, že se vypočítá bodový součin mezi vložením uživatele a vložením nepozorovaných položek a doporučí se položky s nejvyšším bodovým součinem. Jednou z výhod použití NMF v doporučovacích systémech je to, že může pomoci vypořádat se s řídkostí matic hodnocení uživatelů.

Obrázek 31. NMF Algoritmus

```
[ ]  reader=Reader()
     data = Dataset.load_from_df(hodnoceni[["userID", "movieID", "rating"]], reader)

     algoritmus_nmf = NMF()

     result_nmf = cross_validate(algoritmus_nmf, data, measures=['MAE','RMSE'], cv=5, verbose=True)
     print(result_nmf)
```

Zdroj: vlastní tvorba

K otestování algoritmu NMF použijeme úryvek kódu zobrazený na obrázku 31. Je konstrukčně dosti podobný algoritmu na obrázku 29, rovněž používá křížové ověřování, byl změněn algoritmus na funkce *NMF(),* kterou nám znovu poskytnula knihovna Surprise.

Tento úryvek ukázka kódu provádí křížovou validaci na souboru dat s hodnocením filmů. Účelem tohoto kódu je vyhodnotit výkonnost doporučovacího algoritmu, konkrétně algoritmu NMF, měřením jeho hodnot MAE (Mean Absolute Error) a RMSE (Root Mean Squared Error) pomocí křížové validace. Algoritmus NMF je vytvořen pomocí metody *NMF().* Následně se provede pětinásobná křížová validace na souboru dat. V tomto případě se funkce *cross_validate()* používá k rozdělení datové sady na 5 složek a k vyhodnocení výkonnosti algoritmu NMF na každé složce. Parametr measures je nastaven na ['MAE', 'RMSE'], aby se pro každou složku vypočítaly hodnoty MAE i RMSE. Parametr "verbose" je nastaven na hodnotu True, aby se zobrazil průběh procesu křížového ověřování.

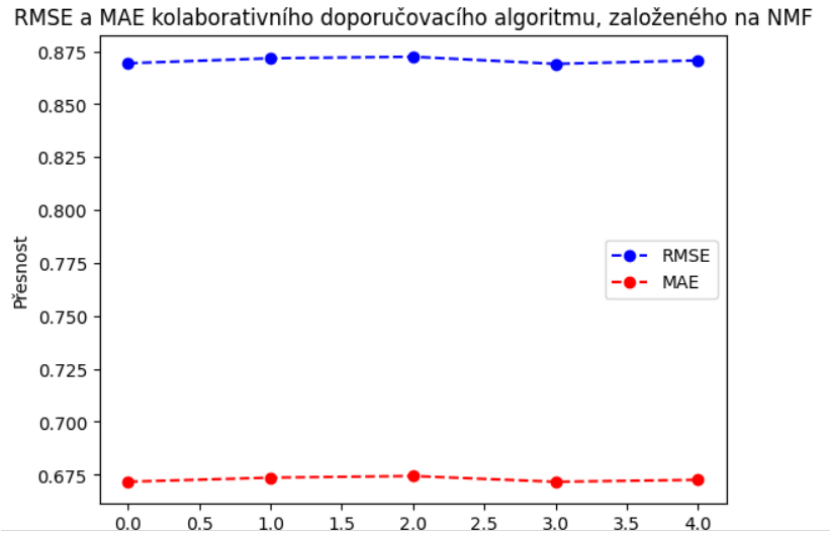
Výsledky procesu křížové validace, které zahrnují průměrné hodnoty MAE a RMSE pro všech 5 složek, byly zařazeny do tabulky 4 a pro přehlednost byly převedeny do grafu 6, který byl vygenerován v Pythonu díky knihovně Matplotlib, a do grafu 7.

Obrázek 32. Ukázka kódu pro vytvoření grafu 5

```
matplt.plot(result_nmf['test_rmse'], 'bo--', label='RMSE')
matplt.plot(result_nmf['test_mae'], 'ro--', label='MAE')
matplt.xlabel('k-folds')
matplt.ylabel('Přesnost')
matplt.legend()
matplt.title('RMSE a MAE kolaborativního doporučovacího algoritmu, založeného na NMF')
```

Zdroj: vlastní tvorba

Graf 6 . RMSE a MAE doporučovacího algoritmu (NMF)



