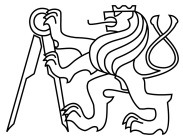
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ



ZADÁNÍ DIPLOMOVÉ PRÁCE

Název:	Doporučovací systém pro výběr volitelných předmětů
Student:	Bc. Ondřej Nový
Vedoucí:	Ing. Magda Friedjungová
Studijní program:	Informatika
Studijní obor:	Znalostní inženýrství
Katedra:	Katedra teoretické informatiky
Platnost zadání:	Do konce letního semestru 2017/18

Pokyny pro vypracování

V rámci doporučeného průchodu studijním programem má každý student povinnost zvolit si volitelné předměty daného programu. Tyto předměty pak tvoří 20% z celkového objemu předmětů, které student musí absolvovat. Volitelných předmětů je pro bakalářské obory v nabídce přes 50, pro magisterské pak přes 90. Student je tak zahlcen nabídkou předmětů a často neví, které si zvolit. Možným řešením je vytvoření doporučovacího systému, který studentovi poskytne personalizované doporučení předmětů k zápisu.

Pokyny pro vypracování:
1) Seznamte se s problematikou doporučovacích systémů využívaných ve školství.
2) Navrhněte a implementujte několik doporučovacích modelů v závislosti na možných typech doporučení.
3) Vyhodnoťte přesnost doporučení jednotlivých modelů na reálných datech (zdroj DWH ČVUT).
4) Výsledky doporučení prezentujte jako prototyp.
5) Proveďte uživatelské testování výsledných doporučení.
6) Proveďte analýzu možností integrace vašeho řešení do portálu s reporty (tzv. EBIE).

Seznam odborné literatury

Dodá vedoucí práce.

doc. Ing. Jan Janoušek, Ph.D.
vedoucí katedry

prof. Ing. Pavel Tvrdík, CSc.
děkan

V Praze dne 17. prosince 2016

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra teoretické informatiky

Diplomová práce

Doporučovací systém pro výběr volitelných předmětů

Bc. Ondřej Nový

Vedoucí práce: Ing. Magda Friedjungová

8. května 2017

Poděkování

Děkuji vedoucí práce **Ing. Magdě Friedjungové** za velmi cenné rady, a za to, že mi byla kdykoliv v případě problémů k dispozici.

Dále děkuji **Ing. Michalu Valentovi, Ph.D.** za iniciativnost při snaze umožnit používání doporučovacího systému studenty.

Dále děkuji všem **studentům**, kteří byli ochotni vyzkoušet doporučovací systém a poskytnout zpětnou vazbu.

Nakonec děkuji své přítelkyni **Věře Procházkové** a své **rodině** za trpělivost a podporu.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 8. května 2017 .....

České vysoké učení technické v Praze

Fakulta informačních technologií

Odkaz na tuto práci

Nový, Ondřej. *Doporučovací systém pro výběr volitelných předmětů.* Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2017.

Abstrakt

Cílem práce bylo vytvořit doporučovací systém volitelných předmětů pro studenty Fakulty informačních technologií ČVUT. Bylo vyzkoušeno více metod doporučování, konkrétně kolaborativní filtrování založené na matici a kolaborativní filtrování založené na dopředné neuronové síti. K uživatelskému testování byla vybrána jednodušší metoda založená na matici, jelikož dokázala lépe odhadnout budoucí známku studenta, což autor považoval za důležitý aspekt toho, aby studenti doporučení důvěřovali. Výsledný doporučovací systém je možné dále vylepšovat, tato práce obsahuje detailní analýzu problematiky i dostatečně obecně předzpracovaná vstupní data, a veškerou další práci v této oblasti výrazně usnadňuje.

Klíčová slova doporučovací systém, doporučování volitelných předmětů, educational data mining, kolaborativní filtrování

Abstract

The aim of the thesis was to create a recommendation system of optional courses for students of the Faculty of Information Technology of CTU. Several recommendation methods have been tried, specifically memory-based

collaborative filtering and collaborative filtering based on forward neural network. The simpler, memory-based method was chosen for user testing, as it was better in predicting the student's future mark, which was considered by the author to be an important aspect to receive trust from students. The resulting recommendation system can be further improved, the work provides a detailed analysis of the topic as well as sufficiently pre-processed input data, and any further work in this area is therefore facilitated.

Keywords recommendation system, optional courses recommendation, educational data mining, collaborative filtering

Obsah

Úvod **1**
Cíl práce 1

I Teoretická část **3**

1 Rešerše **5**
1.1 Společné faktory FITu a škol ve zkoumaných pracích 5
1.2 Specifikum FITu oproti školám ve zkoumaných pracích 5
1.3 Kritéria doporučování podle studenta 6
1.4 Doporučování předmětů podle použitých dat 6
1.5 Modelování doporučovacích systémů 9
1.6 Nejvíce používané metody k doporučování předmětů 12
1.7 Shrnutí 15

2 Použitá metodika **17**
2.1 Porozumění byznysu 17
2.2 Porozumění datům 18
2.3 Předzpracování dat 19
2.4 Modelování 19
2.5 Ohodnocení 19
2.6 Nasazení 20

3 Porozumění byznysu **21**
3.1 Porozumění problematice FITu 21
3.2 Plán projektu 27

4 Porozumění datům **29**
4.1 Zdroje dat 29
4.2 Shrnutí potenciálně užitečných informací 32

4.3 Předměty 32
4.4 Student a studium 37
4.5 Studium předmětu 38
4.6 Zpětná vazba studentů 43
4.7 Zpětná vazba studentů na předměty 44

II Praktická část **47**

5 Předzpracování dat **49**
5.1 Struktura denormalizované tabulky 49
5.2 Sjednocení podobných předmětů 49
5.3 Role předmětu 51
5.4 Zahrnuté semestry 53
5.5 Student a studium 53
5.6 Známky 54
5.7 Využití dat z ankety 55
5.8 Výsledná denormalizovaná tabulka 57
5.9 Předzpracování pro konkrétní modely 57

6 Modelování **61**
6.1 Data 61
6.2 Měření kvality modelu 61
6.3 Optimalizace parametrů modelů 62
6.4 Parametry modelu vycházející z dat 63
6.5 Jednoduché doporučování založené na průměru 64
6.6 Kolaborativní filtrování založené na matici 67
6.7 Dopředná neuronová síť 73
6.8 Model který bude použit k doporučování 82

7 Ohodnocení **85**
7.1 Testovací data 85
7.2 Výsledky na testovacích semestrech 85
7.3 Uživatelské testování 87

8 Nasazení **91**
8.1 Architektura doporučovacího systému 91
8.2 Integrace s reportovacím portálem 92
8.3 Miniaplikace v portálu ČVUT 92

Závěr **95**

Literatura **97**

A Seznam použitých zkratek **101**

B Obsah přiloženého CD **103**

Seznam obrázků

2.1 Fáze metodiky CRISP-DM a jejich návaznost 18

6.1 Jaké hodnoty *průměrné absolutní chyby* se vyskytují v různých variantách jednoduchého doporučování založeného na průměru . . 65
6.2 Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 5 doporučených* se vyskytují v různých variantách jednoduchého doporučování založeného na průměru 66
6.3 Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 10 doporučených* se vyskytují v různých variantách jednoduchého doporučování založeného na průměru 66
6.4 *Odhad správné známky* versus *poměr skutečně zapsaných předmětů v prvních 5 doporučených předmětech* pro různé varianty kolaborativního filtrování 70
6.5 Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 5 doporučených* se vyskytují v různých variantách kolaborativního filtrování využívající přínos 70
6.6 Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 10 doporučených* se vyskytují v různých variantách kolaborativního filtrování 71
6.7 *Odhad správné známky* versus *poměr skutečně zapsaných předmětů v prvních 5 doporučených předmětech* pro různé varianty neuronové sítě 78
6.8 Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 5 doporučených* se vyskytují v různých variantách neuronové sítě využívající přínos 78
6.9 Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 10 doporučených* se vyskytují v různých variantách neuronové sítě . . 79

8.1 Podoba testovací verze portálu ČVUT. 93

Seznam tabulek

3.1 Klasifikační stupnice [1] 25
3.2 Analýza toho, jak jednotlivé metody splňují vytyčené požadavky. . 28

4.1 Četnosti předpon kódů předmětů. 33
4.2 Předměty se stejným názvem lišící v kódu nejen příponou kódu. . 34
4.3 Navzájem zaměnitelné předměty. 36
4.4 Četnost hodnot sloupce *typ_ects*. 37
4.5 Četnost počtu zápisů studenta do bakalářského studia. 38
4.6 Četnost počtu zápisů studenta do magisterského studia. 38
4.7 Počty zapsaných předmětů pro jednotlivé semestry. 39
4.8 Četnost kombinací hodnot sloupců týkajících se klasifikace. 40
4.9 Počty úspěšné a neúspěšně absolvovaných a stále studovaných v jednotlivých semestrech. 42
4.10 Četnost známek. 42
4.11 Názvy oddílů ankety Fakulty informačních technologií. 43
4.12 Otázky v anketě Fakulty informačních technologií. 44

5.1 Mapování hodnot sloupce *typ_ects* na vlastní kódy role předmětu. 52
5.2 Ukázka denormalizované tabulky která je výsledkem Předzpracování dat. 58

6.1 Doporučení pro semestr *B131* vytvořené jednoduchým doporučováním založeným na průměru. 68
6.2 Příklad odhadu známek vytvořený vybranou optimální verzí kolaborativního filtrování. 73
6.3 Příklad doporučení vytvořeného vybranou optimální verzí kolaborativního filtrování. 74
6.4 Příklad odhadu přínosu vybranou optimální verzí neuronové sítě. . 81
6.5 Příklad doporučení vytvořeného vybranou optimální verzí neuronové sítě. 83

7.1 Výsledky optimálního modelu. 86
7.2 Výsledky upraveného optimálního modelu, pokud je měřeno stejným způsobem jako byl optimální model. 86
7.3 Výsledky upraveného optimálního modelu, pokud uvažujeme všechny volitelné bakalářské předměty. 87
7.4 Výsledky upraveného optimálního modelu, pokud uvažujeme všechny volitelné magisterské předměty. 87

Úvod

Současným trendem je stálý růst množství dat která jsou generována různými systémy či chytrými zařízeními. Také je dostupné stále větší množství informací které jsou určeny přímo pro člověka, ovšem místo toho aby přinášela užitek, často způsobují jen informační slepotu, protože jich je zkrátka příliš mnoho.

V obou případech je potřeba data či informace zpracovat různými metodami a přinést tak informace s vyšší užitnout hodnotou

Na Fakultě informačních technologií ČVUT v rámci doporučeného průchodu studijním programem má každý student povinnost zvolit si volitelné předměty. Tyto předměty pak tvoří 20 % z celkového objemu předmětů, které student musí absolvovat. Volitelných předmětů je pro bakalářské obory v nabídce přes 50, pro magisterské pak přes 90. Student je tak zahlcen nabídkou předmětů a často neví, které si zvolit. Možným řešením je vytvoření doporučovacího systému, který studentovi poskytne personalizované doporučení předmětů k zápisu.

Fakulta informačních technologií, jako hlavní fakulta zaměřená na informatiku, je pochopitelně v rozvoji zpracování dat vzorem pro celé ČVUT. Příkladem je datový sklad ČVUT, který vznikl původně jako datový sklad FIT a následně rozšířil působnost na celou univerzitu. Tento potenciál je zde i v oblasti doporučování volitelných předmětů.

Cíl práce

Cílem práce je seznámit se s problematikou doporučovacích systémů využívaných ve školství, navrhnout a implementovat několik doporučovacích modelů v závislosti na možných typech doporučení a vyhodnotit přesnost jednotlivých modelů na reálných datech z datového skladu ČVUT. Doporučovací systém

by měl cílit na prezenční studenty bakalářských a magisterských studijních programů, kterých je na Fakultě informačních technologií většina. Výsledný doporučovací systém by měl být prezentován jako prototyp a mělo by být provedeno uživatelské testování vytvářených doporučení. Dále je potřeba provést analýzu možností integrace doporučovacího systému do portálu s reporty (tzv. EBIE).

Část I

Teoretická část

Kapitola 1

Rešerše

V rámci rešerše bylo prostudováno celkem 18 článků a závěrečných prací [2–20] majících za úkol studentům usnadit práci při výběru volitelných předmětů.

Seznam těchto prací není ani zdaleka vyčerpávající. Prostudovány byly především práce nejnovější, které by měly obsahovat nejnovější výzkum v této oblasti.

1.1 Společné faktory FITu a škol ve zkoumaných pracích

Každá prozkoumaná práce [2–12] se zabývá doporučováním, které má svoje specifika, typicky se ale zabývají doporučováním vysokoškolských předmětů pro bakalářské nebo magisterské programy.

1.2 Specifikum FITu oproti školám ve zkoumaných pracích

V [2, 3, 5, 8, 10] je explicitně zmíněno, že student si musí na dané univerzitě již na začátku studia vybrat obor studia.

Společný obor je přirozeně dobrý indikátor podobnosti studentů, jenž může doporučovací systém výhodně využít, jak učinili autoři [3, 5, 10, 14].

Na FITu mají studenti tu výhodu, že si svůj obor nemusí vybírat předem, ale až v pokročilé fázi studia (viz Porozumění byznysu). Díky tomu si obor mohou zvolit až na základě absolvování určitých předmětů a zjištění, jestli je bude bavit. Pro doporučovací systém z toho vyplývá, že u velké části studentů nezná obor, nebo ho může jen odhadovat na základě absolvovaných předmětů.

1.3 Kritéria doporučování podle studenta

Doporučovací systém by měl pro každého studenta vybrat takové předměty, které by si on sám vybral, kdyby měl všechny relevantní informace.

1.3.1 Průzkum kritérií studenta při výběru předmětu

Tím co je pro studenty relevantní při výběru předmětů se zabývá výzkum [13] týkající se studentů z akademických roků 2009-2010 a 2011-2012. Podle něj někteří studenti očekávají od předmětů dobrý základ pro svou budoucí kariéru, jiní mají za cíl, aby je předmět co nejvíce bavil. Také mají od ideálních předmětů jiná očekávání studenti s dobrými známkami než ti se špatnými. Výzkum říká, že v oblasti informatiky se studenti zaměřují více na svůj profesní základ. Další faktory, které studenti uvedli ve výzkumu jako relevantní, je časová náročnost, vyučující a každý bere v potaz termín ve kterém bude výuka probíhat.

1.3.2 Shrnutí

Pro studenta relevantní informace tedy jsou:

1. jak ho připraví na budoucí profesy,
2. jak ho předmět bude bavit,
3. jak bude časově náročný,
4. jak mu sednou vyučující,
5. jaké místnosti a časy paralelek budou dostupné.

přičemž ale lze tušit, že pro každého studenta je každý bod jinak důležitý.

1.4 Doporučování předmětů podle použitých dat

1.4.1 Profesní základ

Byl nalezen jen jeden doporučovací systém s ohledem na toto kritérium [8]. Přestože toto kritérium by mělo být nejdůležitější nejen z pohledu studenta, ale i z pohledu vzdělávacího systému jako takového, není vůbec jednoduché na něm doporučovací systém postavit, jelikož je potřeba zjistit jaké dovednosti skutečně přináší konkrétní předměty a jak tyto dovednosti namapovat na reálné profese.

CareerMaker

CareerMaker [8] je unikátní doporučovací systém který je postaven právě na tomto kritériu.

Zavádí takzvané Core Values, což jsou dovednosti jako například *Programování*, *Práce v teamu*, *Schopnost hovořit více jazyky*. Dále je potřeba určit množinu pracovních pozic a předmětů. Pro každou pracovní pozici expert určí z kolika procent je pro ní relevantní jaká dovednost. Tedy například pro pracovní pozici *Software engineer* je relevantní na 50 % *Programování*, na 35 % *Práce v teamu* a na 15 % *Schopnost hovořit více jazyky*. Analogicky pro každý předmět je určeno z kolika procent se absolvováním předmětu student zlepšuje v konkrétní dovednosti.

Na základě toho, jakou si student vybere kariéru, pak systém doporučí nejlepší předměty ke zlepšení potřebných dovedností. Navíc je možné poskytovat zpětnou vazbu ohledně předmětů a dovedností které poskytují, takže se nestane, že v praxi předmět poskytuje úplně jiné dovednosti než deklaruje.

Shrnutí

Doporučování na základě dovedností užitečných pro budoucí profesy je možné, ovšem vyžaduje expertní znalosti, a proto se v praxi příliš nepoužívá.

1.4.2 Vyučující

V [3] je jeden typ doporučovaní založen na oblíbenosti vyučujícího. Oblíbenost je spočítána jako součet jím vedených oblíbených předmětů. Lepší výsledky ale ukázalo doporučování založené na podobnosti studentů, jelikož mělo vyšší přesnost.

V [5] byly z výsledků naopak odfiltrovány předměty vedené neoblíbenými vyučujícími, což ale nemělo významný vliv na výsledek.

Shrnutí

Doporučování na základě vyučujících předmětu má smysl jen jako doplňkové kritérium.

1.4.3 Podobnost studentů na základě známek

Zdaleka nejčastěji používanou metodou v prozkoumaných pracích je doporučování na základě známek. Používají ji [2,4–7,11,14–16]. To ovšem nemusí být známka toho, že tento přístup je nejlepší. Tento přístup totiž požaduje data, jenž jsou zdaleka nejlépe dosažitelná.

Na základě studentů, kteří mají podobné známky a známek které tito podobní studenti získali z ještě studentem neabsolvovaných předmětů se predikuje jaké známky nejspíš dostane a jaké předměty by si měl tudíž zapsat.

Výhody:

- tato metoda je obvykle velice efektivní v predikování známek, takže student má velice jistý odhad jaké známky dostane [16].

Hypotetické výhody:

- student kterého předmět baví mu věnuje více energie, a proto má lepší známku. Z toho důvodu vlastně metoda předpovídá předměty, které studenta budou bavit.

Hypotetické nevýhody:

- metoda nahrává předmětům, které jsou snadno absolvovatelné, což jsou často úplně jiné předměty než ty, ze kterých získá nejvíce užitečných profesních znalostí,
- také nahrává zapisování předmětů ve kterých se probírá látka kterou už student zná odjinud.

V [3] se tato metoda používá čistě k predikci známek. Známka kterou student pravděpodobně dostane je zobrazena jen jako přídavná informace k navrženým doporučením - navržený předmět je obarven zeleně, oranžově nebo červeně.

Práce provedená na Masarykově univerzitě v Brně [20] vyslovuje a potvrzuje hypotézu, že dovednosti studenta lze rozpoznat na základě absolvovaných předmětů a výsledků z nich a na základě těchto dovedností/výsledků určit, jak bude student úspěšný v nějakém ještě neabsolvovaném předmětu.

Shrnutí

Přestože má podobnost na základě známek významné hypotetické nevýhody, bylo dokázáno, že sdružuje studenty s podobnými dovednostmi. V praxi je využívaná především díky dostupnosti vstupních dat.

1.4.4 Podobnost studentů na základě oblíbených předmětů

V [3] byla použita metoda podobnosti na základě oblíbených předmětů, která ze všech vyzkoušených měla nejlepší výsledky při doporučování čistě volitelných předmětů.

Na rozdíl od podobnosti na základě známek je zde explicitně určeno, že se studentovi předmět líbil. To co to v praxi znamená záleží čistě na kritériích konkrétního studenta, což může být výhoda i nevýhoda této metody.

Nevýhodou této metody je, že data o oblíbenosti předmětů vyžadují explicitní ohodnocení studentem, tedy zvyšují vytížení studenta.

Shrnutí

Doporučování na základě oblíbených předmětů je hypoteticky velmi flexibilní metoda vzhledem k individuálním požadavkům každého studenta, vyžaduje ovšem data které není jednoduché získat a v praxi se příliš nevyužívá.

1.4.5 Sociální vazby

V [20] byla potvrzena hypotéza, že účast přátel v předmětu zvyšuje šanci na jeho úspěšné zakončení. Přátelé zde byli rozpoznáni podle kooperace v univerzitním informačním systému.

Lze navíc předpokládat, že předmět bude studenta i více bavit, pokud ho budou zároveň absolvovat jeho přátelé.

Shrnutí

Předměty má smysl doporučovat na základě toho, že se do nich zapisují studenti, které dotyčný student zná.

1.5 Modelování doporučovacích systémů

1.5.1 Obecná klasifikace doporučovacích systému

Pro začátek je nutné zmínit, že klasifikace doporučovacích systémů rozhodně není jednotná a je v podstatě v každé literatuře trochu jiná. Proto je potřeba brát následující kapitolu s nadsázkou, jen jako přehled nejčastějších metod v doporučování a vysvětlení jejich principů.

Pokud není uvedeno jinak jsou všechny následující myšlenky v této části přejaty z [21].

Základní členění doporučovacích systému je na **doporučování založené na znalostech**, kde doporučovací systém je vytvořen čistě odlitím znalostí do-

ménového experta o tom, jak souvisí preference uživatelů s doporučovanými objekty [6], a na metody které jsou založeny na datech a na přístupech data miningu, tedy učení z dat.

Nevýhoda doporučování založeného na znalostech je potřeba získání znalostí od experta, který ani nemusí existovat nebo nemusí mít dostatek relevantních informací. Tyto znalosti navíc mohou časem zastarávat. [6]

Data miningové přístupy pro doporučovací systémy jsou **kolaborativní filtrování** a **doporučování založené na obsahu**.

Kolaborativní filtrování je založené na podobnosti uživatelů v preferencích k doporučovaným objektům a je nezávisle na doméně pro kterou je určeno. Při doporučování je potřeba seznam doporučovaných objektů a ohodnocení objektů uživatelem i ostatními uživateli, pokud se s objekty již setkali.

Typickými problémy, kterými trpí kolaborativní filtrování je sparsity problem a cold-start problem.

