MENDELU
Provozně
ekonomická
fakulta

Záměr diplomové práce
PROVOZNĚ EKONOMICKÁ FAKULTA

Jméno a příjmení: **Adam Prchal**
Číslo studenta (UID): **82161**
Kontaktní email: **xprchal@mendelu.cz**

Program a obor / specializace:
N-OI-ZNOI
Forma: ☒ prezenční
☐ kombinovaná

Ústav, kde má být práce zadána: 116 - Ústav informatiky
Uvažovaný vedoucí práce: Ing. Pavel Turčínek, Ph.D.
Konzultant práce (nepovinné): Konzultant práce

.....
podpis konzultanta

Navrhovaný název práce:

Detekce duplicit v geoprostorových datech společnosti CleverMaps

Vymezení řešeného problému v kontextu současného stavu poznání řešené problematiky s využitím odborné literatury, příp. (je-li to relevantní) formulace výzkumných otázek a hypotéz:

Společnost CleverMaps využívá svou Location Intelligence platfomu k poskytování komplexních geoprostorových analýz, což je klíčové pro efektivní rozhodování v různých oblastech podnikání. Jako součást jejich platformy CleverMaps nabízí službu zvanou Data Marketplace, která umožňuje uživatelům získávat a integrovat různorodé datové sady. Tento marketplace poskytuje širokou škálu dat. Příkladem takových dat mohou být demografické informace, které klientům umožňují získat přesné informace například o jejich cílové skupině. (CleverMaps, 2024)

Vzhledem k tomu, že při práci (nejen) s geodaty se často zapracovávají informace získané z různých zdrojů, nastává to, že se totožné údaje ve shromážděných datech mohou objevovat vícekrát. Tento jev se označuje jako duplicity. Tyto duplicity mohou nastat z mnoha důvodů a mohou mít různé podoby. Příkladem může být situace, kdy jsou v různých databázích stejná místa nebo objekty zaznamenány pod různými. Například, jedna databáze může obsahovat bod zájmu (point of interest) jako "Velký městský park", zatímco druhá jej může uvádět jako "Park v centru města". Obě tato označení se mohou vztahovat ke stejnému místu, ale různé názvy komplikují jejich rozpoznání jako duplicit. Další běžnou příčinou je chybné zadání dat, kde se například dva různé záznamy pro stejnou ulici mohou lišit pouze malými typografickými chybami nebo odlišnými způsoby zkrácení názvů. (Felix Naumann 2010)

Pro zajištění integrity a přesnosti geodat je detekce a řešení duplicit nezbytná. K identifikaci duplicit mohou být využity různé metody, od jednoduchých porovnání řetězců až po sofistikovanější techniky, jako jsou algoritmy strojového učení a neuronové sítě. Tyto pokročilé metody umožňují rozpoznávat a porovnávat podobnosti na

základě kontextu a pravděpodobnosti, což vede k efektivnějšímu a přesnějšímu detekování duplicit. (Christen 2012)

Detekce duplicit bez využití strojového učení zahrnuje například algoritmus pro výběr párů Sorted-Neighborhood-Method (SNM). Je možné provést porovnání všech možných kombinací záznamů, ale SNM efektivně snižuje počet nutných porovnání tím, že seřadí data a porovnává pouze sousední záznamy. Pro následné měření podobnosti mezi záznamy se často používají metody jako Jaro-Winkler a Levenshtein, které posuzují, do jaké míry se záznamy podobají. Na základě takové podobnosti lze určit, zda jsou porovnané záznamy duplicity. Tyto metody jsou obzvláště užitečné pro identifikaci menších rozdílů a překlepů ve zpracovávaných datech. (Draisbach 2013)

Metody strojového učení nabízí pokročilejší možnosti, které umožňují identifikovat složitější vzory, které by metody bez využití strojového učení nemuseli odhalit. Například, hluboké neuronové sítě mohou být trénovány na rozsáhlých datových sadách a díky tomu mohou nalézt v datech složité vzory. Takto natrénované neuronové sítě mohou označit za duplicity záznamy, které na první pohled jako duplicity nevypadají. (Pašek 2022)

V oblasti detekce duplicit existuje řada služeb, které nabízejí pokročilé řešení tohoto problému. Tyto služby jsou často založené na cloud technologiích a strojovém učení. Jsou schopné efektivně identifikovat a eliminovat duplicity v rozsáhlých datových sadách. Mezi poskytovatele těchto služeb patří například služby jako Data Ladder, Tilores nebo Melissa, které nabízejí různé nástroje pro automatickou detekci duplicitních záznamů.

Ačkoliv tyto služby přinášejí významné výhody v podobě úspory času a zlepšení kvality dat, mohou být spojeny s vysokými náklady, zejména v případech, kdy je zapotřebí zpracovat velké objemy dat. Finanční zátěž z těchto služeb může být pro některé společnosti značná, což může motivovat k implementaci vlastního řešení.

Předpokládaný cíl práce:

Cílem této práce je prozkoumat a otestovat různé metody detekce duplicit na geoprostorových datech, včetně metod založených na strojovém učení. Na základě analýzy výsledků testů doporučit nejvhodnější metody pro konkrétní sady geoprostorových dat poskytnutých společností CleverMaps.

Výsledná doporučení by měla společnosti CleverMaps pomoci v rámci zvyšování automatizace a zkvalitnění procesů kontrol kvality dat.