Zdroj: vlastní tvorba

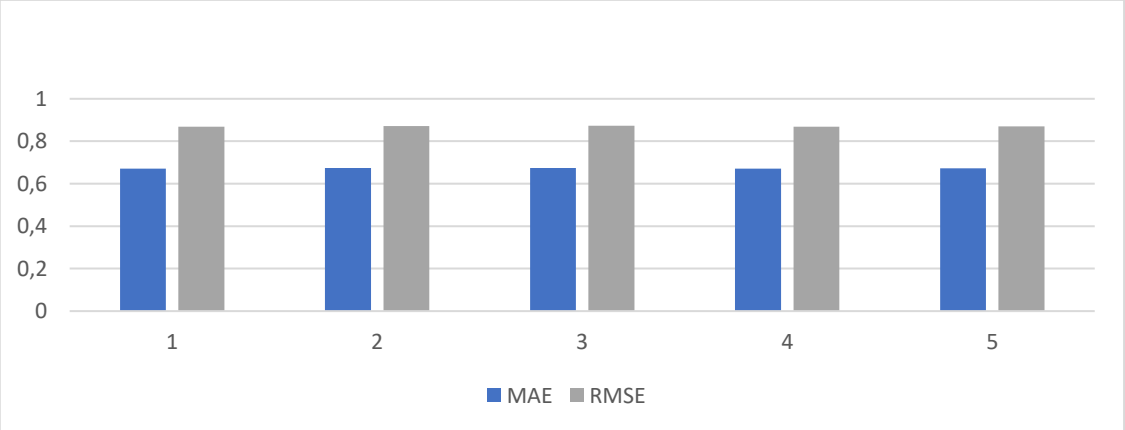
Tabulka 3. Hodnoty MAE a RMSE NMF algoritmu

	Složení 1	Složení 2	Složení 3	Složení 4	Složení 5	Průměr	STD
MAE	0.6717	0.6736	0.6744	0.6717	0.6726	0.6728	0.0011
RMSE	0.8693	0.8716	0.8724	0.8690	0.8706	0.8706	0.0013

Zdroj: vlastní vypočet

Jak je vidět z výsledků algoritmu zaznamenaných v tabulce 2, algoritmus NMF je také dost přesný, protože jeho průměrná hodnota RMSE je 0,8706. Není to perfektní hodnota, ale je stále dobrá. Obecně platí, že RMSE 0,87 není špatná, ale v závislosti na kontextu může, ale nemusí být dobrá. Pracuji jsem na doporučovacím algoritmu, který má velký počet položek a uživatelů a hodnocení nejsou velmi řídká, může být RMSE 0,87 považována za normální.

Graf 7. RMSE a MAE doporučovacího algoritmu, založeného na NMF



Zdroj: vlastní tvorba

4.5. PMF Algoritmus

Pravděpodobnostní maticová faktorizace (PMF) je doporučovací algoritmus, který se používá k předvídání, jak by uživatel hodnotil položku, kterou ještě nehodnotil. PMF předpokládá, že v pozadí matice hodnocení uživatele a položky jsou určité latentní faktory nebo vlastnosti, které nejsou přímo pozorovatelné. Tyto latentní faktory se pak používají k rozkladu matice hodnocení na matice nižších rozměrů, které reprezentují vlastnosti uživatele a položky v latentním prostoru.

PMF předpokládá, že pozorovaná hodnocení se řídí Gaussovým rozdělením, a k odhadu parametrů modelu používá přístup maximální věrohodnosti. Snaží se najít hodnoty latentních faktorů, které nejlépe odpovídají pozorovaným hodnocením, a to tak, že minimalizuje rozdíl mezi předpovídanými hodnoceními a skutečnými hodnoceními.

Probabilistic Matrix Factorization je jedním z oblíbeným algoritmem pro doporučování, protože si poradí s chybějícími údaji a dokáže škálovat na velké soubory dat. Byl použit v řadě aplikací, včetně doporučování filmů a personalizované medicíny.

Obrázek 33. PMF Algoritmus

```
reader=Reader()
data = Dataset.load_from_df(hodnoceni[["userID", "movieID", "rating"]], reader)

algoritmus_pmf = SVD(biased=False)

result_pmf = cross_validate(algoritmus_pmf, data, measures=['MAE','RMSE'], cv=5, verbose=True)
print(result_pmf)
```

