FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

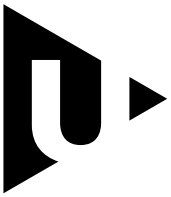
Bakalářská práce

Vytvoření Wordpress pluginu pro vyhledávání předků pro Czech-American TV

Jan Čácha

PLZEŇ

2025



FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

Bakalářská práce

Vytvoření Wordpress pluginu pro vyhledávání předků pro Czech-American TV

Jan Čácha

Vedoucí práce
Ing. Martin Dostal, Ph.D.

PLZEŇ 2025

Citace v seznamu literatury:
ČÁCHA, Jan. *Vytvoření Wordpress pluginu pro vyhledávání předků pro Czech-American TV*. Plzeň, 2025. Bakalářská práce. Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Katedra informatiky a výpočetní techniky. Vedoucí práce Ing. Martin Dostal, Ph.D.

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2024/2025

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Jan ČÁCHA**
Osobní číslo:	**A22B0019P**
Studijní program:	**B0613A140015 Informatika a výpočetní technika**
Specializace:	**Informatika**
Téma práce:	**Vytvoření Wordpress pluginu pro vyhledávání předků pro Czech-American TV**
Zadávající katedra:	**Katedra informatiky a výpočetní techniky**

Zásady pro vypracování

1. Prostudujte existující plugin s názvem Genealogy a související problematiku s vyhledáváním předků z jiného kontinentu.
2. Analyzujte návrhy a informace od zástupce Czech-American TV.
3. Navrhněte nový plugin a vyberte vhodné nástroje pro jeho realizaci.
4. Plugin implementujte. Při implementaci dbejte na možnosti budoucího rozšiřování.
5. Vytvořený plugin otestujte reprezentativní sadou testů.
6. Kriticky zhodnoťte vytvořené řešení včetně vyhodnocení názorů zástupců Czech-American TV na vytvořené dílo.

Rozsah bakalářské práce: **doporuč. 30 s. původního textu**
Rozsah grafických prací: **dle potřeby**
Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam doporučené literatury:

Dodá vedoucí bakalářské práce.

Vedoucí bakalářské práce: **Ing. Martin Dostal, Ph.D.**
Katedra informatiky a výpočetní techniky

Konzultant bakalářské práce: **Ing. Miroslav Krýsl**
Czech-American TV

Datum zadání bakalářské práce: **30. září 2024**
Termín odevzdání bakalářské práce: **5. května 2025**

L.S.

Doc. Ing. Miloš Železný, Ph.D.
děkan

Doc. Ing. Přemysl Brada, MSc., Ph.D.
vedoucí katedry

V Plzni dne 25. října 2024

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného akademického titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., zákon o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) v platném znění, a zejména skutečnost, že Západočeská univerzita v Plzni má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Plzni dne 1. ledna 2025

. . . . . .
Jan Čácha

V textu jsou použity názvy produktů, technologií, služeb, aplikací, společností apod., které mohou být ochrannými známkami nebo registrovanými ochrannými známkami příslušných vlastníků.

Abstrakt

Tato bakalářská práce se zaměřuje na vývoj WordPress pluginu, který má usnadnit vyhledávání předků pro diváky Czech-American TV. S rostoucím zájmem o genealogii mezi českou diasporou se potřeba efektivních nástrojů pro pátrání po předcích stala zásadní. Projekt začíná analýzou existujících řešení, následovanou shromážděním požadavků od zúčastněných stran, včetně zástupců Czech-American TV. Na základě této analýzy je navržen a implementován nový plugin, který zahrnuje uživatelsky přívětivé funkce pro efektivní vyhledávací možnosti. Plugin je důkladně testován, aby se zajistila jeho spolehlivost a jednoduchost použití. Tato práce přispívá k vylepšení online zážitku pro diváky, kteří se zajímají o zkoumání svého dědictví, a podporuje hlubší spojení s jejich kořeny.

Abstract

This bachelor's thesis focuses on the development of a WordPress plugin designed to facilitate ancestor searches for viewers of Czech-American TV. With the increasing interest in genealogy among the Czech diaspora, the need for effective tools to trace ancestry has become paramount. The project begins with an analysis of existing solutions, followed by gathering requirements from stakeholders, including representatives from Czech-American TV. Based on this analysis, a new plugin is proposed and implemented, incorporating user-friendly features for efficient search functionalities. The plugin is rigorously tested to ensure reliability and ease of use. This work contributes to enhancing the online experience for viewers interested in exploring their heritage, promoting a deeper connection with their roots.

Klíčová slova

Bakalářská práce • wordpress • plugin • genealogy

Poděkování

Rád bych vyjádřil upřímné poděkování svému vedoucímu práce, Ing. Martinu Dostalovi, Ph.D., za cenné rady, odborné vedení a trpělivost, kterou mi věnoval během celé práce. Jeho podpora a připomínky mi pomohly posunout projekt na vyšší úroveň.

Dále bych chtěl poděkovat zástupcům Czech-American TV za ochotu spolupracovat a poskytnout cenné podklady a zpětnou vazbu, která mi pomohla lépe pochopit potřeby uživatelů.

Velké díky patří také mé rodině a přátelům za trpělivost, podporu a motivaci během celého procesu psaní této bakalářské práce.

V neposlední řadě děkuji všem, kteří mi jakkoli pomohli, ať už radou, technickou pomocí nebo jen slovy povzbuzení.

Obsah

Úvod 1

Czech-American TV je nezisková organizace působící ve Spojených státech amerických, která se zaměřuje na podporu, uchovávání a propagaci české kultury, historie a tradic zejména mezi česko-americkou komunitou. Hlavním posláním organizace je budovat most mezi generacemi českých imigrantů a jejich potomky v USA a původními kořeny v České republice a bývalém Československu. Czech-American TV pravidelně tvoří a vysílá televizní pořady, které se věnují české kultuře, významným městům, historickým událostem, památkám a lidovým tradicím.

Vedle televizního vysílání poskytuje organizace také rozsáhlé online zdroje, které napomáhají uživatelům při hledání informací o jejich českých předcích. Mezi tyto zdroje patří genealogické databáze, digitalizované archivní dokumenty, informační weby, interaktivní mapy a vzdělávací programy. Tyto nástroje slouží nejen k objevování vlastních kořenů, ale i k prohlubování znalostí o české historii a tradicích.

V tomto kontextu vzniká tato bakalářská práce, jejímž cílem je vytvořit WordPress plugin, který bude součástí online platformy Czech-American TV a bude dále rozšiřovat možnosti genealogického vyhledávání. Plugin má za úkol zpřístupnit uživatelům různé nástroje sloužící k hledání předků, a to prostřednictvím intuitivního uživatelského rozhraní a propojení s existujícími databázemi a zdroji informací.

Motivace pro realizaci tohoto projektu vychází z rostoucího zájmu o genealogii v digitálním věku. V posledních letech roste počet uživatelů, kteří se zajímají o své předky, rodinnou historii a původ, a to jak v rámci akademického výzkumu, tak i z osobní potřeby. Digitalizace historických záznamů a rozvoj online nástrojů tento trend ještě více podporují.

Česko-americká komunita je přitom skupinou, která často čelí specifickým výzvám — především jazykovým bariérám a roztříštěnosti dostupných zdrojů. Proto je cílem této práce vytvořit centralizovaný, přístupný a uživatelsky přívětivý nástroj, který umožní snadnější vyhledávání informací, propojení s historickými materiály a hlubší porozumění kulturnímu dědictví.

Navrhovaný plugin bude obsahovat moderní prvky interakce, jako jsou dynamické vyhledávání, vizualizace dat nebo propojení s mapovými službami, které zpříjemní a zefektivní uživatelský zážitek. V práci bude kladen důraz jak na tech-

nickou implementaci, tak i na využitelnost výsledného řešení v reálném prostředí Czech-American TV.

Teoretická část 2

2.1 Genealogie a její význam v rámci Czech-American TV

Genealogie, tedy studium rodokmenů a rodinné historie, má hluboký význam nejen pro historiky, ale i pro jednotlivce, kteří hledají své kořeny a snaží se porozumět svému původu.

Na platformě Czech-American TV je patrný rostoucí zájem o propojení mezi českou a americkou historií. Tento projekt si klade za cíl vytvořit nástroj integrovaný do této platformy, který usnadní uživatelům vyhledávání jejich předků a pochopení jejich historických souvislostí. Interaktivní nástroj umožní filtrování dat, vizualizaci informací na mapě a možnost překladu mezi češtinou a angličtinou.

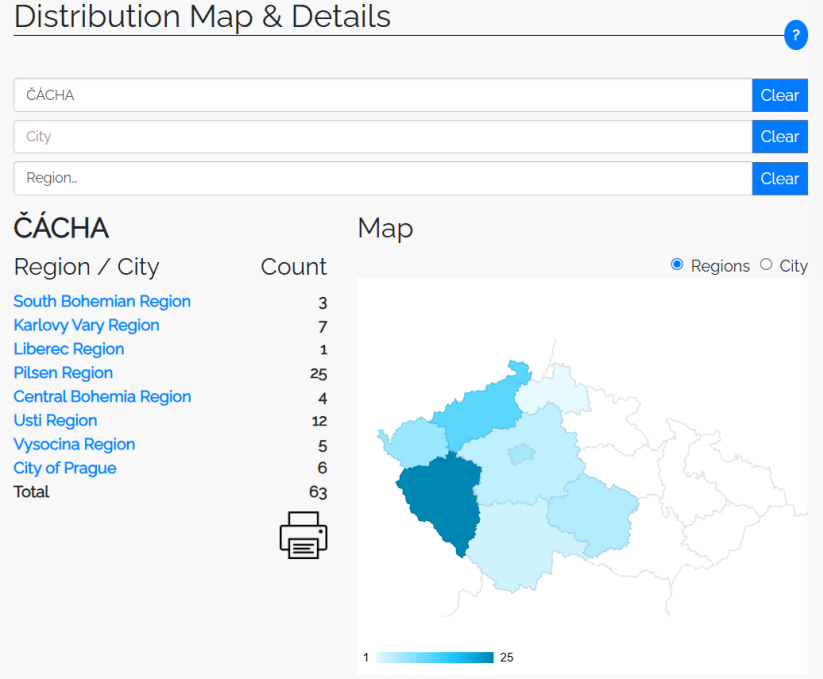
2.2 Genealogická mapa České republiky

Genealogická mapa Czech-American TV České republiky je nástroj umožňující vizualizaci četnosti příjmení v jednotlivých regionech. Tato funkce je užitečná zejména pro zájemce o rodokmeny, kteří chtějí zjistit historické rozložení svého příjmení a případné vazby na určité oblasti.

Princip fungování mapy spočívá v zadání konkrétního příjmení, na základě kterého je provedeno vyhledání jeho výskytu v dostupných databázích. Výsledky jsou poté zobrazovány na interaktivní mapě, kde intenzita zbarvení jednotlivých regionů odpovídá četnosti výskytu daného příjmení.

Zdrojová data pro tuto funkcionalitu pocházejí z veřejně dostupných databází obyvatelstva, historických sčítání lidu a dalších genealogických zdrojů. Výzvou při tvorbě této mapy bylo nejen správné zpracování dat, ale také efektivní způsob jejich vizualizace tak, aby bylo možné výsledky snadno interpretovat.

Díky tomuto nástroji mohou uživatelé získat cenné informace o původu svého příjmení a jeho historickém rozšíření v rámci České republiky.



Obrázek 2.1: Princip fungování mapy

2.3 Německo-české názvosloví historických měst

Historicky měla řada českých měst a obcí své ekvivalenty v německém jazyce, což bylo dáno dlouhodobým soužitím českého a německého obyvatelstva na našem území. Tento dvojjazyčný charakter byl zvláště patrný v pohraničních oblastech a v obdobích, kdy české země byly součástí Rakouska-Uherska.

Mnoho historických dokumentů, map a genealogických záznamů obsahuje názvy měst v jejich německé podobě, což může být matoucí pro dnešní uživatele, kteří hledají české ekvivalenty. Proto vznikla databáze německých a českých názvů, která umožňuje rychlé a snadné přiřazení odpovídajících míst.

Tento nástroj pracuje tak, že uživatel zadá historický německý název města, načež systém vyhledá a zobrazí jeho odpovídající český ekvivalent. Výsledky jsou vizualizovány na mapě, což usnadňuje orientaci v historických geografických údajích.