Návrh metodiky řešení včetně identifikace zkoumaného vzorku:

Na začátku práce bude provedena literární rešerše. Ta bude zaměřena na identifikaci typů duplicit v datech a metod, pro jejich detekci. Rešerše zahrne jak metody založené na strojovém učení, tak jiné relevantní přístupy. Společně s tím budou popsány i nástroje, které lze použít pro implementaci těchto metod.

V rámci analýzy a přípravy dat bude proveden rozbor datových sad poskytnutých společností CleverMaps, pro lepší pochopení struktury a charakteristik konkrétních datových sad.

Na základě zjištění z literární rešerše a rozboru datových sad budou vybrány metody pro testování. Takto vybrané metody se následně naimplementují za použití příslušných algoritmů a nástrojů.

Naimplementované metody budou testovány na poskytnutých datových sadách. V této části tak budou zajištěny relevantní podklady pro analýzu výsledků a tvorbu doporučení.

Po implementaci a testování metod se provede analýza výsledků testů a vyhodnotí se, které metody jsou pro detekci duplicit v poskytnutých datových sadách nejúčinnější na základě relevantních metrik.

V závěrečné části práce budou shrnuty všechny testované metody, s důrazem na jejich efektivitu a aplikovatelnost pro konkrétní datové sady společnosti CleverMaps. Budou také diskutovány výhody a možná omezení každé z metod. Na základě analýzy a výsledků testů budou formulována specifická doporučení pro společnost.

Návrh literárních pramenů pro vypracování práce:

ALBAYRAK, Osman Semih, Tevfik AYTEKIN a Tolga Ahmet KALAYCI, 2022. Duplicate product record detection engine for e-commerce platforms. Expert Systems with Applications [online]. 193, 116420. ISSN 0957-4174. Dostupné z: doi:10.1016/j.eswa.2021.116420

CHRISTEN, Peter, 2012. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Berlin Heidelberg: Springer. Data-centric systems and applications. ISBN 978-3-642-43001-5.

HUYEN, Chip, 2022. Designing machine learning systems: an iterative process for production-ready applications. First edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly. ISBN 978-1-09-810796-3.

NAUMAN, Felix a Melanie HERSCHEL, 2010. An Introduction to Duplicate Detection. B.m.: Springer Nature. ISBN 978-3-031-01835-0.

Zamýšlený rozsah samostudia, zejména doplňující literatura pro prohloubení znalostí v oboru práce:

V rámci samostudia budou prohlubovány znalosti v rámci praktického zpracování dat a implementace neuronových sítí.

KERAS, 2024. Keras documentation: Getting started with Keras [online]. Dostupné z: https://keras.io/getting_started/

MCGREGOR, Susan E., 2022. Practical Python data wrangling and data quality. Sebastopol, CA: O'Reilly Media. Inc. ISBN 978-1-4920-9150-9.

TENSORFLOW, 2024. Machine learning education | TensorFlow [online]. Dostupné z: https://www.tensorflow.org/resources/learn-ml

Dále budou využívány i další online zdroje dle potřeby.

Předpokládaná struktura práce s členěním na kapitoly a nástin jejich obsahu:

- **Úvod a cíl práce**
    - Seznámení čtenáře s tématem práce a stanovení cílů.
- **Literární rešerše**
    - Zavedení pojmů a seznámení s problematikou.
- **Metodika práce**
    - Popis postupu práce.
- **Analýza a příprava dat**
    - Zkoumání datových sad od CleverMaps, jejich úprava a optimalizace pro další zpracování.
- **Výběr a implementace metod**
    - Rozhodování o nejvhodnějších metodách detekce duplicit a jejich aplikace na zvolené datové sady.
- **Testování a evaluace metod**
    - Provádění testů zvolených metod na datech a měření jejich výkonu pomocí relevantních metrik.
- **Analýza výsledků a formulace doporučení**
    - Interpretace získaných dat z testů, odvození závěrů a navrhování konkrétních kroků pro implementaci v CleverMaps.
- **Závěr a shrnutí**
    - Souhrn klíčových nálezů, ověření, zda byly splněny cíle práce, a úvaha nad případnými budoucími směry výzkumu v této oblasti.

Harmonogram řešení práce:

Květen 2024 – Zpracování literární rešerše

Červen 2024 – Zpracování úvodních analýz

Září 2024 – Dokončení implementace metod

Říjen 2024 – Testování metod a porovnání výsledků

Listopad 2024 – Dokončení práce a sepsání doporučení pro společnost CleverMaps

Zdroje k části „Vymezení řešeného problému“:

CHRISTEN, Peter, 2012. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Berlin Heidelberg: Springer. Data-centric systems and applications. ISBN 978-3-642-43001-5.

CLEVERMAPS, 2024. Location Insights for Business. CleverMaps [online]. Dostupné z: https://www.clevermaps.io

DRAISBACH, Uwe a Felix NAUMANN, 2013. On Choosing Thresholds for Duplicate Detection [online]. Dostupné z: https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/2013/On_Choosing_Thresholds_for_Duplicate_Detection.pdf

NAUMAN, Felix a Melanie HERSCHEL, 2022. An Introduction to Duplicate Detection. B.m.: Springer Nature. ISBN 978-3-031-01835-0.

PAŠEK, Jan, Jakub SIDO, Miloslav KONOPÍK a Ondřej PRAŽÁK, 2022. MQDD -- Pre-training of Multimodal Question Duplicity Detection for Software Engineering Domain. In: [online]. Dostupné z: https://www.researchgate.net/publication/359518098_MQDD_--_Pre-training_of_Multimodal_Question_Duplicity_Detection_for_Software_Engineering_Domain#fullTextFileContent