Zdroj: vlastní tvorba

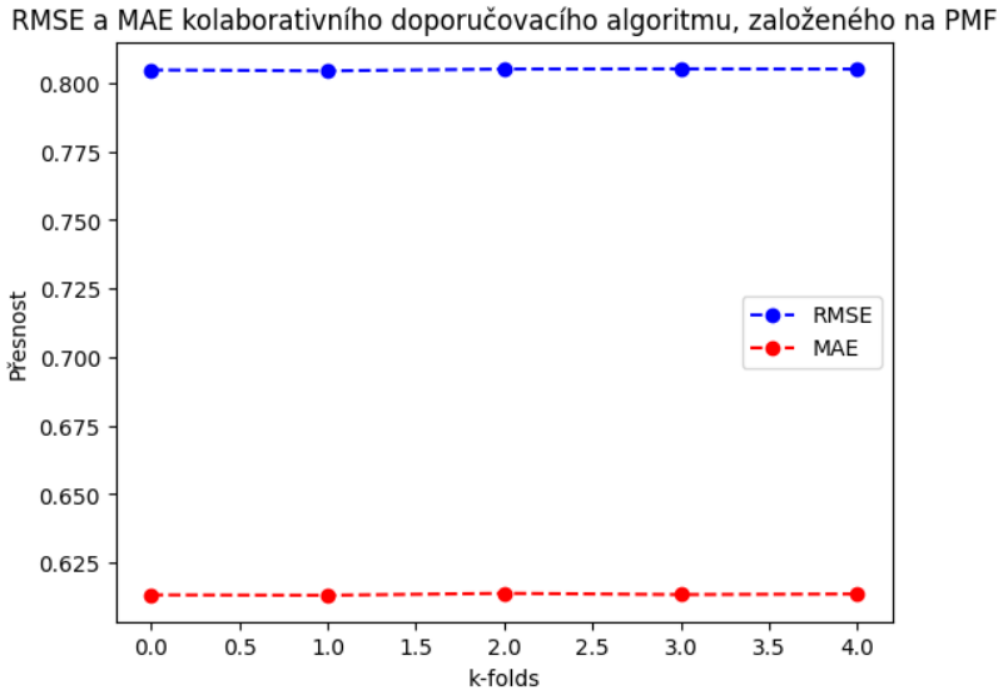
Úryvek kódu na obrázku 32 ukazuje, jak vytvořit a vyhodnotit výkon algoritmu PMF v knihovně Surprise. V něm je objekt *algoritmus_pmf* vytvořen pomocí algoritmu *SVD()* s parametrem "*biased"* nastaveným na hodnotu False, čímž je vlastně ekvivalentní algoritmu PMF. Měřením hodnot MAE a RMSE pomocí křížové validace můžeme posoudit účinnost algoritmu při předpovídání hodnocení filmů a v případě potřeby jej vylepšit.

Obrázek 34. Ukázka kódu pro vytvoření grafu 8

```
matplt.plot(result_pmf['test_rmse'], 'bo--', label='RMSE')
matplt.plot(result_pmf['test_mae'], 'ro--', label='MAE')
matplt.xlabel('k-folds')
matplt.ylabel('Přesnost')
matplt.legend()
matplt.title('RMSE a MAE kolaborativního doporučovacího algoritmu, založeného na PMF')
```

Zdroj: vlastní tvorba

Graf 8. RMSE a MAE doporučovacího algoritmu (PMF)



Zdroj: vlastní tvorba

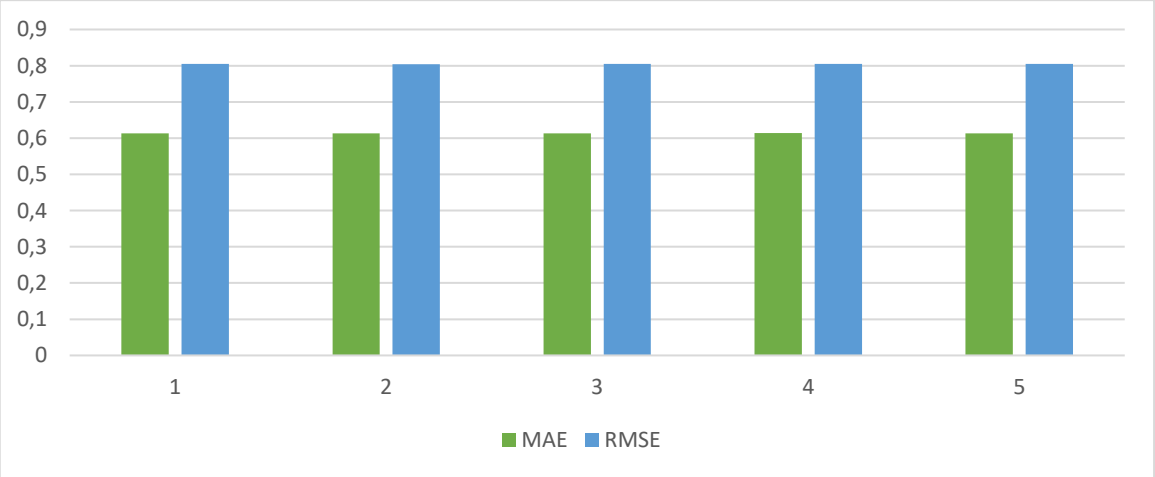
Tabulka 4. Hodnoty MAE a RMSE PMF algoritmu