Tento projekt je cenným přínosem nejen pro genealogické bádání, ale i pro historiky a všechny zájemce o kulturní dědictví českých zemí. [5, 6]

2.4 Machine Learning v genealogii

Strojové učení představuje moderní technologii, která má v genealogii významný potenciál. Umožňuje analyzovat rozsáhlé soubory historických dat, identifikovat vzory a navrhovat rodinné vztahy na základě pravděpodobnosti. Implementace technik strojového učení v rámci genealogického výzkumu může výrazně zlepšit přesnost a rychlost vyhledávání informací, zejména v případech, kdy se jedná o historicky změněná jména, neúplné záznamy nebo geograficky vzdálené zdroje dat.

Cílem je vytvořit nástroj, který bude využívat strojové učení k posílení genealogického výzkumu. Součástí tohoto projektu je implementace algoritmu Word2Vec, který bude využit při překladech z češtiny, němčiny a latiny do angličtiny. Word2Vec zde hraje klíčovou roli při rozpoznávání významově příbuzných slov a rozšiřování překladů o relevantní varianty. To zlepší přesnost a srozumitelnost překladů historických dokumentů a umožní uživatelům získat lepší výsledky při vyhledávání genealogických informací.

2.4.1 Word2Vec a jeho využití při překladu

Algoritmus Word2Vec je jednou z klíčových metod pro zpracování textových dat v oblasti strojového učení. Tento algoritmus převádí slova do matematických vektorů v mnoharozměrném prostoru, což umožňuje měřit podobnosti mezi nimi pomocí geometrických operací. Model byl poprvé představen v roce 2013 vědci z Google Brain pod vedením Tomáše Mikolova [1] a od té doby se stal základním nástrojem v oblasti zpracování přirozeného jazyka (NLP).

Word2Vec využívá dva hlavní modely:

- **Skip-Gram**: Tento model se snaží předpovědět okolní slova na základě aktuálního slova. Je vhodnější pro malé datové sady, protože je schopný efektivně generalizovat.
- **CBOW (Continuous Bag of Words)**: Opačný přístup, který se snaží předpovědět aktuální slovo na základě okolních slov. Tento model je často rychlejší než Skip-Gram a lépe funguje s velkými objemy dat.

Použití Word2Vec v procesu překladu přináší několik výhod:

- **Rozšíření překladů o příbuzná slova**: Pokud daný slovník přeloží konkrétní slovo nebo frázi, Word2Vec umožní rozšíření překladu o slova s podobným významem, čímž se zlepší přesnost a přirozenost překladu.
- **Překonání omezení běžných slovníků**: Historické texty mohou obsahovat méně běžné nebo archaické výrazy, které se nemusí nacházet v dostupných

slovnících. Word2Vec pomůže najít jejich modernější nebo příbuzné ekvivalenty.

- **Podpora kontextového překladu**: Word2Vec umožňuje zohlednit kontext, ve kterém se slovo vyskytuje, což je užitečné při překladu víceznačných termínů.

2.4.2 Matematický model Word2Vec

Algoritmus Word2Vec převádí slova na vektory $\vec{w}$ v n-rozměrném prostoru. Pro optimalizaci těchto vektorů využívá neuronovou síť s jednou skrytou vrstvou, která se učí minimalizovat chybu predikce. Model lze formalizovat následovně:

2.4.2.1 Skip-Gram

Skip-Gram model se snaží předpovědět okolní slova w_{t+j} daného středového slova w_t. Pravděpodobnost výskytu slova w_{t+j} v kontextu w_t je definována jako:

$$P(w_{t+j}|w_t) = \frac{e^{v'_{w_{t+j}} \cdot v_{w_t}}}{\sum_{w \in V} e^{v'_w \cdot v_{w_t}}} \tag{2.1}$$

kde v_w a v'_w jsou vektory slov v různých vrstvách modelu a V je velikost slovníku. Tento výraz odpovídá softmax funkci aplikované na skalární součin vektorů slov.

2.4.2.2 CBOW (Continuous Bag of Words)

CBOW model funguje opačně než Skip-Gram – snaží se předpovědět aktuální slovo w_t na základě jeho kontextu C (okolních slov). Formálně je pravděpodobnost $P(w_t|C)$ definována jako:

$$P(w_t|C) = \frac{e^{v_{w_t} \cdot \sum_{w_c \in C} v'_{w_c}}}{\sum_{w \in V} e^{v_w \cdot \sum_{w_c \in C} v'_{w_c}}} \tag{2.2}$$

Tento vzorec znamená, že pro určení cílového slova w_t se sčítají vektory jeho okolních slov w_c a výsledek se transformuje pomocí softmax funkce.

Vzhledem k výpočetní náročnosti softmaxu se často používají optimalizační techniky jako:

- **Negativní vzorkování (Negative Sampling)**: Místo výpočtu pravděpodobnosti pro všechna slova se trénuje model na správná slova a několik negativních příkladů.
- **Hierarchický softmax**: Nahrazuje klasický softmax binárním stromem, což umožňuje efektivnější výpočty.

Kromě učení vektorů slov se měří jejich podobnost. Pro měření podobnosti mezi dvěma slovy w_1 a w_2 se nejčastěji používá kosinová podobnost:

$$\text{cosine\_similarity}(\vec{w}_1, \vec{w}_2) = \frac{\vec{w}_1 \cdot \vec{w}_2}{\|\vec{w}_1\| \cdot \|\vec{w}_2\|} \tag{2.3}$$

kde $\vec{w}_1 \cdot \vec{w}_2$ je skalární součin a $\|\vec{w}_1\|, \|\vec{w}_2\|$ jsou velikosti vektorů. Kosinová podobnost nabývá hodnot od -1 do 1, přičemž hodnoty blízké 1 znamenají vysokou podobnost mezi slovy. [4, 3, 2, 13]

2.5 Rešerše existujících řešení pro genealogický výzkum

2.5.1 Přehled dostupných genealogických platforem

V současné době existuje několik významných platforem, které se zaměřují na genealogický výzkum. Mezi nejznámější patří:

- **Ancestry.com** – Jedna z největších genealogických platforem na světě. Nabízí rozsáhlé databáze historických záznamů, DNA testy a nástroje pro vytváření rodokmenů. Výhodou je obrovské množství dat, nevýhodou vysoká cena předplatného.
- **MyHeritage** – Platforma zaměřená na propojení rodin po celém světě. Nabízí DNA testy a nástroje pro vizualizaci rodokmenů. Výhodou je uživatelsky přívětivé rozhraní, ale některé funkce jsou dostupné pouze v placené verzi.
- **FamilySearch** – Bezplatná platforma provozovaná Církví Ježíše Krista Svatých posledních dnů. Obsahuje velké množství digitálních záznamů, vhodná pro začátečníky i pokročilé genealogy. Nevýhodou je menší podpora pro neanglicky mluvící uživatele.
- **Geni.com** – Platforma zaměřená na spolupráci mezi uživateli při vytváření rodokmenů. Výhodou je sociální aspekt, nevýhodou menší podpora pro samostatný výzkum.

2.5.2 Srovnání s projektem

Projekt se od těchto platforem liší v několika klíčových aspektech:

- **Lokalizace** – Zaměřuje na propojení české a americké historie, což umožňuje hlubší analýzu lokálních dat.

- **Jazyková podpora** – Překlady mezi češtinou, němčinou, latinou a angličtinou poskytují výhodu při práci s historickými dokumenty.
- **Interaktivní nástroje** – Genealogická mapa a vizualizace dat usnadňují interpretaci výsledků.

2.6 Analýza potřeb uživatelů

Cílovou skupinou projektu jsou genealogičtí nadšenci, historici a rodinní badatelé, kteří se snaží propojit historická data se svými rodokmeny. Klíčovými potřebami těchto uživatelů jsou překlady historických dokumentů, vizuální reprezentace dat, efektivní vyhledávání v databázích a podpora pro neanglicky mluvící uživatele. Projekt tyto potřeby řeší prostřednictvím jazykových nástrojů, genealogických map a využití strojového učení pro vyhledávání v historických datech.

2.7 Analýza technologických možností

2.7.1 Výběr Word2Vec a jeho alternativy

Word2Vec byl vybrán díky své efektivitě při práci s historickými texty, schopnosti rozpoznávat podobnosti mezi slovy a relativně nízké výpočetní náročnosti. Alternativy zahrnují modely jako BERT, GPT, FastText a TF-IDF, z nichž každý má své výhody a nevýhody.

- **BERT** je založen na transformerové architektuře a umožňuje lepší pochopení kontextu, což lze vyjádřit vzorcem:

$$h_t = \text{Transformer}(X) \tag{2.4}$$

  kde h_t je kontextový vektor slova X, generovaný hlubokými vrstvami modelu. Nevýhodou BERTu je vysoká výpočetní náročnost.

- **GPT** pracuje na principu autoregresivního generování textu, kde pravděpodobnost dalšího slova vychází z Bayesova pravidla:

$$P(w_t|w_{1:t-1}) = \frac{P(w_{1:t})}{P(w_{1:t-1})} \tag{2.5}$$

  Tento přístup je vhodný pro generování textu, ale méně efektivní při hledání podobností mezi historickými výrazy.

- **FastText** rozšiřuje Word2Vec tím, že rozkládá slova na n-gramy, což lze matematicky vyjádřit jako:

$$v(w) = \sum_{i=1}^{n} v(g_i) \tag{2.6}$$

kde g_i jsou jednotlivé podslovy slova w. Tento model se lépe hodí pro češtinu díky práci s ohýbanými tvary slov, ale složitější implementace brání jeho využití v projektu.

- **TF-IDF** je jednodušší metoda pro zpracování textu, která se řídí vzorcem:

$$\text{TF-IDF}(w) = \text{TF}(w) \times \log \frac{N}{\text{DF}(w)} \tag{2.7}$$

kde N je počet dokumentů a $\text{DF}(w)$ je počet dokumentů obsahujících slovo w. Tento přístup je rychlý, ale nezachycuje význam slov v kontextu.

Vzhledem k tomu, že Word2Vec dokáže efektivně pracovat s historickými dokumenty a má nízkou výpočetní náročnost, byl zvolen jako nejvhodnější řešení.

2.8 Analýza datových zdrojů

Projekt využívá data z historických sčítání lidu, digitálních archivů a veřejných genealogických databází. Kvalita těchto dat je ovlivněna neúplností historických záznamů a regionálními rozdíly v dostupnosti informací. Mezi hlavní výzvy patří normalizace různých formátů, jazykové bariéry u německých a latinských dokumentů a řešení chybějících záznamů, které mohou komplikovat rekonstrukci rodokmenů. [20, 19, 18]

2.9 Analýza rizik a omezení

Technická rizika zahrnují chyby v překladech způsobené omezeními strojového učení a omezenou dostupností historických záznamů. Etická problematika zahrnuje ochranu osobních údajů a citlivost některých historických záznamů. Výsledky projektu jsou také omezeny kvalitou dat a schopností Word2Vec modelu zachytit složitější jazykové kontexty.

Analytická část 3

Na základě teoretické analýzy byl navržen plugin, který kombinuje výhody WordPressu pro správu obsahu s pokročilými funkcemi strojového učení a geografické vizualizace. Tato kapitola detailně popisuje, jak byly identifikované požadavky převedeny do funkčního řešení. Zaměřuje se na tři klíčové oblasti:

- Architekturu pluginu a integraci s WordPress API
- Zpracování genealogických dat v Pythonu
- Implementaci překladačů a vizualizací

3.1 Analýza požadavků

Tato kapitola se zaměřuje na podrobnou analýzu požadavků vyplývajících z potřeb neziskové organizace Czech-American TV, která si klade za cíl zprostředkovat genealogické nástroje členům česko-americké komunity. Výsledkem této analýzy je specifikace funkcionalit, které by měl výsledný WordPress plugin obsahovat, aby naplnil očekávání zadavatele i cílové skupiny uživatelů.

3.1.1 Současný stav a motivace ke změně

Organizace Czech-American TV již v současnosti nabízí na svém webu několik online nástrojů zaměřených na podporu genealogického výzkumu a uchovávání českého kulturního dědictví. Mezi dostupné funkcionality patří například interaktivní genealogická mapa, nástroj pro překlad českých příjmení, informační stránky o změnách jmen po emigraci a instruktážní videa.

Interaktivní mapa využívá geolokační data k vizualizaci výskytu příjmení v jednotlivých oblastech České republiky, což může uživatelům pomoci s lokalizací jejich rodových kořenů. Další funkcí je nástroj pro identifikaci změn v příjmeních, které byly často po příchodu do Spojených států amerických foneticky přizpůsobeny nebo

zjednodušeny. Czech-American TV také poskytuje videonávody, které uživatele provádějí základními kroky genealogického výzkumu a využíváním dostupných online nástrojů.

