Systém na odporúčanie predmetov pre študentov FMFI UK

Course recommender for FMPH CU students

Patrícia Vnenčáková

Vedúci: Mgr. Askar Gafurov, PhD.

CIELE PRÁCE

- pochopiť problematiku
- vykonať experimenty pre rôzne metódy podobnosti
- vytvoriť model systému na odporúčanie predmetov pre študentov FMFI UK

Odporúčací systém

Recommender system

- typ softvérového systému, ktorý poskytuje používateľom personalizované odporúčania rôznych položiek na základe ich predchádzajúceho správania, preferencií a záujmov

zalando

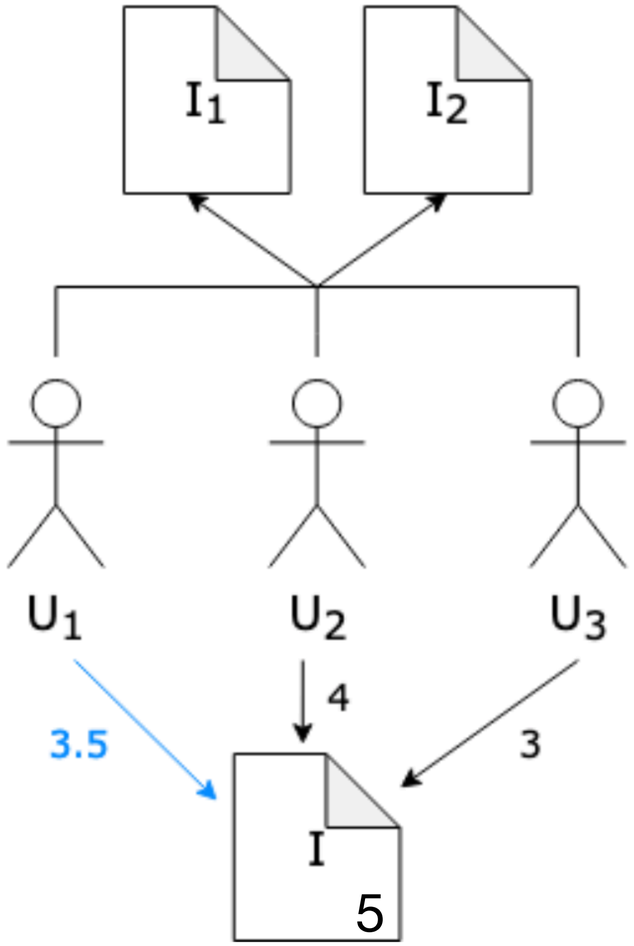
Odporúčacie systémy

Základné modely

- **kolaboratívne filtrovanie (collaborative filtering)**
  - založené na užívateľoch (user-based)
  - založené na položkách (item-based)
- **filtrovanie založené na obsahu (content-based filtering)**

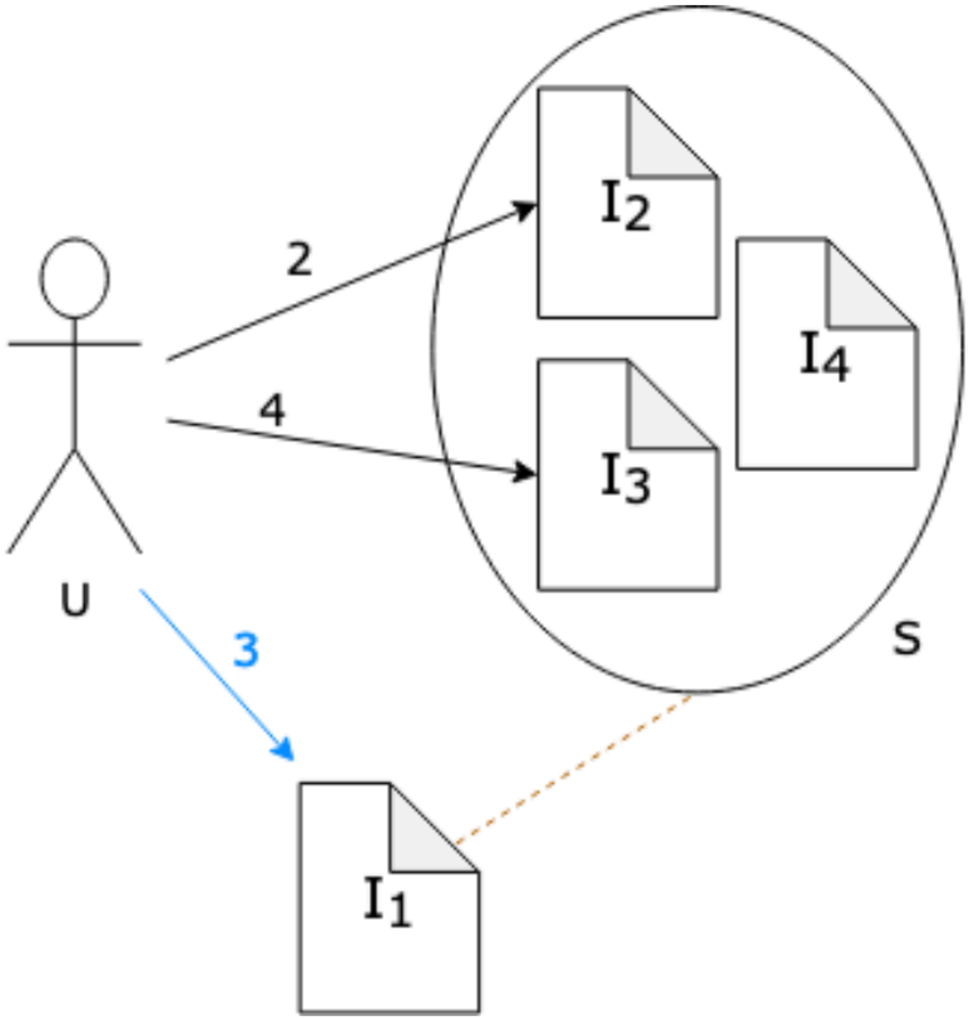
Základné modely

- **collaborative filtering** -> založené na skutočnosti, že podobní používatelia majú podobné správanie a podobné položky dostávajú podobné hodnotenia
  - **user-based**



