

POTENCIÁL WEBCRAPINGU PROSTOROVĚ LOKALIZOVANÝCH DAT

diplomová práce

Cíle práce

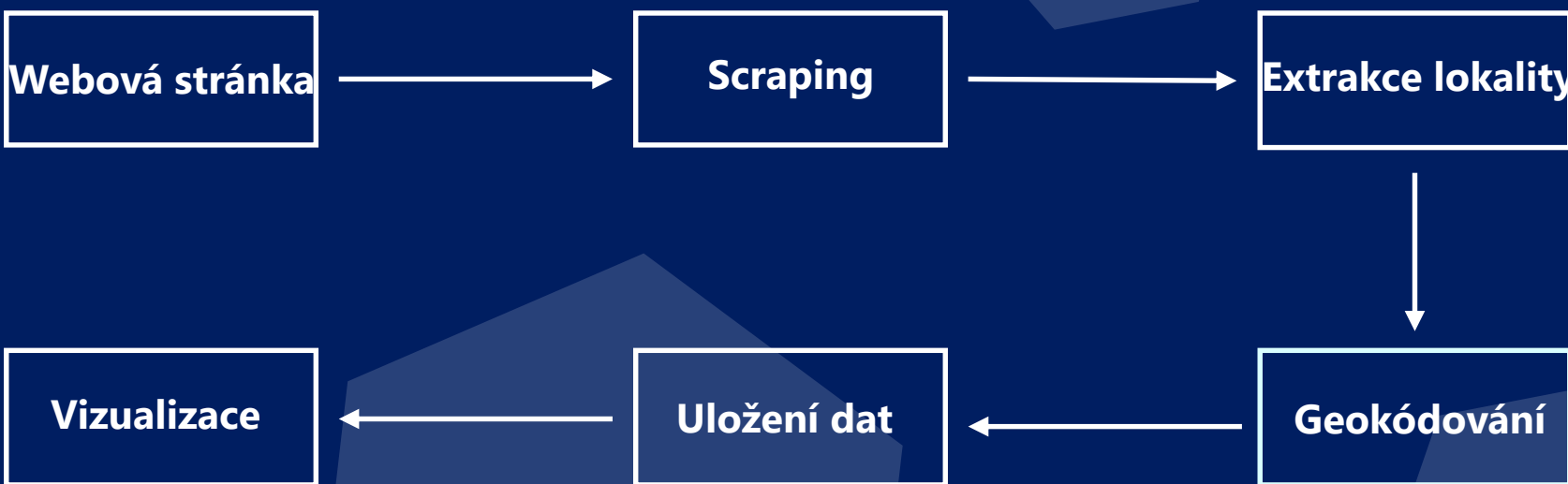
Cílem práce je analyzovat potenciál web scrapingu pro získávání prostorově lokalizovatelných dat a jejich následné využití. Zaměřuje se na identifikaci a analýzu vhodných webových zdrojů a návrh metod automatizovaného sběru dat, přičemž zohledňuje i alternativní přístupy. Součástí je také extrakce a interpretace geografické informace z různých typů dat, včetně zhodnocení právních a etických aspektů. Výsledky jsou ověřeny na praktických příkladech a vizualizovány ve webových mapách.

Použité metody

Zpracování dat je založeno především na web scrapingu, který je realizován různými způsoby podle typu zdroje – od přímého získávání HTML obsahu webových stránek až po využití XML sitemap pro efektivní výběr relevantních článků. V případě dynamického obsahu je využita simulace chování uživatele pro získání kompletního vykresleného obsahu stránky. Scraping je doplněn o alternativní přístupy, konkrétně získávání dat z API a RSS kanálů, které poskytují strukturovanější data.