	Složení 1	Složení 2	Složení 3	Složení 4	Složení 5	Průměr	STD
MAE	0.6132	0.6131	0.6138	0.6133	0.6133	0.6134	0.0002
RMSE	0.8048	0.8044	0.8051	0.8052	0.8051	0.8049	0.0003

Zdroj: vlastní vypočet

Průměrné hodnoty MAE 0,6134 a RMSE 0,8049 ukazují přesnost algoritmu PMF při předpovídání hodnocení filmů. Nižší hodnoty těchto ukazatelů naznačují lepší přesnost, takže tyto hodnoty naznačují, že algoritmus PMF je ve svých předpovědích středně přesný.

Graf 9. RMSE a MAE doporučovacího algoritmu, založeného na PMF



Zdroj: vlastní tvorba

4.6. Porovnávání algoritmů

Díky tomu, že vše algoritmy byly úspěšně spuštěny a hodnoty chyb byly získaný, můžeme je porovnat na základě jejich hodnot MAE a RMSE, abychom našli ten nejpřesnější algoritmus, který použiji pro doporučování filmů v další kapitole.

Pokud chceme porovnat výkonnost a přesnost SVD, NMF a PMF v doporučovacím algoritmů, aby najit nejpřesnější algoritmus, můžeme k vyhodnocení přesnosti předpovídaných hodnocení použít metriku RMSE. Jedná se o běžně používanou metriku hodnocení doporučovacích algoritmů. RMSE je užitečná metrika, protože bere v úvahu velikost chyb a větší chyby jsou více penalizovány. Kromě toho se poměrně snadno interpretuje a její hodnoty lze porovnávat mezi různými modely a soubory dat.

Porovnáním hodnot RMSE různých modelů lze posoudit, které z nich jsou pro daný problém a datovou sadu výkonnější. To vám může pomoci vybrat nejlepší model pro váš doporučovací systém a zlepšit jeho přesnost.

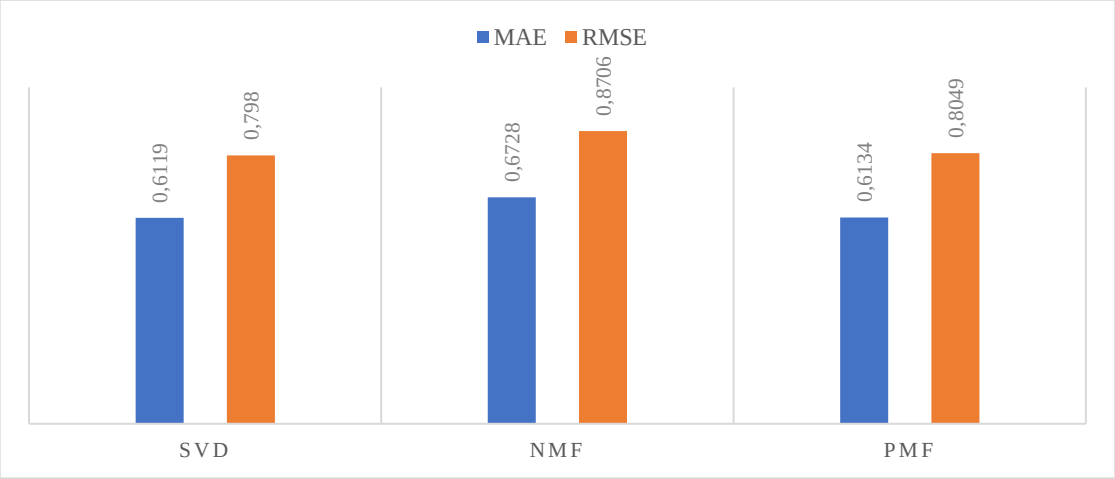
Pro porovnání algoritmů musím nejdříve vytvořit tabulku, která bude obsahovat průměrné hodnoty MAE a RMSE u obou použitých algoritmů. Tyto údaje můžete vidět v tabulce 5, pomocí které je vytvořen graf 10.