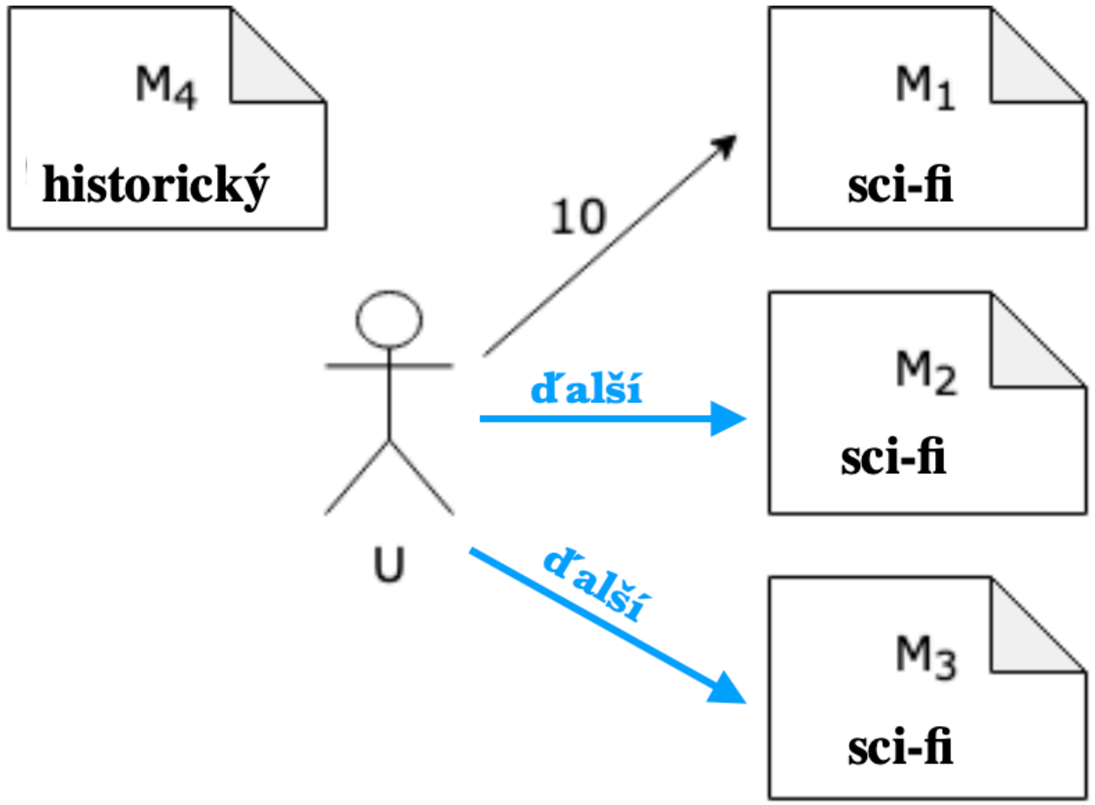
Základné modely

- collaborative filtering
  - **item-based**



Základné modely

- **content-based filtering** -> používajú podrobnejší opis položiek (vlastnosti, atribúty a špecifikácie) na odporúčanie položiek používateľom



Určenie podobnosti

kNN algoritmus (k- Nearest Neighbours)

- algoritmus strojového učenia
- k novému bodu pomocou zvolenej metódy podobnosti nájde v trénovacej množine **k** najbližších bodov (**susedov**) a použije tieto body na vytvorenie predpovede
- metódy podobnosti:
  - **Prienik**
  - **Hammingová vzdialenosť**
  - **Jaccardov index**

Metódy podobnosti

- **Prienik**
  - porovnanie dvoch množín A, B
  - výsledná hodnota je spoločná množina prvkov
  - čím väčšia je mohutnosť spoločnej množiny, tým sú množiny podobnejšie

$$I(A, B) = A \cap B = \{x \mid x \in A \text{ and } x \in B\}$$

Metódy podobnosti

- **Hammingová vzdialenosť**
  - porovnanie dvoch reťazcov a, b
  - výsledná hodnota je celé číslo, ktoré reprezentuje počet pozícií, v ktorých sa reťazce líšia
  - čím nižšia je Hammingová vzdialenosť, tým sú reťazce podobnejšie
  - napr. d(‘001’, ‘100’) = 2 (líšia sa na pozícii 1 a 3)

$$d(a,b) = \left| \left\{ i \in \{1,\ldots,n\} : a_i \neq b_i \right\} \right|$$

Metódy podobnosti

- **Jaccardov index**
  - porovnanie dvoch množín A, B
  - podiel mohutnosti prieniku A, B a zjednotenia A, B
  - výsledná hodnota je z rozsahu <0;1>, pričom 0 = žiadna podobnosť, 1 = množiny sú identické
  - čím väčší Jaccardov index, tým sú dve množiny podobnejšie

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Evaluácia

- potrebná na určenie “skóre” systému a následné vylepšovanie
- zahŕňa porovnanie predikcií so skutočnými hodnotami v testovacej množine pomocou zvolenej metriky presnosti
- metriky presnosti:
  - **RMSE**
  - **F1 skóre**

Metriky presnosti

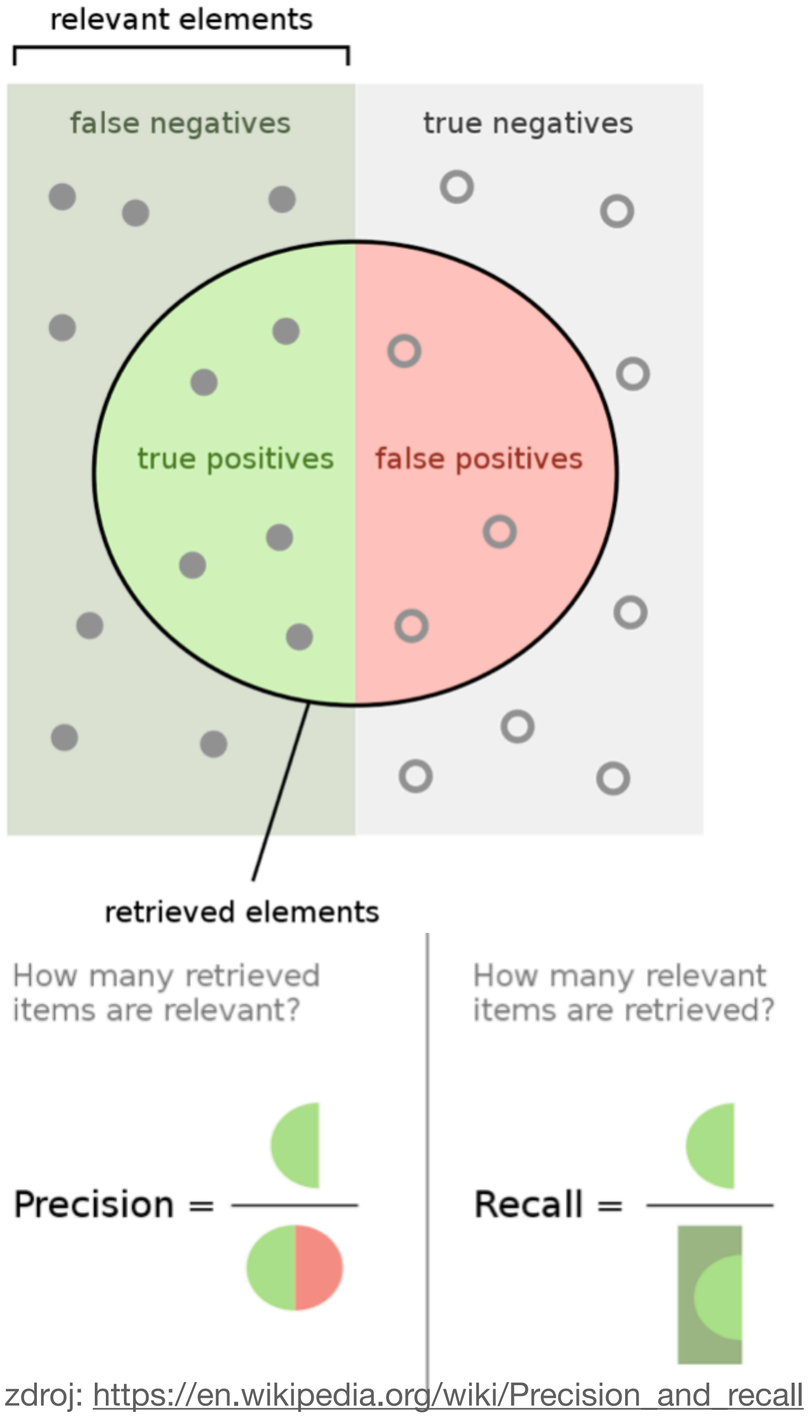
- **RMSE (Root Mean Squared Error)**
  - meranie chybovosti systému
  - výsledná hodnota je z rozsahu <0; inf), pričom záleží na rozsahu dát
  - čím menšie RMSE, tým je systém presnejší
  - E = testovacia množina, r_{uj} = reálna hodnota, r'_{uj} = predikovaná hodnota, $e_{uj} = r'_{uj} - r_{uj}$ vyjadruje chybu