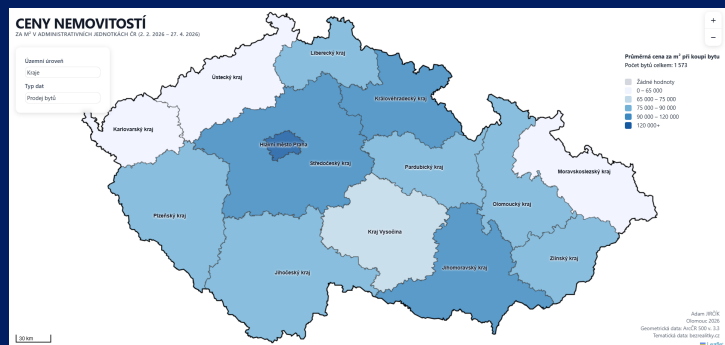
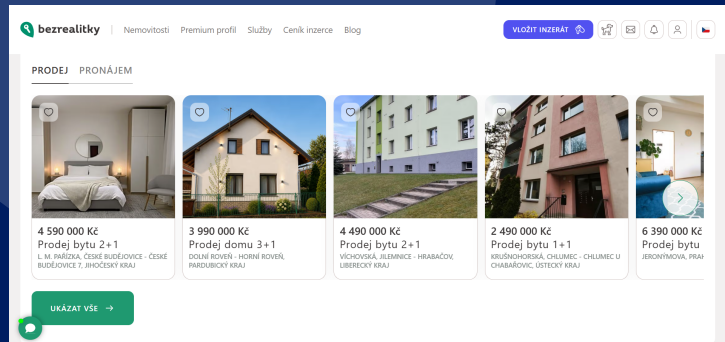
Z textových částí získaných dat jsou následně extrahovány geografické informace třemi hlavními způsoby – pomocí pravidlového přístupu založeného na vzorech v textu, metod zpracování přirozeného jazyka pro rozpoznání geografických entit a přístupů využívajících velké jazykové modely. Získané lokality jsou následně převedeny na geografické souřadnice prostřednictvím geokódování.

Výsledná data jsou ukládána do prostorové databáze PostgreSQL/PostGIS, GeoJSON. Pro publikaci dat je využit v jendom případě GeoServer. Vizualizace je realizována pomocí knihovny Leaflet, kde jsou data zobrazena v podobě interaktivních map.



Legislativní a etický aspekt

Web scraping se pohybuje v právně nejednoznačném prostředí, protože neexistuje samostatná legislativa, která by jej přímo upravovala. Jeho posouzení proto vychází z autorského práva, ochrany databází, ochrany osobních údajů (např. GDPR) a smluvních podmínek webových služeb. Zásadní roli hraje zejména to, zda dochází k neoprávněnému kopírování chráněného obsahu nebo systematickému využívání databází. Určitou výjimku představují vědecké a výzkumné účely, které mohou být za splnění podmínek legální. Vedle právních aspektů je důležitá i rovina etická – zejména ochrana soukromí, dopady na provozovatele webů a férové využití dat. Klíčové je proto vždy posuzovat konkrétní případ podle typu dat a účelu jejich využití. Obecně je scraping lépe obhajitelný při omezeném, nekomerčním a analytickém využití dat, zatímco rozsáhlé nebo komerční využití může představovat významné právní i etické riziko.



Bezrealitky.cz

První skript slouží k automatizovanému získávání dat o reálných inzerátech z webu bezrealitky.cz s možností parametrizace podle typu nemovitosti a typu transakce (prodej/pronájem). Na základě těchto vstupů je dynamicky sestavena URL adresa výpisu inzerátů, které jsou postupně procházeny po jednotlivých stránkách. Z HTML obsahu je extrahována strukturovaná JSON část obsahující detailní informace o inzerátech, jako je cena, plocha, adresa nebo ID. Skript zahrnuje mechanismy pro detekci duplicit, kontrolu neúplných dat a ukončení procházení při dosažení konce výpisu inzerátů.

Pro získání prostorové informace jsou primárně využívány souřadnice dostupné přímo ve zdrojových datech, v opačném případě je aplikováno geokódování adres pomocí služby Nominatim. Výsledná data jsou ukládána do databáze PostgreSQL/PostGIS, kde dochází k jejich aktualizaci pomocí mechanismu upsert a k prostorovému zpracování. Následně jsou vytvářeny agregované výstupy (ceny za m² podle územních jednotek), které jsou publikovány prostřednictvím GeoServeru jako WFS a vizualizovány pomocí webové mapy.

VYPRACOVANÉ SKRIPTY

Deutsche Welle (RSS)

Třetí skript využívá RSS kanál zpravodajského webu Deutsche Welle k extrakci článků a jejich prostorové lokalizaci pomocí metod zpracování přirozeného jazyka. Vstupem jsou položky RSS feedu, ze kterých jsou získány základní informace o článcích: Titulek, popis a datum publikace. Tyto textové části jsou následně analyzovány pomocí knihovny spaCy, konkrétně metodou rozpoznávání pojmenovaných entit (NER), která identifikuje geografické názvy v textu.