Doporučování založené na obsahu vychází z metadat doporučovaných objektů a je tak závislé na doméně. K rozpoznání relevantních objektů k doporučení může používat metadata uživatele (explicitní zájmy a preference) nebo metadata uživatelem dříve ohodnocených objektů. [22]

Nevýhoda doporučování založené na obsahu je, že je závislé na dostatečných a správných metadatech. Navíc vede k tomu, že nejsou doporučovány objekty nějakým způsobem nové nebo inovativní.

Shrnutí

Každý typ doporučování (doporučování založené na znalostech, kolaborativní filtrování, doporučování založené na obsahu) má odlišné silné a slabé stránky, proto bývají někdy s úspěchem kombinovány. Takovým doporučovacím systémům se pak říká **hybridní**.

1.5.2 Kolaborativní filtrování

Tato metoda, dále jen KF, funguje na základě matice uživatel-objekt, kde konkrétní pozice v matici udává preferenci uživatele k danému objektu.

Rozdělujeme dva typy KF podle toho, zda je matice uživatel-objekt držena v paměti a používaná přímo při vybavování tj. **KF založené na matici** (dále dělíme na přístup založený na uživateli nebo objektu), nebo zda je použita jen jako vstup pro model při trénování **KF založené na modelu**.

KF založené na uživateli

Přístup **založený na uživateli** je postaven na myšlence *Zjisti jaké preference bude mít uživatel k objektu na základě toho jaké preference k němu mají uživatelé s podobnými preferencemi.*

Příklad jednoduchého algoritmu založeného na uživateli je následující [23]:

1. spočítej podobnost všech ostatních uživatelů k uživateli, kde podobnost je obvykle kosínová podobnost,
2. pro každý objekt predikuj preferenci uživatele, tím, že uděláš vážený průměr přes preference ostatních uživatelů k danému předmětu, kde váha je podobnost uživatele; predikovanou preferenci lze ještě normalizovat součtem podobností všech ostatních uživatelů k uživateli,
3. doporučuj sestupně podle predikovaných preferencí.

Vylepšením jednoduchého algoritmu může být, že k vytvoření predikce se uvažuje jen k nejbližších sousedů uživatele. [16]

Jiným vylepšením může být to, že se uvažují jen ti sousedé, kteří překročí minimální povolenou (prahovou) podobnost. [16]

KF založené na objektu

Přístup **založený na objektu** je postaven na myšlence *Zjisti jaké objekty se podobají objektu na základě preferencí které jim dávají ostatní uživatelé a na základě preferencí, které k nim uživatel má odhadni preferenci k objektu.*

Příklad jednoduchého algoritmu založeného na objektu je následující [24]:

1. spočítej matici podobností objekt-objekt, kde podobnost je obvykle Pearsonova korelace nebo kosínová podobnost,
2. pro každý objekt predikuj preferenci uživatele, tím, že uděláš vážený průměr již přidělených preferencí, kde váha je podobnost objektu s preferencí a objektu bez preference; predikovanou preferenci lze ještě normalizovat součtem podobností všech objektů s preferencemi k objektu,
3. doporučuj sestupně podle predikovaných.

KF založené na modelu

Tato technika používá matici uživatel-objekt k naučení modelu. Vypočet doporučení je tak obvykle rychlejší, protože ve vybavovací fázi se již nemusí pracovat se vstupní maticí.

Sparsity problem

Problém nastává, pokud uživatele celkově udělili málo preferencí. V takovém případě dochází ke špatným doporučením, jelikož není možné najít uživatele s podobnými zájmy.

Cold-start problem

Pokud do systému přibude nový uživatel nemůže mu být nic doporučeno / kvalitně doporučeno, dokud nevytvoří dostatečný počet preferencí.

1.5.3 Asociační pravidla

Doporučovací systém založený na asociačních pravidlech, označován v literatuře často jako RARE [11,12,15], lze klasifikovat jako KF založené na modelu.

Modelem jsou zde pravidla ve tvaru:

Pokud se uživatel setkal s objektem A a B, doporuč mu objekt C.

Učení tohoto modelu je na základě databáze takzvaných transakcí, kde transakcí je v tomto kontextu myšlena množina objektů, se kterými se uživatel setkal.

Pro každé pravidlo lze měřit dvě následující čísla, která vyjadřují jeho kvalitu:

Support Je procentuální zastoupení transakcí, které obsahují objekty A, B i C.

Confidence Je poměr transakcí které obsahují A i B, ku transakcím které obsahují A, B i C.

K vytvoření pravidel se používá *Apriori algoritmus*. Přesná podoba tohoto algortimu není příliš důležitá, a proto zde nebude uvedena. Důležité je že vytvoří jen taková pravidla, která mají support a confidence větší, než jsou pro model nastavené **minimální support** a **minimální confidence**. [19]

1.6 Nejvíce používané metody k doporučování předmětů

V následující části budou zmíněny příklady aplikace různých přístupů k doporučovacím systémům na univerzitách ve světě.

Obecné pojmy *uživatel*, *objekt* zde budou zaměněny za pojmy *student*, *předmět*.

1.6.1 Doporučování založené na znalostech

Národní univerzita v Jeju, Jižní Korea

Systém založený na externích znalostech místo dat byl použit v [10], kde expert nastavoval důležitost oborů, relevanci předmětů pro jednotlivé obory a závislosti předmětů mezi sebou.

Čínská univerzita kultury v Taipeji, Taiwan

Další takový systém byl použit v [8], kde byl nutný expert k nastavení mapování předmětů na dovednosti a dovedností na profese.

1.6.2 Asociační pravidla

Univerzita v Montréalu

V [12] je základem pro asociační pravidla informace o absolvovaných předmětech. Tato technika je zde použita pro start doporučovacího systému, který ještě nemá data od studentů a který postupně začne fungovat čistě na základě jejich zpětné vazby.

Univerzita v Solapuru, Indie

V [19] se snaží doporučovat kurzy pro elektronický vzdělávací systém Moodle. Základem pro asociační pravidla je zde informace o absolvovaných předmětech.

Předpříprava dat zde spočívá v odstranění kurzů, které byly absolvovány málo studenty a odstranění studentů, kteří absolvovali málo kurzů. To kvůli tomu, že jinak vznikala jen pravidla, které nic nedoporučovala (Na pravé straně pravidla bylo jen samé *NE*).

Jordánská univerzita

V [4] se na základě známek pomocí k-means s Manhatanskou vzdáleností určí skupiny podobných studentů. Pro každou skupinu se pak vytvářejí zvlášť asociační pravidla. Položky transakce zde nejsou jen předměty, ale dvojice *(předmět, známka)*.

1.6.3 Kolaborativní filtrování založené na paměti

Indický institut managementu v Indore

V [11] byla snaha předpovídat známky z volitelných předmětů na základě již absolvovaných předmětů a na základě toho vytvářet doporučení.

Vstupní dataset obsahoval 255 studentů a 25 předmětů. Podle autorů, metody kolaborativního filtrování založené na paměti byly v tomto velmi úspěšné, přičemž se významně nelišili přístupy založené na studentovi nebo na předmětu.

Nizozemské střední školy

V [16] bylo použito 20 % holandských studentů středních škol k otestování různých metod kolaborativního filtrování založeného na paměti.

Byly vyzkoušeny přístupy založené na předmětu s různým měřením podobnosti.

Také byly vyzkoušeny přístupy založené na studentovi s různým měřením podobnosti a s různým omezením na k nejbližších sousedů nebo na minimální podobnost.

Lepší výsledky měly metody založené na studentovi. Jejich průměrná absolutní chyba v odhadu známky byla necelých 0.6, tedy necelý jeden stupeň.

Masarykova univerzita v Brně

Při predikci známek na Masarykově univerzitě v Brně [20] bylo použito kolaborativní filtrování založené na studentovi s minimální podobností, která ovšem nebyla konstantou, ale byla vypočítána jako podobnost k jednomu z následujících hypotetických prahových studentů:

Průměrný student Známka z každého předmětu je průměrnou známkou z tohoto předmětu.

Uniformní student Známka z každého předmětu je 2.5.

Tento přístup byl lepší než kolaborativní filtrování s 25 nejbližšími sousedy a než jiné metody založené na modelu, přičemž oba prahový studenti vytvářeli téměř stejné výsledky.

Univerzita v Lisabonu

Práce vycházející z dat 20letého bakalářského a magisterského programu *Informační systémy a počítačové inženýrství* na Univerzitě v Lisabonu [6] testovala metodu kolaborativního filtrování založeného na studentovi pomocí k-sousedů s předchozí redukcí počtu studentů a předmětů pomocí metody **SVD**.

Ukázalo se, že pokud je počet sousedů menší než 300, průměrná absolutní chyba je výrazně menší, pokud se nejdříve použije redukce pomocí SVD. Pomocí SVD, tak lze hypoteticky vylepšit výpočetní náročnost metody kolaborativního filtrování založené na paměti.

1.7 Shrnutí

Jak se ukázalo v rešerši, existuje mnoho vědeckých prací o doporučování volitelných předmětů na vysokých školách. Nejvíce z nich používá jako vstupní data výsledné známky studentů z předmětů, nejspíše proto, že jsou nejlépe a všude dostupná. Mnoho toho bylo již prozkoumáno o modelech pracujících právě se známkami studentů, ale méně bylo zjištěno o modelech pracujících s jinými typy dat, které mají potenciál vytvářet stejná nebo lepší doporučení.

Kapitola 2

Použitá metodika

V této práci bude použita metodika **CRISP-DM** (Cross Industry Standard Process for Data Mining), jelikož se dnes jedná v podstatě o standard pro data miningové projekty, tj. projekty s cílem získávání užitečných znalostí z dat. [25]

Skládá se z následující částí:

1. porozumění byznysu,
2. porozumění datům,
3. předzpracování dat,
4. modelování,
5. ohodnocení,
6. nasazení.

Kapitoly v této práci budou přímo odpovídat částem v CRISP-DM metodice.

Návaznost jednotlivých částí na sebe ukazuje obrázek 2.1.

Následuje detailní popis rozpis jednotlivých částí CRISP-DM.

2.1 Porozumění byznysu

Určení byznysových cílů Pochopení toho co je byznysovým cílem, tedy co ten pro kterého má být hodnota vytvářena opravdu chce.

Obrázek 2.1: Fáze metodiky CRISP-DM a jejich návaznost

Posouzení situace Detailní popis zdrojů, omezení, hypotéz a dalších faktorů, které by měli být brány v potaz.

Určení cílů dolování dat Převedení byznysových cílů na cíle které mohou být dosaženy pomocí modelů pro dolování dat.

Vytvoření plánu projektu Popis zamýšleného plánu pomocí kterého by mělo být dosaženo vytyčených cílů.

2.2 Porozumění datům

Sběr počátečních dat Sesbírání dat ze všech datových zdrojů, případná integrace.

Popis dat Prozkoumání základních vlastností dat.

Prozkoumání dat Průzkum dat vzhledem k data miningovému cíli. Může vyžadovat jednoduché statistické analýzy, vizualizaci, zjišťování vztahů mezi atributy. Již v této fázi je teoreticky možné splnit data miningové cíle.

Ověření kvality dat Zodpovězení otázek: *Jsou data kompletní? Jsou v nich chybějící hodnoty?*

2.3 Předzpracování dat

Výběr dat Rozhodnutí o tom, jaké vzorky budou použity. Výběr zohledňuje data miningové cíle, kvalitu dat, a to, jak velké množství vzorků je technicky možné zpracovat.

Čištění dat Typicky řešení chybějících hodnot.

Konstrukce dat Odvození nových atributů, vytvoření nových vzorků či tranformace atributů.

Integrace dat Sloučení dat z více tabulek do jedné.

Formátování dat Změna formátu dat, který je vhodnější pro konkrétní modelovací techniku.

2.4 Modelování

Výběr konkrétního modelu Výběr teoreticky vhodného modelu nebo modelů na základě data miningových cílů.

Vytvoření testovacího mechanismu Před samotným naučením modelu vytvořit mechanismus, pomocí kterého je možné otestovat jeho úspěšnost.

Vytvoření modelu Naučení modelu nebo modelů na předzpracovaných datech, Výběr ideálních parametrů modelu.

Posouzení modelu Interpretace modelu. Posouzení kvality modelu z více technického hlediska.

2.5 Ohodnocení

Ohodnocení modelu Otestování na testovacích datech. Zjištění, jak dobře model splňuje byznysové cíle - například uživatelským testováním.

Posouzení kvality procesu Zamyšlení, zda proces byl celý proces byl dosud proveden správně.

Vyhodnocení dalšího postupu Rozhodnutí o tom, zda dojde k nasazení, nebo zda se celý proces bude opakovat. Je potřeba zhodnotit i finanční/časové prostředky dostupné pro projekt.

2.6 Nasazení

Získané znalosti musí být prezentovány nebo organizovány tak, aby bylo možné z nich čerpat byznysovou hodnotu. V závislosti na požadavcích se může jednat buď o jednorázový report nebo implementaci opakovatelného data miningového procesu, který je potřeba v budoucnu udržovat a aktualizovat.

Kapitola 3

Porozumění byznysu

3.1 Porozumění problematice FITu

V této kapitole bude vysvětleno, jak vše na fakultě funguje ohledně předmětů a studentů, tedy věcí zásadních pro doporučovací systém.

3.1.1 Studijní programy

Prakticky existuje na FITu jeden program pro bakalářské, jeden pro magisterské studium a jeden pro doktorské studium. Všechny jsou označeny jako *Studijní program Informatika*. Doporučená doba studia pro bakalářské programy je tři roky, pro magisterské dva roky, pro doktorské potom čtyři roky. [26, 27]

Fakulta existuje od roku 2009. První studijní program Informatika měl akreditaci, která skončila. Od akademického roku 2015/2016 se studenti bakalářského programu a od akademického roku 2016/2017 studenti magisterského programu hlásí do nového programu Informatika. Cílem nové akreditace byla aktualizace obsahu předmětů tak, aby odrážely požadavky praxe.

3.1.2 Obory a zaměření

Každý bakalářský i magisterský studijní program má několik oborů ze kterých si student může vybrat. Některé obory se navíc mohou dělit do různých zaměření.

Student nemusí zvolit obor a zaměření ihned, ale kdykoliv v průběhu studia, nejpozději však při schválení tématu závěrečné práce, protože ta se musí týkat konkrétního oboru. To dovoluje studentovi dynamicky si měnit obor podle toho co ho ve skutečnosti začne bavit.

K tomu, aby byl student připuštěn k státní závěrečné zkoušce, musí mít splněny mimo jiné všechny povinné předměty vybraného oboru a zaměření (dále

už jen jako oboro-zaměření). [26]

Následuje výčet oborů a zaměření u jednotlivých studijních programů. Novou akreditací došlo ke změně oborů a zaměření u bakalářského studijního programu.

Bakalářské obory a zaměření původního programu Informatika

V době psaní dostupné z adresy: http://fit.cvut.cz/student/bakalarsky-program/informatika09_14.

- Informační systémy a management
- Informační technologie
- Počítačové inženýrství
- Softwarové inženýrství
- Teoretická informatika
- Web a multimédia

Bakalářské obory a zaměření nového programu Informatika

V době psaní dostupné z adresy: http://fit.cvut.cz/student/bakalarsky-program/informatika15.

- Bezpečnost a informační technologie
- Informační systémy a management
- Počítačové inženýrství
- Teoretická informatika
- Webové a softwarové inženýrství - zaměření Počítačová grafika
- Webové a softwarové inženýrství - zaměření Softwarové inženýrství
- Webové a softwarové inženýrství - zaměření Webové inženýrství
- Znalostní inženýrství

Magisterské obory a zaměření nového a původního programu Informatika

V době psaní dostupné z adresy: http://fit.cvut.cz/student/magistersky-program.

- Počítačová bezpečnost
- Počítačové systémy a sítě
- Návrh a programování vestavných systémů
- Webové a softwarové inženýrství, zaměření Informační systémy a management
- Webové a softwarové inženýrství, zaměření Softwarové inženýrství
- Webové a softwarové inženýrství, zaměření Webové inženýrství
- Znalostní Inženýrství
- Systémové programování, zaměření Systémové programování
- Systémové programování, zaměření Teoretická informatika

3.1.3 Předměty podle studijního programu a kódy předmětů

Dělení podle typu programu

Předmět obvykle spadá do bakalářského, magisterského nebo doktorského programu. To lze poznat podle předpony kódu předmětu. Například: bakalářský kód - *BI-PA1*, magisterský kód - *MI-PAR* a doktorský kód - *PI-SCN*.

Student studuje výhradně předměty svého programu až na několik výjimek.

Přestože si obvykle nelze standartní cestou zapsat předmět z jiného studijního programu, lze o to požádat u děkana. U bakalářských předmětů BI-3DT (3D Tisk), BI-AND (Programování pro operační systém Android), BI-SQL.1 (Jazyk SQL, pokročilý) bylo jejich zapisování magisterskými studenty tak časté, že jim bylo umožněno jejich zapisování standartní cestou, a tak na ně lze hledět jako na standartní volitelné magisterské předměty.

Výjimkou jsou také povinně-volitelné humanitní předměty, například FI-MPL (Manažerská psychologie), které si lze zapisovat v bakalářském i magisterském programu.

Různé verze předmětů

Verze předmětů speciálně určené pro kombinované studium mají například kód *BIK-PA1*.

Verze předmětů vyučované v angličtině mají například kód *BIE-PA1*.

Verze předmětů pro novou akreditaci mají speciální příponu, například *MI-MVI.16*, z původního *MI-MVI*. Stejně tak předměty jejichž obsah se nějak výrazně změní mohou získat příponu, například *BI-SI1.2*, z původního *BI-SI1*.

3.1.4 Systém kreditů

Na FITu se využívá kreditový systém ECTS (European Credit Transfer and Accumulation System) [28]. Každý předmět je v něm ohodnocen počtem kreditů podle celkové časové náročnosti, kde 1 kredit odpovídá 30 hodinám zátěže pro průměrného studenta. Potřeba je ke splnění bakalářského i magisterského programu potřeba získat při rovnoměrném rozložení studijní zátěže 30 kreditů za semestr.

Minimální počet zapsaných kreditů v semestru je 20, pokud studentovi ke splnění programu nestačí méně.

Maximální počet zapsaných kreditů v semestru je 40.

3.1.5 Role předmětu

Informace o rolích předmětů byly v době psaní dostupné z adresy: http://bk.fit.cvut.cz/cz/role.html.

PP (povinné předměty programu) Jsou povinné, a tedy společné pro všechny studenty stejného programu; některé jsou označovány také jako PT (povinná tělesná výchova), PE (povinné ekonomické), PH (povinné humanitní), PJ (povinné jazykové).

PO (povinné předměty oboru) Jsou povinné pro všechny studenty daného oboru, pro studenty jiného oboru fungují jako volitelné.

PZ (povinné předměty zaměření) Jsou povinné pro všechny studenty daného zaměření, pro studenty jiného oboro-zaměření fungují jako volitelné.

VH (povinně-volitelné humanitní předměty) V bakalářském i magisterském programu je povinné splnit alespoň jeden takto označený humanitní předmět.

Bodové hodnocení	100-90	89-80	79-70	69-60	59-50	<50
Klasifikační stupeň	A	B	C	D	E	F
Číselná klasifikace	1	1,5	2	2,5	3	4

Tabulka 3.1: Klasifikační stupnice [1]

VE (povinně-volitelné ekonomické předměty) Týká se pouze bakalářského programu, je povinné splnit alespoň jeden takto označený ekonomický předmět.

V (volitelné předměty) Jsou předměty, které jsou volitelné pro všechny, tedy pro žádné oboro-zaměření nejsou povinné.

3.1.6 Klasifikační stupnice

3.1.7 Způsob zakončení předmětu

Informace o způsobu zakončení předmětu byly v době psaní převzaty z popisů předmětů dostupných na adrese: http://bk.fit.cvut.cz/cz/prehled.html.

ZK (pouze zkouškou) Student je připuštěn ke zkoušce automaticky, výsledná známka závisí jen na zkoušce.

Z,ZK (zápočtem a zkouškou) Student je ke zkoušce připuštěn po udělení zápočtu a body ze semestru se sčítají s body ze zkoušky a vytváří konečnou známku.

KZ (klasifikovaným zápočtem) Student neabsolvuje zkoušku, ale stačí mu získat zápočet, známku určují jen body ze semestru.

Z (pouze zápočtem) Student neabsolvuje zkoušku, ale stačí mu získat zápočet, z předmětu není známka.

3.1.8 Kódy semestrů

Semestry jsou na FITu, jakožto celé univerzitě ČVUT, označovány kódem.

A v kódu označuje roky ve tvaru *19xy*, **B** v kódu označuje roky ve taru *20xy*. V kódu pak následuje *xy* a poslední číslo v kódu označuje zimní semestr (1) a letní semestr (2).

Příklady:

B091 je zimní semestr akademického roku 2009/2010.

B092 je letní semestr akademického roku 2009/2010.

B152 je letní semestr akademického roku 2015/2016.

A731 je zimní semestr akademického roku 1973/1974.

Speciální kódy semestrů

Existují speciální kódy, které se nepoužívají pro žádný konkrétní semestr. Jsou to pomocné kódy**A00** a**A01** a kódy pro uznávání předmětů**A000-A011**.

3.1.9 Studijní plán

Každé oboro-zaměření má vlastní studijní plán, který určuje, které předměty jsou pro studenta povinné a kolik musí dohromady nastřádat kreditů, aby mohl být připuštěn k státní závěrečné zkoušce.

3.1.10 Doporučený průchod studijním plánem

Doporučený průchod studijním plánem doporučuje studentovi jak si pro konkrétní oboro-zaměření zapsat povinné a povinně-volitelně předměty v jakém semestru a za kolik kreditů si zapsat volitelných předmětů.