Tabulka 5. Průměrné hodnoty MAE a RMSE algoritmů SVD, NMF a PMF

	SVD	NMF	PMF
MAE	0,6119	0,6728	0,6134
RMSE	0,7980	0,8706	0,8049

Zdroj: vlastní vypočet

Graf 10. Porovnání průměrných hodnot MAE a RMSE algoritmů SVD, NMF a PMF



Zdroj: vlastní tvorba

Po vyhodnocení výkonnosti dvou algoritmů faktorizace matic, SVD, NMF a PMF, jako doporučujícího algoritmu a jejich srovnávaní pomoci tabule a grafu jsem zjistila, že SVD a PMF je přesnější než NMF. Je to z důvodu, že průměrné hodnoty RMSE pro SVD a PMF byly 0,7980, resp. 0,8049, když NMF má hodnotu 0,8706. Je známo, že čím menší je hodnota RMSE, tím je algoritmus přesnější. Také můžeme vidět, že hodnota MAE algoritmu SVD je nižší, pokud se podíváme na graf 8. Ona také znamená, že čím nižší bude, tím že pravděpodobnost chyby je menší. ale hodnoty RMSE jsou obvykle přesnější než, a z toho důvodu to oni jsou srovnáváni.

Metrika RMSE měří průměrnou vzdálenost mezi předpovídanými hodnoceními a skutečnými hodnoceními v testovací množině. Nižší hodnota RMSE znamená lepší výkonnost modelu, protože odráží, jak dobře dokáže algoritmus předpovídat hodnocení uživatelů. Hodnoty RMSE 0,7980, 0,8049 a 0,8706 tedy naznačují, že SVD v tomto konkrétním problému a datové sadě funguje lépe než NMF a PMF.

4.7. Doporučení filmů

V této závěrečné části kódu můžeme použít algoritmus, který se v předchozí podkapitole ukázal jako nejpřesnější, a to algoritmus založený na SVD. Použijeme jej jako algoritmus pro funkci, která na základě ID daného uživatele vytváří doporučení filmů. Na obrázku níže vidíte úryvek kódu, který byl použit k implementaci tohoto doporučení.

Obrázek 35. Ukázka kódu pro finální doporučení

```
def topdoporuceni(userId):
    df=pnd.DataFrame()

    filmidlist=hodnoceni['movieID'].unique().tolist()
    odhadhod=[]
    nazevfilmulist=[]
    existhodlist=[]
    zanrlist=[]
    for i in filmidlist:
        pred = algoritmus.predict(userId, i, verbose=False)
        odhadhod.append(pred[3])
        nazevfilmu = film[film.movieID==i]['title'].values
        nazevfilmulist.append(nazevfilmu)
        existhod = hodnoceni[(hodnoceni.movieID==i)&(hodnoceni.userID==userId)]['rating'].values
        existhodlist.append(existhod)
        zanr = film[film.movieID==i]['genres'].values
        try:
            zanrlist.append(zanr)
        except:
            zanrlist.append('Unavailable')
    df['movie_id']=filmidlist
    df['estimated_rating']=odhadhod
    df['title']=nazevfilmulist
    df['actual_rating']=existhodlist
    df['genres']=zanrlist
    df=df.sort_values(['estimated_rating'], ascending=[False])
    return df
```

Zdroj: vlastní tvorba

Funkce přijímá argument userId, který slouží k identifikaci uživatele, pro kterého se generují doporučení. První řádek funkce vytvoří prázdný objekt Pandas DataFrame, který bude sloužit k ukládání doporučení. Několik dalších řádků funkce vytvoří seznamy, do kterých se vloží důležité informace o filmu. Uvnitř smyčky algoritmus provede předpověď pro daného uživatele – film, a odhadované hodnocení se připojí do seznamu nazvaného odhadhod. Název filmu je extrahován z datové sady nazvané film a přidán do seznamu nazvaného nazevfilmulist. Uživatelovo skutečné hodnocení (pokud existuje) filmu je extrahováno z data setu hodnoceni a přidáno do seznamu nazvaného existhodlist. Žánr filmu je rovněž extrahován z data setu filmů a přidán do seznamu nazvaného zanrlist.

Nakonec jsou všechny seznamy spojeny do objektu Pandas DataFrame, seřazeny podle odhadovaného hodnocení sestupně a vráceny jako výstup funkce.

