Mendelova univerzita v Brně
Provozně ekonomická fakulta

---

# Detekce duplicit v geoprostorových datech

**Diplomová práce**

Vedoucí práce:
Ing. Pavel Turčínek, Ph.D.

Bc. Adam Prchal

Brno 2024

**Poděkování**

To-udělat: Poděkování

**Čestné prohlášení**

Prohlašuji, že jsem práci **Detekce duplicit v geoprostorových datech** vypracoval samostatně a veškeré použité prameny a informace uvádím v seznamu použité literatury. Souhlasím, aby moje práce byla zveřejněna v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách ve znění pozdějších předpisů a v souladu s platnou Směrnicí o zveřejňování závěrečných prací.

Jsem si vědom, že se na moji práci vztahuje zákon č. 121/2000 Sb., autorský zákon, a že Mendelova univerzita v Brně má právo na uzavření licenční smlouvy a užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

Dále se zavazuji, že před sepsáním licenční smlouvy o využití díla jinou osobou (subjektem) si vyžádám písemné stanovisko univerzity, že předmětná licenční smlouva není v rozporu s oprávněnými zájmy univerzity, a zavazuji se uhradit případný příspěvek na úhradu nákladů spojených se vznikem díla, a to až do jejich skutečné výše.

Brno 2024 .....

podpis

**Abstract**

To-udělat: EN abstract

**Key words:**

To-udělat: EN klíčová slova

**Abstrakt**

To-udělat: CZ abstract

**Klíčová slova:**

To-udělat: CZ klíčová slova

# Obsah

# 1 Úvod a cíl

## 1.1 Úvod

Proč by bylo potřeba z konkrétního dokumentu rozpoznávat, zda jsem či nejsem autorem textu? V mnoha případech může jít pouze o zvědavost, zda-li jsem schopný uspět v testu autorství u vlastního dokumentu. Jindy může jít o soutěž v tom, kdo dokáže nejlépe vyplnit daný test pro cizí práci a ukázat, že je schopný adaptovat se cizímu stylu psaní. Nicméně podstatnějším důvodem pro rozpoznávání autorství může být případ, kdy se snažíme obhájit práci, která byla označena jako *plagiát*.

Přesnou definici plagiátu lze nalézt v normě ČSN ISO 5127–2003, která říká: „Představení duševního díla jiného autora půjčeného nebo napodobeného v celku nebo z části, jako svého vlastního“. Plagiátorství jako činnost popisuje například Masarykova univerzita na svých internetových stránkách jako „úmyslné kopírování cizího textu a jeho vydávání za vlastní, nedbalé nebo nepřesné citování použité literatury, opomenutí citace (byť neúmyslné) některého využitého zdroje.“ (Masarykova univerzita, 2022). Rozpoznání autorství pomoci testu by mohlo sloužit jako nástroj k dokázání autorství v případech, kdy se nejedná o nedbalé, nepřesné či zapomenuté citování, ale v případě, že osoba vydává ukradený obsah úmyslně za svůj.

Další uplatnění rozpoznání autorství lze nalézt v případě, že si osoba nechala napsat práci na zakázku. Na trhu existuje několik firem, které nabízejí služby jako napsání seminárních, bakalářských a diplomových prací. Analýze tohoto trhu se ve své diplomové práci s názvem Analýza trhu s podvodnými seminárními a závěrečnými pracemi v ČR věnovala Ing. Veronika Králíková. Ta ve své práci nechala vypracovat od dvou různých společností esej a srovnávala kvalitu těchto služeb.

Kvalita výsledných prací se velice lišila. Jedna práce byla kvalitativně srovnatelná s prací, kterou běžně studenti odevzdávají, a u druhé nebylo dodrženo zadání práce. (Králíková, 2017)

Existují nástroje sloužící pro rozpoznání plagiátu, ale v případě prací na zakázku mají malou šanci na odhalení. Jak uvádí ve své práci Králíková, v ČR se pro detekci plagiátorství používají systémy jako například Theses.cz od Masarykovi univerzity, které srovnávají textový obsah a hledají shodu s ostatními pracemi (Králíková, 2017). Z toho vyplývá, že pokud je obsah práce na zakázku správně ocitovaný, nástroj práci nemusí označit jako plagiát. Pokud by však existovalo alespoň podezření na takové podvodné jednání, mohlo by rozpoznání autorství posloužit jako částečný důkaz při zkoumání, zda osoba práci napsala sama či ji napsal někdo jiný. (Albayrak, Aytekin, Kalaycı 2022)

Současná nabídka webových API[1] pokrývá velký rozsah služeb, které by mohl kdokoliv potřebovat k naplnění svého cíle. I přesto nelze pokrýt úplně všechny přípa-

[1] Application Programming Interface (rozhraní pro programování aplikací).

dy a je zapotřebí vytvořit si vlastní webové API, které bude řešit specifické problémy a nabízet službu, kterou zatím nikdo nenabízí.

# 2 Pojmy

To-udělat: Napsat

# 3 Metodika

To-udělat: Napsat

# 4 Analýza problému

To-udělat: Napsat

# 5 Diskuze

To-udělat: Napsat

# 6 Závěr

To-udělat: Napsat

# 7 Literatura

ALBAYRAK, Osman Semih, AYTEKIN, Tevfik a KALAYCI, Tolga Ahmet, 2022. Duplicate product record detection engine for e-commerce platforms. *Expert Systems with Applications*. Online. květen 2022. Vol. 193, p. 116420–116421. DOI 10.1016/j.eswa.2021.116420. [Accessed 28 duben 2024].

Having a clean product catalog and keeping it complying with the standards of the industry is one of the primary concerns of e-commerce companies. Integrating product data from multiple providers confronts the companies with a challenging issue: duplicate product records. Since it is possible to describe a product with a variety of different words, images and attributes, detecting duplicate product records is a difficult task to overcome. In this work, a novel duplicate record detection engine is proposed for an e-commerce company, Hepsiburada. The engine is developed based on a real-world dataset. In order to build a training set we use text similarity and domain-specific distance metrics for generating candidate duplicate product pairs which are then labeled by human experts. We performed extensive feature engineering and state-of-the-art classification models to determine whether any two products are duplicated or not. The experimental results show that our engine is able to detect duplicate product records with high precision and outperforms the accuracy of non-adaptive methodologies.