Doporučený průchod studentovi usnadňuje zapisování předmětů, jelikož nemusí řešit které předměty si musí povinně zapsat, jaké předměty jsou prerekvizity jiných a podobně.

Doporučené průchody studijním plánem byly v době psaní práce dostupné z adresy: http://bk.fit.cvut.cz/cz/prehled.html .

3.1.11 KOS (Komponenta studium)

KOS je systém společný pro celé ČVUT, kde si studenti zapisují předměty, paralelky, mají přehled o studijních výsledcích, zapisují se na zkoušky atd. Tento systém je dostupný z adresy: https://www.kos.cvut.cz.

3.1.12 Anketa

Anketa je aplikace ČVUT, která realizuje zákonem danou povinnost provádět pravidelné hodnocení činnosti VŠ a zveřejňovat jeho výsledky. V době psaní práce je dostupná z adresy: https://anketa.cvut.cz.

V aplikaci Anketa si každá fakulta vytvoří vlastní dotazník (nebo také anketu, dále jen anketa). Aktuální otázky pro FIT byly v době psaní práce dostupné na: https://anketa.cvut.cz/otazky/otazky_anketa_fit.html .

Obsahem ankety je hodnocení předmětů, učitelů a studia obecně.

Anketa FIT

Obsah ankety se na FITu mezi semestry prakticky nemění.

Studentovi nabídnuto ohodnotit předměty které v předchozím semestru absolvoval. Kromě toho může vyplnit **doplňující otázky**.

U každého absolvovaného předmětu se pak hodnotí **předmět** jako takový, **přednášející**, **cvičící** a **zkouška**.

3.2 Plán projektu

Byla provedena rešerše problematiky ve světě a popis toho, jak funguje Fakulta informačních technologií ČVUT.

Nyní je možné vytyčit si **cíle** doporučovacího systému a na základě **dostupných dat** určit plán jak bude vytvořen.

3.2.1 Cíle

Navržená metoda by měla k doporučování použít informace které jsou pro studenta při výběru předmětů relevantní.

Zároveň ale není žádoucí studenty zatěžovat dalšími dotazníky a otázkami, proto bude potřeba vystačit si s daty, která jsou k dispozici ve stávajících systémech.

3.2.2 Dostupná data

Jsou k dispozici data ze systémů KOS a Anketa.

V systému **KOS** jsou informace o studijních programech, předmětech, zapsaných předmětech, studijních výsledcích, vyučujících, paralelkách apod.

V **Anketě** jsou hodnocení a názory studentů na předměty, vyučující, zkoušky a studium obecně.

3.2.3 Výběr metod

Aby bylo možné zvolit vhodnou metodu pro doporučování, bylo stanoveno několik subjektivních kritérií, na základě kterých bude metoda vybrána.

Stanovená kritéria:

Data jsou v získaných zdrojových datech potřebné informace pro danou metodu doporučování?

metoda	data	nerizikové
Profesní základ		
Vyučující	X	
Podobnost studentů na základě známek	X	X
Podobnost studentů na základě oblíbených předmětů	X	
Sociální vazby		

Tabulka 3.2: Analýza toho, jak jednotlivé metody splňují vytyčené požadavky.

Nerizikové byla metoda použita minimálně ve dvou pracích zmíněných v rešerši a riziko, že doporučování nebude dobře fungovat, je tak menší?

Tabulka 3.2 ukazuje, jak jednotlivé typy algoritmů splňují stanovená kritéria.

Vybrané typy metod

Na základě vytyčených požadavků byla vybrána metoda používající **podobnost studentů na základě známek**.

Dále byla vybrána metoda **podobnost studentů na základě oblíbených předmětů**. Přestože nebyla ještě dostatečně vyzkoušena v praxi, jedná se o dobrý způsob, jak využít poměrně cenná data z Ankety.

3.2.4 Výběr algoritmů

Vybrané metody jsou definované spíše používanými daty než konkrétním způsobem implementace, proto je ještě potřeba určit, jak budou konkrétně implementovány.

Metody *podobnost studentů na základě známek* i *podobnost studentů na základě oblíbených předmětů* lze z pohledu doporučovacích systémů klasifikovat jako *kolaborativní filtrování*. Obě lze implementovat pomocí jednoduchého matice uživatel-objekt nebo pomocí modelu.

Bude použit algoritmus *kolaborativního filtrování založeného na matici*, jelikož je jednoduchý na implementaci, a jak již v této fázi autor zjistil, dat není tolik, aby to pro tento algoritmus byl problém.

Dále bude jako experiment použito kolaborativní filtrování založené na neuronové síti. Tento přístup nebyl v žádné prozkoumané práci použit a bude tak zjištěno, zda se složité modely jako je neuronová síť hodí k doporučování předmětů.

Kapitola 4

Porozumění datům

Byla provedena rešerše problematiky doporučování předmětů ve světě a bylo popsáno fungování Fakulty informačních technologií ČVUT z hlediska předmětů a studentů.

V této části již budou zkoumána konkrétní data, která byla poskytnuta k implementaci doporučovacího systému.

Jedná se o data ze školních systémů KOS a Anketa. Nejdříve bude prozkoumáno jaké všechny informace se v datech nachází. Následně budou zkoumána ta data, která autor posoudil jako užitečná pro doporučování.

4.1 Zdroje dat

Zdrojem dat pro vytvoření doporučovacího systému bude export ze systémů **KOS** a **Anketa** provedený v první polovině ledna 2017. Přesněji, nejedná se export přímo z těchto systémů, ale z datového skladu ČVUT, kam jsou data z těchto systémů nahrávána.

Přestože Anketa a KOS jsou celouniverzitní systémy, data obsažená v exportu jsou pouze ta relevantní pro Fakultu informačních technologií.

Soubory exportu mají převážně formát CSV. Každý soubor představuje tabulku z datového skladu, z relační SQL databáze.

4.1.1 Vysvětlení pojmů použitých vzhledem ke zdrojům dat

Soubory versus tabulky

Soubory exportu jsou fyzicky převážně ve formátu CSV, přesto o nich bude hovořeno spíše jako o tabulkách, jelikož představují tabulky z relační databáze.

NULL versus None versus NaN

Nevyplněná hodnota atributu je v SQL obvykle reprezentována pomocí **NULL**. V této práci jsou tabulky zpracovány pomocí knihovny jazyka Python, kde je taková hodnota obvykle reprezentována jako **None**. Navíc je použita knihovna NumPy, která takovou hodnotu reprezentuje jako **NaN**

Kdekoliv se tedy v této práci vyskytnou pojmy **None** či **NaN**, jedná se v původních datech o hodnotu **NULL**.

4.1.2 Seznam exportovaných souborů z KOSu

t_orgj_organizacni_jednotka.csv Organizační jednotka univerzity jako je *fakulta*, *katedra* nebo *ústav*. Jednotka může mít nadřízenou jednotku jako například katedra spadá pod fakultu.

t_osob_osoba.csv Studenti a zaměstnanci fakulty. Jejich datum narození, pohlaví, tituly.

t_arok_semestr.csv Překlad *Kódu semestru* na *Název semestru*. Informace kdy semestry začínají a končí.

t_behpredmetu_ucitel_rel.csv Informace o tom, které osoby zastávají jaké funkce v konkrétním běhu předmětu.

t_para_paralelka.csv Seznam paralelek pro jednotlivé běhy předmětů. Jejich kapacita.

t_pred_beh_predmetu.csv Konkrétní běh předmětu v nějakém semestru, jeho kapacita a organizační jednotka která ho má na starost.

t_pred_predmet.csv Informace o předmětu. Jeho kód, název, způsob zakončení, počet kreditů za jeho dokončení.

t_pred_skupina_predmetu.csv Shromažďování předmětů do skupin.

t_predmet_predmet_vztah_rel.csv Popis vztahů mezi předměty. Například, pokud předmět nahrazuje jiný, nebo pokud je jeden prerekvizitou druhého.

t_predmet_skupinapredmetu_rel.csv Přiřazení role k předmětu.

t_prih_prihlask.xlsx Podrobnější informace o studentovi, které byly získány z přihlášky ke studiu. Například vystudovaná střední škola a výsledky, přechozí studium a výsledky, způsob přijetí, bydliště, místo narození.

t_stpl_studijni_obor.csv Obory a zaměření.

t_stpl_studijni_program.csv Bakalářské, magisterské i doktorandské studijní programy.

t_stud_studium.csv Studentovo studium konkrétního studijního programu. Například student absolvující bakalářský i magisterský program má tedy v tabulce dva záznamy. Informace o zahájení, ukončení či přerušení studia.

t_studijniobor_studijniprogram_rel.csv Informace, pod které studijní programy spadají jaké obory.

t_szzk_statni_zkouska.csv Informace o závěrečné zkoušce pro konkrétní studium. Datum absolvování, výsledky.

t_ucit_ucitel.csv Existence osoby v této tabulce dává najevo, že se jedná o učitele. Organizační jednotka pod kterou učitel spadá (typicky katedra).

t_ucitel_paralelka_rel.csv Informace o tom, kteří učitelé učí na konkrétních paralelkách.

t_zapi_klasifikace.csv Klasifikace studentova zapsaného předmětů - datum udělení a osoba která udělila zápočet, známku nebo uznala zapsaný předmět.

t_zapi_zapsany_predmet.csv Předmět zapsaný v konkrétním studentově studiu v nějakém semestru.

4.1.3 Seznam exportovaných souborů z Ankety

anke_anketa.csv Název ankety a semestr ve kterém se konala a organizační jednotka která ji vytvořila.

anke_hodnoceni.csv Konkrétní předmět hodnocený v nějaké anketě anonymním studentem, o kterém je zde pouze jeho známka z předmětu a studijní průměr.

anke_hodnoceni_cislo.csv Studentem vybrané odpovědi na otázky.

anke_oddil.csv Oddíly v konkrétní anketě.

anke_odpoved.csv Možné odpovědi na otázku v anketě.

anke_otazka.csv Otázky položené studentovi v oddílu ankety, pořadí otázky, informace, zda je povinná či nikoliv.

anke_vyplnil.csv Záznam zde ukazuje, že student při svém konkrétním studiu ohodnotil v daný semestr konkrétní předmět v rámci konkrétní ankety.

fakulta.csv Seznam fakult a ústavů ČVUT.

katedra.csv Seznam kateder.

predmet.csv Informace duplicitní s **t_pred_predmet.csv**.

studium.csv Informace duplicitní s **t_stud_studium.csv**.

4.2 Shrnutí potenciálně užitečných informací

4.2.1 Informace z KOSu

V exportu z KOSu se nachází informace o:

- existujících **předmětech**, jejich náročnosti, způsobu zakončení, obsahově stejných předmětech a prerekvizitách,
- **studentech** jejich původu, studiích mimo FIT i na FITu,
- **účasti studentů a učitelů** na konkrétních bězích předmětu, stejně jako na konkrétních paralelkách přednášek a cvičení,
- **studijních výsledcích** studentů v předmětech i v závěrečné zkoušce,
- **struktuře** studijních programů, oborů a kateder.

4.2.2 Informace z Ankety

V exportu z KOSu se nachází informace o:

- **názorech studentů** na předměty, učitele a studium obecně,
- **studijních výsledcích respondentů**.

4.3 Předměty

Předmět je v kontextu této práce subjekt doporučování.

Informace o předmětech jsou dostupné v tabulce *t_pred_predmet*, konkrétně například jeho kód, název, způsob zakončení, počet kreditů za jeho dokončení.

	Četnost	Relativní četnost
BI	207	0.25875
MI	161	0.20125
BIE	130	0.16250
MIE	119	0.14875
BIK	107	0.13375
PI	34	0.04250
PIK	19	0.02375
FI	10	0.01250
FIT	4	0.00500
R	2	0.00250
U3V	1	0.00125
X000802	1	0.00125
X000801	1	0.00125
FIE	1	0.00125
X000621	1	0.00125
AKCE	1	0.00125
VYJEZD	1	0.00125

Tabulka 4.1: Četnosti předpon kódů předmětů.

4.3.1 Relevantní předměty

Bylo rozhodnuto že budou doporučovány jen bakalářské a magisterské předměty prezenčního studia v češtině.

V tabulce 4.1 jsou vidět četnosti různých předpon všech unikátních kódů předmětů, získaných z tabulky *t_pred_predmet*.

Z tabulky lze vyčíst, že nejvíce zastoupeny jsou bakalářské a magisterské předměty prezenční (*BI*, *MI*), následované anglickými variantami (*BIE*, *MIE*) variantami pro kombinované studium (*BIK*, *MIK*). Následují doktorské předměty (*PI*, *PIK*) a humanitní předměty (*FI*). Zbytek přípon se vyskytuje už jen velice řídce.

V práci bude dále pracováno už jen s bakalářskými a magisterskými předměty prezenčního studia v češtině (*BI*, *MI*).

4.3.2 Předměty se stejným názvem ale jiným kódem lišící se jen příponou kódu

Jak již bylo zmíněno v Porozumění byznysu, existují předměty, u kterých se částečně změní jejich obsah, například novou akreditací, a tak získají nový kód předmětu. Typicky se jen za původní kód přidá přípona, například *MI-MVI*

nazev_cs	kod
Hardwarová bezpečnost	MI-HWB.16, BI-HWB
Jazyk C# - přístup k datům	BI-PCS, BI-CS2
Programovací praktika	BI-ACM, BI-ACM2, BI-ACM3
Řízení podnikové informatiky	MI-RIC, MI-MBI.16

Tabulka 4.2: Předměty se stejným názvem lišící v kódu nejen příponou kódu.

se změní na *MI-MVI.16*.

U takových předmětů se obvykle nezmění název, systémově se ale vytvoří nový záznam v tabulce *t_pred_predmet*.

Na základě unikátních kódů předmětů z tabulky *t_pred_predmet* bylo zjištěno, že existuje 99 kolizí v názvech předmětů, kde se kódy kolizních předmětů liší jen v příponě.

4.3.3 Předměty se stejným názvem ale jiným kódem lišící se nejen příponou kódu

Na základě unikátních kódů předmětů z tabulky *t_pred_predmet* byly nalezeny 4 kolize v názvu předmětu, kde se kód duplicitních předmětů nelišil pouze v příponě. Tyto duplicity jsou vypsány v tabulce 4.2.

Na základě dat z tabulky *t_pred_predmet* byly porovnány kolizní předměty, aby bylo zjištěno, jak moc jsou stejné. Tabulka *t_pred_predmet* obsahuje mimo jiné sloupec *schvaleno*, pomocí kterého lze rozpoznat, který z kolizních předmětů vzniknul dříve.

Analýza kolizi (Šipka znamená, že předmět nalevo vznikl dříve než předmět napravo):

MI-RIC -> MI-MBI.16 Snížila se náročnost.

BI-PCS -> BI-CS2 Změnil se semestr výuky z letního na zimní, jinak předmět vypadá stejně. Vzhledem k rozdílu data schválení 6 let lze předpokládat že se jedná o novou verzi předmětu, nikoliv navazující předměty se stejným jménem.

BI-HWB a MI-HWB.16 Deset měsíců po BI-HWB vytvořen MI-HWB. BI-HWB je vyučován v letním, MI-HWB naopak v zimním semestru. MI-HWB je doporučen pro 1. ročník magisterského program, BI-HWB pro 3. ročník bakalářského. To nahrává tomu, že se jedná o jeden předmět vyučovaný jedním vyučujícím pro oba typy programů.

BI-ACM, BI-ACM2, BI-ACM3 Předměty popisem stejné, konající se popořadě v letním, zimním a letním semestru. Schválené postupně zhruba rok po sobě. To napovídá tomu, že se jedná o jeden 'předmět' rozdělený do třech semestrů.

4.3.4 Prakticky stejné předměty s jiným kódem i názvem

Může se stát, že při změně obsahu předmětu se kromě kódu změní i název. V této části bude diskutováno, jak se takové předměty poznají a zda má cenu takto změněný předmět považovat stále za ten stejný.

Tabulka *t_predmet_predmet_vztah_rel* obsahuje vztahy mezi předměty. Předměty které jsou v nějakém vztahu jsou ve sloupcích *fk_predmet_prvni* a *fk_predmet_druhy*. Typ vztahu je obsažen ve sloupci *vztah*.

Náhrada za předmět

Jedním ze vztahů je *náhrada předmětu* (*vztah=R*). Náhrada za předmět je asymetrický vztah.

Jak je uvedeno v souboru *tvztpred_ciselnik.txt*, který je součástí exportu:

> Nechť student má ve svém studijním plánu předmět B. Jestliže je nastaven vztah "předmět A je náhradou za předmět B", potom při kontrole studijních plánů studenta je předmět B považován za studentem absolvovaný i tehdy, když student neabsolvoval předmět B, ale absolvoval předmět A.

Předměty A a B jsou tedy řekněme plně zaměnitelné, pokud A nahrazuje B a B nahrazuje A. Pokud jen A nahrazuje B, pak nejspíše A bude těžší předmět a vedení fakulty nechtělo, aby si studenti ulehčovali studium studiem lehčí varianty.

Zaměnitelné předměty

Dva předměty tedy lze považovat za jeden, pokud se navzájem nahrazují.

V tabulce *t_predmet_predmet_vztah_rel* se nachází 11 takových případů, které jsou vypsané v tabulce 4.3.

4.3.5 Povinné versus volitelné předměty (Role předmětu)

Doporučovací systém by se měl zabývat pouze doporučováním volitelných předmětů. V této části bude provedena diskuze toho, jaké předměty lze považovat za volitelné.

	kod1	kod2
0	BI-EKP	BI-EPD
1	BI-EKP	BI-EPD.2
2	BI-EPD	BI-EPD.2
3	BI-SI1	BI-SI1.2
4	BI-SI1	BI-ZSI
5	BI-SI1.2	BI-ZSI
6	BI-SI2	BI-SI2.2
7	BI-TWA	BI-WT2
8	BI-WMM	BI-WT1
9	MI-PAR	MI-PAR.1
10	MI-TES	MI-TES.1

Tabulka 4.3: Navzájem zaměnitelné předměty.

Data o roli předmětu

Tabulka *t_pred_predmet* ve sloupci *typ_ects* podle hodnot kterých nabývá evidentně určuje zda je předmět volitelný, povinný či povinný pro nějaký obor.

Informace o roli předmětu, vypadá to, lze získat také z tabulky *t_pred_skupina_predmetu*, která obsahuje skupiny, do kterých lze předměty sdružovat, přičemž tyto skupiny jsou obvykle například *Volitelné předměty bakalářského programu Informatika, kombinovaná forma, verze 2015* a z Tabulka *t_predmet_skupinapredmetu_rel*, pak mapuje na sebe předměty a tyto skupiny. Tato cesta je ovšem složitější, a proto v této práci použijeme výše zmíněnou, tedy sloupec *typ_ects* v tabulce *t_pred_predmet*.

Hodnoty sloupce typ_ects

Četnost hodnot sloupce *typ_ects* je zobrazena v tabulce 4.4.

Jak je vidět, vyskytují se předměty, které jsou povinné pro určité obory a volitelné pro ostatní obory či předměty které jsou povinně volitelné a zároveň volitelné. Obě tyto kombinace se zdají být logické z hlediska fungování fakulty.

Předmětů, které tuto hodnotu nemají vyplněnou je 39. Z těchto předmětů, například *MI-IOS* je volitelný předmět, ostatní jsou většinou nové verze povinných předmětů, u kterých možná ještě tato informace ještě nebyla datově ošetřena.

	Četnost	Relativní četnost
PO	112	0.304348
V;	75	0.203804
P	48	0.130435
NaN	39	0.105978
V	37	0.100543
PO;	20	0.054348
P;	9	0.024457
PZ	7	0.019022
PV;	5	0.013587
PV	4	0.010870
PO;V;	4	0.010870
PZ;	3	0.008152
PO;PZ;	2	0.005435
PV;V;	2	0.005435
PZ;V;	1	0.002717

Tabulka 4.4: Četnost hodnot sloupce *typ_ects*.

4.4 Student a studium

Studium představuje studenta zapsaného v konkrétním studijním programu. Studentem pak můžeme chápat kohokoliv kdo absolvoval či právě absolvuje nějaké studium.

Tato část se bude zabývat analýzou studenta a studií, přičemž budeme brát v úvahu pouze bakalářské a magisterské programy.

4.4.1 Data o studentech a studiích

Tabulka *t_stud_studium* obsahuje studia studentů. Ukládají se zde informace o zahájení, ukončení či přerušení studia. Sloupec *fk_studijniprogram* určuje k jakému studijnímu programu se studium vztahuje. Sloupec *fk_osoba_peridno* zase určuje k jaké fyzické osobě se studium vztahuje.

Tabulka *t_osob_osoba* obsahuje informace o fyzických osobách, a to i jiných než studentech, například učitelích.

Tabulka *t_stpl_studijni_program* obsahuje informace o studijních programech. Například sloupec *kod* studijního programu má podobný tvar jako kód předmětu. Z předpony kódu studijního programu lze určit, zda se jedná o bakalářský (BI), magisterský (MI) či jiný program. V této práci se omezíme na bakalářské a magisterské programy.

	Četnost	Relativní četnost
1	4938	0.876153
2	628	0.111427
3	55	0.009759
4	14	0.002484
5	1	0.000177

Tabulka 4.5: Četnost počtu zápisů studenta do bakalářského studia.

	Četnost	Relativní četnost
1	1375	0.913621
2	117	0.077741
3	13	0.008638

Tabulka 4.6: Četnost počtu zápisů studenta do magisterského studia.

4.4.2 Počty studentů a studií

Z výše zmíněných tabulek bylo zjištěno, že za celou historii fungování fakulty existuje 6420 bakalářských studií a 5636 bakalářských studentů, 1648 magisterských studií a 1505 magisterských studentů.