$$RMSE = \sqrt{\frac{1}{|E|} \sum_{(u,j) \in E} e_{uj}^2}$$

Metriky presnosti

- **F1 skóre**
  - meranie celkového skóre systému
  - binárne hodnotenia
  - výsledná hodnota je z rozsahu <0; 1>
    a na výpočet sa používajú hodnoty **precision** a **recall**
  - čím je F1 väčšie, tým je systém presnejší

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$$



zdroj: https://en.wikipedia.org/wiki/Precision_and_recall

PRAKTICKÁ ČASŤ

Dáta

- anonymizované dáta z akademického informačného systému AIS2 od rokov **2009/10** po **2021/22**
- neobsahujú dáta o doktorandoch
- potrebný GDPR tréning
- dôležité tabuľky:
  - **export** = údaje o zápise študenta (kto, čo, rok, semester, pass, program)
  - **studprog** = štúdijné programy
  - **predmet** = ponúkané kurzy
- veľa nadbytočných záznamov => potreba čistenia

Proces čistenia

- **spojenie identických predmetov**
- **odstránenie nadbytočných záznamov**
  - predmety, ktoré sú doktorandské
  - predmety, ktoré boli zapísane menej ako dvakrát
  - programy, ktoré nik nenavštevoval

kód predmetu	skratka	názov predmetu
RKCMBF.CD.BA/K-KT61-114/20	K-KT61-114	Cirkevné dejiny - novovek
FiF.KAA/A-buAN-430/20	A-buAN-430	Dejiny komiksu v USA
FM.KIS/370B/19	370B	Obchodná čínština I

Proces čistenia

počet predmetov pred spojením identických	4688
počet predmetov po spojení identických	**4349**
počet doktorandských predmetov	1032
počet predmetov zapísaných menej ako dvakrát	1641
celkový počet nadbytočných predmetov	1649
celkový počet predmetov po odstránení nadbytočných	**2700**
počet programov pred odstránením nenavštevujúcich	148
počet programov po odstránení nenavštevujúcich	**68**

Implementácia

- model: **user-based collaborative filtering**
- hodnotenia: **binárne (0/1)**
- metódy podobnosti:
  - **Prienik**
  - **Hammingová vzdialenosť**
  - **Jaccardov index**
- metriky presnosti (evaluátor):
  - **RMSE**
  - **F1 skóre**

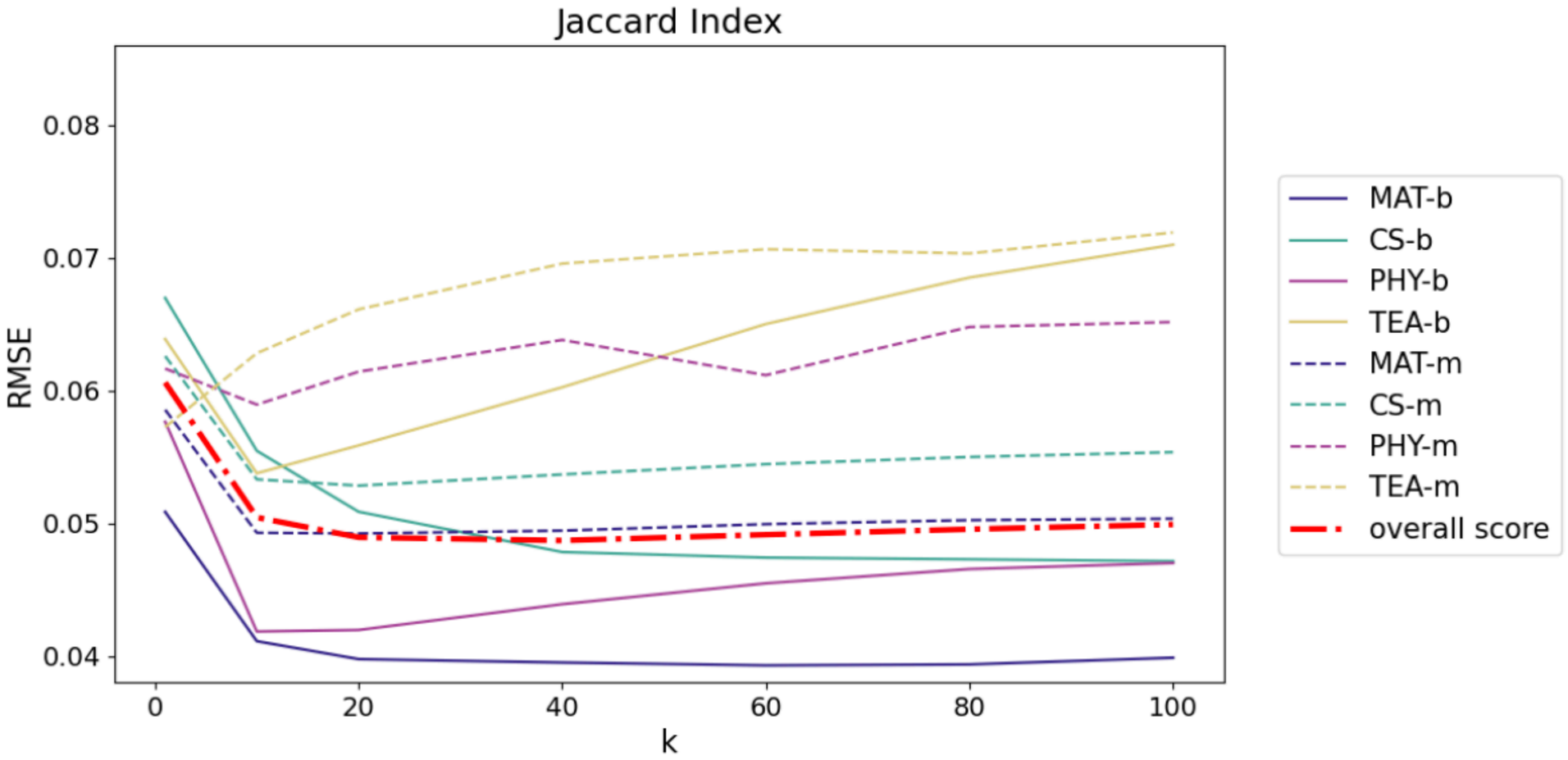
Experimenty

- vykonané experimenty:
  - Vplyv veľkosti k v kNN algoritme
  - Evaluácia roku 2021/22 s postupným pridávaním trénovacích dát
  - Vplyv subsamplingu trénovacích dát
- vyhodnotenie pre jednotlivé odbory
- celkové hodnotenie