Celkově tato funkce využívá kolaborativní filtrování k vytváření doporučení filmů pro daného uživatele na základě jeho předchozích hodnocení a hodnocení ostatních uživatelů, kteří hodnotili podobné filmy. Výstupem je setříděný seznam filmů s odhadovanými hodnoceními, který lze použít k navržení nových filmů, které by měl uživatel sledovat.

Obrázek 36. Ukázka kódu pro spuštění funkce

```
userid=20
print('Doporučení na základě žánrů pro uživatele :', userid)
topdoporuceni(userid).head(10)
```

Zdroj: vlastní tvorba

Pro zobrazení výsledku, na obrázku 36 vidíte, jak byla funkce spuštěna a jaké ID uživatele bylo použito. Dále jsem se rozhodla získat 10 nejlepších doporučení pro tohoto uživatele. Obrázek 37 zobrazuje, že odhadovaná hodnota každého z nich je poměrně vysoká, což vysvětlí, proč byly algoritmy z předchozích podkapitol porovnávány z hlediska přesnosti.

Můžeme, pokud chceme, vypsat libovolný počet doporučených filmů pomocí zadaní hodnoty v závorce funkci *head()*. Pokud vás také zajímáji nejhorší doporučení, můžete také zobrazit doporučené filmy s nejhorším možným hodnocením (přibližně 2,7), které se uživatelům pravděpodobně nebude líbit. K tomu můžeme použít funkci *tail()*, která je přesným opakem té, kterou jsme použili.

Obrázek 37. Výsledek kolaborativního algoritmu

Doporučení na základě žánrů pro uživatele : 20

	movie_id	estimated_rating	title	actual_rating	genres
1293	318	4.419858	[Shawshank Redemption, The (1994)]	[]	[Drama]
133	527	4.355029	[Schindler's List (1993)]	[]	[Drama|War]
213	50	4.342639	[Usual Suspects, The (1995)]	[]	[Crime|Mystery|Thriller]
8631	8684	4.332486	[Man Escaped, A (Un condamnÃ© Ã mort s'est Ã©...	[]	[Adventure|Drama]
4689	44555	4.323449	[Lives of Others, The (Das Leben der Anderen) ...	[]	[Drama]
5632	6896	4.320666	[Shoah (1985)]	[]	[Documentary|War]
2056	2324	4.310301	[Life Is Beautiful (La Vita Ã¨ bella) (1997)]	[]	[Comedy|Drama|Romance|War]
1926	4973	4.305226	[Amelie (Fabuleux destin d'AmÃ©lie Poulain, Le...	[]	[Comedy|Romance]
34	858	4.301807	[Godfather, The (1972)]	[]	[Crime|Drama]
8142	8516	4.293884	[Matter of Life and Death, A (Stairway to Heav...	[]	[Comedy|Fantasy|Romance]

Zdroj: vlastní tvorba

V tomto seznamu doporučení to není možné vidět, ale teoreticky se může stát, že jeden z doporučených filmů je již hodnocen. několikrát se to stalo během testování kódu, že se někdy v nejlepších doporučeních objevili filmy, které již mají existující hodnoceni od uživatele. Z mého hlediska není to velký problém, protože takto fungují například algoritmy Netflixu. Na některé uživatele možná zafunguje připomenutí dříve zhlédnutého oblíbeného filmu a rozhodnou se jej zhlédnout znovu – v tomto případě to není nevýhoda algoritmu.

5. Závěr

V závěre lze konstatovat, že všechny cíle této bakalářské práce byly úspěšně splněny. Získané výsledky splnily očekávání stanovená na počátku studie. Autorkou prací byla provedena analýza doporučovacích algoritmů a jejich typu s použitím dat, získaných z odborné literatury.

Jedním z cílů této studie bylo poskytnout přehled o použití doporučovacích algoritmů v elektronickém obchodě. Prostřednictvím hloubkové analýzy literatury a webových stránek služeb bylo prokázáno, že tyto algoritmy se staly nezbytnou součástí platforem elektronického obchodování a poskytují uživatelům personalizované a relevantní návrhy produktů. Výsledky společností Amazon, Netflix a Spotify, tří nejúspěšnějších společností, které využívají doporučovací algoritmy, poskytují důkazy o jejich účinnosti při zvyšování uživatelského komfortu a podpoře prodeje. Poznatky získané z této časti studia mohou být užitečné pro podniky, které chtějí zlepšit své strategie elektronického obchodování a poskytnout personalizovanou a uspokojivou uživatelskou zkušenost.