4.4.3 Opětovné zápisy

Z výše uvedených počtů je jasné, že někteří studenti museli být zapsání na jeden studijní program opakovaně. Tabulka 4.5 uvádí četnosti toho kolikrát byly studenti zapsány na bakalářský studijní program. Tabulka 4.6 uvádí stejné četnosti pro magisterský studijní program.

4.4.4 Navazující studium

Navazujícím studiem se zde myslí studium magisterského programu, s předchozím studiem bakalářského programu na Fakultě informačních technologií, počet navazujících studentů zjištěný na základě výše zmíněných tabulek je 832. Tedy z celkových 1505 magisterských studentů je to zhruba 55 %.

4.5 Studium předmětu

4.5.1 Data o zapsaných předmětech

Tabulka *t_zapi_zapsany_predmet* obsahuje informace o předmětech zapsaných/studovaných studenty. Sloupec *fk_predmet* určuje jaký předmět je za-

	Četnost	Relativní četnost
A000	7041	0.047240
B091	3185	0.021369
B092	1426	0.009567
B101	7272	0.048790
B102	5625	0.037740
B111	10342	0.069388
B112	8741	0.058646
B121	11519	0.077284
B122	8838	0.059297
B131	12536	0.084108
B132	8820	0.059176
B141	12328	0.082712
B142	8971	0.060189
B151	13257	0.088945
B152	8038	0.053929
B161	12279	0.082383
B162	8829	0.059236

Tabulka 4.7: Počty zapsaných předmětů pro jednotlivé semestry.

psán, sloupec *fk_studium* pak studium v rámci kterého je zapsán. Sloupec *fk_semestr* určuje pro jaký semestr zápis platí.

4.5.2 Data o klasifikaci

Tabulka *t_zapi_klasifikace* obsahuje dodatečné informace o klasifikaci. Sloupec *fk_zapsanypredmet_bk*, říká, k jakému zapsanému předmětu se klasifikace vztahuje, sloupce *zakonceno*, *zapocteno*, *znamka* pak dávají ucelený obraz o výsledku studia předmětu.

4.5.3 Počty zapsaných předmětů pro jednotlivé semestry

Jak je vidět v tabulce 4.7, počet zapsaných předmětů první dva roky existence fakulty prudce rostl, a poté se víceméně ustálil.

Dále je vidět velký rozdíl mezi zimním a letním semestrem, kde v zimním semestru je počet zapsaných předmětů vždy výrazně vyšší.

Semestr *A000*, ve kterém bylo zapsáno značné množství předmětů, zastřešuje předměty, které byly uznány, tedy nebyly fyzicky absolvovány.

	zakonceno	zapocteno	znamka	počet
0	None	None	None	30695
1	None	N	None	7939
2	None	Z	None	1457
3	A	Z	None	9421
4	A	None	A	2939
5	A	Z	A	20001
6	A	None	B	1988
7	A	Z	B	15761
8	A	None	C	2013
9	A	Z	C	16191
10	A	None	D	2141
11	A	Z	D	16327
12	A	None	E	1713
13	A	Z	E	10983
14	None	None	F	933
15	None	N	F	3261
16	None	Z	F	5284

Tabulka 4.8: Četnost kombinací hodnot sloupců týkajících se klasifikace.

4.5.4 Analýza klasifikace

Rozpoznání úspěšně absolvovaného, neúspěšně absolvovaného a stále studovaného předmětu

V tabulce 4.8 můžeme vidět četnosti kombinací hodnot sloupců *zakonceno*, *zapocteno* a *znamka* z tabulky *t_zapi_klasifikace*.

V závislosti na způsobu zakončení předmětu lze vysledovat následující případy:

1. předmět má pouze zkoušku. V datech se zřejmě jedná o předměty, kde *zapocteno=None*. Předmět je úspěšně absolvován, pokud *zakonceno=A*, které se vyskytuje jen se *znamka!=F*,

2. předmět má pouze zápočet bez klasifikace. Úspěšné absolvování takového předmětu se v datech projevuje jako *zakonceno=A*, *zapocteno=Z* a *znamka=None*,

3. pokud *zakonceno=A*, *zapocteno=Z* a *znamka!=F*, jedná se o úspěšné absolvování předmětu se zápočtem a zkouškou nebo jen s klasifikovaným zápočtem,

4. všechny ostatní případy v tabulce výše lze považovat za neúspěšné absolvování předmětu. Tedy student buď nedostal zápočet, nebo ve všech pokusech při zkoušce či klasifikovaném zápočtu získal klasifikaci F.

Z výše uvedených bodů vyplývá, že pokud *zakonceno=A*, pak byl předmět úspěšně absolvován, jinak ne.

Pokud nebyl předmět úspěšně absolvován, může to znamenat neúspěch nebo to, že je předmět stále studován.

Teoreticky by předmět, který byl neúspěšně absolvován a má klasifikaci měl být poznat tak, že *znamka=F*, ovšem podíváme-li se na tabulku, takových předmětů je pouze zhruba 9 tisíc (řádek 0,1,2 v tabulce 4.8), zatímco předmětů *zakonceno=None* a *znamka=None* je zhruba 40 tisíc (řádek 13,14,15 v tabulce 4.8), což rozhodně není počet právě studovaných předmětů v probíhajícím semestru.

Možné varianty jsou, že oněch 40 tisíc zahrnuje studenty, kteří si nechají zápis předmětu zrušit v systému KOS nebo, že v některých případech vyučující nepřidělí známku F, přestože by měl, jelikož výsledek je pak stejný - student nedostane za předmět kredity.

Ať je pravda jakákoliv, z výše uvedeného plyne, že nelze bez informace, jaký probíhá semestr jednoznačně rozlišit, zda je předmět stále studován nebo byl neúspěšně absolvován.

Rozdělení úspěšné a neúspěšně absolvovaných a stále studovaných přes jednotlivé semestry

Z tabulky 4.9 můžeme vyčíst, že v prvním semestru vůbec *B091* je jen malá část předmětů označena jako neúspěšně absolvované.

Také je zde patrný trend, že v zimním semestru bývá neúspěšné absolvovaných předmětů zhruba stejně jako předmětů N, zatímco v letním je neúspěšně absolvovaných předmětů zhruba dvakrát méně než předmětů N. Tyto fakta se v rámci této práce nepodařilo interpretovat.

Naopak je evidentní, že semestr B162 stále probíhá, jelikož žádné předměty ještě nebyly absolvovány.

Úspěšnost

Pokud ignorujeme i semestr B161, kde jsou ještě některé zapsané předměty stále studovány. Úspěšnost absolvování předmětu je 72.24 %.

STATE fk_semestr	N	NA	UA
B162	8829.0	NaN	NaN
B161	2628.0	2591.0	7060.0
B152	1587.0	936.0	5515.0
B151	1886.0	2481.0	8890.0
B142	2357.0	785.0	5829.0
B141	1677.0	1969.0	8682.0
B132	1882.0	818.0	6120.0
B131	1958.0	2028.0	8550.0
B122	1687.0	822.0	6329.0
B121	1351.0	1696.0	8472.0
B112	1573.0	833.0	6335.0
B111	1081.0	1054.0	8207.0
B102	866.0	477.0	4282.0
B101	1166.0	659.0	5447.0
B092	321.0	200.0	905.0
B091	1302.0	68.0	1815.0
A000	1.0	NaN	7040.0

Tabulka 4.9: Počty úspěšné a neúspěšně absolvovaných a stále studovaných v jednotlivých semestrech.

	Četnost	Relativní četnost
A	22940	0.153911
B	17749	0.119083
C	18204	0.122136
D	18468	0.123907
E	12696	0.085181
F	9478	0.063591
NaN	49512	0.332191

Tabulka 4.10: Četnost známek.

Četnost známek

Četnost hodnot pro sloupec *znamka* je zobrazena v tabulce 4.10. Z tabulky lze vyčíst, že četnost známek B, C a D je zhruba stejná, četnost horších známek E a F je nižší, četnost A je naopak o něco vyšší.

	id_oddilu_bk	nazev_oddilu
0	1037	PREDMET
1	1038	PŘEDNÁŠEJÍCÍ
2	1039	2. PŘEDNÁŠEJÍCÍ
3	1040	CVIČÍCÍ / VEDOUCÍ PROSEMINÁŘE
4	1041	2. CVIČÍCÍ
5	1042	CVIČÍCÍ V LABORATOŘI / POČ. UČEBNĚ
6	1043	2. CVIČÍCÍ V LABORATOŘI
7	1044	ZKOUŠKA A ZKOUŠEJÍCÍ
8	1045	Doplňující otázka
9	1037	PŘEDMĚT
10	1045	DOPLŇUJÍCÍ OTÁZKY

Tabulka 4.11: Názvy oddílů ankety Fakulty informačních technologií.

4.6 Zpětná vazba studentů

4.6.1 Data o anketách

Tabulka *anke_anketa* obsahuje seznam anket v aplikaci Anketa. Primárním klíčem tabulky je kombinace *anketa_id_bk* a *arok_semestr_id_bk*. Pro Fakultu informačních technologií je relevantní *anketa_id_bk=522*. Vzorový název ankety (sloupec *nazev_ankety*) je *FIT - letní semestr 2013/14*.

4.6.2 Data o oddílech

Anketa obsahuje oddíly, což jsou tématicky oddělené skupiny otázek. Seznam oddílů je dostupný z tabulky *anke_oddil*. Oddíl je navázaný na konkrétní anketu pomocí sloupců *anke_anketaanketa_id* a *anke_anketaarok_semestr_id_bk*. V tabulce 4.11 je vidět unikátní kombinace hodnot sloupců *id_oddilu_bk*, *nazev_oddilu*. Z této tabulky lze usoudit, že oddíly se v průběhu let prakticky nemění až na interpunkci v názvu oddílu a podobné drobnosti.

4.6.3 Typy otázek

Jak je vidět patrné z informací výše, oddíly odpovídají částem ankety popsaných z uživatelského hlediska v části Porozumění byznysu. Otázky na studenty se tedy dělí na otázky o předmětu, přednášejících, cvičících, zkoušce a na doplňující otázky.

text_otazka	slovni_otazka	povinnost	poradi_otazky
Předmět byl pro mne přínosem	None	A	1
Studijní materiály byly kvalitní	None	A	2
Souhlasila specifikace předmětu v KOSu s obsahem...	N	A	3
Duplicity nebo nekonzistence s ostatními předměty	A	N	4
Co se vám na předmětu líbilo	A	N	5
Co by se dalo zlepšit	A	N	6
Jak hodnotíte obtížnost předmětu	N	N	7
Můj vztah k předmětu vystihuje možnost	None	None	8
Připravoval(a) jsem se pravidelně během semestru	None	None	9

Tabulka 4.12: Otázky v anketě Fakulty informačních technologií.

4.7 Zpětná vazba studentů na předměty

V této části se omezíme na analýzu zpětné vazby na předměty, tedy na otázky z oddílu *PŘEDMĚT*, tedy *id_oddilu_bk=1037*.

Data o otázkách ankety jsou obsaženy v tabulce *anke_otazka*. Každá otázka je přiřazena do oddílu pomocí sloupce *anke_oddilid_oddilu*.

4.7.1 Otázky o předmětu

V tabulce 4.12 je seznam otázek z tabulky *anke_otazka*, které jsou pokládány studentovi pro každý absolvovaný předmět. Tyto otázky se v průběhu let neměnily.

Sloupec *slovni_otazka* udává zda se jedná o otevřenou otázku (*slovni_otazka=A*) nebo výběr z možností (*slovni_otazka=N*). Sloupec *povinnost_otazky* udává, zda je nutné otázku vyplnit. Sloupce *poradi_otazky* udává, v jakém pořadí jdou otázky za sebou.

Hodnota *None* ve sloupcích *slovni_otazka* a *povinnost_otazky* je v obou případech ekvivaletní hodnotě N.

4.7.2 Možné odpovědi na uzavřené otázky

Tabulka *anke_odpoved* obsahuje seznam odpovědí na uzavřené otázky.

Seznam uzavřených otázek a možných odpovědí na ně:

1. Předmět byl pro mne přínosem
   - rozhodně ano

- spíše ano
- spíše ne
- rozhodně ne
- nevím, neumím se vyjádřit

2. Studijní materiály byly kvalitní
   - rozhodně ano
   - spíše ano
   - spíše ne
   - rozhodně ne
   - nevím, neumím se vyjádřit
3. Souhlasila specifikace předmětu v KOSu s jeho obsahem. V případě negativní odpovědi konkretizujte.
   - rozhodně ano
   - spíše ano
   - spíše ne
   - rozhodně ne
   - nevím, neumím se vyjádřit
4. Jak hodnotíte obtížnost předmětu
   - příliš vysoká
   - trochu vysoká
   - přiměřená
   - mohlo to být těžší
   - zvládl by to "deváťák"
5. Můj vztah k předmětu vystihuje možnost
   - zajímavé téma dobře odprednášené
   - zajímavé téma, ale špatně odprednášené
   - nezajímavé téma, ale dobře odprednášené
   - nezajímavé téma špatně odprednášené
   - ani jedna z těchto variant, neumím se rozhodnout
6. Připravoval(a) jsem se pravidelně během semestru
   - ano
   - občas
   - ne, jen ve zkouškovém období

4.7.3 Anonymní hodnocení

Tabulka *anke_hodnoceni* obsahuje informace o vyplnění ankety studentem. Sloupce *anke_anketaanketa_id* a *anke_anketaarok_semestr_id_bk* určují jaké ankety se vyplnění týká.

Jelikož je anketa anonymní není možné spojit konkrétního studenta s hodnocením předmětu. V této tabulce je tedy alespoň sloupec *znamka*, která prozrazuje známku studenta z hodnoceného předmětu, sloupec *prumer_studenta*, který sděluje průměr studenta, pravděpodobně z dotčeného semestru a sloupec *obor_studenta*, který prozrazuje jeho obor.

Tabulka obsahuje ještě sloupec *predmet_id*, kterým lze identifikovat hodnocený předmět

4.7.4 Odpovědi studentů na uzavřené otázky

Tabulka *anke_hodnoceni_cislo* nese informace o konkétních odpovědích na uzavřené otázky. Sloupec *anke_otazkaid_otazky* říká, jaké otázky se odpověď týká, sloupec *anke_odpovedporadi_odpoved* potom říká, jakou odpověď na otázku student vybral (odpovídá sloupci *poradi_otazky* v tabulce *anke_otazka*). Sloupec *anke_hodnoceniid_hodnoceni* je referencí do tabulky *anke_hodnoceni*, kde lze zjistit jakého předmětu se hodnocení týkalo a anonymizované informace o studentovi.

4.7.5 Odpovědi studentů na otevřené otázky

Data o odpovědích na otevřené otázky nejsou součástí exportu. Jednak by bylo poměrně náročné je exportovat, a za druhé textová analýza otevřených odpovědí, by byla příliš náročná a nad rámec této práce.

4.7.6 Počty odpovědí na uzavřené otázky

Medián počtu odpovědí pro otázku je 48, průměr pak 179.

Některé otázky mají velmi malý počet odpovědí. 784 otázek má jen 8 odpovědí, 801 otázek jen 16, což je dohromady zhruba 30 % ze všech otázek.

Část II

Praktická část

Kapitola 5

Předzpracování dat

V této kapitole budou data, která byla autorem shledána užitečná pro doporučování předzpracována. Předzpracováním je myšleno například změna formátu hodnot ve sloupcích, vygenerování odvozených sloupců či odfiltrování neužitečných záznamů.

Výsledkem předzpracování bude jedná denormalizovaná tabulka, a to proto, aby bylo co nejjednoduší s daty dále pracovat.

Výsledná denormalizovaná tabulka bude následně transformovaná do podoby, kterou vyžadují už konkrétní modely na svém vstupu.

5.1 Struktura denormalizované tabulky

Ze sloupců v původních tabulkách budou vytvořeny nové předzpracované, se kterými rovnou bude možné pracovat v doporučovacím systému. Tabulka bude obsahovat jak tyto nové sloupce, tak původní sloupce, aby bylo možné v případě potřeby pracovat i s původními hodnotami, nebo případně jednoduše odhalit chybu ve vytváření nových odvozených sloupců.

Denormalizovaná tabulka bude obsahovat řádky s granularitou absolvovaného předměty.

Sloupce tabulky lze rozdělit na skupiny o předmětu, o studentovi a hodnocení této kombinace na základě kterého lze doporučovat (budoucí hodnota v matici pro kolaborativní filtrování).

5.2 Sjednocení podobných předmětů

Jak bylo popsáno v Porozumění datům, existuje několik kategorií podobných předmětů. Různé předměty se stejným názvem lišící se jen příponou kódu

předmětu, předměty se stejným názvem lišící se celým kódem, a pak předměty lišící se v kódu i názvu, které lze ovšem identifikovat jako podobné díky tabulce *t_predmet_predmet_vztah_rel*.

5.2.1 Realizace sjednocení předmětů

Sjednocení bude pro jednoduchost realizováno tak, že pro každou množinu podobných předmětů, bude celou množinu reprezentovat kód předmětu který je při abecedním seřazení všech kódů předmětů z množiny první.

5.2.2 Předměty se stejným názvem ale jiným kódem lišící se jen příponou kódu

Jak bylo popsáno v Porozumění datům, předměty se stejným názvem se v drtivé většině liší jen číselnou příponou kódu, například, například *MI-MVI* a *MI-MVI.16*. Tyto přípony se na Fakultě informačních technologií používají k označení nové verze předmětu, například *.16* označuje předměty v nové akreditaci.

Obsah předmětu se nejspíše může trochu změnit (obsahem nové akreditace je ostatně aktualizace osnov předmětů), avšak budeme předpokládat že ne natolik, že by se jednalo o úplně nový předmět, a proto budeme považovat takové předměty za stejné a budou spojeny.

5.2.3 Předměty se stejným názvem ale jiným kódem lišící se nejen příponou kódu

Byly identifikovány 4 kolize tohoto druhu:

MI-RIC -> MI-MBI.16 Náročnost předmětu se snížila. Předměty budou považovány za rozdílné, s tím, že pokud systém doporučí *MI-RIC*, je možné na úrovni uživatele dojít k závěru, že systém doporučil vlastně *MI-MBI.16*, na úrovni doporučovacího systému by se ovšem jednalo o příliš odvážný krok.

BI-PCS -> BI-CS2 Jak bylo odůvodněno v [Porozumění datům], *BI-CS2* je zřejmě nová verze *BI-PCS*, proto budou předměty spojeny.

BI-HWB a MI-HWB.16 Jakkoliv jsou si předměty podobné, jsou vypsány pro rozdílné programy, a proto budou považovány za odlišné.

BI-ACM, BI-ACM2, BI-ACM3 Jedná se zřejmě o jeden řekněme **velký předmět** rozdělený do více semestrů. Zapisovány pravděpodobně bude probíhat postupně, proto budou předměty považovány za odlišné.

5.2.4 Stejné předměty s jiným kódem i názvem

Předměty se mohou lišit kódem i názvem, pokud se ovšem, na základě tabulky *t_predmet_predmet_vztah_rel*, navzájem nahrazují, pak budeme pro účely doporučování předměty považovat za stejné. V praxi to totiž znamená, že lze předměty obousměrně uznávat, což je poměrně dobrá známka toho, že se vlastně jedná o jeden předmět.

5.3 Role předmětu

Doporučovány by měly být pouze předměty které si student může volit. Doporučovat povinné předměty nedává úplně smysl, jelikož si je student stejně musí dříve nebo později zapsat.

Rozdělení na povinné a volitelné na Fakultě informačních technologií ale není jednoznačné. Jak je zmíněno v rešerší existují předměty povinné oboru a povinné zaměření, které jsou povinné pro určité oboro-zaměření, ale volitelné pro jiné. Obor a zaměření můžou být zvoleny až ke konci studia, a to, jestli je předmět pro studenta povinný nebo volitelný je tedy často poměrně relativní.

5.3.1 Data o roli předmětu

Jak bylo zmíněno v porozumění datům, role předmětu bude určována podle sloupce *typ_ects* tabulky *t_pred_predmet*.

Hodnoty ve sloupci významově odpovídají hodnotám v části Role předmětu v kapitole Porozumění byznysu. Například hodnota *PO;V;* odpovídá povinnému předmětu oboru, protože jak již bylo zmíněno, povinný předmět oboru je pro některé obory volitelný. Hodnota *PV* zahrnuje všechny povinně-volitelné předměty tedy ekonomické i humanitní.

5.3.2 Rozdělení

Pro zjednodušení problému budou předměty rozděleny do těchto kategorií. P - povinné, V - volitelné, VP - nerozhodnuté, N - bez informace o roli.

Rozřazení předmětů do těchto kategorií bude následující:

P povinné,

V volitelné, povinně volitelné,

VP povinné oboru, povinné zaměření,

N v datech není informace o roli předmětu.

	typ_ects	OPTIONAL
0	V	V
1	P	P
2	PO	VP
3	PO;	VP
4	PO;PZ;	VP
5	PZ	VP
6	V;	V
7	PV	V
8	PV;	V
9	PO;V;	VP
10	P;	P
11	PZ;	VP
12	null	N
13	PV;V;	V
14	PZ;V;	VP

Tabulka 5.1: Mapování hodnot sloupce *typ_ects* na vlastní kódy role předmětu.

Povinně volitelné předměty byly zařazeny do volitelných, jelikož přestože si student volí z omezené množiny, stále si může zvolit konkrétní předmět.

Jak bylo řečeno u povinných předmětů oboru či zaměření je situace složitější, proto jsou v separátní kategorii.

V tabulce 5.1 je zobrazeno mapování všech existující hodnot sloupce *typ_ects*, na tyto hodnoty.