- **RMSE ->** čím menšie, tým presnejší systém
- **F1 skóre ->** čím väčšie, tým presnejší systém

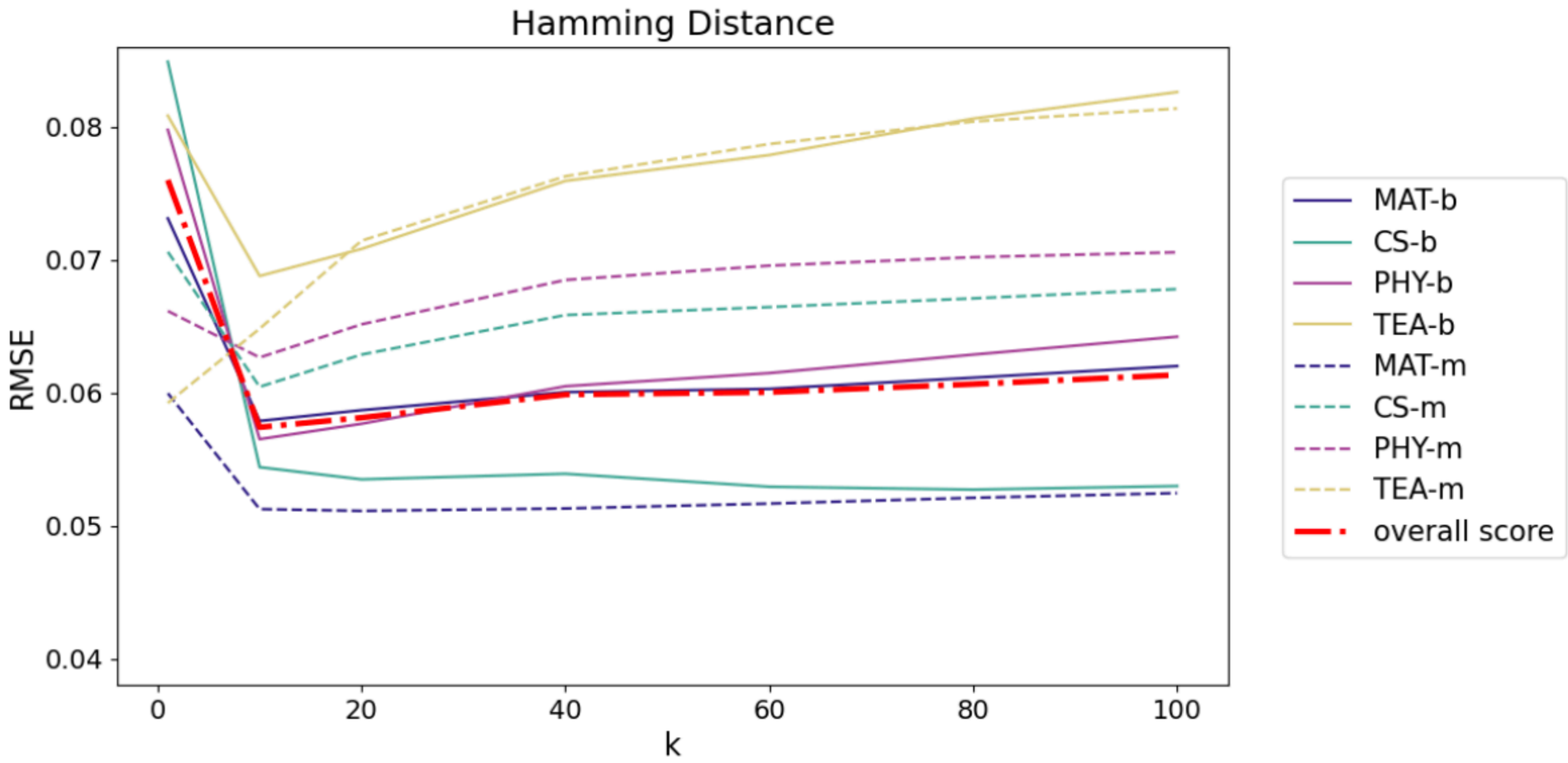
Vplyv veľkosti k v kNN algoritme

- očakávanie: čím väčšie k, tým presnejší systém



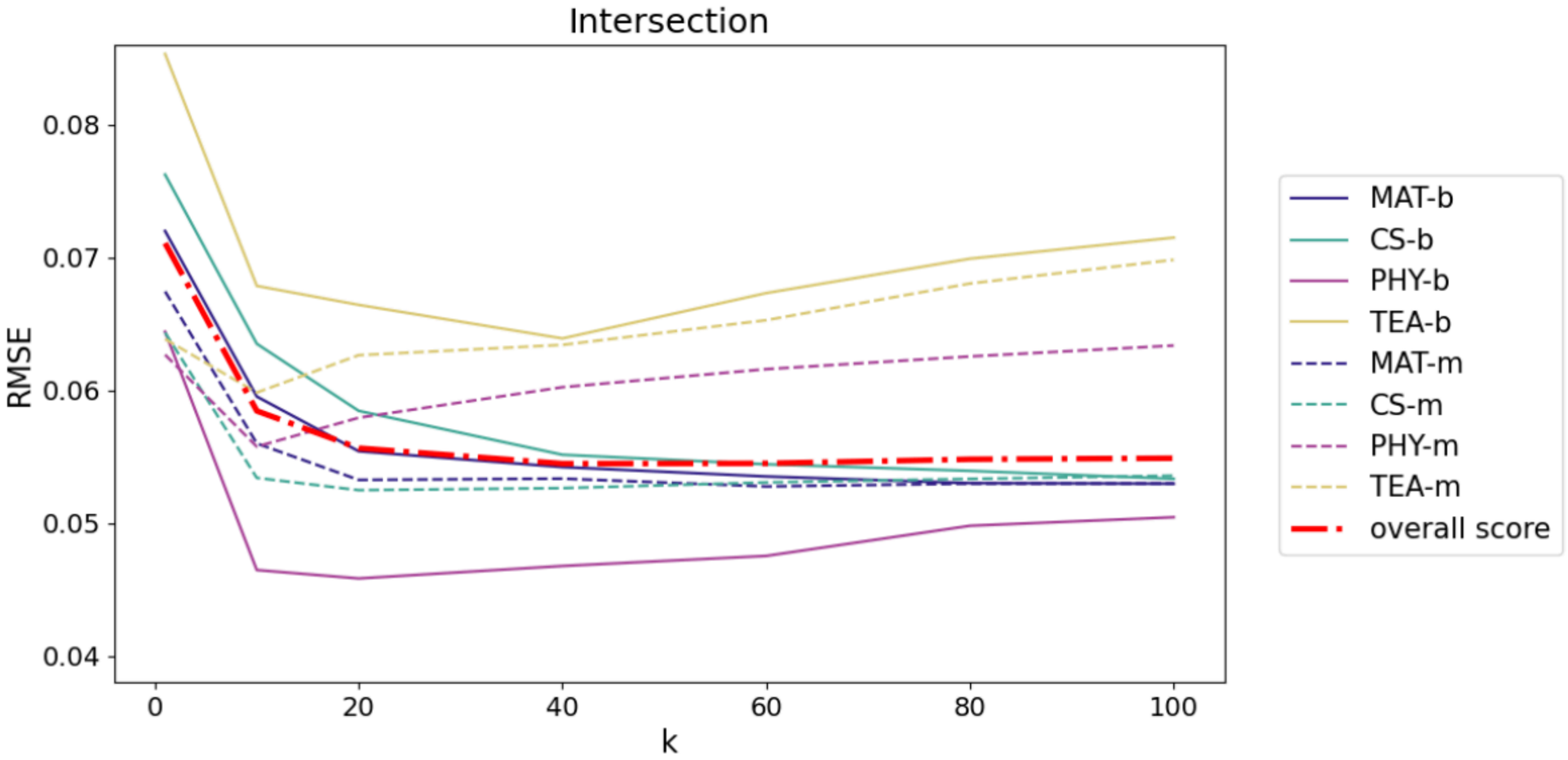