Přestože tyto nástroje tvoří solidní základ, některé aspekty systému by mohly být dále rozvinuty. Například oblast překladů genealogických termínů je omezena pouze na příjmení a chybí podpora běžně používaných latinských nebo německých výrazů z historických matrik. Dále chybí možnost sémantického vyhledávání, které by umožnilo uživatelům najít významově podobné výrazy nebo archaické tvary. Také správa obsahu není plně integrovaná do administračního rozhraní WordPressu, což omezuje možnosti rozšíření a editace dat neprogramátory.

Na základě zpětné vazby od uživatelů a identifikovaných mezer vznikla potřeba doplnit existující řešení o specializovaný plugin pro WordPress, který rozšíří současné funkce o další překladové databáze, interaktivní prvky, moderní vyhledávací algoritmy (např. Word2Vec) a nástroje pro správu obsahu dostupné přímo z administrace webu. Cílem je vytvořit nástroj, který bude přívětivý i pro uživatele bez pokročilých technických znalostí a zároveň nabídne kvalitní podporu pro genealogické bádání.

Zadavatel na základě zkušeností s provozem stávající genealogické sekce a zpětné vazby od uživatelů identifikoval několik oblastí, kde je prostor pro zlepšení a rozšíření funkcionality:

- Omezené interaktivní prvky – současná genealogická mapa sice zobrazuje výskyt příjmení a některá místa původu, nicméně chybí například propojení s historickými názvy měst, nebo možnost rozšířeného vyhledávání podle různých jazykových podob jmen (např. česká, německá, latinská verze).
- Nedostatečné pokrytí překladů genealogických termínů – aktuálně dostupné nástroje se zaměřují převážně na příjmení. V historických dokumentech však často figurují i latinské nebo německé výrazy (např. „uxor“, „defunctus“, „geboren“), které mohou být pro běžného uživatele nesrozumitelné.
- Omezené možnosti správy a úprav dat – stávající systém neumožňuje snadnou aktualizaci překladových databází nebo přidávání nových položek bez přímého zásahu do kódu, což komplikuje udržitelnost projektu z pohledu běžného redaktora.
- Absence sémantického vyhledávání – současný systém nenabízí funkci, která by uživateli pomohla nalézt významově podobné výrazy, archaické formy slov nebo termíny s podobným významem.

Z těchto důvodů vznikl návrh na vývoj specializovaného pluginu pro redakční systém WordPress, který rozšíří stávající možnosti Czech-American TV o nové

interaktivní nástroje pro genealogické vyhledávání, správu překladových databází a sémantické vyhledávání. Plugin bude navržen tak, aby byl snadno použitelný jak pro koncové uživatele, tak pro administrátory bez nutnosti technických znalostí.

3.1.2 Cílová skupina a specifika

Uživatelskou základnu tvoří především američtí potomci českých imigrantů, kteří často neumí česky, případně mají omezené historické znalosti o českých reáliích. Z tohoto důvodu je nutné, aby byl systém jazykově přívětivý a intuitivní. Důraz musí být kladen na přehlednost a snadnou použitelnost, přičemž plugin by měl fungovat jako prostředník mezi uživatelem a komplexními datovými zdroji.

3.1.3 Požadavky na funkčnost pluginu

Na základě požadavků zadavatele a rešerše existujících řešení byly definovány následující klíčové funkcionality:

1. Interaktivní mapa historických českých měst – Pomocí knihovny Leaflet.js a dat z OpenStreetMap bude zobrazena mapa s lokalitami, které mají genealogický význam. Uživatel bude moci kliknout na jednotlivá města a zobrazit historické informace, případně přejít na relevantní záznamy.

2. Překlad genealogických výrazů – Modul umožní překlad klíčových pojmů mezi angličtinou, češtinou a latinou. Překlady budou čerpány z databáze vytvořené na základě historických pramenů (např. matrik), s možností rozšíření.

3. Integrace modelu Word2Vec – Pro účely sémantického vyhledávání bude integrován model Word2Vec, který navrhuje významově podobná slova. To umožní např. vyhledání záznamu i v případě, že je výraz napsán v archaické či variantní podobě.

4. Administrace překladů – Plugin umožní spravovat databázi překladů přímo z administračního rozhraní WordPressu. Uživatelé s příslušnými oprávněními budou moci přidávat, upravovat či mazat výrazy bez nutnosti zásahu do zdrojového kódu.

3.1.4 Technické požadavky a omezení

Plugin bude implementován v prostředí WordPress 6.x a musí být kompatibilní s nejčastěji používanými šablonami a verzemi PHP (minimálně 7.4). Dále bude využívat následující technologie:

- PHP pro logiku pluginu a práci s databází.

- JavaScript (včetně Leaflet.js) pro práci s mapou a interaktivními prvky.
- REST API pro komunikaci s backendem a případnou integraci se vzdálenými službami (např. Word2Vec API).
- MySQL databáze pro ukládání překladů a souvisejících dat.

3.1.5 Možnosti realizace a návrhové rozhodnutí

Z hlediska realizace bylo posouzeno několik přístupů. Jednou z alternativ byla tvorba samostatné webové aplikace mimo WordPress, což by však znamenalo vyšší náklady na správu a nižší integraci s existujícím ekosystémem Czech-American TV. Na základě konzultací se zadavatelem bylo rozhodnuto o tvorbě pluginu přímo pro WordPress, neboť tato platforma již spravuje hlavní web a nabízí dostatečnou flexibilitu a rozšiřitelnost.

Podrobnosti o konkrétním návrhu architektury a implementačních rozhodnutích jsou uvedeny v následující kapitole *Návrh řešení*.

3.2 Návrh řešení

Na základě identifikovaných potřeb a slabých míst v existujícím řešení (viz kapitola Analýza požadavků) bylo navrženo vytvoření sady nástrojů, které budou implementovány formou několika samostatných, ale vzájemně propojitelných WordPress pluginů. Tyto pluginy budou sloužit nejen ke zpřístupnění překladových databází, ale také k vizualizaci geografických údajů, návrhu významově podobných slov a efektivní správě obsahu z uživatelského prostředí WordPressu.

Cílem návrhu je rozdělit funkcionalitu do samostatných modulů tak, aby bylo možné jednotlivé části udržovat a rozvíjet nezávisle, a zároveň zajistit maximální přehlednost a flexibilitu při nasazení na web Czech-American TV.

Přehled navržených pluginů

Navržené řešení zahrnuje následující komponenty:

- **Překladový plugin (Translation Plugin)** – nástroj umožňující překlad výrazů mezi několika jazykovými kombinacemi, konkrétně:
  - Čeština → Angličtina
  - Němčina → Angličtina
  - Latina → Čeština → Angličtina

Plugin kombinuje databázový překlad s online službou MyMemory API a umožňuje rozšiřování databáze přímo z administračního rozhraní.

- **Mapový plugin – historická města (German City Names Map)** – plugin zobrazující interaktivní mapu s historickými názvy měst (německé a české verze), využívající knihovnu Leaflet.js místo Google Maps. Mapa umožňuje kliknutí na město a zobrazení jeho historických názvů nebo přesměrování na externí mapové služby.
- **Genealogická mapa (Genealogy Map Plugin)** – přepracovaná verze stávající genealogické mapy Czech-American TV, která byla původně postavena na Google Maps. Nově využívá knihovnu Leaflet.js, zobrazuje lokalizovaná příjmení nebo původní rodová místa a slouží jako interaktivní nástroj pro hledání předků.
- **Sémantický modul s Word2Vec (Word2Vec REST API)** – externí serverová služba běžící na FastAPI, která načítá předtrénovaný model Word2Vec (Google News) a poskytuje návrhy významově podobných výrazů. WordPress pluginy s tímto API komunikují přes AJAX a mohou výsledky zobrazovat nad přeloženými výrazy.

Takto navržená architektura zajišťuje modularitu, přehlednost a snadné rozšiřování celého systému v budoucnosti. Následující podkapitoly se věnují jednotlivým komponentám detailněji: zpracování jazykových dat, výběru překladových služeb, nasazení Word2Vec a vizualizaci dat v mapových komponentách.

Při návrhu řešení byl kladen důraz na propojení uživatelsky přívětivého rozhraní (WordPress plugin) s výkonnými backendovými službami.

V první fázi vývoje bylo nutné získat relevantní jazyková data. Firma Czech-American TV poskytla dva dokumenty obsahující:

- Česko-anglické překlady ve formátu RTF (cca 5000 kB)
- Latinsko-české překlady ve formátu TXT (cca 500 kB)

Oba soubory bylo nutné zpracovat – odstranit formátovací znaky, očistit data od nadbytečných mezer a nesrovnalostí a následně převést do jednotného formátu vhodného pro databázový import. K tomuto účelu byly vytvořeny Python skripty, které data připravily ve formě CSV a SQL.

3.2.1 Výběr překládacích služeb

Pro zajištění dynamických překladů byl zvažován způsob, jak kombinovat lokální databázové vyhledávání s online překladatelskými službami. Po testování různých

možností padla volba na **MyMemory API**, které poskytuje překlady zdarma a zároveň umožňuje ukládat získané výsledky do vlastní databáze pro další offline použití.

Při překladu probíhá nejprve vyhledání v lokální databázi. Pokud se překlad nenajde, systém odešle požadavek na MyMemory API a výsledek uloží do databáze pro budoucí využití. Tímto způsobem dochází ke kombinaci online i offline překladu.

3.2.2 Integrace Word2Vec

Jednou z největších výzev byl výběr a nasazení vhodného Word2Vec modelu. Byla provedena rozsáhlá rešerše veřejně dostupných modelů a existujících API služeb. Mnoho z těchto řešení však neumožňovalo jednoduchou integraci do prostředí WordPress pluginu, nebo nebyla volně dostupná pro samostatné nasazení.

Po testování několika knihoven a modelů (např. Gensim, Flair, Spacy) bylo zohledněno zejména:

- zda je knihovna open-source a bez licenčního omezení,
- jak snadno lze model načíst a využít v samostatném serveru,
- dostupnost předtrénovaných modelů (např. Google News, Glove),
- podpora hledání významově podobných slov (metoda `most_similar` apod.).

Na základě těchto kritérií bylo rozhodnuto vytvořit vlastní **API server pomocí frameworku FastAPI**, který načítá předtrénovaný Word2Vec model (Google News) a poskytuje významově podobné výrazy pomocí jednoduchého REST API. Tato služba běží nezávisle na WordPressu a plugin s ní komunikuje přes AJAX požadavky. [15, 12]

3.2.3 Zpracování geografických dat

Pro vizualizaci historických názvů měst v rámci České republiky byla použita knihovna Leaflet.js, která načítá geolokační data z databáze a zobrazuje je v interaktivní mapě. Po kliknutí na konkrétní město se zobrazí jeho historické názvy, případně se otevře odkaz na Google Maps.

Data byla importována z otevřených datasetů a dále doplněna o překlady názvů měst do němčiny, češtiny a angličtiny, pokud byly k dispozici.

3.3 Technická specifikace systému

Celý systém je navržen jako modulární, přičemž jednotlivé komponenty komunikují pomocí jasně definovaných rozhraní (API). Klíčovým prvkem je oddělení fronten-

dové části (plugin ve WordPressu) od backendových služeb (překladače, Word2Vec server).

- WordPress plugin zajišťuje správu obsahu, práci s uživatelem a integraci do administrace.
- Překladatelská vrstva kombinuje lokální databázi a MyMemory API. Překlady jsou cachovány pro optimalizaci výkonu.
- Samostatný FastAPI server hostuje předtrénovaný Word2Vec model. Komunikace probíhá přes REST API s výstupem ve formátu JSON.

V rámci návrhu byly definovány vstupní a výstupní formáty každé komponenty, včetně validace vstupů, autentifikace API dotazů a struktury odpovědí.

3.4 Výkonové a bezpečnostní hledisko

Při návrhu řešení byl kladen důraz na výkonnost systému a bezpečnost uživatelských dat. Byla přijata následující opatření:

- **Cachování překladů** pomocí databáze – eliminuje nadbytečné API požadavky a zrychluje odezvu systému.
- **Omezení API dotazů** (rate limiting) – na úrovni FastAPI serveru, aby se předešlo zneužití služby.
- **Validace vstupních dat** – kontrola dotazovaných slov, jazykových kódů a velikosti vstupních řetězců.
- **Bezpečná komunikace** – přenos dat mezi pluginem a backendem probíhá přes HTTPS.