5.3.3 Role po sjednocení předmětů

V předchozí části o sjednocení předmětů bylo vysvětleno, že podobné předměty budou považovány za jeden. V případě informace o roli předmětu je toto problém, protože podobné předměty mohou mít různou roli. Tato situace není nutně chyba v datech, jelikož role předmětu se může v průběhu let změnit.

Jedna varianta řešení tohoto problému by byla, že by takové předměty za podobné považovány nebyly. Autor této práce se ale rozhodl že budou, a proto je potřeba určit, jak se k této kolizí rolí postavit.

Pokud bude v množině podobných předmětů nějaký VP, bude skupina označena jako VP. Pokud v množině podobných předmětů budou P i V, bude skupina taktéž označena jako VP. V ostatních případech není problém s kolizí. Jen v případě, že žádný z předmětů ve skupině nebude mít informaci o

roli, tedy všechny budou N, bude skupina označená jako N, tedy bez informace o roli.

Předměty bez role

Předměty, které i po spojení nemají informaci o roli jsou celkem 4.

Jedná se o předměty *Povinně volitelný ekonom.-man.*, *Povinně volitelný humanitní*, *Volitelný předmět*, *Pokročilé techniky v iOS aplikacích*.

Jak je evidentní první tři nejsou opravdové předměty ale spíše role předmětů z neznámého důvodu datově reprezentované jako předměty.

Čtvrtý by měl být podle bílé knihy volitelný, a proto bude jako výjimka z výše uvedených pravidel označený jako *V*.

5.4 Zahrnuté semestry

5.4.1 Uznané předměty

Jak bylo zmíněno v Porozumění datům, datově uznávání předmětů probíhá tak, že se vytvoří absolvovaný předmět zařazený do semestru *A000*.

Uznané předměty budou odfiltrovány.

5.4.2 Neukončené předměty

Jak je patrné z Porozumění datům, semestr *B162* neobsahuje žádné neúspěšné nebo úspěšně absolvované předměty, z čehož je evidentní že stále ještě probíhá. Tento semestr bude tedy odfiltrován.

Stejně tak bude odfiltrován semestr *B161*. U toho nelze jednoduše z dat poznat že stále probíhá, avšak export byl proveden ještě v probíhajícím zkouškovém období tohoto semestru, a tak nemusí být klasifikace u všech předmětů finální.

5.5 Student a studium

V Porozumění datům bylo vysvětleno, že student může mít více studií. V rámci stejného studijního programu je tato situace poměrně vzácná, kolem 10 %. Navazujících studentů, tedy těch, co mají více studií v rámci různých studijních programů je ale nezanedbatelných 55 %.

Z toho důvodu by mohlo být užitečné nedělat doporučování na základě studia, ale na základě studenta, což ale vyžaduje další předzpracování dat.

5.5.1 Výhody

Použití studenta místo studia, by mělo zvýšit množství dostupných dat o absolvovaných předmětech. To může být užitečné, jelikož někdy se doporučovací systémy kolaborativního filtrování potýkají se *sparsity problémem*, tj. dat je příliš málo na to, aby bylo doporučování kvalitní.

Důležitější výhoda ale je, že se vyřeší *cold-start* problém pro magisterské studenty. To znamená, že bude možné doporučovat předměty již pro první semestr magisterského studia, a to na základě předchozího bakalářského studia.

5.5.2 Realizace použití studenta místo studia

Všechny informace o předmětech jsou napojeny na studium. Je tedy potřeba vytvořit tabulku, která bude mapovat studium na studenta.

Tato informace je obsažena v tabulce *t_stud_studium*, která obsahuje informace o studiích, kde sloupec *fk_osoba_peridno* říká jakého studenta se studium týká.

5.5.3 Shrnutí

Použití studenta místo studia při doporučování má mnoho potenciálních výhod. Byla tedy vytvořena struktura, která toto umožnuje.

5.6 Známky

Nejčastěji bylo v pracích prozkoumaných v rešerši doporučování předmětů založené na známkách. Tato část se bude zabývat předzpracování známek a klasifikace tak, aby byly použitelné k doporučování.

5.6.1 Předměty pouze se zápočtem

Ne všechny předměty jsou po ukončení hodnocené známkou. Pokud je předmět zakončen **pouze zápočtem**, pak známku nemá.

Převod na známku

Aby mohly i předměty pouze se zápočtem být použity v doporučování, musí být převedeny na známku, respektive na číselnou hodnotu, která má pak nějaký význam vzhledem ke klasifikační stupnici.

Odpověď na otázku, na jakou známku úspěšné ukončený předmět pouze se zápočtem převést není jednoznačná. Může to být A, jelikož studentovi nelze nic vytknout. Nebo to může být C, což je medián ze všech hodnot na klasifi-

kační stupnici, které indikují úspěšné dokončení předmětu. Nebo to může být průměrná známka ze všech úspěšně absolvovaných předmětů se známkou.

Jelikož tuto otázku autor nepovažuje za zásadní pro doporučování, bude zvolena, bez hlubší analýzy, varianta *A*. Tedy úspěšný zápočet bude považován za známku A, číselně 1.0.

5.6.2 Rating

Jelikož v doporučovacích systémech je obvyklé používat hodnoty typu čím větší tím více doporučeno, budou číselné hodnoty známek převedeny tak aby A bylo číselně nejvyšší a F naopak nejnižší. Tyto nové číselné hodnoty známek budou obsaženy ve sloupci *RATING* výsledné denormalizované tabulky.

Číselná hodnota známek typu větší je lepší

0 je hodnota, která reprezentuje stav kdy student předmět neabsolvovat, a to ani úspěšně ani neúspěšně.

A, B, C, D a E by pak měly odpovídat hodnotám větším než 0. Pokud ponecháme původní číselné rozestupy z klasifikační stupnice, mohlo by očíslování být popořadě 3, 2.5, 2, 1.5, 1.

Známka F by mohla být buď záporná, tedy například -1 či -3, nebo by mohla být rovná 0.

Lze předpokládat, že čím menší hodnota, tím větší odrazující efekt bude mít neúspěch studenta v předmětu na doporučení předmětu podobnému studentovi.

Jelikož volba není jednoznačná, budou všechny tři varianty vyzkoušeny ve fázi předzpracování. Datově bude F v denormalizované tabulce reprezentována jako -1, s tím že hodnotu lze bez problému ve fázi modelování přeložit na jinou, jelikož žádná jiný případ nenabývá hodnoty -1.

5.7 Využití dat z ankety

Jak je uvedeno v porozumění datům, studenti odpovídají v rámci ankety na 6 uzavřených otázek ohledně předmětu.

Bohužel odpovědi jsou anonymizované, lze je tak použít například k zjištění toho, jak si který předmět v konkrétních oblastech vede. Tato informace je taková že nelze na základě ní provádět personalizované doporučování, které je cílem této práce, ale jen doporučování obecné, které každému doporučuje stejné položky.

Naštěstí součástí dat je informace o známce, kterou student, který předmět hodnotil, z předmětu dostal. Na základě toho lze vytvořit doporučování, které doporučuje na základě zpětné vazby studentů se stejnou známkou.

5.7.1 Předzpracování sloupce známka

Sloupec *znamka* v tabulce *anke_hodnoceni* obsahuje známky jak ve formátu s desetinnou čárkou, tak s desetinnou tečkou. Všechny hodnoty budou převedeny na formát s desetinnou tečkou, která se ve zdrojích dat jinak vyskytuje ve sloupcích obsahujících známky.

5.7.2 Přínos předmětu

Studenti v anketě odpovídají kromě jiného na otázku, zda byl pro ně předmět přínosem, která je hypoteticky velmi užitečná pro doporučovací systém.

Číselná hodnota odpovědi

Možné odpovědi na tuto otázku jsou *rozhodně ano*, *spíše ano*, *spíše ne*, *rozhodně ne*, *nevím, neumím se vyjádřit*. Tabulka s odpověďmi *anke_odpoved* obsahuje mimo jiné sloupec *poradi_odpoved*, jehož hodnota odpovídá pořadí odpovědí, v jakém jsou zde uvedené.

Číselná hodnota přínosu bude upravena tak, aby měla význam pro doporučování. Konkrétně *rozhodně ano* bude odpovídat číselná hodnota 5, *spíše ano* bude odpovídat číselná hodnota 4, *spíše ne* bude odpovídat číselná hodnota 2, *rozhodně ne* bude odpovídat číselná hodnota 1 a *nevím, neumím se vyjádřit* bude odpovídat číselná hodnota 3.

Přínos předmětu pro studenty se stejnou známkou

Jak je uvedeno výše, data z ankety lze využít, tak, že ke každému absolvovanému předmětu uvedeme, jaký byl pro studenta předmět zhruba přínosem. Toto číslo lze spočítat jako průměr z číselné hodnoty přínosu z odpovědí na otázku přínosu od studentů se známkou stejnou jako byla známka z absolvovaného předmětu daného studenta.

Tento průměrný přínos je samozřejmě jen odhadem. Navíc pro předměty pouze se zápočtem bude tento odhad pro všechny studenty vždy stejný. Je to ale pravděpodobně maximum, které lze z anonymních odpovědí vytěžit.

Nedostatek odpovědí studentů se stejnou známkou

Může se stát, že přestože předmět má na otázku přínosu respondenty, nemá žádné respondenty s konkrétní známkou. V takovém případě je otázkou, jaký by měl být odhad přínosu. Nejpřesnější odhad by mohl být odhad přínosu pro

nejbližší známku, která respondenty má. Z důvodu implementační složitosti ovšem bude použit průměrný přínos předmětu, přes všechny známky, který je pravděpodobně méně přesný.

Počet respondentů

Pro každý absolvovaný předmět bude do tabulky přidána informace o počtu respondentů pro vypočítaný průměrný přínos. Jak je uvedeno v průzkumu dat, zhruba 15 % předmětů má bez ohledu na známku jen 8 respondentů, 15 % má jen 16 respondentů. Bude tak ve fázi modelování možné odfiltrovat hodnocení přínosu s příliš malým počtem respondentů, které by nemuselo být reprezentativní.

5.8 Výsledná denormalizovaná tabulka

Výsledkem veškerého předchozího předzpracování je denormalizovaná tabulka, kde řádek představuje předmět absolvovaný studentem.

Tato tabulka je společná pro všechny modely a bude následně dále předzpracována podle potřeb konkrétních modelů.

5.9 Předzpracování pro konkrétní modely

V této části bude denormalizovaná tabulka jakožto produkt veškerého předchozího předzpracování dále upravena, aby splňovala požadavky vybraných modelů.

5.9.1 Dataset pro kolaborativní filtrování založené na matici

Kolaborativní filtrování založené na matici vyžaduje jako vstup matici, kde jedna dimenze představuje subjekt doporučování a druhá dimenze představuje osoby kterým je doporučováno. V tomto konkrétním případě tedy jde o matici student, předmět.

Hodnoty v matici mohou být buď přímo známky, nebo odhad přínosu, přičemž odhad přínosu je závislý na známkách.

Co představuje řádek v matici

Řádek v matici bude představovat známky (případně přínos) z předmětů absolvovaných studentem. V případě, že jako student bude použito pouze jedno konkrétní studium, budou to jen předměty absolvované v rámci tohoto studia. Pokud jako student bude použit skutečně student, datově údaj *peridno* z tabulky *t_osob_osoba*, bude se jednat o všechny předměty absolvované v rámci fakulty.

kod	nazev_cs	fk_semestr	fk_studium	zakonceno	zapocteno	znamka
MI-TES.2	Teorie systémů	B151	11653206	A	Z	A
MI-VMW	Vyhledávání multimediálního obsahu na webu	B151	11653206	A	Z	A
MI-PAR.2	Paralelní algoritmy a systémy	B151	11653206	A	null	B
MI-MVI.16	Metody výpočetní inteligence	B161	11653206	A	Z	A
MI-MVI.16	Metody výpočetní inteligence	B161	11653206	null	null	null
MI-DSP	Databázové systémy v praxi	B152	11653206	A	Z	B
BI-PRP	Právo a podnikání	B162	11653206	null	null	null
MI-PAA	Problémy a algoritmy	B161	11653206	A	Z	D

username_id	studium_id	peridno	COMM_SUB_CODE	OPTIONAL	GRADE	RATING	AVG_BENEFIT	RESPONSE_CNT
NOVYOND2	11653206	4773606	MI-TES	VP	1.0	3.0	2.808511	752.0
NOVYOND2	11653206	4773606	MI-VMW	VP	1.0	3.0	3.818182	88.0
NOVYOND2	11653206	4773606	MI-PAR	P	1.5	2.5	3.769231	104.0
NOVYOND2	11653206	4773606	MI-MVI	VP	1.0	3.0	4.600000	80.0
NOVYOND2	11653206	4773606	MI-MVI	VP	NaN	0.0	NaN	NaN
NOVYOND2	11653206	4773606	MI-DSP	V	1.5	2.5	4.111111	72.0
NOVYOND2	11653206	4773606	BI-PRP	VP	NaN	0.0	NaN	NaN
NOVYOND2	11653206	4773606	MI-PAA	VP	2.5	1.5	4.148148	864.0

Tabulka 5.2: Ukázka denormalizované tabulky která je výsledkem Předzpracování dat.

5.9.2 Dataset pro neuronovou síť

Převedení problému doporučování na neuronovou síť

Dopředná neuronová síť je velmi univerzální model. Doporučování pomocí ní bude realizováno tak, že vstupními i výstupními neurony budou všechny existující předměty. Hodnota vstupního neuronu bude známka či přínos z odpovídajícího předmětu. Tyto hodnoty musí být typu větší je lepší. Hodnota na výstupním neuronu bude pak předpověď hodnoty stejného atributu v budoucím semestru.

Neuronová síť potřebuje k naučení, jak vstup, tak ideální správný výstup. Pokud budeme uvažovat doporučování pro konkrétní semestr, pak vstupem budou známky/přínos z již absolvovaných předmětů v předchozích semestrech a výstupem budou skutečně získané známky/přínos z předmětů zapsaných v doporučovaném semestru.

Podoba datasetu

Dataset bude tedy podobný jako matice pro kolaborativní filtrování, jen dimenze studenta bude obsahovat i informaci o tom, jakého doporučovaného semestru se řádek týká a dimenze předmětů bude obsahovat dvojnásobné množství předmětů - v předchozích semestrech absolvované a v doporučovaném semestru skutečně zapsané.

5.9.3 Vícekrát zapsaný předmět

V případě vytváření datasetu pro kolaborativní filtrování i neuronovou síť může dojít ke kolizi, a to, pokud student absolvoval předmět vícekrát.

Typicky může student mít známku F, a v budoucím semestru získá lepší známku a předmět úspěšně absolvuje, jelikož se na předmět lépe připraví.

Řešení

Otázkou je, jak se k této situaci postavit, je potřeba totiž soubor známek získaných z předmětu převést na jednu.

Zde jsou tři možné způsoby:

Nejhorší známka Pokud se studentovi předmět nepovede absolvovat, na jeho další pokusy už není brát zřetel.

Nejlepší známka Student si může známku opravit, to že předmět na poprvé neudělal nebude vůbec uvažováno.

Průměr Uvažováno je jak to, že student předmět neudělal, tak že si ho následně opravil.

Jelikož tato otázka není autorem práce považována za nijak zásadní, bude k agregaci více známek použit průměr.

Poznámka

U povinných předmětů lze předpokládat, že student bude opakovat, dokud neabsolvuje, aby splnil studijní plán. U volitelných naopak tato snaha není potřeba a student si místo předmětu nejspíše zapíše nějaký jiný.

Kapitola 6

Modelování

6.1 Data

Ve fázi modelování bude použita jen podmnožina dostupných dat (tzv. trénovací data), a to proto aby mohla být kvalita modelů otestována vzhledem k novým datům (tzv. testovací data) a zabránilo se tak zvolení přeučeného modelu.

Jako testovací data budou sloužit předměty ze semestrů *B151* a *B152* představující akademický rok 2015/2016. Jako trénovací data budou sloužit všechny dřívější semestry, tedy semestry do a včetně akademického roku 2014/2015.

6.2 Měření kvality modelu

Aby bylo možné vyhodnotit, jak jsou modely implementující doporučování kvalitní, je potřeba určit, jak se bude kvalita měřit.

6.2.1 Skutečně zapsané předměty

Úspěšnost doporučení bude validována na základě výsledků z předmětů, které si student skutečně zapsal. Hypoteticky mohou existovat mnohem lepší předměty které kdyby si student zapsal, dosáhl by lepších výsledků nebo by pro něj mohli být větším přínosem, to ale není jak ověřit a poněkud výhodou pro předmět z hlediska doporučování je už to, že si ho student zapsal.

6.2.2 Jsou zapsané předměty v prvních n doporučených

První hodnota vypovídající o kvalitě doporučení bude vyjadřovat kolik procent skutečně zapsaných předmětů se nacházelo v prvních n doporučených předmětech.

Pokud je n menší, než počet skutečně zapsaných předmětů, je v takovém případě n zvětšeno na počet skutečně zapsaných předmětů.

n použité k testování bude 5 a 10. Z logiky věci by takto nastavené meze mohly našemu vyhodnocení postačovat.

6.2.3 Jak moc se liší odhadnuté hodnoty od získaných

Druhá hodnota vypovídající o kvalitě doporučení bude vyjadřovat, jak moc se odhadnutá hodnota blíží skutečné.

Toto bude mít zejména význam u odhadu známek. Pořadí doporučených předmětů sice nezávisí na tom, jak moc jsou známky podobné těm skutečně získaným, avšak pro studenta se může jednat o zajímavou informaci, jelikož pak získá vzhled do toho, jak doporučování funguje a může mu více důvěřovat.

Konkrétně bude to, jak se známky v doporučení podobají skutečně získaným známkám měřit pomocí průměrné absolutní chyby přes všechny zapsané předměty.

Pokud předmět nezíská žádnou hodnotu pro doporučení, což se může stát u kolaborativního filtrování založeném na matici, bude tento předmět z měření absolutní chyby vynechán.

6.3 Optimalizace parametrů modelů

Oba implementované modely mají svoje parametry, jimiž lze určitým způsobem měnit vlastnosti modelu. Dále jsou tu parametry vycházející z dat, jimiž lze měnit podobu podkladových dat, a tudíž zároveň výstupy obou modelů.

Všechny tyto parametry lze měnit, čímž se bude měnit i doporučení generované modelem. Cílem je zvolit takové parametry, aby doporučování bylo co nejlepší, tedy mělo co nejlepší hodnoty funkcí *Jsou zapsané předměty v prvních n doporučených* nebo *Jak moc se liší odhadnuté hodnoty od získaných.*

6.3.1 Generování kombinací parametrů a ohodnocením jejich kvality

Počet možných kombinací parametrů je veliký, konkrétně u kolaborativního filtrování založeném na matici zhruba 4 tisíce a u neuronové sítě přes 4 miliony. Kromě toho lze každou kombinaci vyzkoušet na každém semestru. Místo toho, aby byly vyzkoušeny tyto všechny možné kombinace, budou kombinace náhodně vygenerovány.

Pro každou takto vygenerovanou kombinaci pak bude přes všechny studenty vypočítána průměrná kvalita pomocí ukazatelů zmíněných v kapitole *Měření*

kvality modelů. Tedy průměrné hodnoty *Jsou zapsané předměty v prvních n doporučených* a *Jak moc se liší odhadnuté hodnoty od získaných*, případně další pomocné ukazatele kvality jako průměrná standartní odchylka generovaných doporučení.

6.4 Parametry modelu vycházející z dat

Z kapitoly Předzpracování dat vzešlo několik otázek, jakým způsobem vytvořit vstupní data pro modely.

Tyto otázky tedy budou figurovat v této části v podstatě jako parametry jednotlivých modelů a budou stejné pro všechny modely.

6.4.1 Přínos versus známky

Kritériem pro doporučování předmětů bude buď známka, nebo přínos pro studenta, který je odhadnutý na základě Ankety ČVUT (dále jen Anketa) a hodnocení udělených studenty s podobnou známkou na otázku *Předmět byl pro mne přínosem*.

6.4.2 Student versus studium

Studentem bude buď myšleno jedno konkrétní studium studijního programu, nebo celá studijní historie osoby na fakultě.

6.4.3 Omezení se na předměty s rolí

Doporučovány budou buď opravdu jen čistě volitelné předměty, nebo i předměty které mohou být pro některé obory povinné.

Přestože to není cílem doporučování, bude vyzkoušeno doporučování i všech předmětů včetně povinných. U povinných předmětů totiž může být zajímavé předpovídání známky, které může najít hypoteticky uplatnění i mimo oblast doporučování. Navíc pak budou hlouběji pochopeny vlastnosti modelů vzhledem k pravidelně zapisovaným a méně pravidelně zapisovaným. Lze předpokládat, že doporučování bude úspěšnější v případě povinných předmětů.

6.4.4 Reprezentace známky F

Pro reprezentaci známky F byly ve fázi předzpracování navrženy tři možné hodnoty, a to 0, -1 a -3 a všechny budou otestovány.

6.4.5 Minimální počet respondentů

Pokud bude počet respondentů Ankety na otázku *Předmět byl pro mne přínosem* menší než tento parametr, nebude považován vypočítaný průměrný přínos za vypovídající a nebude použit.

Tento parametr je zdánlivě relevantní jen pokud je parametr výše nastaven na *přínos*. Přesto bude otestován i pro doporučování založené na známkách. Jeho zvyšováním dochází k odfiltrovávání předmětů, které nemají dostatečnou odezvu v Anketě, což může být například, tím že jsou zapisovány jen zřídka.

6.4.6 Definice absolvovaných předmětů

Otázkou je, zda k doporučování předmětů jsou relevantní nedávno absolvované předměty, které pravděpodobně vypovídá o nejčerstvějších zájmech a schopnostech studenta, nebo naopak všechny absolvované předměty, které poskytují bohaté množství informací o zájmech a schopnostech, přestože již mohou být zastaralé.