Vplyv veľkosti k v kNN algoritme

- očakávanie: čím väčšie k, tým presnejší systém



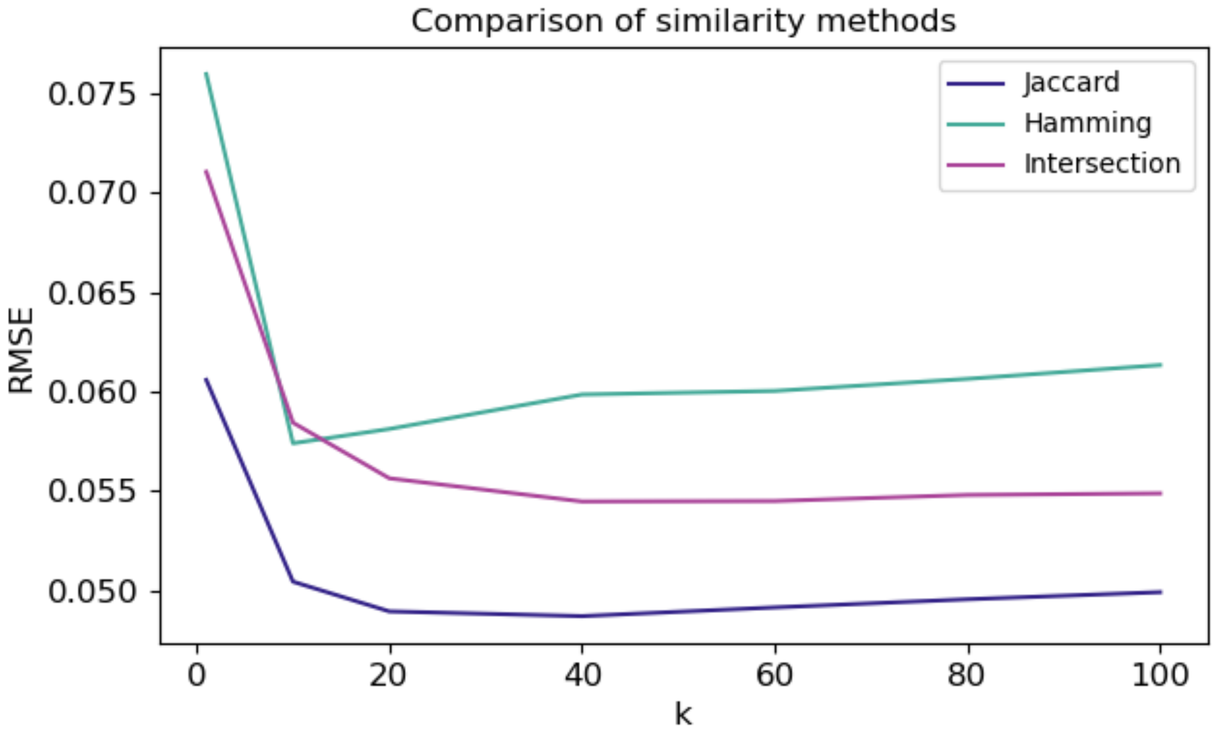
Vplyv veľkosti k v kNN algoritme

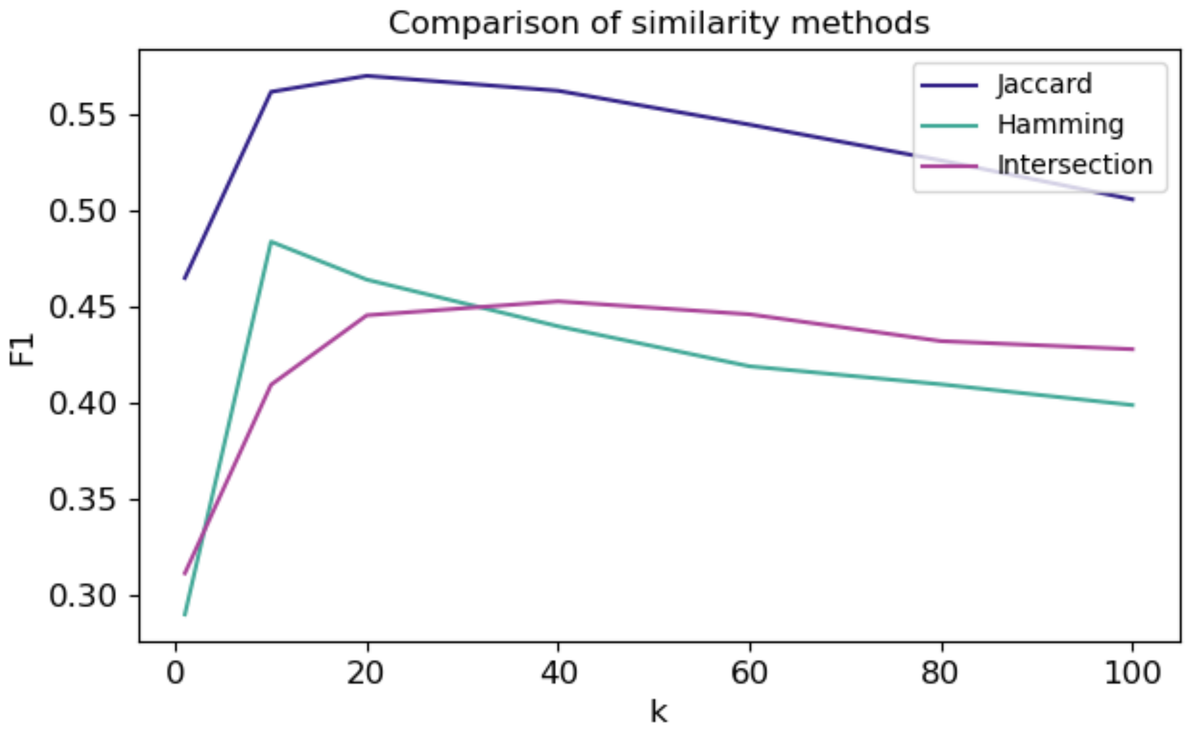
- očakávanie: čím väčšie k, tým presnejší systém