Protože NER model nerozlišuje význam jednotlivých lokalit v kontextu článku, je nad extrahovanými entitami aplikován systém pravidel pro výběr nejrelevantnější lokality. Ten zohledňuje například výskyt v titulku, délku názvu nebo četnost výskytu. Výsledkem je jedna reprezentativní lokalita, která nejlépe vystihuje geografický kontext článku.

Tato lokalita je následně převedena na geografické souřadnice pomocí geokódování, přičemž skript pracuje s více kandidáty a vybírá nejvhodnější výsledek na základě úrovně prostorové přesnosti. Výstupem je datová vrstva ve formátu GeoJSON, kde je každý článek reprezentován jedním bodovým prvkem s atributy obsahujícími základní informace o článku i v vybrané lokalitě.

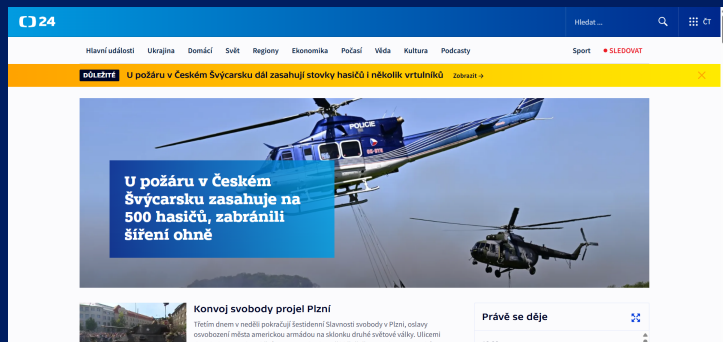


ČT 24

Skript pro portál ČT24 slouží k automatizovanému získávání zpravodajských článků, identifikaci hlavních geografických lokalit v jejich obsahu a převodu těchto informací na prostorová data. Na základě uživatelem zadaného časového rozsahu skript využívá XML sitemapu k získání seznamu článků, které následně filtruje podle data publikace. Jednotlivé články jsou načítány pomocí Selenium. Z textu článků jsou extrahovány základní atributy, jako je titulek, text, kategorie a datum publikace.

Klíčovou částí řešení je identifikace hlavních geografických lokalit pomocí velkých jazykových modelů, které z textu vybírají pouze relevantní místa, o nichž článek skutečně pojednává. Výstup modelu je následně normalizován, očištěn a převeden do strukturované podoby. Tyto lokality jsou dále geokódovány. Výsledná data jsou ukládána ve formátu GeoJSON.

Nad tímto skriptem je vytvořena jednoduchá webová aplikace, která umožňuje interaktivní vizualizaci výsledků. Backend zajišťuje spuštění celého procesu na základě uživatelského vstupu, zatímco frontend zobrazuje geokódované lokality v mapě s možností filtrování a detailního zobrazení informací o jednotlivých článcích.



Výsledky a přínos práce

Vyhodnocení ukazuje, že jednotlivé přístupy ke sběru a zpracování dat se výrazně liší kvalitou i náročností. Nejlepších výsledků dosahuje scraping reálního portálu, kde jsou data již strukturovaná a obsahují přímo souřadnice, což umožňuje přesnou lokalizaci. U zpravodajských dat je situace složitější – přístupy založené na AI modelech a NER dosahují podobné úspěšnosti přibližně 80–85 %, přičemž hlavním zdrojem chyb je samotné geokódování a nejednoznačnost lokalit. RSS feed představuje jednodušší zdroj dat, ale s omezenou dostupností. Nejjednodušší řešení nabízí API, bývá ale většinou výrazně omezené.

Přínosem práce je návrh a ověření obecného postupu převodu webových textových dat na prostorová data a jejich následné využití v geografické analýze. Bylo prokázáno, že web scraping a související metody mohou sloužit jako efektivní nástroj pro získávání aktuálních prostorových informací, a to i v případech, kdy data nejsou primárně určena pro geografické zpracování. Práce zároveň ukazuje silné a slabé stránky jednotlivých přístupů.