Vyzkoušeny budou varianty, kdy za absolvované předměty budou považovány předměty z posledního semestru, z posledních dvou semestrů a ze všech předchozích semestrů.

6.4.7 Definice předmětů které by měl systém doporučit

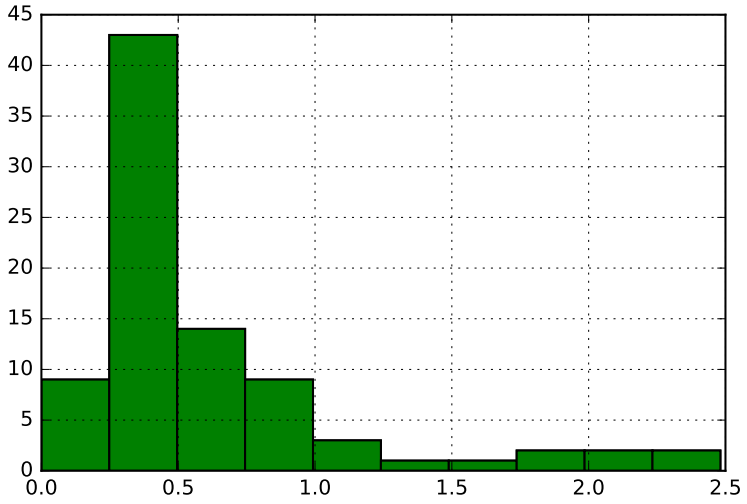
Za správně doporučené předměty lze jistě považovat takové, které si student skutečně zapsal v semestru, pro který je doporučení určené.

Na druhou stranu, pokud by za správně doporučené předměty byly považovány všechny, o kterých víme, že si je student zapíše v budoucnu, mohlo by to být pro některé modely výhodné, protože se tím zmírňuje problém toho, že student je schopen za semestr *vyzkoušet* jen několik málo předmětů.

Vyzkoušeny budou varianty, kdy správný výstup doporučování je validován na základě předmětů zapsaných v následujícím semestru, zapsaných v následujících dvou semestrech a zapsaných kdykoliv v budoucnu.

6.5 Jednoduché doporučování založené na průměru

Aby bylo možné lépe vyhodnotit kvalitu navržených modelů, bude vytvořen i *jednoduchý* doporučovací systém, vůči kterému bude moci být jejich kvalita posouzena.



Obrázek 6.1: Jaké hodnoty *průměrné absolutní chyby* se vyskytují v různých variantách jednoduchého doporučování založeného na průměru

6.5.1 Popis

Tento systém bude fungovat tak, že předpovídaná hodnota známky pro předmět bude průměrná známka z tohoto předmětu s ohledem na všechny jeho absolvování. Odhad přínosu, jelikož je závislý na předpovězené známce, bude také spočítán jako průměr.

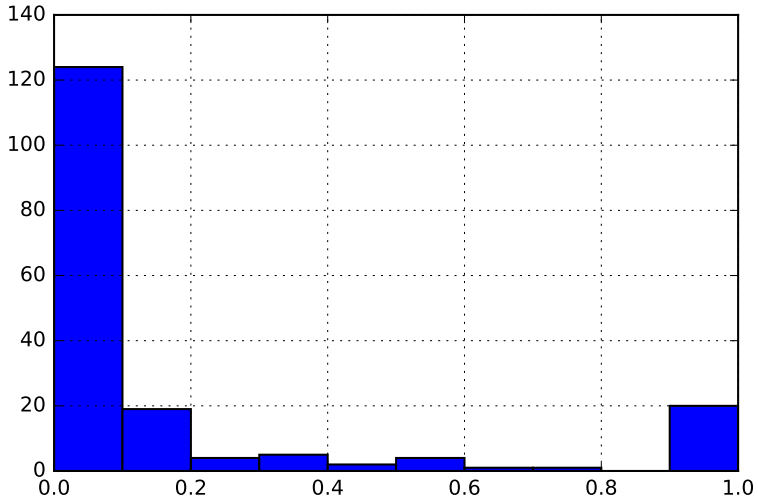
Podoba doporučení tedy bude stejná pro všechny studenty. Díky tomu bude možné například posoudit, zda doporučování na míru má vůbec smysl, nebo zda by stačilo obecné doporučení stejné pro všechny studenty.

6.5.2 Vyhodnocení kvality modelu

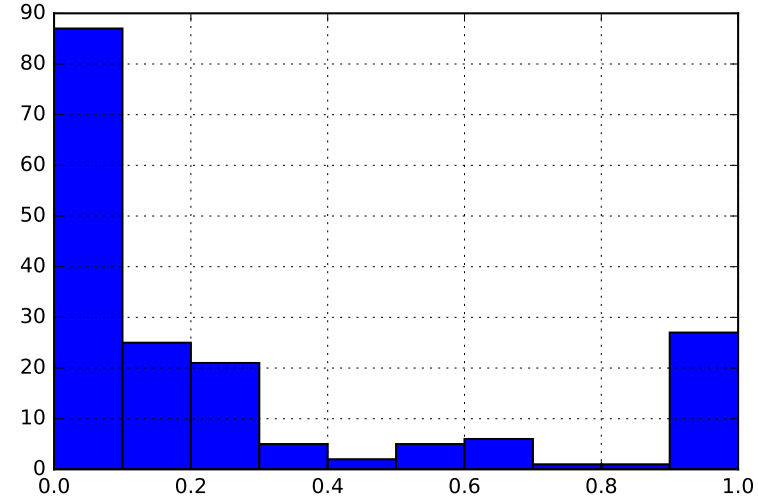
Pro jednoduchý model byly vygenerovány náhodné kombinace hodnot parametrů. U tohoto modelu se jednalo jen o parametry vstupních dat, jelikož model samotný žádné parametry nemá. Těchto kombinací bylo 180.

Na histogramu 6.1 je vidět, že průměrná absolutní chyba pro odhad známky byla většinou do 0.5, tedy odhad se průměrně lišil maximálně o jeden klasifikační stupeň.

Na dalších dvou histogramech 6.2, 6.3 je vidět, že poměr skutečně zapsaných předmětů byl v prvních 5 a prvních 10 doporučeních v obou případech zhruba jen 10 %.



Obrázek 6.2: Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 5 doporučených* se vyskytují v různých variantách jednoduchého doporučování založeného na průměru



Obrázek 6.3: Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 10 doporučených* se vyskytují v různých variantách jednoduchého doporučování založeného na průměru

Shrnutí

Jednoduchý model je poměrně v úspěšný v předpovídání známky z pohledu průměrné absolutní chyby. Na druhou stranu v *Jsou zapsané předměty v prvních n doporučených* jednoduchý model příliš úspěšný není.

6.5.3 Příklad doporučení

Pro názornost toho, jaké typy doporučení generuje jednoduchý model byl použito doporučení pro semestr *B131*, které je, jak bylo vysvětleno, stejné pro všechny studenty. Doporučení je vidět v tabulce 6.1.

6.6 Kolaborativní filtrování založené na matici

V této části bude popsáno fungování modelu - kolaborativní filtrování založené na matici. Bude odůvodněno, proč bylo vybráno jako jeden z modelů a bude popsáno jaké parametry modelu lze měnit.

6.6.1 Důvody pro použití modelu

Kolaborativní filtrování může být implementováno různými modely jako například oblíbenými asociačními pravidly. Kolaborativní filtrování založené na modelu, jako právě asociační pravidla, se používají spíše než základní přístup založený na matici, jelikož je výpočetně méně náročný.

Matice která zde bude použita má necelých 50 tisíc řádků. Jak bylo ozkoušeno, je matice natolik malá, že výpočet doporučení pro studenta proběhne na několik let starém kancelářském počítači prakticky okamžitě. Z toho důvodu bude použito kolaborativní filtrování založené na matici kvůli jednoduchosti implementace.

6.6.2 Popis

Model využívá matice, kde jedna dimenze představuje studenty a druhá předměty. Hodnota v matici na průsečíku studenta a předmětu, představuje buď známku, kterou student z předmětu získal, nebo přínos, kterým pro něj předmět byl, odhadnutý na základě Ankety a hodnoceních udělených studenty s podobnou známkou na otázku *Předmět byl pro mne přínosem.*

Přístup založený na předmětu versus na studentovi

Kolaborativní filtrování založené na matici může matici student-předmět využívat dvěma způsoby:

Kód předmětu	Průměrná známka
BI-A1L	3.000000
BI-MIK	3.000000
MI-RRI	3.000000
MI-PSL	3.000000
BI-ZRS	3.000000
BI-ZPI	3.000000
BI-XML	3.000000
BI-PYT	3.000000
BI-A2L	3.000000
BI-PMA	3.000000
MI-SAS	3.000000
BI-EP2	3.000000
BI-APJ	3.000000
BI-MEK	3.000000
BI-EP1	3.000000
BI-DNP	2.884354
BI-IOS	2.833333
BI-PHP	2.800562
BI-PJV	2.793981
MI-AIT	2.693396
MI-RUB	2.671429
BI-PJS	2.563158
BI-DAN	2.550459
MI-DSP	2.234043
BI-ZUM	1.710526
BI-PRR	1.655738
BI-ZWU	1.546875

Tabulka 6.1: Doporučení pro semestr *B131* vytvořené jednoduchým doporučováním založeným na průměru.

Přístup založený na podobnosti studentů Na základě absolvovaných předmětů spočítá podobnost ostatních studentů. Na základě známek z neabsolvovaných předmětů ostatních studentů pak spočítá předpokládanou známku/přínos pro neabsolvaných předmětů.

Přístup založený na podobnosti předmětů Na základě všech studentů spočítá podobnost všech předmětů. Známku/přínos pro neabsolvované předměty spočítá na základě podobnosti s již studentem absolvovanými předměty.

6.6.3 Vyhodnocení kvality modelu

V této části bude vyhodnoceno, jaké kvality doporučování může dosahovat kolaborativní filtrování založené na matici, pokud je vybrána optimální kombinace hodnot parametrů modelu.

Toto vyhodnocení proběhne na základě náhodných kombinací parametrů modelu, kterých bylo vygenerováno 160.

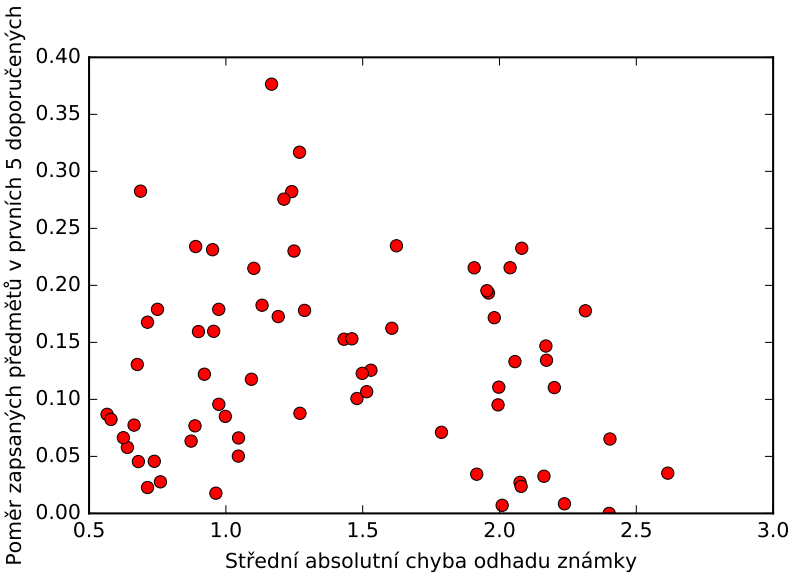
Přestože některé kombinace v nějakých ohledech mohou být nadprůměrné, pokud více kombinací nebude dosahovat podobné kvality, bude tento ojedinělý případ považován za nerelevantní. Může se totiž jednat o náhodu nebo o přeučení.

Analýza výsledků

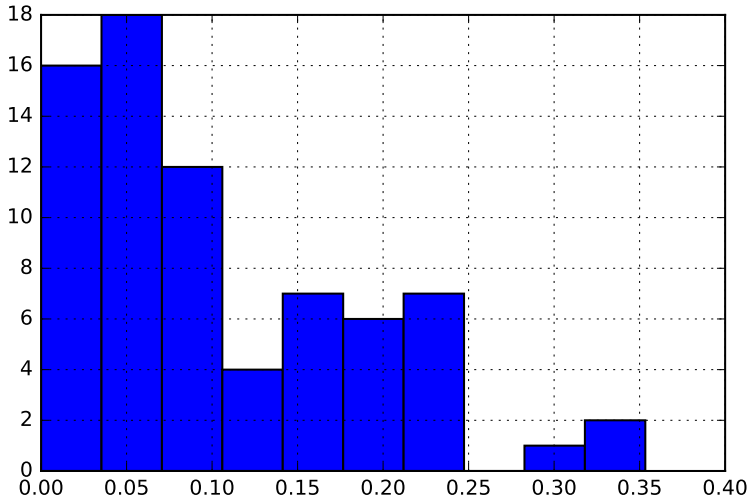
Bod v grafu 6.4 představuje konkrétní kombinaci parametrů modelu, přičemž jsou zobrazeny jen kombinace používající k doporučení známky. Nejkvalitnější kombinace jsou umístěny vlevo nahoře. Jak je vidět, není problém aby doporučování vytvořené kolaborativním filtrováním na základě matice trefilo v prvních pěti doporučených předmětech 20 % skutečně zapsaných předmětů. Také není problém, aby se průměrně odhad známky lišil od skutečnosti maximálně o dva klasifikační stupně. Zajistit obě tyto podmínky zároveň už je trochu větší problém, protože kritéria na sobě nejsou nezávislá a možná jdou proti sobě.

Na histogramu 6.5 je zobrazena úspěšnost toho, že v prvních pěti doporučených předmětech je co největší počet skutečně zapsaných předmětů změřený pro přínos, a jak je vidět v podstatě je stejná jako když použijeme známky.

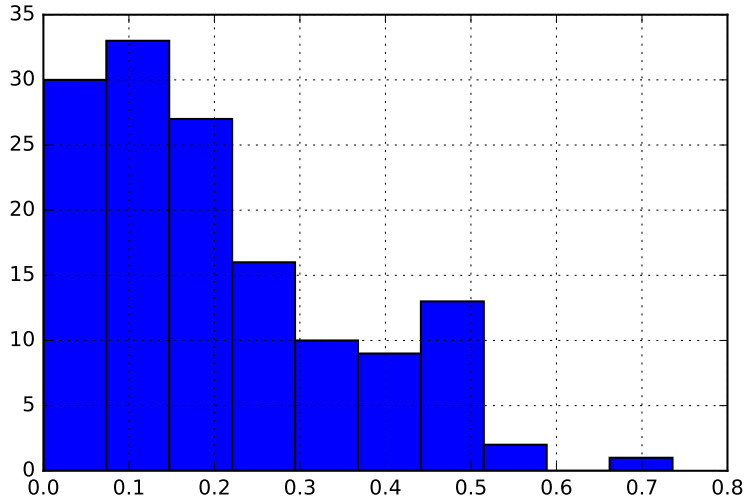
Na histogramu 6.6 je pak vidět stejná úspěšnost pro prvních deset doporučených předmětů, a to pokud uvažujeme kombinace používající známky i přínos. V tomto ohledu lze dosáhnout až 50% přesnosti. Tedy v prvních deseti doporučených předmětech je až polovina skutečně zapsaných předmětů.



Obrázek 6.4: *Odhad správné známky* versus *poměr skutečně zapsaných předmětů v prvních 5 doporučených předmětech* pro různé varianty kolaborativního filtrování



Obrázek 6.5: Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 5 doporučených* se vyskytují v různých variantách kolaborativního filtrování využívající přínos



Obrázek 6.6: Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 10 doporučených* se vyskytují v různých variantách kolaborativního filtrování

6.6.4 Optimální nastavení parametrů

Ze všech 160 vyzkoušených kombinací bude vybrána jedna dosahující vzhledem k ostatním kombinacím téměř optimální kvality.

Ze 160 byly vybrány ty které mají *Jsou zapsané předměty v prvních 5 doporučených* větší než 20 % a *Jsou zapsané předměty v prvních 10 doporučených* větší než 40 %. Takových kombinací bylo celkem 14, což je necelých 9 %.

Všechny tyto kombinace jsou k určené doporučování čistě volitelných předmětů.

Z těchto nejlepších vybraných 14 kombinací převažovaly kombinace, kde za absolvované předměty jsou považovány úplně všechny v minulosti absolvované (71 % nejlepších kombinací).

Z nejlepších vybraných 14 kombinací také převažovali kombinace, kde byl použit přístup založený na studentovi (78 % nejlepších kombinací)

Ostatní atributy se tedy nezdají být sami o sobě rozhodující o úspěšnosti kombinace.

Vybraná optimální kombinace hodnot parametrů

Vybrána jako optimální ze 14 nejlepších kombinací parametrů byla kombinace s nejnižší průměrnou absolutní chybou (0.69). Poměr skutečně zapsaných předmětů v prvních 10 doporučeních je průměrně 56 %, poměr skutečně zapsaných předmětů v prvních 5 doporučeních je průměrně 28 %.

Vybraná optimální kombinace je následující:

Reprezentace známky F -1,

Definice absolvovaných předmětů Všechny předměty v minulosti absolvovaní,

Definice předmětů které by měl systém doporučit Všechny předměty které student absolvuje v budoucnu,

Minimální počet respondentů v Anketě 50,

Známky vs přínos Použity k doporučování budou známky,

Omezení se na předměty s rolí Čistě volitelné předměty,

Student vs studium Uvažováno bude pouze konkrétní studium,

Kolaborativní filtrování založené na Studentovi (v terminologii doporučovacích systémů *přístup založený na objektu*).

6.6.5 Příklad doporučení

Na základě optimálního nastavení hodnot parametrů zvoleného v předchozí části bude vytvořeno doporučení pro autora této práce na základě části jeho bakalářského studia, pro další semestr bakalářského studia. Konkrétně bude doporučení vygenerováno pro semestr *B141*, což je 5 semestr autorova bakalářského studia.

Jak vyplývá z parametrů modelu, doporučování je založené na známkách, tedy hodnoty v tabulkách 6.2 a 6.3 představují známky. A je reprezentováno jako 3, B je reprezentováno jako 2.5, C je reprezentováno jako 2.0 atd.

Predikce známek

V tabulce 6.2 jsou vidět skutečně získané známky versus známky předpovězené modelem. Z parametrů modelu vyplývá, že skutečné známky nejsou jen z předmětů ze semestru *B141*, ale také ze semestru *B142*, což byl poslední semestr bakalářského studia toho studenta.

Kód předmětu	BI-IOS	BI-PJV	BI-ZUM
Odhad známek	3.0	2.87418	2.85131
Skutečné známky	3.0	3.00000	3.00000

Tabulka 6.2: Příklad odhadu známek vytvořený vybranou optimální verzí kolaborativního filtrování.

Vektor doporučení

Tabulka 6.3 pak představuje přímo doporučení, tak jak by mohlo být předložené studentovi. Nejvíce nalevo jsou předměty ze kterých by měl student získat nejlepší známky, a tudíž by si je měl zapsat.

Jak je vidět doporučení se ve většině případů blíží známce A, což nejspíše bude dáno tím, že známky z volitelných předmětů bývají obecně dobré. Jen u předmětu *BI-PMA* je předpovězená známka horší, konkrétně C.

Dále je zde mnoho předmětů s hodnotou 0, což nejspíše znamená, že nebylo dostatek podobných studentů, kteří by tyto předměty někdy absolvovali. Alternativně by to mohlo znamenat, že se přesně vyvážil počet podobných studentů se známkami značícími úspěch (A-E) a počet studentů s F, což je ale velmi nepravděpodobné.

Z parametrů modelu vyplývá, že v tomto doporučení jsou uvažovány jen předměty, které měly alespoň 50 respondentů v Anketě.

6.7 Dopředná neuronová síť

K doporučování předmětů bude použita také dopředná neuronová síť.

6.7.1 Důvody pro použití modelu

V žádné z v rešerši prozkoumaných prací zabývajících se doporučováním předmětů nebyla neuronová síť použita. Jedná se o velmi flexibilní model s velkým potenciálem, který je ale zároveň náročnější na implementaci než například kolaborativní filtrování založené na matici.

Zde bude vyzkoušena a bude zjištěno, jestli stojí za to neuronovou síť na tento typ úlohy používat, nebo zda jsou jednodušší modely plně dostačující.

6.7.2 Co je dopředná neuronová síť

Neuronová síť je model inspirovaný biologií, konkrétně tím, jak funguje mozek.

Kód předmětu	Predikovaná známka
BI-IOS	3.000000
BI-ZPI	3.000000
BI-ZRS	3.000000
BI-APJ	3.000000
BI-EP2	2.962202
BI-XML	2.930761
BI-DAN	2.923865
BI-DNP	2.875683
BI-PJV	2.874180
BI-PYT	2.870490
BI-3DT	2.869855
BI-PHP	2.859754
BI-ZUM	2.851310
BI-PRR	2.826317
BI-PJS	2.799065
BI-PMA	2.000000
MI-AIT	0.000000
MI-RUB	0.000000
MI-RRI	0.000000
MI-PSL	0.000000
MI-IVS	0.000000
MI-DSP	0.000000
BI-ATS	0.000000
BI-MEK	0.000000
BI-MIK	0.000000
BI-UVM	0.000000
BI-ST2	0.000000
BI-ST1	0.000000
MI-SAS	0.000000

Tabulka 6.3: Příklad doporučení vytvořeného vybranou optimální verzí kolaborativního filtrování.