Vplyv veľkosti k v kNN algoritme

- očakávanie: čím väčšie k, tým presnejší systém

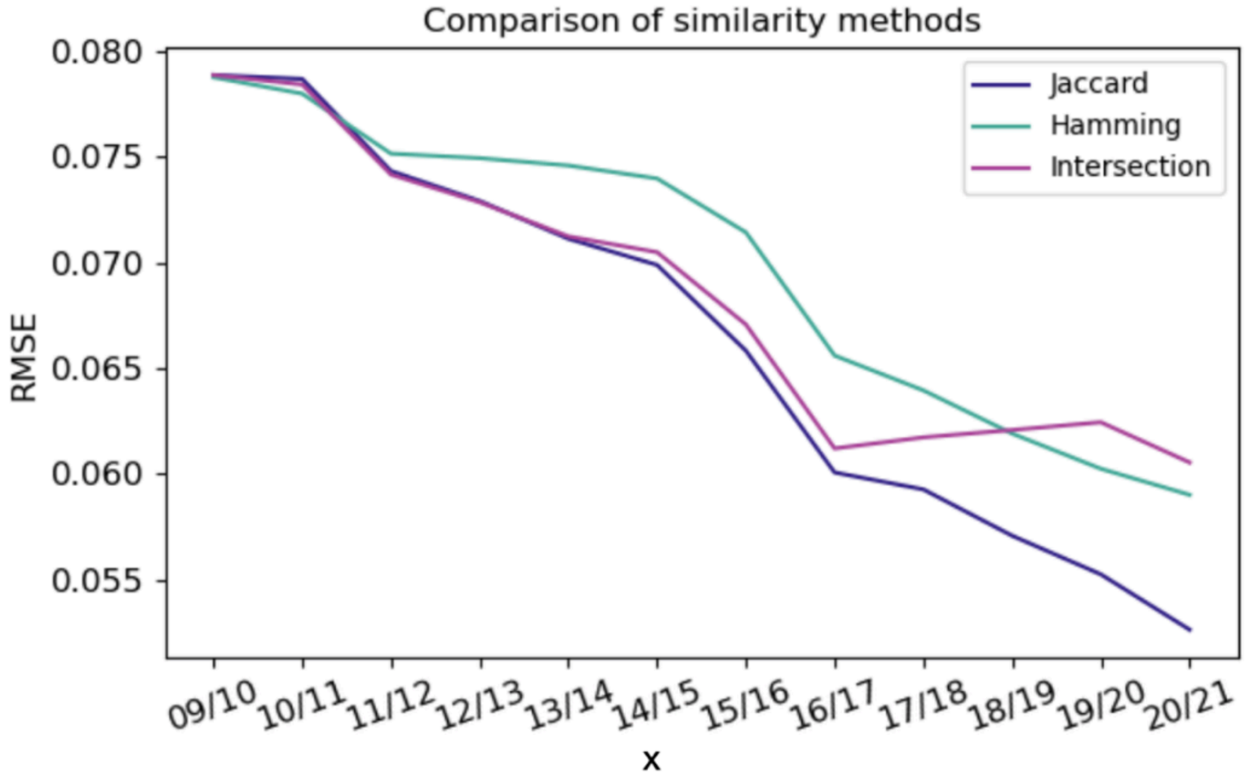


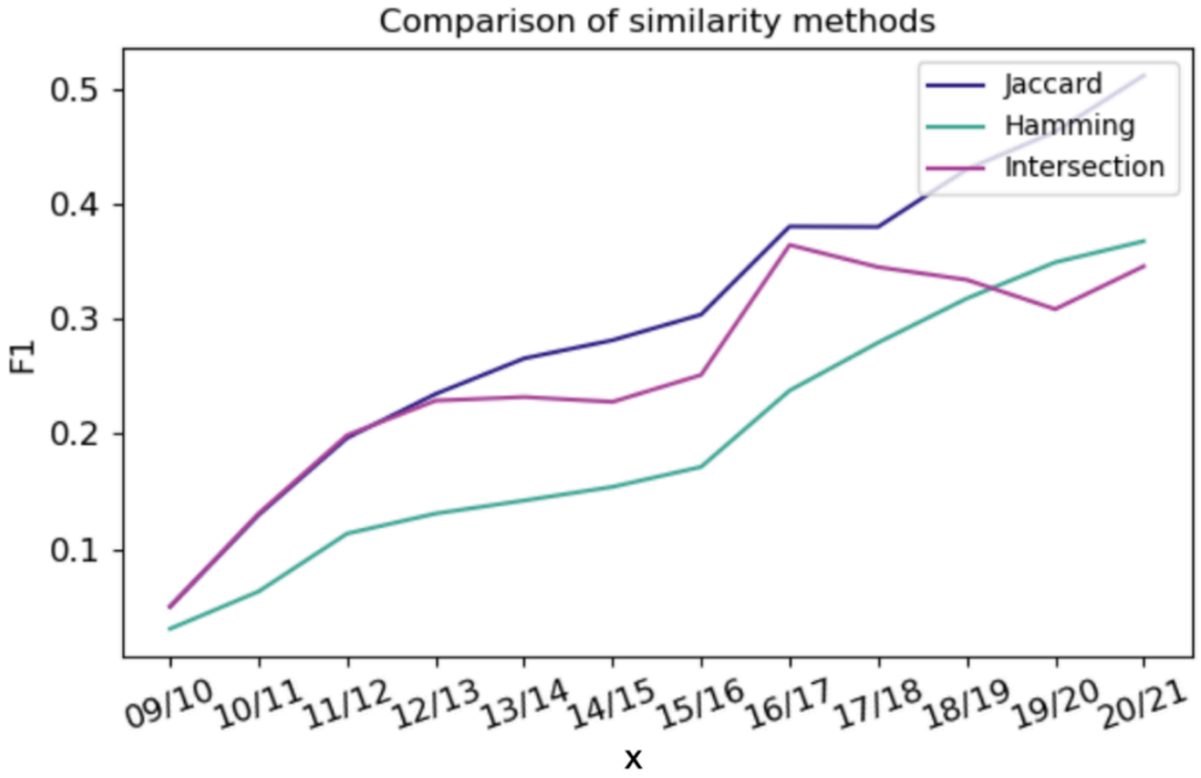


- k = 15, najlepšia metóda: **Jaccardov index**

Evaluácia roku 2021/22 s postupným pridávaním trénovacích dát

- očakávanie: čím viac trénovacích dát, tým presnejší systém

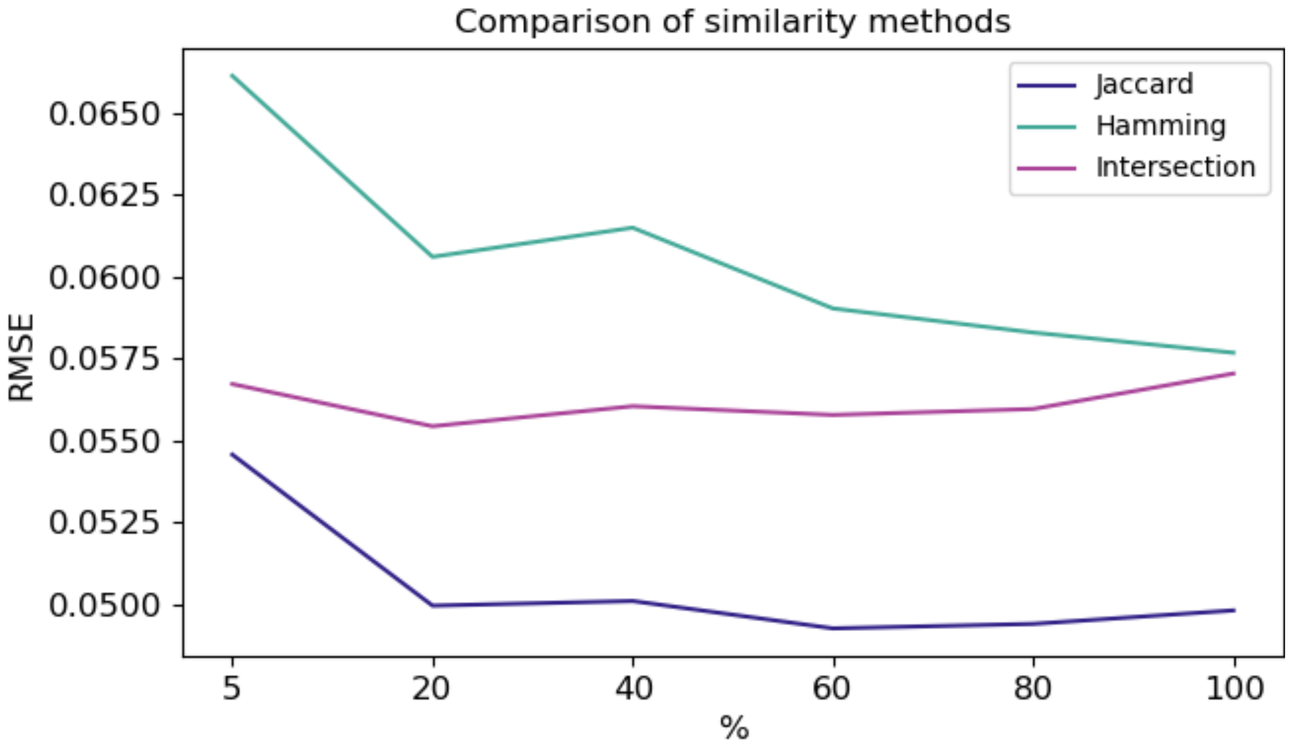


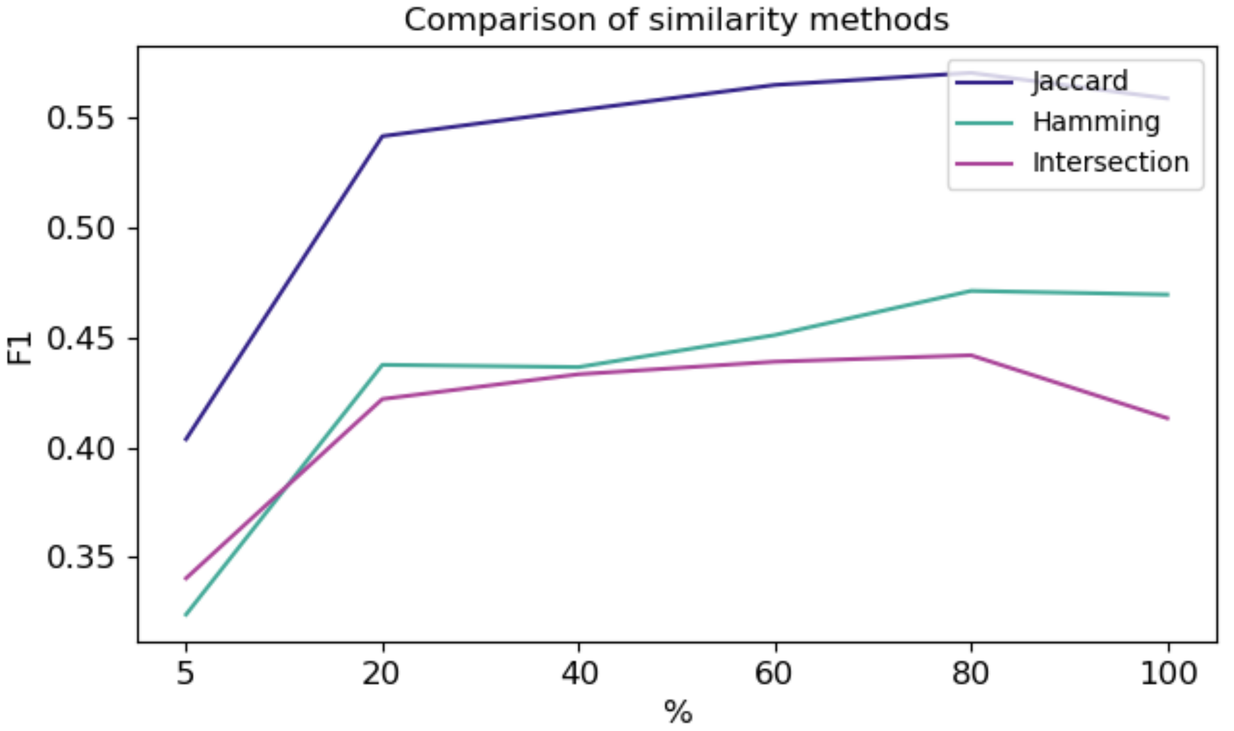


- najlepšia metóda: **Jaccardov index**

Vplyv subsamplingu trénovacích dát

- očakávanie: čím väčšie percento trénovacích dát ponechám, tým presnejší systém





- najlepšia metóda: **Jaccardov index**

Výsledok

- najlepšia metóda podobnosti: **Jaccardov index**
- optimálne **k = 15**
- vykonaný test na mojich osobných dátach

Zapísaný predmet	predikcia
Algoritmy a dátové štruktúry	1.0
Databázové praktikum	0.533
Formálne jazyky a automaty (1)	1.0
Linux - princípy a prostriedky	0.333
Princípy tvorby softvéru	0.933
Programovanie (3)	0.933
Ročníkový projekt (1)	1.0
Telesná výchova a šport (3)	0.666
Úvod do databázových systémov	1.0
Operačné systémy	0.933
Počítačové siete (1)	0.866
Ročníkový projekt (2)	0.933
Spoločenské aspekty informatiky	0.933
Telesná výchova a šport (4)	0.466
Tvorba efektívnych algoritmov	0.933
Úvod do matematickej logiky	1.0