Cílem, stanoveným pro praktickou část bylo vytvořit doporučovací algoritmus se záměrem na přesnost parametru a doporučení. Za účelem splnění tohoto cíle jsem vytvořila a analyzovala kolaborativní algoritmy pomocí knihovny Surprise v jazyce Python, které využívají techniky křížového ověřování. Těmito algoritmy jsou SVD – Singulární rozklad či Singular Value Decomposition, NMF – Faktorizace nezáporných matic či Non-negative Matrix Factorization a PMF – Faktorizace pravděpodobnostních matic či The Probabilistic Matrix Factorization. Všechny tyto algoritmy jsou jedinečné, mají své silné a slabé stránky, určité zvláštnosti. Pro určení přesnosti algoritmů byly použity hodnoty MAE – střední absolutní chyba a RMSE – kořenová střední kvadratická odchylka. Čím nižší je tato hodnota, tím je algoritmus přesnější. Po porovnání všech průměrných hodnot RMSE a MAE těchto 3 algoritmů bylo zřejmé, že nejpřesnějším algoritmem je pro nás algoritmus SVD. Ten pomáhá doporučení filmů bylo provedeno pro konkrétního uživatele.

Potenciál pro budoucí výzkum existuje v každé práci nebo studijním oboru vzhledem k neustálému vývoji znalostí a chápání, stejně jako ke vzniku nových otázek a oblastí zkoumání. Například, v oblasti doporučovacích algoritmů v kontextu umělé inteligence má obrovský potenciál pro další výzkum vzhledem k tomu, roste poptávka po sofistikovanějších a přesnějších doporučovacích algoritmech, které dokáží zpracovávat

větší objemy dat, lépe předpovídat a poskytovat personalizovanější uživatelské prostředí. Je také možnost zkoumat nové metody sběru dat, jako jsou aktivní metody učení, které umožňují algoritmu učit se od uživatelů během interakce s nimi.

Bez ohledu na přesnost algoritmu je v každé úloze prostor pro další vylepšení. V mém případě lze v budoucí práci dosáhnout zlepšení zvýšením přesnosti výpočtů, například použitím novějších a rozsáhlejších verzí databáze, konkrétně současných MovieLens 20M a MovieLens 25M. Kromě toho by s výše uvedenými databázemi mohlo lze bychom bylo bez problémů pracovat s použitím modernějšího a výkonnějšího počítače, který by data zpracovával mnohonásobně rychleji. V budoucnu lze tento algoritmus také kombinovat s existujícími algoritmy s otevřeným zdrojovým kódem či s použitím open source zdrojů, nebo vyvíjet k tomu nové algoritmy a rozšířit takém způsobem doporučení.

6. Summary

The purposes of this thesis were to analyse what a recommendation algorithm is, create an overview of its' usage in electronic commerce, mostly in online shops and streaming platforms. With all the knowledge gained, the goal was to create a recommendation algorithm of our own with the focus on its accuracy.

Our analysis has shown that recommendation algorithms have become an essential part of e-commerce platforms, providing users with relevant product suggestions. The results of analysis of platforms such as of Amazon, Netflix and Spotify provide evidence of algorithms' effectiveness in enhancing the user experience and driving sales. The insights gained can be useful for businesses looking to improve their e-commerce strategies and provide personalized and satisfying user experience.

The goal set for the practical part was to develop a recommendation algorithm with a focus on parameter and recommendation accuracy. I created and analysed collaborative algorithms using the Surprise library in Python that use cross-validation techniques. These algorithms are SVD or Singular Value Decomposition, NMF or Non-negative Matrix Factorization and PMF or Probabilistic Matrix Factorization. All these algorithms are unique and have their own advantages and peculiarities. Mean Absolute Error and Root Mean Square Error were used to determine the accuracy of the algorithms. The lower the value, the more accurate the algorithm is. After comparing all the average RMSE and MAE values of these 3 algorithms, it was clear that the most accurate algorithm for us is the SVD algorithm. This algorithm helps to recommend movies to a unique user. Improvements can be made in future work by increasing the accuracy of the calculations, for example by using newer and larger versions of the database, specifically the current MovieLens 20M and MovieLens 25M. In addition, the above databases could easily be handled using a more modern and powerful computer that would process the data faster. In the future, this algorithm could also be combined with existing open-source algorithms or using open-source resources, or new algorithms could be developed to extend the recommendations.

Key words: recommendation algorithm, recommendations, e-commerce, Python.

7. Seznam použitých zdrojů

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734–749. DOI: 10.1109/TKDE.2005.99.

Aggarwal, C. C. (2016). Recommender Systems – The Textbook. Springer. ISBN: 978-3-319-29659-3.