Dopředná neuronová síť je jeden z jednodušších typů neuronové sítě. Skládá se z vrstvy vstupních neuronů, vnitřních vrstev neuronů a výstupní vrstvy neuronů, přičemž výstupy z předchozí vrstvy slouží jako vstupy do vrstvy následující. V této práci bude použita varianta, kdy každý neuron z předchozí vrstvy slouží jako vstup pro každý neuron z následující vrstvy. [29]

6.7.3 Aplikace problému doporučování na neuronovou síť

Problém doporučování předmětů zde bude modelován tak, že každý neuron vstupní vrstvy bude představovat jeden z možných absolvovaných předmětů. Hodnota na vstupním neuronu bude známka, kterou student z předmětu dostal, nebo přínos kterým pro něj předmět byl. Pokud student předmět neabsolvoval, bude hodnota neuronu 0.

Každý neuron výstupní vrstvy bude na druhou stranu představovat předměty, které si může student zapsat. Hodnota přivedená na výstupní neuron je vlastně předpověď známky, kterou může student očekávat nebo přínos, kterým by pro něj mělo studium předmětu být.

6.7.4 Implementace

Pro implementaci neuronové sítě byl použit framework Keras, jelikož je v něm vytvoření libovolné neuronové sítě relativně jednoduché a obecně je v době psaní práce tento framework poměrně populární. Jako výpočetní backend lze použít knihovny TensorFlow či Theano. [30]

6.7.5 Parametry modelu

Počet vnitřních vrstev

Se zvyšujícím počtem vnitřních vrstev se obvykle zvyšuje učící potenciál neuronové sítě.

Zde budou vyzkoušeny varianty s žádnou, jednou a třemi vnitřními vrstvami. Více vnitřních vrstev je pravděpodobně k složitosti řešeného problému zbytečných.

Počet neuronů ve vnitřní vrstvě

Počet neuronů vnitřní vrstvy obvykle také určuje učící potenciál neuronové sítě.

Bude vyzkoušena varianta s 10 neurony ve vnitřní vrstvě, což znamená že neuronů je mnohem méně než předmětů.

Dále bude vyzkoušena varianta, kdy počet je stejný jako počet předmětů, tedy stejný jako je velikost vstupní a výstupní vrstvy.

Aktivační funkce

Pro všechny neurony kromě vstupních bude použita stejná aktivační funkce.

Vyzkoušené aktivační funkce budou dvě.

První *relu*, je spojitá aproximace funkce

$$\max(0, x)$$

Vybrána byla kvůli tomu, aby model mohl predikovat přímo výsledné známky, což se sigmoidou, která se jinak typicky používá, možné není.

Druhá *sigmoid*, je spojitá funkce, která nabývá hodnot od -1 do 1 a typicky se pro tento účel používá. Nevýhodou je, že výstup sítě nemůže být vyšší než 1, tedy tato funkce nebude určena pro odhadování přesné známky, ale jen pro určení pořadí doporučovaných předmětů.

Počáteční váhy

Byla vybrána varianta *uniform* pro počáteční nastavení vah. To znamená že váha je vybrána z rovnoměrného rozdělení na intervalu (0; 0,05).

Tento parametr sítě v této práci vlastně není parametr, protože jeho hodnota je vždy stejná.

Ztrátová funkce

Tato funkce je zásadní pro to, jak dobré poskytuje síť výsledky. Funkce musí být diferenciovatelná, aby mohla být síť efektivně učena za pomocí zpětné propagace.

Byly vyzkoušeny následující ztrátové funkce:

mse střední kvadratická chyba

kld Kullback–Leiblerova divergence

cosine kosínová podobnost,

poisson průměr z ($predikovaná\_hodnota - skutečná\_hodnota * \log(predikovaná\_hodnota + \epsilon)$),

categorical_crossentropy kategorická křížová entropie.

Zkratky funkcí odpovídají zkratkám ve frameworku Keras. Přesné definice dostupné na: https://github.com/fchollet/keras/blob/master/keras/losses.py.

Velikost batche

Velikost batche je počet vzorků které jsou síti vystaveny než dojde k úpravě vah.

Budou vyzkoušeny varianty 100, 1000 a 10000.

Optimalizátor

Framework Keras nabízí na výběr při učení neuronové sítě jeden z optimalizátorů, který řídí její učení.

Jen pro případ, že by některý optimalizátor byl výrazně lepší než jiný, bude vyzkoušeno více variant, a to *adam*, *sgd* a *adagrad*.

6.7.6 Vyhodnocení kvality modelu

V této části bude vyhodnoceno, jaké kvality doporučování může dosahovat dopředná neuronová síť, pokud je vybrána optimální kombinace hodnot parametrů modelu.

Toto vyhodnocení proběhne na základě náhodných kombinací hodnot parametrů modelu, kterých bylo vygenerováno 890.

Přestože některé kombinace v nějakých ohledech mohou být nadprůměrné, pokud více kombinací nebude dosahovat podobné kvality, bude tento ojedinělý případ považován za nerelevantní. Může se totiž jednat o náhodu nebo o přeučení

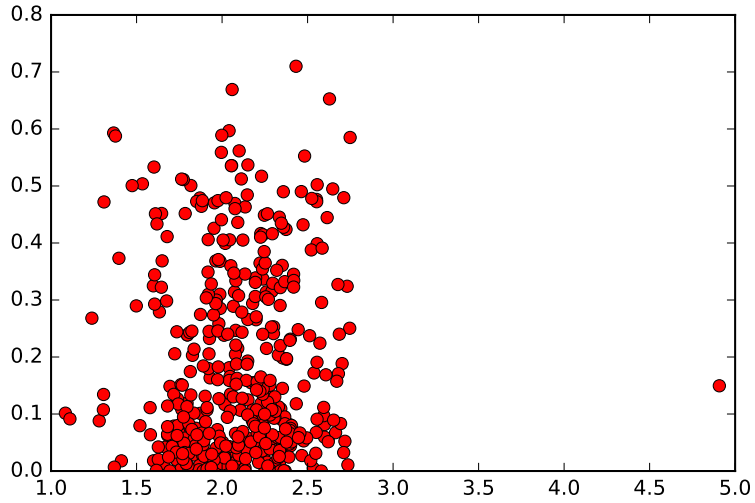
Analýza výsledků

Bod v grafu 6.7 představuje kombinaci parametrů modelu, přičemž jsou zobrazeny jen kombinace používající k doporučení známky.

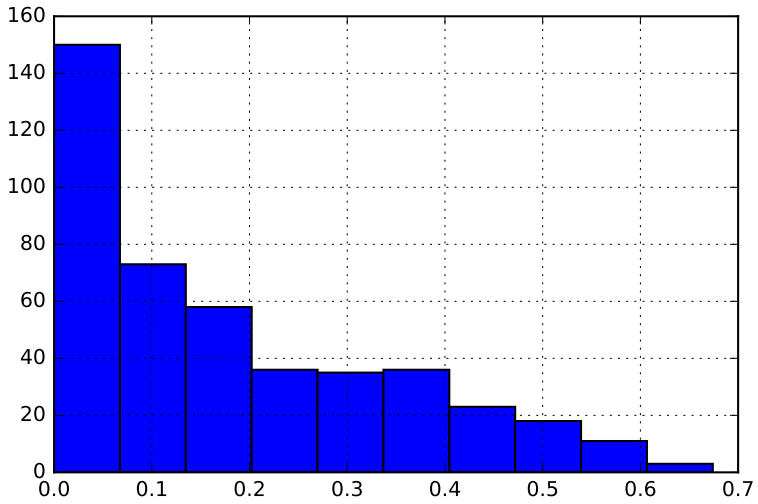
Nejkvalitnější kombinace jsou umístěny vlevo nahoře. Jak je vidět, není problém, aby doporučování vytvořené dopřednou neuronovou sítí trefilo v prvních pěti doporučených předmětech 50 % skutečně zapsaných předmětů. Odhad známky se ovšem obvykle liší průměrně o tři klasifikační stupně, což se zdá být hodně.

Na histogramu 6.8 je zobrazena úspěšnost toho, že v prvních pěti doporučených předmětech je co největší počet skutečně zapsaných předmětů změřený pro přínos, a jak je vidět je velmi podobná tomu, když použijeme známky.

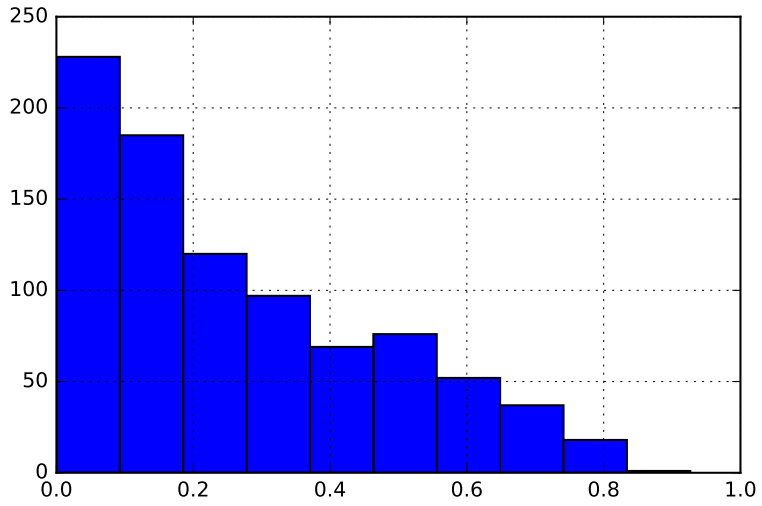
Na histogramu 6.9 je pak vidět stejná úspěšnost pro prvních deset doporučených předmětů, a to pokud uvažujeme kombinace používající známky i přínos. V tomto ohledu lze dosáhnout až 80 % přesnosti. Tedy v prvních deseti doporučených předmětech jsou téměř všechny skutečně zapsané předměty.



Obrázek 6.7: *Odhad správné známky* versus *poměr skutečně zapsaných předmětů v prvních 5 doporučených předmětech* pro různé varianty neuronové sítě



Obrázek 6.8: Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 5 doporučených* se vyskytují v různých variantách neuronové sítě využívající přínos



Obrázek 6.9: Jaké hodnoty *poměru skutečně zapsaných předmětů v prvních 10 doporučených* se vyskytují v různých variantách neuronové sítě

6.7.7 Optimální nastavení parametrů

Ze všech 890 vyzkoušených kombinací parametrů bude vybrána jedna dosahující vzhledem k ostatním kombinacím téměř optimální kvality.

Ze 890 byly vybrány ty které mají *Jsou zapsané předměty v prvních 5 doporučených* větší než 50 % a *Jsou zapsané předměty v prvních 10 doporučených* větší než 70 %. Takových kombinací bylo celkem 23, což je zhruba 2,5 %.

Všech 23 nejlepších kombinací je určeno k doporučování čistě volitelných předmětů.

Všechny tyto nejlepší kombinace v odhadu známek dosahují průměrné absolutní odchylky kolem 2, to znamená odhad se liší běžně o více než 4 klasifikační stupně, což je hodně a neuronovou síť určitě nelze použít k přesnému odhadu budoucích známek.

Odfiltrování předmětů s malou odezvou v Anketě má obecně úspěch - ve 30 % případů byly odfiltrovány předměty s méně jak 20 respondenty, ve zbylých 70 % případů byly odfiltrovány předměty s méně jak 50 respondenty. To je dáno nejspíše tím, že se eliminují méně frekventované předměty, na které se model nezaměřuje, jelikož jeho úspěch je dán většinou. Také předmětů je potom celkově méně a je větší šance, že se trefí.

Ve 23 nejlepších kombinacích také převažovaly kombinace, kde byl jako opti-

malizátor použit *adagrad* (70 % kombinací).

Ostatní výše nezmíněné parametry se nezdají být sami o sobě rozhodující o úspěšnosti kombinace hodnot parametrů, přestože neuronová síť má parametrů opravdu mnoho. V podstatě to může znamenat, že o nejlepších kombinacích nerozhodují individuální hodnoty parametrů, ale jejich specifická kombinace.

Z toho plyne například poměrně zajímavý fakt, že neuronové síti stačí ve vnitřní vrstvě pro doporučování předmětů pouze deset neuronů.

Vybraná optimální kombinace hodnot parametrů

Vybrána jako optimální ze 23 nejlepších kombinací parametrů byla kombinace s nejvyšší průměrnou standartní odchylkou (0,73) ve vytvářených doporučeních. To v podstatě znamená že hodnoty v doporučování nejsou příliš podobné.

Toto kritérium bylo zvoleno, protože mnoho naučených neuronových sítí mělo problém s tím, že generovala velmi podobná čísla. S tím není problém, pokud uživatel nevidí přímo hodnoty doporučení. Pokud je ovšem vidí, může nabýt dojmu, že v jednotlivých předmětech není vlastně téměř žádný rozdíl a v doporučování prvním předmětům nemusí klást takovou váhu.

Poměr skutečně zapsaných předmětů u této kombinace v prvních 10 doporučeních je průměrně 77 %, počet skutečně zapsaných předmětů v prvních 5 doporučeních je průměrně 53 %.

Vybraná optimální kombinace je následující:

Reprezentace známky F 0,

Definice absolvovaných předmětů předměty absolvované v předchozích dvou semestrech,

Definice předmětů které by měl systém doporučit předměty jen z následujícího semestru,

Minimální počet respondentů v Anketě 50,

Známky vs přínos použit k doporučování bude přínos předmětu,

Omezení se na předměty s rolí čistě volitelné předměty,

Student vs studium uvažovány budou všechny studia daného studia,

Aktivační funkce relu,

Velikost batche 10000,

Kód předmětu	BI-IOS	BI-PJV
Odhad přínosu	0.000000	7.441568
Skutečný přínos	4.857143	4.608108

Tabulka 6.4: Příklad odhadu přínosu vybranou optimální verzí neuronové sítě.

Počet epoch 20,

Počet vnitřních vrstev 3,

Počet neuronů vnitřní vrstvy 10,

Ztrátová funkce kosínová podobnost,

Optimalizátor *sgd*.

6.7.8 Příklad doporučení

Na základě optimálního nastavení hodnot parametrů zvoleném v předchozí části bude vytvořeno doporučení pro autora této práce na základě části jeho bakalářského studia, pro další semestr bakalářského studia. Konkrétně bude doporučení vygenerováno pro semestr *B141*, což je 5 semestr autorova bakalářského studia.

Jak vyplývá z hodnot parametrů modelu, doporučování je založené na přínosu, kterým by pro studenta předmět měl být, a to na základě Ankety. Hodnoty v tabulkách 6.4 a 6.5 představují právě přínos. Na otázku *Předmět byl pro mne přínosem* 5 představuje *rozhodně ano*, 4 představuje *spíše ano*, 3 představuje *nevím, neumím se vyjádřit*, 2 představuje *spíše ne*, 1 představuje *rozhodně ne*.

Predikce přínosu

Jak je vidět v tabulce 6.4, která ukazuje odhad skutečného přínosu versus hodnoty doporučení, odhad přesného přínosu není příliš dobrý. Na to ale ani model s konkrétními parametry není optimalizovaný. Cílem je u něj jen to, aby v prvních n doporučených předmětech byly předměty, které si student opravdu zapsal.

Vektor doporučení

Tabulka 6.5 představuje přímo doporučení, tak jak by mohlo být předložené studentovi. Nejvíce nalevo jsou předměty, které by pro studenta měly být největším přínosem, a tudíž by si je měl zapsat.

Mezi prvních deset doporučených předmětů se dostal i magisterský předmět, přestože se jedná o doporučování pro bakalářské studium. To je důsledek toho,

že tato konkrétní kombinace hodnot parametrů uvažuje přímo fyzické osoby, a nikoliv jen konkrétní studium.

6.8 Model který bude použit k doporučování

- Přestože neuronová síť ukázala lepší výsledky v *Jsou zapsané předměty v prvních n doporučených*, kolaborativní filtrování založené na matici ukázalo mnohem lepší schopnost odhadu známek. Autor předpokládá, že odhad známek lze považovat za důležitý aspekt pro to, aby student chápal, jak doporučování funguje, a více mu proto důvěřoval.
- Neuronová síť má nevýhodu v tom, že se pokaždé naučí jinak, tedy proces učení je nedeterministický, a mohla by generovat nekonzistentní doporučení pro stejného studenta.

Kvůli těmto důvodům bude jako doporučovací systém pro studenty Fakulty informačních technologií použit model kolaborarativní filtrování založené na matici s optimálním nastavením hodnot parametrů popsaným výše. Jak bude zmíněno ve fázi Ohodnocení, model bude na základě uživatelské testování a praktickým problémům které vyvstanou ještě mírně upraven.

Kód předmětu	Predikce přínosu
BI-A2L	3.535941
BI-PRR	2.344818
BI-PHP	1.033876
BI-EP1	1.029642
BI-PJV	0.630999
MI-AIT	0.554534
BI-DAN	0.256182
MI-RRI	0.195165
BI-ZWU	0.082385
BI-XML	0.064469
BI-MEK	0.046298
BI-PJS	0.025926
BI-DNP	0.013573
MI-IVS	0.001214
BI-UVM	0.000000
MI-PSL	0.000000
MI-RUB	0.000000
BI-ZUM	0.000000
BI-ZRS	0.000000
BI-ZPI	0.000000
MI-DSP	0.000000
BI-3DT	0.000000
BI-ST2	0.000000
BI-ST1	0.000000
BI-PYT	0.000000
BI-A1L	0.000000
BI-PMA	0.000000
BI-MIK	0.000000
BI-IOS	0.000000
BI-EP2	0.000000
BI-ATS	0.000000
BI-APJ	0.000000
BI-A2Z	0.000000
MI-SAS	0.000000

Tabulka 6.5: Příklad doporučení vytvořeného vybranou optimální verzí neuronové sítě.

Kapitola 7

Ohodnocení

V této části bude vyhodnoceno, jak má vybraný optimální model kvalitní doporučování na ještě neviděných datech, tedy jaká se dá očekávat jeho kvalita při praktickém použití.

Dále bude provedeno uživatelské testování, ze kterého vyplynou dodatečné úpravy doporučovacího systému, tak, aby byl opravdu prakticky použitelný. Také se zjistí, jaká vylepšení do budoucna by bylo dobré provést.

7.1 Testovací data

Data, která byla dostupná ve fázi Průzkumu dat a Předzpracování dat byl export databázových tabulek vytvořený v první polovině ledna 2017.

K uživatelskému testování a ohodnocení však byl dodatečně dodán export stejných tabulek z první poloviny dubna 2017. Díky tomu bude k ohodnocení použito kromě původně zamýšlených semestrů *B151* a *B152*, také semestr *B161*. Z hlediska uživatelského testování pak bude moci být vytvářeno aktuálnější doporučení.

7.2 Výsledky na testovacích semestrech

Tabulka 7.1 ukazuje výsledky vybraného optimálního modelu na semestrech *B151*, *B152* a *B161*.

Poměr skutečně zapsaných předmětů v prvních 10 doporučení na trénovacích datech byl průměrně 56 %, na testovacích datech je úspěšnost podobná.

Poměr skutečně zapsaných předmětů v prvních 5 doporučení na trénovacích datech byl průměrně 28 %, na testovacích datech je úspěšnost také podobná.

Průměrná absolutní chyba	Semestr	Top 10	Top 5
1.061449	B151	0.605356	0.293525
0.977629	B152	0.513610	0.172953
1.135548	B161	0.636713	0.322197

Tabulka 7.1: Výsledky optimálního modelu.

Průměrná absolutní chyba	Semestr	Top 10	Top 5
0.791421	B151	0.605200	0.146206
0.722412	B152	0.487339	0.162920
0.712324	B161	0.664126	0.253811

Tabulka 7.2: Výsledky upraveného optimálního modelu, pokud je měřeno stejným způsobem jako byl optimální model.

Průměrná absolutní chyba v odhadu známek byla na trénovacích datech 0.69. Na testovacích datech je to zhruba kolem 1, tedy o něco horší. Průměrně se tedy odhadnutá známka liší o dva klasifikační stupně.

7.2.1 Vyhodnocení změn provedených ve fázi uživatelského testování

Jak bude zmíněno v nadcházející části, na základě uživatelského testování byly provedeny drobné změny v modelu.

Konkrétně bylo kolaborativní filtrování provedeno se všemi předměty, nikoliv jen s volitelnými, což zlepšilo doporučení pro některé studenty a vyřešilo tak nejspíše cold-start problem či sparsity problem, které jsou popsané v rešerši.

Pokud z doporučení produkovaných takto upraveným modelem vybereme jen předměty s minimální počtem 50 respondentů v Anketě a jen volitelné předměty, což bylo nastavení původního modelu, získáme výsledky v tabulce 7.2.

Jak je vidět, odhad skutečně zapsaných předmětů v prvních n doporučeních se úpravou lehce zhoršil, na druhou stranu se zlepšil odhad známek.

Byly provedeny i změny v tom, že jsou doporučovány všechny volitelné předměty, nejen ty s minimálním počtem respondentů a doporučení bylo rozděleno na doporučení pro bakalářské studenty a pro magisterské studenty.

Tyto úpravy se možná dají spíše považovat vzhledem k měření kvality za změnu v interpretaci doporučení než změnu v modelu, proto už tyto výsledky nelze přímo porovnávat s původně vybraným optimálním modelem. Přesto pro představu jsou v tabulce 7.3 výsledky pro bakalářská doporučení a v tabulce

Průměrná absolutní chyba	Semestr	Top 10	Top 5
1.056521	B151	0.255957	0.048458
0.937083	B152	0.274904	0.114067
0.847702	B161	0.275893	0.101042

Tabulka 7.3: Výsledky upraveného optimálního modelu, pokud uvažujeme všechny volitelné bakalářské předměty.

Průměrná absolutní chyba	Semestr	Top 10	Top 5
0.868699	B151	0.706972	0.278867
0.692992	B152	0.704932	0.355867
0.483701	B161	0.820261	0.330065

Tabulka 7.4: Výsledky upraveného optimálního modelu, pokud uvažujeme všechny volitelné magisterské předměty.