Výsledok

- najlepšia metóda podobnosti: **Jaccardov index**
- optimálne **k = 15**
- vykonaný test na mojich osobných dátach

nezapísané predmety	**predikcia**
Webové aplikácie (2)	0.066
Matematická analýza (3)	0.066
Kryptológia (1)	0.066
Algebra (3)	0.133
Formálne jazyky a automaty (2)	0.333
Základy reverzného inžinierstva	0.066
Technológie digitálnej výroby	0.066
Matematika (2) - Matematická analýza	0.066
Vývoj mobilných aplikácií	0.066
Úvod do informačnej bezpečnosti	0.333
Kvantové spracovanie informácie	0.066
Programovanie (2)	0.066

Budúca práca

- **získať podprobnejší opis predmetov**
  - content-based filtering model
  - zamerať sa na voliteľné a povinne voliteľné premdety
- **získať dáta zo študentskej ankety**
  - recenzie a hodnotenia predmetov
- **vytvoriť webové rozhranie na používanie**
- **implementovať do aplikácie Votr**

ĎAKUJEM ZA POZORNOSŤ

DISKUSIA

Napriek tomu, že konečná množina kľúčových slov tu nie je, nemožno považovať príslušnosť k programu a kategorizáciu predmetov na povinné/povinne voliteľné/ výberové čiastočne za takýto popis?

Áno.

Takáto kategorizácia by bola vhodná, priam žiadúca. V poskytnutej databáze sa však tieto informácie nenachádzali. Pre zistenie príslušnosti predmetu do kategórie povinný/ povinne voliteľný/ výberový by bolo potrebné scrape-ovanie informačných listov z webu. Zvažovali sme informácie takto získať, no keďže som sa s touto činnosťou nikdy predtým osobne nestretla, tak z dôvodu časovej tiesne sme sa rozhodli to nerobiť.

Uniká mi, prečo je študent každý rok iný „user“ ; nestrácajú sa tým možné závislosti?

Dáta boli uložené v nasledujúcej forme:

[ID], [doposiaľ zapísané predmety], [zapísané predmety v danom ak. roku], [ak. rok], [štúdijný program]

prvák:
[doposiaľ zapísané predmety] = {}
[zapísané predmety v danom ak. roku] = {1, 2, 3}
[ak. rok] = 2019/20
[študijný program] = INF

druhák:
[doposiaľ zapísané predmety] = {1, 2, 3}
[zapísané predmety v danom ak. roku] = {4, 5}
[ak. rok] = 2020/21
[študijný program] = INF

tretiak:
[doposiaľ zapísané predmety] = {1, 2, 3, 4, 5}
[zapísané predmety v danom ak. roku] = {6}
[ak. rok] = 2021/22
[študijný program] = INF

Prečo sa využíva len informácia zapísal/ nezapísal si predmet?

Predstava funkcionality nášho systému bola nasledovná:
1.) študent si nechá odporúčiť predmety na základe toho, čo si doposiaľ zapísal
2.) systém porovná množinu doposiaľ zapísaných predmetov tohto študenta s množinami doposiaľ zapísaných predmetov užívateľov v databáze
3.) určí najpodobnejších a odporúči predmety

Keďže sme určovali podobnosť na základe doposiaľ zapísaných predmetov, tak sme pri ukladaní dát použili práve túto informáciu. Je však pravda, že v databáze sa nachádza aj informácia o absolvovaní predmetu. Túto informáciu sme však, bohužial, z nepozornosti nepoužili.

V tabuľke 3.4 sa objavuje aj informácia o študijnom programe študenta. Akým spôsobom bola táto informácia využívaná?

Tabuľka 3.4 predstavuje internú reprezentáciu dát, ktoré sme mali k dispozícii. Túto informáciu sme využívali v prípadoch, keď sme vytvárali predikcie pre prvákov. Keďže prváci majú množinu doposiaľ zapísaných predmetov prázdnu, tak by prváci zo všetkých študijných programoch boli považovaní za rovnakých. Bolo preto potrebné, aby boli prváci z rôznych študijných programov odlíšení.

User ID	so far	this year	academic year	study programme
1200	[]	[1, 2]	2019/20	INF
1201	[1, 2]	[3, 4]	2020/21	INF
1202	[1, 2, 3, 4]	[5, 6]	2021/22	INF

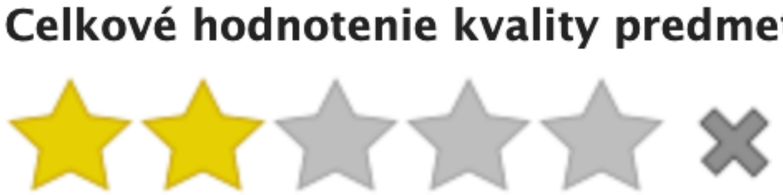
Table 3.4: Internal representation of the data

Pri výpočte podobnosti má (ak správne rozumiem) každá zložka vektora rovnakú váhu. Nebolo by rozumné uvažovať aj váhovanie jednotlivých položiek vzhľadom k významnosti?

Áno, bolo by to rozumné.

Pri dostupných dátach sme však neprišli k žiadnému inému rozumnému váhovaniu ako 0/1.

Jedna z rozumných možností váhovania sa ukazuje využitie ohodnotení predmetov jednotlivými študentmi v študentskej ankete. Vo vektore by tým pádom bola pre daný predmet uvedená hodnota na základe počtu udelených hviezdičiek. Dáta z ankety sme však k dispozícii nemali.



Viete si predstaviť nejaký spôsob zakomponovania študijných programov, kategorizácie predmetov, prerekvizít,.. do reprezentácie dát a vyhodnocovacích funkcií?

Ak by sme mali informáciu o kategorizácií predmetov, tak by model napríklad neodporúčal povinné predmety, keďže je rozumnejšie odporúčať predmety, kde má študent pri výbere “voľnosť”. Pri vytváraní predikcie pre študenta by teda bola pravdepodobnosť zápisu povinných predmetov nulová.

Ak by sme mali informáciu o prerekvizitách, tak by sme napríklad neodporúčali študentovi tie predmety, pri ktorých nemá absolvované prerekvizity. Aj v tomto prípade by teda bola pravdepodobnosť zápisu takýchto predmetov nulová.

Nemali by sa pri odporúčaní brať nejakým (akým) spôsobom do úvahy podmienky na absolvovanie programu?

Jednou z podmienok absolvovania programu, je dosiahnutie 180 kreditov. Pri odporúčaní by sa mohol tento fakt zohľadňovať napríklad tak, že študentovi odporučíme toľko relevantných predmetov, aby súčet ich kreditov bol aspoň 60 (odporúčaný počet kreditov za jeden rok). Ak by teda študent využil naše odporúčanie, tak by bolo zaručené, že na konci tretieho ročníka by mal minimálne 180 kreditov.

Iný prístup k tejto podmienke by mohol byť taký, že pred samotným odporúčaním by systém zistil, koľko kreditov doposiaľ daný študent získal a odporučil relevantné predmety podľa toho, koľko kreditov mu chýba, aby splnil odporúčaný rozsah. Napríklad, ak by si študent nechal odporučiť predmety do druhého ročníka a mal by zatiaľ 80 kreditov, tak systém by mu odporučil predmety tak, aby súčet ich kreditov bol aspoň 40 (na konci druhého ročníka je odporúčaný rozsah 120 kreditov).