Anderson, C. (2006) *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion. ISBN 1401302378.

Chen, T., Liu, J., Wu, Y., Tian, Y., Tong, E., Niu, W., Li, Y., Xiang, Y., & Wang, W. (2021). Survey on Astroturfing Detection and Analysis from an Information Technology Perspective. *Security and Communication Networks*, ID 3294610. DOI: 10.1155/2021/3294610.

Dong, Z., Wang, Z., Xu, J., Tang, R., & Wen, J. (2022). *A Brief History of Recommender Systems*. Cornell University. DOI: 10.48550/arXiv.2209.01860.

Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative Filtering Recommender Systems. *Foundations and Trends in Human–Computer Interaction*, 4(2), 81-173. DOI: 10.1561/1100000009.

Falk, K. (2019). *Practical Recommender Systems*. Manning Publications. ISBN 9781617292705.

Floorwalker, M. (2021). *What It Was Really Like to Subscribe to Netflix In The '90s.* Looper. Dostupné z: https://www.looper.com/327206/what-it-was-really-like-to-subscribe-to-netflix-in-the-90s/.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12), 61-70. DOI: 10.1145/138859.138867.

Harper, F. M., & Konstan, J. A. (2015). The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* (TiiS), 5(4), 19. DOI: 10.1145/2827872.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., Riedl, J. T., & Lui, J. C. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1), 5-53. DOI: 10.1145/963770.963772.

Huang, W., Liu, B., & Tang, H. (2019). *Privacy Protection for Recommendation System: A Survey.* Journal of Physics: Conference Series, 1325, 012087. DOI: 10.1088/1742-6596/1325/1/012087.

Hug, N. (2020). *Surprise: A Python library for recommender systems*. Columbia University. DOI: 10.21105/joss.02174.

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: An introduction.* Cambridge University Press. DOI: 10.1017/CBO9780511763113.

Konstan, J. A., & Riedl, J. (2012). *Recommender systems: From algorithms to user experience. User Modeling and User-Adapted Interaction*, 22(1-2), 101-123. DOI: 10.1007/s11257-011-9112-x.

Langville, A. N., & Meyer, C. D. (2012). *Who's #1? The science of rating and ranking*. Princeton University Press.

Linden, G., Smith, B., & York, J. (2003). *Amazon.com recommendations: item-to-item collaborative filtering*. IEEE Internet Computing, 7(1), 76-80. DOI: 10.1109/MIC.2003.1167344.

Lops, P., de Gemmis, M., & Semeraro, G. (2011). Content-based Recommender Systems: State of the Art and Trends. In L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 73-105). Springer. DOI: 10.1007/978-0-387-85820-3_3.

Maity, A. (2021). *Understanding Netflix's Personalized Recommendation System.* Medium. Dostupné z: https://maity-anirban06.medium.com/understanding-netflixs-personalized-recommendation-system-1d1157478d28.

Melville, P., & Sindhwani, V. (2017). Recommender Systems. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 1156-1160). Springer. DOI: 10.1007/978-1-4899-7687-1_964.

Palmer, A. (2022) *Amazon sues two companies that allegedly help fill the site with fake reviews*. CNBC. Dostupné z: https://www.cnbc.com/2022/02/22/amazon-sues-alleged-fake-reviews-brokers-appsally-rebatest.html.

Pariser, E. (2011). *The filter bubble: How the new personalized Web is changing what we read and how we think*. Penguin.

Pilgrim, M. (2014). *Dive into Python*. Apress.

Plummer, L. (2017). *This is how Netflix's top-secret recommendation system works*. Wired. Dostupné z: https://www.wired.co.uk/article/how-do-netflixs-algorithms-work-machine-learning-helps-to-predict-what-viewers-will-like.

Ricci, F. L., Rokach, B., Shapira, & Kantor, P. B., (Eds.), *Recommender Systems Handbook* (pp. 1-35). Springer US. DOI: 10.1007/978-0-387-85820-3_1.

Schafer, J. B., Konstan, J. A., & Riedl, J. (1999). Recommender Systems in E-Commerce. *The 1st ACM Conference on Electronic Commerce* (pp. 158-166). DOI: 10.1145/336992.337035.

Sunstein, C. R. (2017). *Republic: Divided democracy in the age of social media*. Princeton University Press.

Sunwall, A. (2021). *Recognize strategic opportunities with long-tail data*. Nielsen Norman Group. Dostupné z: https://www.nngroup.com/articles/long-tail/ .

8. Seznam obrázků

9. Seznam tabulek

10.Seznam grafů