7.4 výsledky pro magisterské doporučení. Jak je vidět, kvalita doporučení z hlediska zvolených metrik je mnohem lepší pro magisterské studenty.

7.3 Uživatelské testování

Pro uživatelské testování byly použiti studenti, kteří o doporučení projevili zájem. Na základě studentova školního username byl vytvořen HTML dokument obsahující doporučení. Součástí dokumentu byl i odkaz na dotazník, kde student mohl vyplnit zpětnou vazbu.

7.3.1 Změny v modelu provedené na počátku testování

Již na úplném začátku testování bylo odhaleno několik problémů:

- doporučení obsahuje pohromadě bakalářské i magisterské předměty,
- mnoho doporučení obsahuje v podstatě všechny předměty se stejnou odhadnutou známkou, což je způsobeno nedostatkem absolvovaných volitelných předmětů,
- nedoporučují se všechny volitelné předměty. To může vadit, protože student například nevidí jak je to s předměty nad kterými při zápisu uvažoval. Také se tím snižuje diverzita doporučení. Eliminují se relativně nové předměty, které nemají dostatek respondentů, přestože mohou být velmi přínosné.

Z toho důvodu byl model ještě dodatečně upraven tak, že byly ve fázi kolaborativního filtrování použity všechny (i povinné) předměty. Z produkovaného doporučení pak byly vybrány jen předměty, které jsou volitelné. Tím se problém toho, že všechny předměty pro studenta mají stejnou předpovídanou známky, zmírnil.

Doporučení pro lepší orientaci bylo ještě navíc rozděleno na bakalářské a magisterské předměty, takže každý student v podstatě obdrží doporučení dvě.

Všichni studenti obdrželi výsledky na základě takto upraveného modelu a zpětná vazba je tedy relevantní vzhledem k němu.

7.3.2 Dokument s doporučením

Každému studentovi byl v případě zájmu zaslán dokument s doporučením.

Dokument obsahoval:

- informace o tom, jak interpretovat čísla v doporučení, tedy také jaké známce číslo odpovídá,
- odkaz na dotazník pro zpětnou vazbu,
- doporučení pro bakalářské studium,
- doporučení pro magisterské studium.

Příklad dokumentu lze nalézt na přiloženém médiu.

7.3.3 Dotazník

Otázky v dotazníku byly následující:

1. Věříte tomu, že doporučení které jste dostali od systému je dobré?
   - Ano
   - Ne
   - Jiné:
2. S ohledem na probíhající semestr, myslíte že byste si bývali zapsali lepší předměty, pokud byste měli k dispozici toto doporučení?
   - Ano
   - Ne
   - Jiné:

3. S ohledem na nadcházející semestr, použijete doporučení při zápisu předmětů.
   - Rozhodně
   - Vezmu jej v potaz
   - Ne
   - Jiné:
4. Váš školní username:
5. Nápady na zlepšení, připomínky, pocity:
6. S ohledem na probíhající semestr, mezi jakými předměty jste při zápisu váhali / u jakých jste pochybovali, že jejich zapsání je dobrý nápad.

První tři otázky byly uzavřené, avšak student mohl sdělit svůj konkrétní názor pokud si vybral odpověď *Jiné:*. Tyto otázky byly povinné.

Zbývající tři otázky byly otevřené a nepovinné.

7.3.4 Výsledky dotazníku

Celkový počet respondentů dotazníku byl 21. Dotazník byl proveden pomocí aplikace Google Forms a byl podán v průběhu předběžných zápisů do semestru *B171*.

Konkrétní odpovědi na otevřené otázky lze nalézt na přiloženém médiu.

Důvěra v doporučení

Na otázku *Věříte tomu, že doporučení, které jste dostali od systému je dobré* odpovědělo **47,6 %** respondentů že **ano**, **23,8 %** že **ne** a **28,6 %** odpovědělo **jinak**.

Doporučení tedy věří zhruba polovina respondentů.

Užitečnost doporučení vzhledem k probíhajícímu semestru

Na otázku: *S ohledem na probíhající semestr, myslíte že byste si bývali zapsali lepší předměty, pokud byste měli k dispozici toto doporučení?* odpovědělo **81 %** respondentů že **ne**, **14,3 %** odpovědělo že **ano**, 4,8 % což je jeden respondent odpověděl **jinak**.

Většina studentů si vyhodnotila, že by si lepší předměty nezapsali, pokud by měli k dispozici toto doporučení. To může znamenat, že informovanost studentů je dobrá, a doporučovací systém pro většinu studentů není velkým přínosem, nebo to může znamenat, že kvalita doporučování není dostatečná.

Užitečnost doporučení vzhledem k nadcházejícím semestru

Na otázku: *S ohledem na nadcházející semestr, použijete doporučení při zápisu předmětů.* odpovědělo **61,9 %** respondentů že jej **vezmou v potaz**, **33,3 %** že **ne** a **4,8 %** že **rozhodně ano**.

Přestože většina studentů si nemyslí, že by si bývali zapsali díky doporučování lepší předměty v tomto semestru, naopak většina studentů doporučování vezme v potaz při zapisování do následujícího semestru, které v době dotazování probíhalo.

Připomínky k doporučování

Další položená otevřená otázka byla: *Nápady na zlepšení, připomínky, pocity:.*

Z odpovědí lze zjistit že několika studentům vadí, že odhady všech předmětů se často liší jen poměrně málo a získávají tak pocit, že je jedno co si zapíší.

Také je vidět problém v tom, když v doporučení nejsou žádné předměty s dostatečně dobrou hypotetickou známkou, což studentovi v podstatě dává pocit beznaděje.

V odpovědích vyvstal nápad přidat názvy předmětů, což by jistě zjednodušilo interpretaci studentem.

Studenti také projevili zájem o předpověď známek pro povinné předměty, což ovšem nebylo náplní doporučovacího systému.

Předměty u jejichž zápisu studenti v minulosti váhali

Další položená otevřená otázka byla: *S ohledem na probíhající semestr, mezi jakými předměty jste při zápisu váhali / u jakých jste pochybovali, že jejich zapsání je dobrý nápad.*

Studenti uvedli že váhali u zápisu *BI-GIT*, *BI-EHA*, *BI-PYT*, *BI-PJV*, *BI-PHP*, *BI-ACM*, *BI-TEX*. Žádné uvedené předměty se ovšem neopakovali, a vzhledem k počtu respondentů je lze brát jako náhodný vzorek, nikoliv jako vyčerpávající sadu problémových předmětů se kterými studenti u zápisu váhají.

Jeden student uvedl, že pochyboval u kombinace předmětů *BI-PA2* A *BI-OSY*, druhý pochyboval u kombinace *BI-BEZ* a *BI-OSY* a to z důvodů časové náročnosti těchto kombinací. Přestože všechny tyto předměty jsou povinné, a nejsou tedy primárním subjektem této práce, dostáváme náhled toho, jaké problémy studenti při zápisu řeší.

Kapitola 8

Nasazení

V této části práce se autor bude zabývat tím, jakým způsobem bude možné dodávat výstupy doporučovacího systému studentům automatizovanou cestou. Cílem není přímo implementace tohoto řešení, ale jen jeho hrubý návrh.

8.1 Architektura doporučovacího systému

Doporučovací systém vyžaduje na svém vstupu CSV export z následující tabulek datového skladu ČVUT:

- t_pred_predmet,
- t_predmet_predmet_vztah_rel,
- t_stud_studium,
- t_stpl_studijni_program,
- t_osob_osoba,
- t_zapi_zapsany_predmet,
- t_zapi_klasifikace.

Dále systém vyžaduje školní username studenta a kód semestru, aby věděl, jaké doporučení má vygenerovat

Jako výpočetně nejefektivnější je na základě vstupních tabulek vygenerovat matici doporučení pro následující semestr. Tato matice obsahuje doporučení pro všechny studenty a je tedy platná až do dalšího semestru.

8.2 Integrace s reportovacím portálem

Původní myšlenka byla integrovat doporučovací systém s reportovacím portálem EBIE, který na Fakultě informačních technologií v době psaní práce vznikal a který má sloužit pro vytváření reportů nad datovým skladem ČVUT.

Kromě toho, že výstup doporučovacího systému nelze chápat jako report v tradičním slova smyslu, jak se ale ukázalo, hlavními uživateli portálu mají být zaměstnanci univerzity, nikoliv studenti. Z toho důvodu byla integrace doporučovacího systému do EBIE shledána nevhodnou.

8.3 Miniaplikace v portálu ČVUT

Vhodnější se ukázalo umístit doporučování do Portálu ČVUT. Portál je v době psaní práce dostupný zatím jen v testovacím režimu na adrese: https://portal-test.cvut.cz/portal/.

V tomto portálu student vidí různé užitečné informace. Portál je rozdělen na miniaplikace v podobě dlaždic, kde každá miniaplikace obsahuje informace z konkrétní oblasti, která může studenta zajímat. Například je zde nabídka obědů v menze, seznam plánovaných akcí a podobně.

Do portálu se student přihlašuje pomocí celouniverzitní autentizace a jednotlivé miniaplikace tak dostávají informaci o tom, jaký konkrétní student si právě portál prohlíží.

8.3.1 Integrace doporučovacího systému

K úspěšné integraci je potřeba minimálně následující:

1. části kódu v jazyce Python z Jupyter notebooků dostupných na přiloženém médiu *Předzpracování dat.ipynb*, *Modelování.ipynb* a *Obecné funkce a nastavení.ipynb* se integrují do jednoho skriptu, který bude z exportů tabulek a kódu nadcházejícího semestru generovat matici doporučení,
2. na serveru odkud je možný přístup do datového skladu ČVUT se jednou za semestr spustí tento skript, který vytvoří matici/tabulku doporučení,
3. na serveru se vystaví XML s definicí miniaplikace, kterou bude využívat přes REST protokol Portál ČVUT,
4. na server se umístí aplikace, ideálně v podobě dalšího Python skriptu. Při REST požadavku o XML z portálu ČVUT, obdrží tato aplikace identifikaci studenta. Na základě identifikace a matice doporučení uložené na serveru vytvoří XML v požadovaném formátu, které bude sloužit jako vstup pro Portál ČVUT.

Obrázek 8.1: Podoba testovací verze portálu ČVUT.

Závěr

Cílem práce bylo vytvoření doporučovacího systému, který bude studentům Fakulty informačních technologií doporučovat volitelné předměty k zápisu.

V rešerši byly prozkoumány jak doporučovací systémy z obecného hlediska, tak doporučovací systémy zabývající se přímo doporučováním předmětů na univerzitách. Jak se ukázalo bylo vyzkoušeno mnoho různých přístupů, v praxi se ovšem obvykle používá hlavně doporučování založené na známkách studentů, nejspíše protože je osvědčené a data o známkách jsou snadno dostupná.

Dále bylo popsáno, jak funguje Fakulta informačních technologií z hlediska studentů a předmětů, bylo prozkoumáno, jaké informace jsou obsaženy v systémech KOS a Anketa, jejichž data slouží jako vstup doporučovacímu systému a bylo provedeno předzpracování dat, tak aby byly jednoduše použitelné jednotlivými doporučovacími modely. Všechny tyto kroky byly provedeny dostatečně obecně s cílem, aby bylo možné z je použít i pro jiné účely, než jsou konkrétní použité doporučovací modely.

Jak se ukázalo, potenciál Ankety, která obsahuje velmi užitečná data o názorech studentů nelze plně využít, jelikož data v ní jsou plně anonymizovaná.

Ve fázi modelování bylo vyzkoušeno více metod doporučování, konkrétně kolaborativní filtrování založené na matici a kolaborativní filtrování založené na dopředné neuronové síti. K uživatelskému testování byla vybrána jednodušší metoda založená na matici, jelikož dokázala lépe odhadnout budoucí známku studenta, což autor považoval za důležitý aspekt toho, aby studenti doporučení důvěřovali.

Optimální model vybraný ve fázi modelování byl poté dodatečně upraven na počátku uživatelského testování, jelikož byl konfrontován s praktickými problémy, které nebyly autorem do té doby odhaleny. Pro studenty bakalářského studia byl počet skutečně zapsaných předmětů v prvních deseti doporučených

předmětech zhruba 26 %, pro studenty magisterského studia byl tento počet zhruba 70 %.

Výstupem uživatelského testování provedeného na studentech byl dotazník. Z dotazníků vyplynulo, že zhruba polovina studentů věří tomu, že doporučení je dobré, většina si nemyslí že by si díky doporučení zapsala lepší předměty, ale zároveň větší část studentů by doporučení využila při zapisování předmětů. Doporučovací systém tedy bude některým studentům užitečný má ale ještě velký potenciál ke zlepšení. V budoucích pracích je možné například vyzkoušet metody doporučování založeného na obsahu předmětů, využít k doporučování sociální sítě a mnoho dalších víceméně experimentálních přístupů zmíněných v rešerši.

V poslední části měly být diskutovány možnosti integrace doporučovacího systému do portálu s reporty (EBIE). Ukázalo se ale, že doporučovací systém, z pohledu použitelnosti studentem, bude mnohem lepší zaintegrovat jako miniaplikaci do připravovaného Portálu ČVUT.

Literatura

[1] Studijní a zkušební řád pro studenty Českého vysokého učení technického v Praze ze dne 8. července 2015. Citováno: 2017-04-26. Dostupné z: https://www.cvut.cz/sites/default/files/content/7e72349e-3ea5-4693-9853-5147f1238481/cs/20160901-studijni-a-zkusebni-rad-pro-studenty-cvut-ze-dne-8-7-2015.pdf

[2] Elbadrawy, A.; Karypis, G.: Domain-Aware Grade Prediction and Top-n Course Recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, New York, NY, USA: ACM, 2016, ISBN 978-1-4503-4035-9, s. 183–190, doi:10.1145/2959100.2959133. Dostupné z: http://doi.acm.org/10.1145/2959100.2959133

[3] Bydžovská, H.: Course Enrollment Recommender System. In *Proceedings of the 9th International Conference on Educational Data Mining*, editace M. F. Tiffany Barnes, Min Chi, Raleigh, NC, USA: International Educational Data Mining Society, 2016, s. 312–317.

[4] Al-Badarenah, A.; Alsakran, J.: An Automated Recommender System for Course Selection. In *International Journal of Advanced Computer Science and Applications*, 2016, doi:10.14569/IJACSA.2016.070323.

[5] Chang, P.-C.; Lin, C.-H.; Chen, M.-H.: A Hybrid Course Recommendation System by Integrating Collaborative Filtering and Artificial Immune Systems. *Algorithms*, ročník 9, 2016, ISSN 1999-4893, doi:10.3390/a9030047. Dostupné z: http://www.mdpi.com/1999-4893/9/3/47

[6] Carballo, F. O. G.: *Masters' Courses Recommendation: Exploring Collaborative Filtering and Singular Value Decomposition with Student Profiling*. Diplomová práce, Instituto Superior Tecnico Universidade de Lisboa, Portugal, 2014. Dostupné z: https://fenix.tecnico.ulisboa.pt/downloadFile/563345090413333/Thesis.pdf

[7] Denley, T.: How Predictive Analytics and Choice Architecture Can Improve Student Success. *Journal of Research and Practice in Assessment*, ročník 9, 2014, ISSN EISSN-2161-4210.

[8] Sharon, Y.-J. L.; Lien, K.-W.: CareerMaker: An Adaptive Elective Courses Recommendation System. 2013. Dostupné z: http://research.ctu.edu.tw/vra/resources/101/11206/B02/482/23/f8ed86eb3df7f62e013e0b66b5c90025.pdf

[9] Parameswaran, A.; Venetis, P.; Garcia-Molina, H.: Recommendation Systems with Complex Constraints: A CourseRank Perspective. *Transactions on Information Systems (TOIS) – To Appear*, June 2011. Dostupné z: http://ilpubs.stanford.edu:8090/909/

[10] Youngseok Lee, J. C.: An Intelligent Course Recommendation System. *Journal of The Smart Computing Review*, ročník 1, 2011: s. 69–84. Dostupné z: http://www.dbpia.co.kr/Article/NODE02464936

[11] Ray, S.; Sharma, A.: *A Collaborative Filtering Based Approach for Recommending Elective Courses*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, ISBN 978-3-642-19423-8, s. 330–339, doi:10.1007/978-3-642-19423-8_34. Dostupné z: http://dx.doi.org/10.1007/978-3-642-19423-8_34

[12] Bendakir, N.; Aïmeur, E.: Using association rules for course recommendation. In *Proceedings of the AAAI Workshop on Educational Data Mining*, ročník 3, 2006.

[13] Mattei, N.; Dodson, T.; Guerin, J. T.; aj.: Lessons Learned from Creating a Course Advising Tool. Technická Zpráva arXiv:1312.4113, Dec 2013. Dostupné z: http://cds.cern.ch/record/1637380

[14] Grewal DS, K. K.: Developing an Intelligent Recommendation System for Course Selection by Students for Graduate Courses. *Business and Economics Journal*, ročník 7, č. 2, 2016: s. 1–9, ISSN 2151-6219, doi:10.4172/2151-6219.1000209. Dostupné z: http://www.omicsonline.com/open-access/developing-an-intelligent-recommendation-system-for-course-selectionby-students-for-graduate-courses-2151-6219-1000209.php?aid=73333

[15] Upendran, D.; Chatterjee, S.; Sindhumol, S.; aj.: Application of Predictive Analytics in Intelligent Course Recommendation. *Procedia Computer Science*, ročník 93, 2016: s. 917 – 923, ISSN 1877-0509, doi:http://dx.doi.org/10.1016/j.procs.2016.07.267. Dostupné z: http://www.sciencedirect.com/science/article/pii/S1877050916315137

[16] de Heus, M.: Design and evaluation of a recommender system for high school courses in the Netherlands. December 2013. Dostupné z: http://essay.utwente.nl/65029/

[17] Gulzar, Z.; Leema, A.: Subject Recommendation Using Ontology for Computer Science ACM Curricula. In *International Conference on Innovations in Computer Science and Technology*, 2016.

[18] Lotfy, M. M.; Salama, A. A.; El-ghareeb, H. A.; aj.: Subject Recommendation Using Ontology for Computer Science ACM Curricula. 2013.

[19] Aher, S. B.; Lobo, L.: Best combination of machine learning algorithms for course recommendation system in e-learning. *International Journal of Computer Applications*, ročník 41, č. 6, 2012.

[20] Bydžovská, H.: A Comparative Analysis of Techniques for Predicting Student Performance. In *Proceedings of the 9th International Conference on Educational Data Mining*, editace M. F. Tiffany Barnes, Min Chi, Raleigh, NC, USA: International Educational Data Mining Society, 2016, s. 306–311.

[21] Isinkaye, F.; Folajimi, Y.; Ojokoh, B.: Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, ročník 16, č. 3, 2015: s. 261 – 273, ISSN 1110-8665, doi:http://dx.doi.org/10.1016/j.eij.2015.06.005. Dostupné z: http://www.sciencedirect.com/science/article/pii/S1110866515000341

[22] Řehořek, T.: Recommender Systems. [Nepublikovaná přednáška], 2014, fakulta informačních technologií ČVUT v Praze.

[23] Gorakala, S. K.: Basic recommendation engine using R. http://www.dataperspective.info/2014/05/basic-recommendation-engine-using-r.html, citováno: 2017-01-17.

[24] Gorakala, S. K.: Item Based Collaborative Filtering Recommender Systems in R. https://www.r-bloggers.com/item-based-collaborative-filtering-recommender-systems-in-r/, citováno: 2017-01-17.

[25] Berka, P.: *Dobývání znalostí z databází*. Academia, 2003, ISBN 9788020010629. Dostupné z: https://books.google.cz/books?id=tGvFAAAACAAJ

[26] Směrnice děkana FIT ČVUT č. 13/2015 pro realizaci bakalářského a magisterského studijního programu Informatika na Fakultě informačních technologií ČVUT v Praze. Citováno: 2017-05-01. Dostupné z: http://www.fit.cvut.cz/sites/default/files/STO/SmerniceDekana13-2015-1.pdf

[27] Směrnice děkana FIT ČVUT č. 22/2016 pro přijímací řízení do všech forem doktorského studijního programu Informatika na FIT ČVUT v Praze pro letní semestr akademického roku 2016/17 a pro zimní semestr akademického roku 2017/2018. Citováno: 2017-05-01. Dostupné z: http://www.fit.cvut.cz/sites/default/files/SmerniceDekana22-2016_1.pdf

[28] O novém studijním programu Informatika. Citováno: 2017-05-01. Dostupné z: http://fit.cvut.cz/student/bakalarsky-program/nova_informatika

[29] Kordík, P.: Vytěžování znalostí z dat. [Nepublikovaná přednáška], 2014, fakulta informačních technologií ČVUT v Praze.

[30] Keras: Deep Learning library for Theano and TensorFlow. Citováno: 2017-05-01. Dostupné z: https://keras.io

Příloha A

Seznam použitých zkratek

CRISP-DM Cross Industry Standard Process for Data Mining

ČVUT České vysoké učení technické v Praze

EBIE Extended Business Intelligence Encyclopedia

ECTS European Credit Transfer and Accumulation System

FIT Fakulta informacních technologií

XML Extensible Markup Language

KOS Komponenta studium

CSV comma-separated values

SQL Structured Query Language

KF kolaborativní filtrování

RARE recommender system based on association rules

REST representational state transfer

HTML HyperText Markup Language

Příloha **B**

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src`
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
  - `Jupyter notebooky` . analýza dat, implementace a část textu práce ve formátu Jupyter notebook
  - `Markdown` ..... části textu práce před přepsáním do formátu LaTeX
  - `dotazník` ..... dotazník a odpovědi
  - `miniaplikace` ..... podklady pro implementaci miniaplikace v Portálu ČVUT
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF