Ústav informačních studií a knihovnictví
Filozofická fakulta
Univerzita Karlova

Název práce: **Využití data miningu v analýze filmových anotací**

Typ práce: diplomová

Hodnocení práce: **velmi dobře**

Vedoucí práce: doc. RNDr. Jiří Souček, DrSc.

Oponent/-ka práce: Ing. Petr Koubský, CSc.

Konzultant/-ka práce: Mgr. Josef Šlerka

Řešitel/-ka práce: Bc. Martina Podhůrská

Slovní hodnocení práce

V neformálním úvodu slibuje autorka, že se pokusí „pokusit se na základě textové analýzy anotace a dalších přidružených atributů konkrétního snímku určit, zda bude film vhodný pro naše filmové preference, a současně, zda je možné na stejném základě odhadnout, jaké hodnocení získá snímek od širokého publika“. To je dobře vytyčený cíl; sice poměrně vysoký, přesto skýtající naději na zajímavé odpovědi.

Teoretické části nelze mnoho vytknout, přehledně popisuje potřebné koncepty a pojmy, cituje též přiměřené množství studií zabývajících se právě tématem predikce divácké úspěšnosti filmů. Tato otázka má přirozeně velký komerční význam a tomu odpovídá objem realizovaného výzkumu – diplomová práce z něj uvádí jen malý, přesto poměrně reprezentativní výběr.

U výzkumné části práce si čtenář povšimne ze všeho nejdříve, že se v nezanedbatelné míře odchyluje od slibného názvu. Analýza anotací totiž tvoří jen menší část výzkumu, těžiště spočívá v hodnocení lépe kvantifikovatelných atributů jednotlivých snímků. Z výčtu ověřovaných hypotéz (str. 49) to ještě není patrné, z popisu použité metodiky již ano.

Autorka vyhodnocovala dva vzájemně nesouvisející datové soubory: jednak výsledky dotazníkového šetření na vzorku 327 osob (metodou dostupného výběru, tj. nereprezentativně), jednak data získaná z databáze ČSFD. Kvantitativní dotazníkový průzkum napomohl k formulaci hypotéz (přesněji řečeno, k potvrzení výchozích odhadů). Ty se pak autorka pokusila testovat hledáním korelace mezi proměnnými (tj. atributy popisu filmu) obsaženými právě v databázi ČSFD. (Formulace na str. 59 – „rozhodla [jsem se] využít jako dataset k výzkumu jiný balíček dat, který jsem získala svépomocí z dostupných a adekvátních zdrojů“ je však poněkud matoucí: není jasné, o jaké „adekvátní zdroje“ jde.) Přitom zjistila, že některé atributy vykazují vysokou korelaci s průměrným bodovým ohodnocením filmu, jiné nikoli – konkrétně jde o rok vzniku filmu, jeho stopáž a u některých žánrů příslušnost k nim (u jiných nikoli).

Mimochodem, poměrně bizarním zjištěním je, že „(…) ve vztahu k uživatelskému hodnocení filmu na ČSFD je patrné, že statisticky významnými proměnnými jsou hodnocení (…)“ – ano, to skutečně jsou, korelace kterékoli proměnné vůči sobě samé bývá velmi vysoká ☺.

U textových anotací, k nimž se zde konečně dostáváme, věnuje autorka dost místa vztahu mezi jejich délkou a hodnocením filmu, což je zajímavý (a nikoli nesmyslný) nápad. Analýza dat naznačuje, že tento vztah nese užitečnou informaci, pro jejíž vyhodnocení by však bylo zapotřebí souběžně zkoumat i další proměnné, zejména žánr, rok vzniku filmu a (nejspíš) výskyt některých klíčových slov. Adekvátní metodou by zde byla vícerozměrná regrese se zvážením možných nelineárních závislostí.

Celkem vzato, autorka odvedla lepší práci při vyhodnocování dotazníkového šetření než při analýze databáze, což je pochopitelné vzhledem k velmi odlišné obtížnosti těchto úkolů. Tomu odpovídají i závěry práce – většina z nich je odvozena právě z vyhodnocení dotazníků.

Dojem z práce kazí řada překlepů, gramatických nepřesností a stylistických závad. Mohla je odstranit pečlivější korektura.

K ústní části obhajoby navrhuji zodpovědět následující otázky:

- Vysvětlit formulaci „rozhodla [jsem se] využít jako dataset k výzkumu jiný balíček dat, který jsem získala svépomocí z dostupných a adekvátních zdrojů“ (str. 59).
- Vysvětlit formulaci „pokud je snímek zařazen jako pohádka, získá o 11,7 hodnotícího bodu méně, než kdyby v této kategorii zařazen nebyl“ (str. 64).
- Vysvětlit vztah mezi vzorcem na str. 44 a následným popisem algoritmu.
- Navrhnout (zhruba, bez technických detailů), jak by se dalo statistické hodnocení modifikovat, aby se odstínil vliv faktoru „čím novější snímek, tím kratší text uživatelského hodnocení“ (viz str. 64 – 65)

Práci doporučuji k obhajobě a navrhuji hodnocení známkou 2 – velmi dobře.

Ústav informačních studií a knihovnictví
Filozofická fakulta
Univerzita Karlova

Hodnotící tabulka

Aspekty práce	Vysvětlení	Hodnocení
metodologie a věcné zpracování tématu		20 bodů
přínos a novost práce		10 bodů
citování, korektnost citování, využití inf. zdrojů		20 bodů
slohové zpracování		7 bodů
gramatika textu		4 body
CELKEM		**61 bodů**

V Praze dne 3. 6. 2018

Petr Koubský, v.r.