Zvláštní pozornost byla věnována odolnosti systému proti výpadkům API nebo špatnému připojení – v takových případech systém pracuje výhradně s lokálními daty.

3.5 Uživatelské scénáře a příklady použití

Na základě požadavků zadavatele byly identifikovány hlavní uživatelské scénáře:

- **Vyhledávání genealogického termínu** – uživatel zadá slovo v jednom jazyce, systém zobrazí překlady a významově podobná slova.

- **Zobrazení historických názvů města** – uživatel klikne na mapu, kde se mu zobrazí informace včetně názvů ve více jazycích.
- **Správa překladů administrátorem** – správce může přidávat nové překlady, upravovat stávající nebo kontrolovat automaticky přeložené termíny.

3.6 Rozšiřitelnost a budoucí možnosti integrace

Řešení bylo navrženo s ohledem na budoucí rozšiřitelnost. Mezi plánované nebo potenciální funkce patří:

- **Integrace dalších jazyků** – např. němčina, polština nebo maďarština, které jsou v kontextu středoevropské genealogie relevantní.
- **Podpora více překladatelských API** – jako DeepL nebo Microsoft Translator s vyšší přesností překladu.
- **Automatická extrakce historických názvů z dokumentů** – pomocí NLP technik.
- **Geolokační filtrování** – zobrazování měst v závislosti na období nebo jazykovém vlivu (např. německé názvy v Sudetech).

3.7 Shrnutí analytické části

Na základě dostupných dat a požadavků zadavatele bylo navrženo řešení, které kombinuje:

- databázově uložené slovníky pro rychlé překlady,
- online překlady pomocí MyMemory API jako záložní řešení,
- Word2Vec model běžící na vlastním serveru pro sémantické doplnění překladů,
- mapovou vizualizaci založenou na Leaflet.js a databázi měst a regionů,
- bezpečnostní prvky a validaci vstupů,
- připravenost na budoucí integrace a rozšíření funkcionality.

Tento hybridní přístup umožňuje efektivní a uživatelsky přívětivé vyhledávání genealogických informací napříč různými jazyky a historickými daty.

Implementační část 4

Tato kapitola popisuje praktickou realizaci celého řešení, od výběru technologie až po konkrétní implementaci klíčových funkcionalit. Cílem je přiblížit způsob, jakým byly jednotlivé komponenty systému navrženy a propojeny za účelem dosažení požadované funkcionality pluginu.

V rámci implementace byly řešeny následující úlohy:

- návrh a vytvoření databázové struktury pro efektivní ukládání jazykových dat,
- parsování a zpracování externích dokumentů s překlady do podoby vhodné pro import,
- vývoj serverové části s využitím frameworku FastAPI,
- vytvoření WordPress pluginu pro interakci s uživatelem,
- integrace externích jazykových služeb jako je MyMemory API a Word2Vec,
- implementace funkcí pro vyhledávání, překlad a doplňování výrazů (autocomplete).

Kapitolou provází postupně jednotlivé fáze zpracování dat, návrhu architektury a technické realizace. Popis se zaměřuje jak na konkrétní implementační detaily, tak na motivaci a odůvodnění jednotlivých kroků.

4.1 Obecný popis a architektura

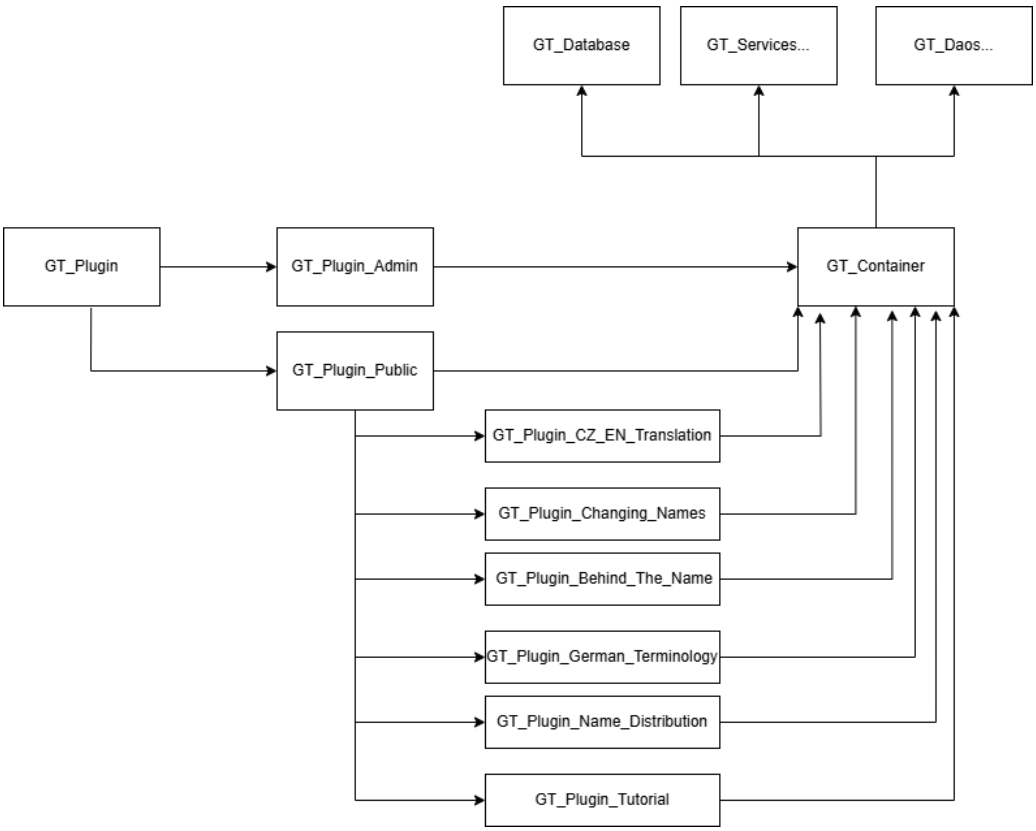
Plugin je postaven na kombinaci PHP, MySQL a JavaScriptu. Komunikace probíhá asynchronně pomocí AJAXu, což umožňuje rychlou odezvu uživatelského rozhraní. Databázová vrstva je navržena tak, aby umožňovala efektivní indexaci slovníku a zároveň podporovala rozšíření o nové překladové modely.

Součástí implementace je také funkce automatického doplňování (autocomplete), která naslouchá vstupu uživatele a dynamicky doplňuje možnosti překladu přímo

z databáze. Tato funkce využívá jQuery UI Autocomplete a AJAX požadavky na backend, kde se provádí vyhledávání relevantních výsledků.

Při zadání alespoň dvou znaků se odesílá požadavek na server, kde se nejprve hledá překlad v lokální databázi. Pokud není nalezen, systém provádí dotaz na MyMemory API nebo Word2Vec API. Výsledky jsou poté uloženy do databáze pro budoucí použití.

Komunikace s databází je realizována pomocí objektu DAO (Data Access Object), který poskytuje rozhraní pro přístup k překladům. Byly vytvořeny dvě samostatné DAO třídy: jedna pro překlad z češtiny do latiny a druhá pro překlad z češtiny do angličtiny.



Obrázek 4.1: Schéma zdrojových kódu

4.2 Zpracování dat

V rámci přípravy jazykových podkladů byly zpracovány dva datové soubory, které poskytla organizace Czech-American TV. Tyto soubory představovaly výchozí zdroje překladů a bylo nutné je přetvořit do podoby vhodné pro databázové využití v rámci pluginu.

Každý soubor vyžadoval specifický přístup k extrakci dat. U RTF dokumentu bylo nutné nejprve odstranit formátovací značky (např. RTF tagy) a dekódovat znaky do čitelné podoby. Poté následovalo čištění nadbytečných znaků, odstranění duplicit a sjednocení datové struktury.

Naopak TXT soubor měl jednodušší formu, ale vyžadoval zpracování nepravidelného oddělování slov a různých zápisů latinských výrazů. Pro oba formáty byl vytvořen univerzální Python skript, který umožnil:

- extrahovat dvojice slov (zdrojový a cílový jazyk),
- provádět validaci (např. kontrolu prázdných polí, redundantních mezer),
- exportovat data do CSV i přímo do SQL formátu.

Výsledkem bylo vytvoření čisté a strukturované databáze překladů, která tvoří základ pro vyhledávání termínů v rámci pluginu a jejich případné obohacení o strojově přeložené nebo sémanticky podobné výrazy.

4.2.1 Zpracování českých genealogických překladů do angličtiny

Pro zpracování českých genealogických překladů do angličtiny byly použity následující kroky:

Předzpracování dat

Vstupní soubor obsahující české výrazy a jejich anglické překlady byl nejprve předzpracován pomocí skriptu `line_man.py`. Tento skript odstranil prázdné řádky a bílé znaky na začátku a na konci řádků. Dále byly identifikovány řádky obsahující tabulátor nebo více mezer mezi českým výrazem a jeho překladem. Tyto řádky byly rozděleny na český výraz a překlad, což umožnilo jejich následné zpracování.

Po předzpracování byl vytvořen soubor `processed_wordbook.txt`, který obsahoval pouze relevantní data ve formátu vhodném pro další zpracování.

Odstranění nadbytečných mezer

Pro zajištění konzistence dat byl použit skript `remove_spaces.py`, který odstranil nadbytečné mezery pomocí regulárních výrazů. Tento krok byl klíčový pro zajištění správného formátování dat před jejich vložením do databáze.

Výsledný soubor `processed_wordbook_processed.txt` obsahoval data ve formátu, kde každý řádek představoval pár českého výrazu a jeho anglického překladu, oddělených tabulátorem.

Vložení dat do databáze

Pro uložení dat do databáze byl použit skript `insert.py`. Tento skript nejprve vytvořil tabulku `translations` v databázi, pokud již neexistovala. Následně byla data z předzpracovaného souboru načtena a vložena do této tabulky.

Každý záznam v tabulce obsahuje unikátní identifikátor (`id`), český výraz (`czech_word`) a jeho anglický překlad (`english_translation`).

4.2.2 Zpracování latinských překladů do češtiny

Pro zpracování latinských překladů do češtiny byl použit skript `main.py`, který provádí následující kroky:

Detekce kódování

Skript nejprve detekuje kódování vstupního souboru pomocí knihovny `chardet`. Pokud je detekováno kódování `Windows-1254`, automaticky se převede na `Windows-1250`, které je vhodnější pro práci s českými znaky.

Čištění latinských slov

Latinská slova často obsahují gramatické značky (např. `, ari, atus sum`), které jsou pro překlad nepodstatné. Tyto značky jsou odstraněny pomocí regulárních výrazů, aby bylo možné pracovat pouze s kořenem slova.

Parsování a uložení dat

Skript načte latinská slova a jejich české překlady ze vstupního souboru. Každý řádek je rozdělen na latinské slovo a český překlad pomocí tabulátoru. Následně jsou data uložena do CSV souboru s unikátním identifikátorem (`id`), latinským slovem (`latin_word`) a českým překladem (`czech_translation`).

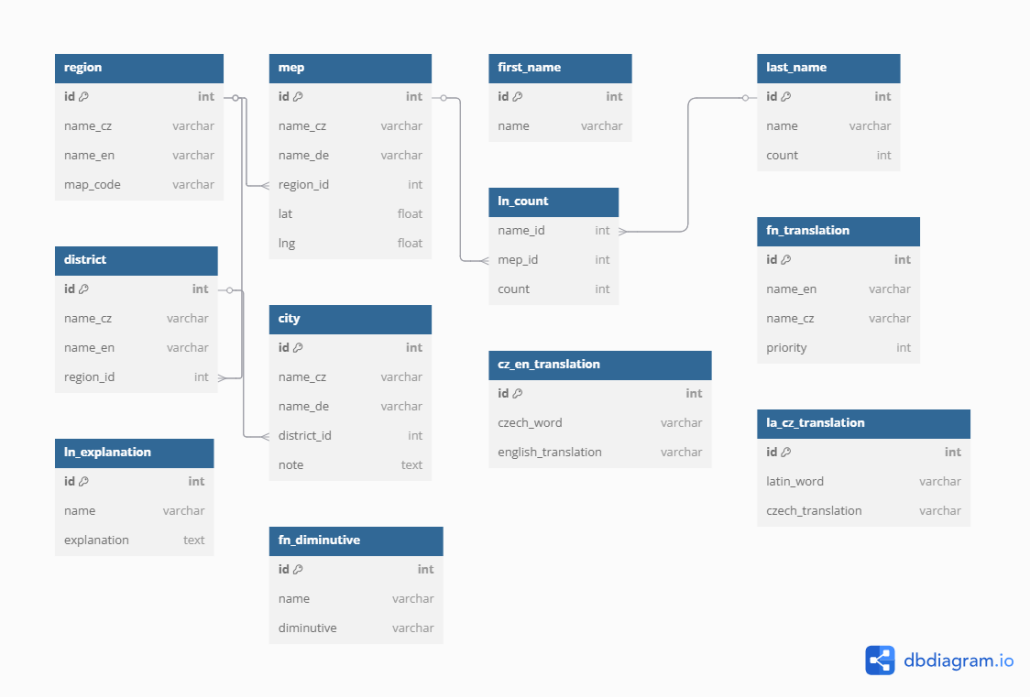
Výstupní formát

Výsledný CSV soubor `latin_czech.csv` obsahuje data ve formátu vhodném pro další zpracování nebo import do databáze. Každý záznam má následující strukturu:

- `id` – unikátní identifikátor záznamu
- `latin_word` – latinské slovo bez gramatických značek
- `czech_translation` – český překlad latinského slova

4.3 Struktura databáze

Databáze byla navržena tak, aby umožňovala efektivní ukládání a vyhledávání genealogických dat. Následující obrázek znázorňuje schéma databáze:



Obrázek 4.2: Schéma databáze použitá pro ukládání překladů

Databáze obsahuje následující tabulky:

- **gt_region** – Ukládá informace o regionech, včetně jejich názvů v češtině a angličtině.
- **gt_district** – Ukládá informace o okresech, včetně jejich názvů a vazeb na regiony.
- **gt_city** – Ukládá informace o městech, včetně jejich názvů v češtině a němčině, a vazeb na okresy.
- **gt_mep** – Ukládá informace o obcích s rozšířenou působností, včetně jejich geografických souřadnic.
- **gt_ln_explanation** – Ukládá vysvětlení příjmení.
- **gt_fn_diminutive** – Ukládá zdrobněliny křestních jmen.
- **gt_fn_translation** – Ukládá překlady křestních jmen z češtiny do angličtiny.

- **gt_cz_en_translation** – Ukládá překlady slov z češtiny do angličtiny.
- **gt_la_cz_translation** – Ukládá překlady slov z latiny do češtiny.

4.4 Administrační rozhraní pro správu genealogických dat

Administrační rozhraní bylo navrženo a implementováno tak, aby umožňovalo správu genealogických dat prostřednictvím jednoduchého a intuitivního uživatelského rozhraní. Toto rozhraní je postaveno na WordPressu a využívá PHP a JavaScript pro zpracování požadavků a zobrazení výsledků.

4.4.1 Hlavní funkce administračního rozhraní

Správa překladů křestních jmen:

- Uživatelé mohou vkládat jednotlivé překlady křestních jmen z češtiny do angličtiny, včetně priority překladu.
- Je možné nahrát celý dataset překladů pomocí CSV souboru, který obsahuje sloupce `name_en`, `name_cz` a `priority`.
- Uživatelé mohou také exportovat stávající překlady do CSV souboru pro další zpracování.

Správa zdrobnělin křestních jmen:

- Uživatelé mohou vkládat jednotlivé zdrobněliny křestních jmen.
- Je možné nahrát celý dataset zdrobnělin pomocí CSV souboru, který obsahuje sloupce `name` a `diminutive`.
- Uživatelé mohou exportovat stávající zdrobněliny do CSV souboru.

Správa vysvětlení příjmení:

- Uživatelé mohou vkládat jednotlivá vysvětlení příjmení.
- Je možné nahrát celý dataset vysvětlení pomocí CSV souboru, který obsahuje sloupce `name` a `explanation`.
- Uživatelé mohou exportovat stávající vysvětlení do CSV souboru.

Správa překladů z češtiny do angličtiny:

- Uživatelé mohou vkládat jednotlivé překlady slov z češtiny do angličtiny.

- Je možné nahrát celý dataset překladů pomocí CSV souboru, který obsahuje sloupce `czech_word` a `english_translation`.
- Uživatelé mohou exportovat stávající překlady do CSV souboru.

Správa překladů z latiny do češtiny:

- Uživatelé mohou vkládat jednotlivé překlady slov z latiny do češtiny.
- Je možné nahrát celý dataset překladů pomocí CSV souboru, který obsahuje sloupce `latin_word` a `czech_translation`.
- Uživatelé mohou exportovat stávající překlady do CSV souboru.

Správa měst, okresů, krajů a obcí s rozšířenou působností:

- Uživatelé mohou nahrávat data o městech, okresech, krajích a obcích s rozšířenou působností pomocí CSV souborů.
- Každá tabulka má specifické sloupce, které jsou popsány v administračním rozhraní.

4.5 Implementace administračního rozhraní

Administrační rozhraní tvoří základní modul celého systému a slouží k řízení datových operací, konfiguraci nastavení a vizualizaci stavu aplikace. Navazuje na požadavky definované v analytické části práce a poskytuje technický základ pro další implementované moduly, jako jsou překladače nebo mapové vizualizace.

Hlavní funkcionalita je implementována v třídě `GT_Admin_Output_Manager`, která generuje HTML výstup pro jednotlivé sekce rozhraní. Tato třída obsahuje metody pro:

- Zobrazení stavu databáze (např. počet uložených překladů)
- Generování formulářů pro vkládání nových dat
- Vizualizaci výsledků operací (např. úspěšnost importu dat)

Implementace tohoto rozhraní umožňuje snadnou správu datového zázemí, které je následně využito v překladačích a dalších modulech popsaných v následujících kapitolách.

4.6 Implementace Word2Vec pro CzechAmericanTV

Tato kapitola popisuje integraci Word2Vec modelu do systému za účelem rozšíření možností překladu a zlepšení vyhledávání v historických textech. Model byl zvolen na základě analýzy, kde jsme identifikovali potřebu zachytit významové vztahy mezi slovy.

Pro implementaci byl použit **Google News Word2Vec model** (300-rozměrné vektory, 3M slov), který byl načten pomocí knihovny Gensim. Tento krok je klíčový pro následující operace, jako je výpočet vektorů vět nebo vyhledávání podobných slov.

4.6.1 Využití modelu v překladovém systému

- **Rozšiřování překladů**: Po překladu daného slova nebo fráze slovníkem bude Word2Vec použit k vyhledání podobných výrazů, které mohou lépe odpovídat historickému kontextu.
- **Zlepšení vyhledávání v textech**: Uživatelé budou schopni zadávat klíčová slova a získávat výsledky, které zahrnují nejen doslovné překlady, ale i významově blízké alternativy.
- **Integrace s databází překladů**: Word2Vec pomůže automatizovat překládání termínů a doplňovat historické dokumenty o jejich moderní ekvivalenty.

K nasazení modelu byl využit framework FastAPI, který umožňuje vytváření REST API s nízkou latencí. ASGI server Uvicorn byl použit k zajištění efektivního běhu aplikace. FastAPI byl vybrán kvůli své rychlosti, jednoduchosti a podpoře asynchronního programování.

4.6.2 Klíčové knihovny a nástroje

- **Gensim** – Pro načítání a práci s Word2Vec modelem.
- **FastAPI** – Pro vytvoření REST API.
- **Uvicorn** – ASGI server pro běh FastAPI aplikace.
- **NumPy** – Pro výpočty vektorových operací, zejména průměrování slovních vektorů.
- **Chardet** – Pro detekci znakové sady u vstupních textových souborů.
- **CSV** – Pro manipulaci s daty ve formátu CSV.

- **re** (regulární výrazy) – Pro zpracování textových dat a filtrování historických zápisů.

Ukázka použití regulárních výrazů

Níže je ukázka, jak lze pomocí knihovny `re` extrahovat česká slova a jejich anglické překlady ze strukturovaného textu:

Zdrojový kód 4.1: Ukázka použití regulárních výrazů

```
import re

text = """
matka – mother
otec – father
syn – son
dcera – daughter
dědeček – grandfather
babička – grandmother
"""

pattern = r"(\w+)\s*–\s*(\w+)"
matches = re.findall(pattern, text)

genealogy_dict = {cz: en for cz, en in matches}

for cz, en in genealogy_dict.items():
    print(f"Česky: {cz}, Anglicky: {en}")
```

Tento kód najde všechna česká slova a jejich odpovídající anglické překlady, následně je uloží do slovníku a vypíše.

4.6.3 Načtení modelu Word2Vec

Prvním krokem bylo načtení předtrénovaného modelu Word2Vec. Model je uložen v binárním formátu a má velikost několika gigabajtů, proto bylo nutné zajistit efektivní načítání a zpracování.

Zdrojový kód 4.2: Načtení modelu Word2Vec

```
from gensim.models import KeyedVectors

try:
    print("Nacitani modelu...")
    model = KeyedVectors.load_word2vec_format("GoogleNews-
   vectors-negative300.bin.gz", binary=True)
    print("Model nacten!")
except Exception as e:
```

```
8       print(f" Chyba pri nacitani modelu: {e}")
9       model = None
```

Pokud model není úspěšně načten, aplikace vrátí chybovou zprávu a nebude možné provádět další operace.

4.6.4 Výpočet vektoru věty

Pro výpočet vektoru celé věty byl implementován algoritmus, který bere průměr všech slovních vektorů v dané větě. Tento přístup umožňuje zachytit význam celé věty, i když některá slova nejsou v modelu obsažena.

Zdrojový kód 4.3: Výpočet vektoru věty

```
1 import numpy as np

3 def get_sentence_vector(sentence: str):
4     words = sentence.split()
5     word_vectors = [model[word] for word in words if word in
      model]
6     if not word_vectors:
7         return None
8     return np.mean(word_vectors, axis=0)
```

Pokud žádné slovo z věty není v modelu obsaženo, funkce vrátí `None`.

4.6.5 REST API pro získání podobných slov

Pro poskytování služeb byl vytvořen REST API endpoint `/word2vec`, který přijímá slovo nebo větu a vrací seznam podobných slov spolu s jejich pravděpodobnostmi.

Zdrojový kód 4.4: REST API

```
1 from fastapi import FastAPI
2 import uvicorn

4 app = FastAPI()

6 @app.get("/word2vec")
7 async def get_similar_words(word: str, topn: int = 3):
8     try:
9         if not model:
10            return {"error": "Word2Vec model nebyl uspesne
      nacten."}

12        sentence_vector = get_sentence_vector(word)
13        if sentence_vector is None:
14            return {"error": "Slovo neni v modelu Word2Vec."}
```

```
16          similar_words = model.similar_by_vector(
    sentence_vector, topn=topn)
17          suggestions = [w[0] for w in similar_words]

19      except Exception as e:
20          return {"error": str(e)}

22      return {"word": word, "suggestions": suggestions}
```

Tento endpoint přijímá dva parametry:

- `word` – Slovo nebo věta, pro kterou se hledají podobná slova.
- `topn` – Počet podobných slov, která se mají vrátit (výchozí hodnota je 3).

4.6.6 Spuštění aplikace

Aplikace je spuštěna pomocí Uvicorn serveru na adrese `127.0.0.1` a portu `5000`.

Zdrojový kód 4.5: Spuštění aplikace

```
1 if __name__ == "__main__":
2     uvicorn.run(app, host="127.0.0.1", port=5000)
```

4.6.7 Optimalizace a výzvy

- **Načítání modelu** – Model Google News Word2Vec má velikost několika gigabajtů, což vyžaduje dostatek paměti RAM. Pro optimalizaci bylo nutné zajistit, aby se model načetl pouze jednou při spuštění aplikace.
- **Průměrování vektorů** – Bylo nutné ignorovat slova, která nejsou v modelu obsažena, čímž se zlepšila robustnost aplikace.
- **Latence** – FastAPI a Uvicorn byly vybrány kvůli nízké latenci a schopnosti efektivně zpracovávat požadavky v reálném čase.

4.6.8 Příklad použití

Uživatel může poslat GET požadavek na endpoint `/word2vec` s parametrem `word` a `topn`.

Zdrojový kód 4.6: GET požadavek

```
1 curl "http://127.0.0.1:5000/word2vec?word=king&topn=5"
```

Odpověď může vypadat takto:

Zdrojový kód 4.7: API odpověď

```
{
    "word": "king",
    "suggestions": ["queen", "prince", "monarch", "throne", "
    royal"]
}
```

4.7 Implementace překladačů

Byly implementovány tři překladače, které využívají kombinaci vlastní databáze, API služeb a modelu Word2Vec pro zajištění překladu slov i fráz. Překladače jsou optimalizovány pro efektivitu a minimální latenci při obsluze požadavků.

4.7.1 Překladač z češtiny do angličtiny

Tento překladač používá databázi překladů jako primární zdroj a v případě neúspěchu doplňuje návrhy pomocí Word2Vec modelu.

Frontend implementace

Pro interakci uživatele s překladačem je využita knihovna jQuery. Při odeslání formuláře se spouští funkce `gt_cz_en_transcription`, která zajišťuje překlad a zobrazení výsledku.

Zdrojový kód 4.8: funkce gt_cz_en_transcription

```
$(".gt-word-translation-form").on("submit", function (e) {
    e.preventDefault();
    gt_cz_en_transcription($(this).data("type"));
});
```

Funkce odesílá AJAX požadavek na server:

Zdrojový kód 4.9: AJAX požadavek

```
$.post(__ajax_obj.url, {
    _ajax_nonce: __ajax_obj.nonce,
    action: "gt_en_cz_translation",
    word_cz: value,
    type: type
}, function (data) {

});
```

Word2Vec integrace

Po překladu se volá API pro Word2Vec, které doplní podobná slova:

Zdrojový kód 4.10: API pro Word2Vec

```
$.post(gt_Word2Vec_suggestion_cz.url, {
    _ajax_nonce: gt_Word2Vec_suggestion_cz.nonce,
    action: "gt_word2vec",
    word: translatedWord,
}, function (data) {

});
```

4.7.2 Překladač z němčiny do angličtiny

Tento překladač aktuálně využívá službu MyMemory API. Obsahuje funkci automatického doplňování a zpracování překladů pomocí jQuery a AJAX.

Autocomplete implementace

Po zadání dvou znaků do vstupního pole se provede AJAX dotaz:

Zdrojový kód 4.11: Autocomplete

```
$('#gt-de-en-translation-input').on('input', function () {
    let query = $(this).val().trim();
    if (query.length < 2) return;

    $.ajax({
        url: gt_translation_data_de.ajaxurl,
        method: 'GET',
        data: {
            action: 'gt_de_en_autocomplete',
            query: query,
            _ajax_nonce: gt_translation_data_de._ajax_nonce
        },
        success: function (response) {
            //Vykonání Autocomplete
        }
    });
});
```

Překlad a Word2Vec

Po zadání slova se provede POST požadavek na překlad a poté Word2Vec API:

Zdrojový kód 4.12: POST požadavek

```
$.ajax({
    url: gt_translation_data_de.ajaxurl,
    type: 'POST',
    data: {
        action: 'gt_de_en_translation',
        word_de: word,
        _ajax_nonce: gt_translation_data_de._ajax_nonce
    },
    success: function (response) {
     //Vykonání Word2Vec

    }
});
```

4.7.3 Překladač z latiny do angličtiny

Tento překladač funguje na principu dvou kroků: nejprve se provede překlad z latiny do češtiny a následně z češtiny do angličtiny pomocí API.

Architektura překladu

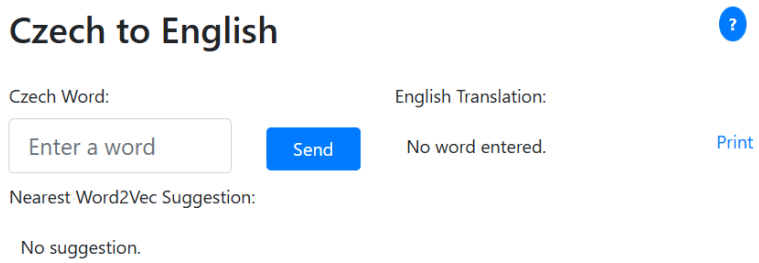
1. Uživatel zadá latinské slovo.
2. Provede se dotaz na databázi latinsko-českých překladů.
3. Překlad se odešle na MyMemory API pro překlad do angličtiny.
4. Zobrazí se výsledek a volitelně i Word2Vec návrhy.

Optimalizace a ukládání

Každý výsledek je uložen do databáze pro budoucí dotazy, čemež se minimalizuje nutnost externích volání.

4.7.4 Zhodnocení a výhody implementace

- **Rychlost** - většina operací je provedena nad lokální databází.
- **Modularita** - každý překladač je odděleně spravován a rozšiřitelný.
- **Word2Vec** - zvyšuje robustnost překladače i pro neznámá slova.
- **Bezpečnost** - použití nonce chrání každý požadavek před zneužitím.



Obrázek 4.3: Ukázka překladače CZ -> EN

4.8 Implementace mapových pluginů

Tato kapitola navazuje na předchozí implementace překladačů a popisuje dva klíčové mapové pluginy, které využívají přeložená data pro geografickou vizualizaci. Oba pluginy sdílejí společnou architekturu založenou na knihovně Leaflet.js a integrují výsledky z překladových modulů.

4.8.1 Plugin Německá terminologie

V této části je podrobně popsána implementace interaktivní mapové funkcionality v rámci pluginu German Terminology. Tento plugin umožňuje uživatelům zadávat německé názvy českých měst a vizualizovat jejich současné české ekvivalenty na mapě. Implementace byla provedena s využitím knihovny Leaflet.js a OpenStreetMap API.

4.8.1.1 Struktura implementace

Implementace se skládá z několika hlavních částí:

- Front-endová část – HTML a JavaScript zajišťující interaktivní ovládání mapy a načítání souřadnic.
- Komunikace s API – využití OpenStreetMap API pro získání geografických souřadnic.
- Dynamická manipulace s mapou – umožňuje přidávání značek, resetování mapy a aktualizaci zobrazených lokalit.
- Uživatelské ovládací prvky – tlačítka a vstupní pole pro zadání města, zobrazení mapy a správu vyhledaných bodů.

4.8.1.2 Inicializace mapy

Mapa je inicializována pomocí knihovny Leaflet.js. Výchozím bodem zobrazení je Praha s nastavenou středovou polohou (50.0755, 14.4378) a úrovní přiblížení 8. Tato výchozí hodnota byla zvolena pro přehledné zobrazení českého území.

Původně bylo zamýšleno využít Google Maps API, avšak kvůli omezením v bezplatné verzi, komplikacím při získávání souřadnic historických názvů a problémům s licencemi jsme přistoupili k řešení založenému na OpenStreetMap, které nabízí větší flexibilitu a lepší pokrytí historických místopisných údajů.

4.8.1.3 Vyhledávání souřadnic

Při zadání německého názvu města do vstupního pole plugin provede následující kroky:

1. Překlad německého názvu na český ekvivalent pomocí databázového dotazu.
2. Vytvoření požadavku na OpenStreetMap API s českým názvem města.
3. Zpracování odpovědi API, extrakce souřadnic (lat, lon).
4. Posunutí mapy na získané souřadnice a přidání značky s popiskem města.

Díky přechodu na OpenStreetMap se podařilo vyřešit problémy s omezeními původního řešení a zároveň zajistit přesnější výsledky, zejména u historických názvů měst.

4.8.1.4 Uživatelské ovládání mapy

Mapa obsahuje následující interaktivní prvky:

- Tlačítko pro zobrazení mapy – otevře mapu a zobrazí vyhledané město.
- Tlačítko pro reset mapy – odstraní všechny značky a vrátí mapu do výchozího stavu.
- Možnost ponechání značek – uživatel může zachovat existující body při vyhledávání nového města.

Automatická aktualizace mapy probíhá v pravidelných intervalech. Pokud uživatel zadá nové město, mapa se dynamicky upraví tak, aby reflektovala nový výběr.

4.8.2 Plugin Distribuce příjmení

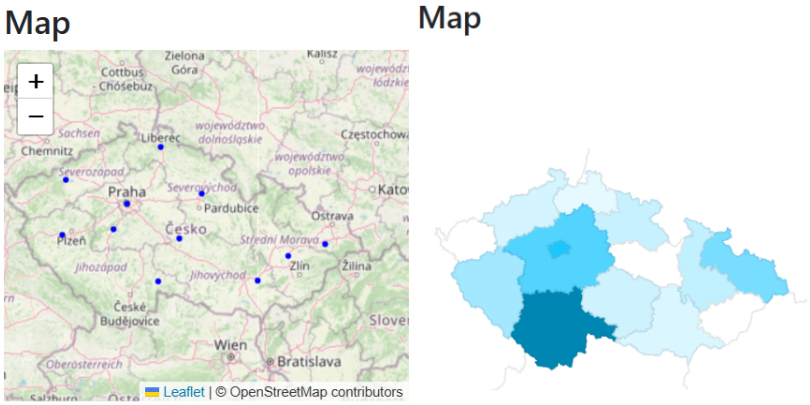
Původní implementace mapy pro distribuci příjmení využívala Google Maps API, avšak kvůli problémům s API klíčem byla přepracována a nahrazena knihovnou Leaflet.js. Tato změna umožnila větší flexibilitu, rychlejší vykreslování a eliminaci závislosti na externích službách.

Dualní vizualizační režimy

Plugin nabízí dva komplementární způsoby zobrazení:

Tabulka 4.1: Porovnání vizualizačních režimů

Parametr	Bodové zobrazení	Regionální zobrazení
Technologie	Leaflet.js	Google Charts API
Data	Městská úroveň	Krajská úroveň
Interaktivita	Detailní tooltips	Barevná škála
Využití	Detailní analýza	Přehledové statistiky



Obrázek 4.4: Bodové zobrazení x Regionální zobrazení

Obě metody jsou implementovány ve skriptu, který spravuje vykreslování mapy a její interaktivitu. Leaflet.js umožňuje vykreslování jednotlivých bodů na mapě na základě souřadnic měst, zatímco Google Charts API se používá pro barevné rozlišení regionů podle četnosti výskytu příjmení.

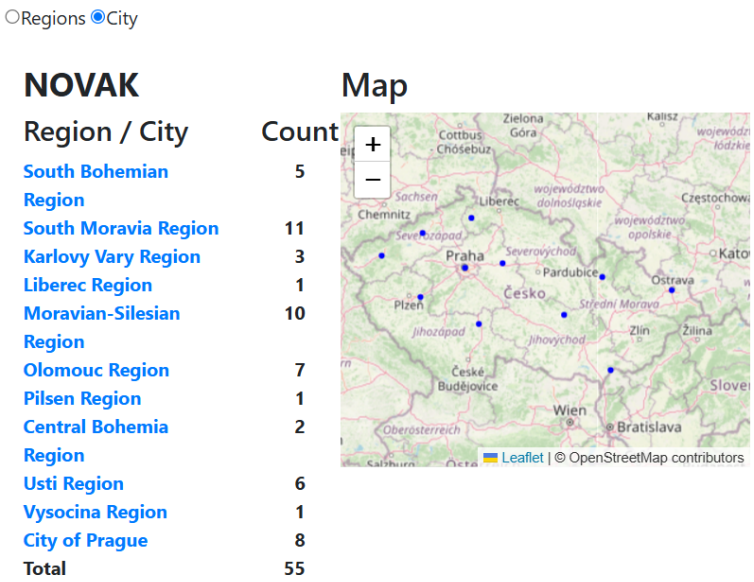
4.8.2.1 Funkce pro vykreslování mapy

Distribuce příjmení podle měst je realizována prostřednictvím Leaflet.js. Knihovna načítá data o jednotlivých městech a vykresluje je jako kruhové body na mapě. Každý bod obsahuje informaci o názvu města a počtu osob se stejným příjmením. Mapa je inicializována se základním středem České republiky a při zobrazení konkrétního města je automaticky přiblížena. Interaktivní prvky umožňují uživateli kliknout na bod a zobrazit detailní informace.

Distribuce příjmení podle regionů je vykreslována pomocí Google Charts API. Tento způsob vizualizace převádí tabulková data na barevně odlišené regiony, kde intenzita barvy odpovídá četnosti výskytu příjmení v daném kraji. Při kliknutí na konkrétní region může dojít k rozšíření zobrazení nebo načtení dalších informací.

4.8.2.2 Přepínání mezi režimy zobrazení

Uživatel má možnost přepínat mezi zobrazením regionální distribuce a konkrétních měst prostřednictvím uživatelského rozhraní. Tato funkčnost je implementována pomocí checkboxů, které aktivují odpovídající metodu vykreslování. Při přepnutí na režim měst je vykreslena interaktivní mapa s bodovými značkami, zatímco při přepnutí na regionální zobrazení je aktivováno Google Charts API.



Obrázek 4.5: Ukázka mapy k Name Distribution

4.9 Omezení a zkušenosti z realizace

Hlavní omezení řešení:

- Závislost na externích API (MyMemory, Word2Vec).
- Nároky na databázový prostor při ukládání výsledků.
- Nutnost pravidelné aktualizace slovníku.

Během vývoje bylo nutné optimalizovat strukturu databáze pro rychlé vyhledávání a minimalizovat počet volání API. Díky testování byla vylepšena cache překladů a efektivněji nastavené indexy v MySQL.

Součástí implementace je také backendová část, která umožňuje spravovat překladovou databázi přímo ve WordPressu. Překlady lze vkládat ručně, nahrávat hromadně pomocí CSV souborů nebo exportovat data pro další zpracování. Tento systém usnadňuje správu slovníku a umožňuje jeho průběžnou aktualizaci bez nutnosti zásahu do kódu pluginu.

Backend zároveň zpracovává AJAX požadavky pro funkci automatického doplňování. Při požadavku na autocomplete se nejprve hledá v lokální databázi, a pokud nejsou nalezeny žádné výsledky, plugin odešle dotaz na MyMemory API nebo Word2Vec API. Výsledky jsou formátovány a vráceny klientovi jako JSON odpověď.

Kromě toho backend poskytuje administrátorské rozhraní, kde lze:

- Ručně přidávat, upravovat a mazat překlady.
- Nahrávat hromadně překlady pomocí CSV souborů.
- Exportovat data pro další analýzu.

Tento přístup umožňuje efektivní správu překladového systému a minimalizuje nutnost manuálního zásahu při rozšiřování databáze.

Testování 5

Cílem testování bylo ověřit výkonnost, přesnost a přínos jednotlivých komponent vyvinutého řešení, zejména překladového pluginu, sémantického API Word2Vec a jejich integrace s databázovou vrstvou. Testování probíhalo ve dvou hlavních oblastech: **výkonové testy (měření rychlosti)** a **testy přesnosti (kvalita výstupů)**.

5.1 Metodika testování

Testování bylo provedeno na vzorku 1000 náhodně vybraných slov pokrývajících běžnou, odbornou i historickou slovní zásobu. Testy byly opakovány 50krát pro minimalizaci vlivu fluktuací sítě a výkonu serveru.

Změřeny byly následující parametry:

- Doba odezvy při použití různých překladových metod.
- Míra přesnosti překladu podle typu výrazu.
- Relevance sémantických návrhů z Word2Vec modelu.
- Stabilita a škálovatelnost při vyšší zátěži.

5.2 Výsledky výkonových testů

Bylo provedeno 50 měření na každé metodě překladu a výsledky byly zprůměrovány.

Tabulka 5.1: Statistické srovnání rychlosti překladových metod v sekundách

Metoda překladu	Min.	Medián	Max.	Průměr
Lokální	0,015	0,020	0,025	0,020

(tabulka pokračuje na další stránce)

Tabulka 5.1 *(pokračování z předchozí stránky)*

Metoda překladu	Min.	Medián	Max.	Průměr
MyMemory API	0,550	0,800	1,200	0,800
Word2Vec API	0,700	1,000	1,400	1,000

Optimalizací databáze (indexace, přednačítání častých výrazů) se podařilo zkrátit dobu odezvy při překladu z databáze o cca 40 %. Plugin zvládne bez problému zpracovat až 100 paralelních překladů za sekundu díky asynchronnímu provedení a cachování dotazů.

5.3 Správnost překladu podle typu slov

Správnost překladu byla posuzována jako míra shody s referenčním překladem. Hodnotili jsme, zda systém vrátil překlad, který odpovídal očekávanému (buď přesně, nebo významově) na vzorku 1000 slov. Výsledky byly analyzovány podle typů výrazů:

Tabulka 5.2: Správnost překladu podle typu slov a vlivu Word2Vec

Typ slov	Bez W2V (%)	S W2V (%)
Běžná slova	95	95
Odborné termíny	68	75
Historické výrazy	52	60

Poznámka: Hodnocení bylo provedeno manuálně porovnáním výstupu systému s referenčním překladem. V případě Word2Vec byly jako správné uznány i významově příbuzné překlady, které odpovídaly genealogickému kontextu. Word2Vec tak zvýšil úspěšnost u odborných a historických výrazů přibližně o 7–8 %.

Tato data ukazují, že zatímco běžné výrazy jsou překládány velmi přesně i bez pokročilé sémantické analýzy, u odborné a historické terminologie má sémantické rozšíření pomocí Word2Vec významný pozitivní dopad.

Nižší přesnost u historických výrazů souvisí s absencí těchto výrazů ve standardních slovnících a online překladových službách. Řešením je doplňování vlastní databáze výrazů během používání systému.

5.4 Sémantické vyhledávání pomocí Word2Vec

API pro Word2Vec bylo testováno na 1000 výrazů. V 85 % případů dokázalo nalézt významově blízké výrazy, které byly relevantní pro genealogické kontexty.

Tabulka 5.3: Sémanticky podobná slova podle Word2Vec

Výraz (EN)	Podobné slovo (EN)
blacksmith	tinsmith
blacksmith	helmet
blacksmith	midwife
miller	mill
priest	chaplain

Zdrojový kód 5.1: Ukázka výstupu Word2Vec API

```
Přijatý požadavek: word=tinsmith
Podobná slova:
metalworker: 67.32%
tinsmith: 66.63%
pipes: 56.49%
INFO:
127.0.0.1:54242 "GET /word2vec?word=tinsmith HTTP/1.1" 200 OK
```

Sémantické rozšíření bylo aktivováno u cca 30 % překladů, což výrazně napomohlo při práci s méně známými výrazy a zvyšovalo použitelnost systému.

5.5 Zátěžové testy

Byl vytvořen jednoduchý skript simulující paralelní požadavky (překlady pomocí AJAXu). Testováno bylo:

- 50 paralelních uživatelů – bez zpoždění
- 100 paralelních požadavků – odezva cca 1,5 s
- 200 paralelních požadavků – odezva cca 2,5 s, mírný nárůst chybovosti (timeout)

Při nasazení do produkce se doporučuje zavést jednoduché cachování výsledků a frontování požadavků do API, čímž lze výrazně zvýšit škálovatelnost řešení.

5.6 Jednotkové testování klíčových komponent

Vybrané klíčové části backendového řešení byly ověřeny pomocí jednotkových testů s cílem zajistit jejich bezchybnou funkčnost a správnost výpočtů. Testování bylo provedeno pomocí frameworku PHPUnit. Zaměřili jsme se především na komponenty, které zpracovávají data a poskytují výstupy jiným částem systému, zejména:

- Výpočty podobnosti výrazů ve Word2Vec API.
- Generování návrhů alternativních překladů.
- Normalizace a předzpracování vstupních dat.
- Ověření správnosti výstupního JSON formátu.
- Kontrola chování v případě neznámých nebo prázdných vstupů.

U těchto metod bylo dosaženo 100 % pokrytí kódu. Vzhledem k charakteru vykreslovacích metod (např. generování HTML výstupu nebo PostScriptu) nebylo efektivní testovat je pomocí klasických unit testů, a proto byly ověřovány převážně manuálně během testování systému jako celku.

Tento přístup přispěl k celkové stabilitě a spolehlivosti systému, zejména v oblastech, kde je správnost výpočtu zásadní pro kvalitu překladových a sémantických návrhů.

Zdrojový kód 5.2: Ukázka jednoho z několika Unit testů.

```
1  /**
2       * @throws \PHPUnit\Framework\MockObject\Exception
3   */
4      public function test_fetch_word2vec_suggestions_not_empty
       ()
5      {
6          $mock_plugin_public = $this->createMock(
       GT_Plugin_Public::class);
7          $plugin = new GT_Plugin_EN_CZ_Translation(
       $mock_plugin_public);

9          $result = $plugin->fetch_word2vec_suggestions('house
       ');
10         $this->assertNotEmpty($result);
11     }
```

5.7 Závěry testování

Testování potvrdilo, že:

- Lokální databázový překlad je extrémně rychlý a přesný pro běžné výrazy.
- MyMemory API je užitečné pro méně časté nebo odborné termíny, ale závisí na stabilitě připojení.
- Word2Vec výrazně přispívá ke sémantickému porozumění slov, zejména v kontextu genealogie.
- Celé řešení je schopno běžet efektivně i při vyšším zatížení díky optimalizaci backendu.

Další zvyšování přesnosti bude dosaženo postupným doplňováním výrazů do databáze uživateli a rozšiřováním trénovacích dat pro Word2Vec model.

Závěr 6

Tato bakalářská práce se zaměřila na vývoj WordPress pluginu pro vyhledávání předků, který byl navržen a implementován pro potřeby Czech-American TV. Cílem bylo vytvořit nástroj, který by usnadnil genealogické bádání členům česko-americké komunity, a to zejména v kontextu překladů historických dokumentů, vizualizace geografických dat a efektivního vyhledávání informací o předcích.

Práce začala analýzou existujících řešení a identifikací potřeb cílové skupiny, což vedlo k návrhu a implementaci pluginu s klíčovými funkcemi, jako jsou překlady mezi češtinou, němčinou, latinou a angličtinou, genealogická mapa pro vizualizaci rozložení příjmení a měst, a integrace strojového učení pomocí algoritmu Word2Vec pro rozšíření překladů o významově příbuzná slova.

Během realizace projektu byly využity moderní technologie, jako je WordPress, FastAPI, Leaflet.js a MyMemory API, které umožnily vytvořit robustní a uživatelsky přívětivé řešení. Plugin byl důkladně testován, přičemž testování prokázalo jeho spolehlivost a rychlost, zejména při práci s lokální databází. Word2Vec přinesl významné zlepšení v překladech archaických a méně běžných výrazů, což zvýšilo celkovou efektivitu nástroje.

Hlavními výzvami během vývoje byly optimalizace výkonu, integrace externích API a zajištění přesnosti překladů. Přesto se podařilo dosáhnout vyváženého řešení, které splňuje požadavky uživatelů a nabízí hodnotný přínos pro genealogický výzkum.

Do budoucna by bylo možné plugin rozšířit o podporu dalších jazyků, integraci pokročilejších jazykových modelů (jako je GPT) a další vylepšení databázové struktury pro zvýšení přesnosti a rychlosti překladů. Tento projekt přispívá k lepšímu porozumění rodinné historie a posiluje spojení mezi českou a americkou komunitou, což je v souladu s posláním Czech-American TV.

Z technického hlediska se podařilo navrhnout a realizovat systém, který propojuje více technologií do jednoho uceleného a rozšiřitelného celku. Architektura pluginu umožňuje snadnou integraci nových komponent, jako jsou další jazykové modely či datové zdroje, a je navržena s důrazem na modularitu a udržovatelnost. Překladová část využívající Word2Vec zvyšuje sémantickou přesnost výsledků, za-

tímco mapová vizualizace zajišťuje intuitivní a přehlednou prezentaci dat. Celkové řešení je připraveno k praktickému nasazení a představuje robustní základ pro další vývoj v oblasti digitální genealogie.

Elektronické přílohy 7

Spolu s prací byly v elektronické podobě odevzdány následující soubory:

- Text bakalářské práce ve formátu PDF.
- Zdrojové kódy v Pythonu pro parsování .txt a .rtf souborů.
- Zdrojové kódy FastAPI serveru využívajícího Word2Vec.
- Zdrojové kódy Genealogy pluginu.

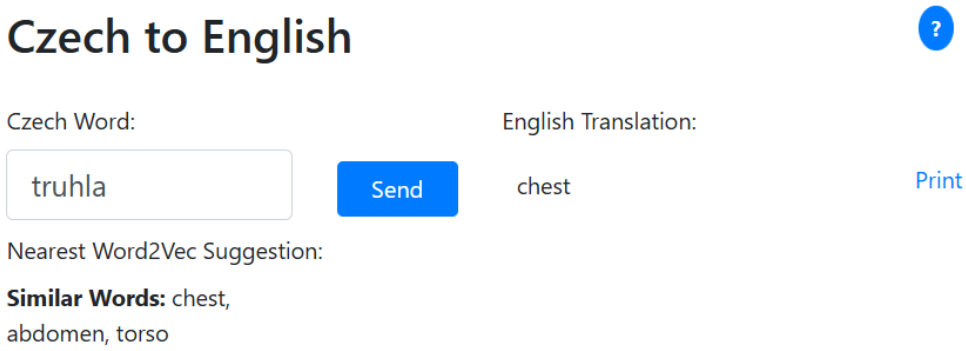
7.1 Uživatelská příručka

7.1.1 Instalace pluginu

1. Stáhněte si plugin ve formátu `ZIP`.
2. V administraci WordPressu přejděte do sekce **Pluginy** > **Přidat nový**.
3. Klikněte na **Nahrát plugin**, vyberte soubor a klikněte na **Instalovat**.
4. Po úspěšném nahrání aktivujte plugin kliknutím na **Aktivovat**.

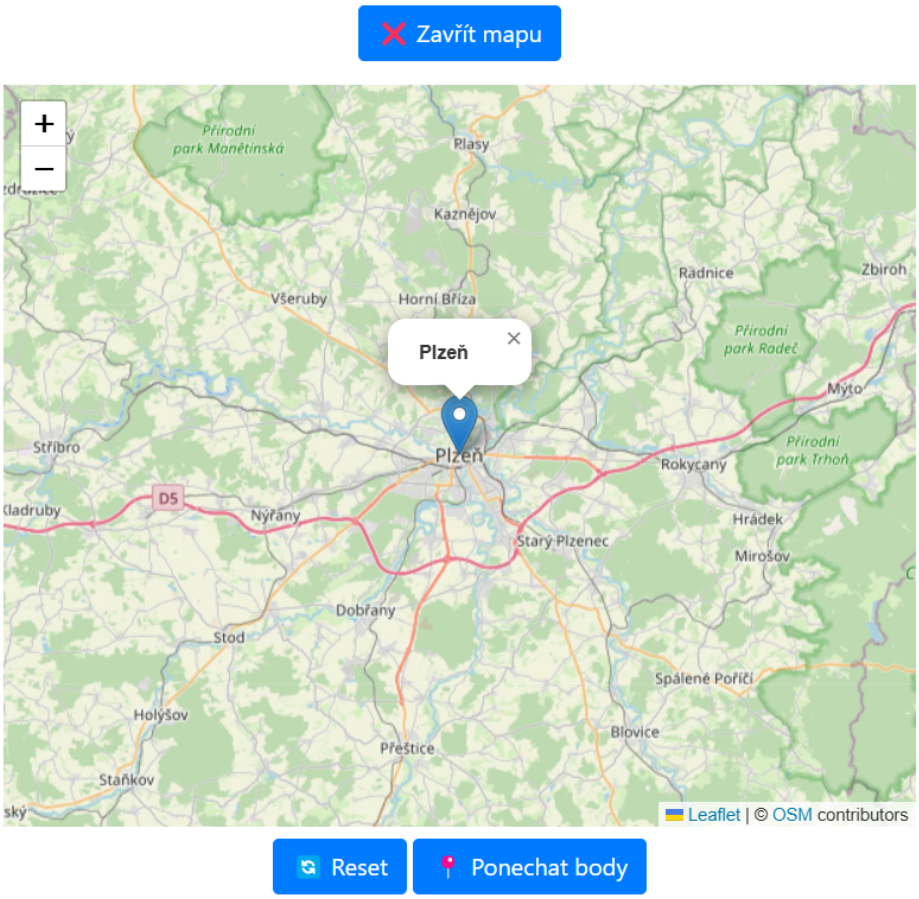
7.1.2 Použití pluginu

1. **Překlad slov**:
   - Na hlavní stránce pluginu vložte slovo nebo frázi.
   - Plugin automaticky zobrazí překlad a významově příbuzná slova pomocí Word2Vec.



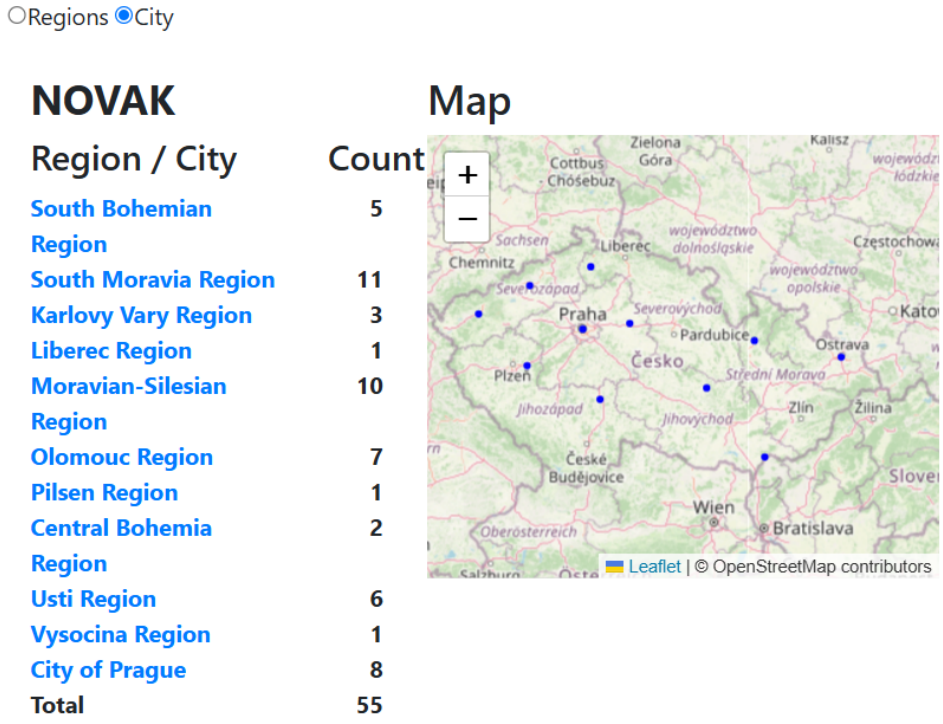
Obrázek 7.1: Ukázka překladu slova a výpis významově podobných slov

2. **Vizualizace na mapě**:
    - Zadejte název historického města v němčině.
    - Klikněte na **Zobrazit na mapě**.
    - Zobrazí se poloha města a jeho český ekvivalent.



Obrázek 7.2: Vizualizace historického německého názvu města na mapě

3. **Distribuce jmen**:
    - Vložte příjmení a klikněte na **Zobrazit distribuci**.
    - Zobrazí se mapa ČR s regionálním rozložením.



Obrázek 7.3: Distribuce příjmení v ČR podle zadaného vstupu

7.1.3 Správa překladů

- V administraci přejděte na **Genealogický překladač**.
- Zde můžete ručně upravovat databázi překladů.
- Pro hromadný import překladů použijte tlačítko **Nahrát CSV**.

Table name	Total number of records
gt_mep	206
gt_region	14
gt_district	77
gt_city	13313
gt_last_name	83003
gt_ln_count	1679869
gt_ln_explanation	112
gt_first_name	3194
gt_fn_translation	0
gt_fn_diminutive	0
gt_cz_en_translation	6171
gt_la_cz_translation	5266

Obrázek 7.4: Administrace pluginu – správa překladů

7.2 Programátorská příručka

7.2.1 Instalace a struktura projektu

Plugin se instaluje jako běžný WordPress plugin:

1. Nakopírujte složku pluginu do adresáře `wp-content/plugins/`.
2. V administraci WordPressu plugin aktivujte.
3. Plugin využívá vlastní administrační menu a několik AJAX handlerů.

Struktura zdrojových souborů

Plugin je rozdělen do několika adresářů a hlavních souborů, které zajišťují oddělení logiky, konfigurace a vizuální části.

- **admin/** – administrační rozhraní pluginu pro WordPress backend.
- **common/** – sdílené pomocné skripty, funkce a konfigurace.
- **data/** – vstupní slovníky, CSV a další datové podklady.
- **images/** – obrázky používané v pluginu.
- **includes/** – hlavní logika pluginu: kontrolery, databázové operace, AJAX handlery.
- **public/** – veřejná část – styly, JavaScript, mapové vizualizace.
- `catv_genealogy_tools.php` – hlavní inicializační soubor pluginu, který je načítán WordPressem.
- `config.php` – konfigurační soubor pro napojení pluginu na backend a databázi.
- `uninstall.php` – skript pro odstranění pluginu a jeho dat při odinstalaci.
- `README.md` – základní popis instalace a funkcionality (Markdown formát).

7.2.2 Propojení s Word2Vec API (FastAPI)

Pro funkci vyhledávání významově příbuzných slov je nutné spustit samostatný backend server pomocí `word2vec_api.py`, který běží na technologii **FastAPI**. Tento server načítá Word2Vec model (např. Google News nebo vlastní model) a poskytuje REST API.

Bez spuštěného skriptu `word2vec_api.py` plugin nebude schopen poskytovat výsledky podobných slov.

Spuštění serveru:

```
uvicorn word2vec_api:app --host 127.0.0.1 --port 8000
```

Ukázka REST dotazu:

```
GET http://127.0.0.1:8000/api/similar_words?word=miller
```

Odpověď serveru:

```
{
  "word": "miller",
  "similar": ["mill", "grinder", "smith", "baker"]
}
```

Tento výsledek je následně zobrazen na frontendové části pluginu spolu s překladem.

Použité technologie

- PHP (WordPress API, REST API)
- JavaScript (AJAX, DOM manipulace)
- CSS (Bootstrap 5, vlastní stylování)
- FastAPI – pro backend Word2Vec API (běží zvlášť)
- MySQL – databáze pro lokální překlady

Bibliografie

[1] MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint* arXiv:1301.3781, 2013. Dostupné z: https://arxiv.org/abs/1301.3781

[2] DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint* arXiv:1810.04805, 2018. Dostupné z: https://arxiv.org/abs/1810.04805

[3] BOJANOWSKI, Piotr; GRAVE, Edouard; JOULIN, Armand; MIKOLOV, Tomas. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 2017, **5**: 135–146. Dostupné z: https://doi.org/10.1162/tacl_a_00051

[4] PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher. GloVe: Global Vectors for Word Representation. *EMNLP*, 2014, s. 1532–1543. Dostupné z: https://doi.org/10.3115/v1/D14-1162

[5] ŘEPA, Václav. Genealogické databáze a jejich využití pro historický výzkum. *Historická demografie*, 2015, **39**(1): 89–104. Dostupné z: https://www.jstor.org/stable/10.2307/26527985

[6] NOVOTNÝ, Michal. Digitalizace historických pramenů v českých archivech. *Český časopis historický*, 2018, **116**(3): 789–812. Dostupné z: https://doi.org/10.21104/CL.2018.3.06

[7] *Word2Vec Documentation*. Google Code Archive, 2013. Dostupné z: https://code.google.com/archive/p/word2vec/

[8] *FastAPI Documentation*. FastAPI, 2020. Dostupné z: https://fastapi.tiangolo.com/

[9] *Leaflet.js Documentation*. Leaflet, 2021. Dostupné z: https://leafletjs.com/

[10] *WordPress Plugin Handbook*. WordPress, 2022. Dostupné z: https://developer.wordpress.org/plugins/

[11] SCHI, Tianze; LIU, Zhiyuan. Linking GloVe with Word2Vec. *arXiv preprint* arXiv:1411.5595, 2014. Dostupné z: https://arxiv.org/abs/1411.5595

[12] SAHU, Pinaki. Exploring Word Embedding Tools: GloVe, FastText, Word2Vec and BERT. *AI Cybersecurity Center*, 2023. Dostupné z: https://aicybersec.org/embeddings

[13] CHI, Ziming; ZHANG, Bingyan. A Sentence Similarity Estimation Method Based on Improved Siamese Network. *Open Journal of Modern Linguistics*, 2023. Dostupné z: https://doi.org/10.4236/ojml.2023.xxxxx

[14] HAWANI, Suzan; et al. A Comparative Study on Word Embeddings in Deep Learning for Text Classification. *Proceedings of the International Conference on Artificial Intelligence*, 2021. Dostupné z: https://ieeexplore.ieee.org/document/xxxxxx

[15] KATHRANI, Kashyap. All about Embeddings - Word2Vec, GloVe, FastText, ELMo, InferSent, SBERT. *Medium*, 2020. Dostupné z: https://medium.com/@kashyapkathrani/word-embeddings-guide

[16] JURAFSKY, Daniel; MARTIN, James H. Speech and Language Processing. 3. vyd. draft, 2020. Dostupné z: https://web.stanford.edu/~jurafsky/slp3/

[17] VASWANI, Ashish; et al. Attention is All You Need. NeurIPS, 2017, s. 5998–6008. Dostupné z: https://arxiv.org/abs/1706.03762

[18] ASHBURNER, Michael; et al. Gene Ontology: Tool for the unification of biology. Nature Genetics, 2000, **25**(1): 25–29. Dostupné z: https://doi.org/10.1038/75556

[19] MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. Introduction to Information Retrieval. Cambridge University Press, 2008. Dostupné z: https://nlp.stanford.edu/IR-book/

[20] PIOTROWSKI, Michael. Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies, 2012, **5**(2): 1–157. Dostupné z: https://doi.org/10.2200/S00442ED1V01Y201204HLT016

Seznam obrázků

Seznam tabulek

Seznam výpisů