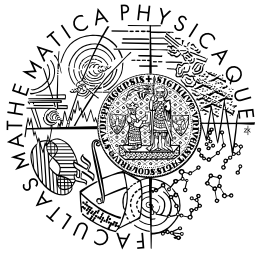

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Adam Turčan

Webová aplikace pro konfigurovatelné zpracování textu s využitím externích služeb

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: doc. RNDr. Pavel Pecina, Ph.D.
Studijní program: Informatika
Studijní obor: Programování a vývoj software

Praha 2026

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Poděkování.

Název práce: Webová aplikace pro konfigurovatelné zpracování textu s využitím externích služeb

Autor: Adam Turčan

Ústav: Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: doc. RNDr. Pavel Pecina, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: Abstrakt.

Klíčová slova: zpracování přirozených jazyků, uživatelské rozhraní, webová aplikace

Title: Web application for configurable text processing using external services

Author: Adam Turčan

Institute: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Pavel Pecina, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Abstract.

Keywords: natural language processing, user interface, web application

Obsah

Červenou označujem skratky v texte, ktoré je potrebné vysvetliť.

Úvod

Vývoj softvérových nástrojov pre špecifické vedecké domény so sebou prináša unikátne výzvy. Vyžaduje nielen hlboké pochopenie doménových dát, ale aj schopnosť navrhnúť intuitívne používateľské rozhranie (UI/UX) pre používateľov, ktorí nemusia byť technicky zdatní. Zároveň však narážame na problém, že riešenia „šité na mieru“ jednej doméne sú často príliš úzko špecializované, a teda ťažko prenositeľné do iných oblastí.

Táto bakalárska práca sa zaoberá kompletným životným cyklom vývoja platformy určenej na spracovanie textu s využitím externých NLP služieb, akými sú preklad, segmentácia, anotácia a semantické značkovanie. Opisuje cestu od prvotného návrhu používateľskej skúsenosti (UX) pre potreby projektu MEMORISE, cez implementáciu počiatočnej architektúry, až po jej následnú evolúciu do podoby univerzálneho, konfigurovateľného systému. Práca tak neprezentuje len finálny stav, ale dokumentuje iteratívny proces, v ktorom prvá implementácia slúžila ako kľúčový prostriedok na identifikáciu architektonických nedostatkov a validáciu interakčného dizajnu.

Projekt MEMORISE a technologické výzvy

Táto práca je zasadená do kontextu medzinárodného výskumného projektu MEMORISE, ktorého cieľom je zachovanie spomienok obetí holokaustu a sprístupnenie ich príbehov pre budúce generácie. Vzhľadom na to, že priami svedkovia a pamätníci holokaustu postupne odchádzajú, projekt sa zameriava na digitalizáciu viac ako 80 000 historických záznamov, vrátane denníkov obetí, svedectiev a multimediálnych materiálov.

V rámci pracovných balíkov projektu zameraných na dátovú infraštruktúru a integráciu (WP2 a WP3) vzniká komplexný proces pre spracovanie týchto dát. Cieľom je transformovať neštruktúrované historické texty do podoby prepojených dát a znalostných grafov, ktoré následne slúžia ako podklad pre vizualizáciu.

Tento proces využíva pokročilé metódy umelej inteligencie a spracovania prirodzeného jazyka (NLP). Automatizované služby zabezpečujú:

- **Segmentáciu textu**: Rozdelenie dlhých textov (napríklad denníkov) na logické celky.
- **Strojový preklad**: Sprístupnenie dokumentov v rôznych jazykoch.
- **Semantické značkovanie**: Identifikáciu tém a kontextu.
- **Rozpoznávanie pomenovaných entít (ďalej ako NER)**: Extrakciu mien osôb, geografických lokácií (napr. getá, tábory) a časových údajov.

Rola kurácie dát v NLP procesoch

Hoci moderné NLP modely dosahujú vysokú presnosť, pri práci s citlivými historickými dátami nie je plná automatizácia postačujúca. Výstupy modelov

často obsahujú chyby v rozpoznávaní entít (napr. zámena mena osoby za názov mesta), nepresnosti v preklade či nevhodnú segmentáciu, ktorá spôsobuje nelogické členenie textu a stratu kontextu. Tieto nedostatky by bez zásahu človeka mohli viesť k nesprávnej interpretácii historických faktov.

Ako uvádza metodika projektu MEMORISE (konkrétne **WP3** [1]), pre dosiahnutie vysokej kvality dát je nevyhnutný prístup *Human-in-the-loop*. To znamená, že automaticky generované metadáta musia byť validované a korigované ľudskými expertmi – kurátormi.

Pôvodná motivácia pre vznik tohto nástroja vychádzala práve z potreby poskytnúť expertom efektívne používateľské rozhranie pre:

- **Vizualizáciu výstupov z NLP API:** Prehľadné zobrazenie segmentov, prekladov a navrhovaných entít.
- **Post-editáciu prekladov a segmentácie:** Možnosť opraviť gramatické a kontextové nepresnosti vzniknuté pri strojovom preklade a segmentácii.
- **Validáciu anotácií:** Rýchlu korekciu nesprávne klasifikovaných entít a úpravu priradených semantických značiek.
- **Semantické prepájanie (Linking):** Manuálne dopĺňanie chýbajúcich väzieb na externé ontológie (napr. prepojenie na holandský tezaurus 2. svetovej vojny).

Potreba takéhoto rozhrania však nie je izolovaná len pre oblasť digitálnej histórie. S masívnym nástupom NLP a generatívnych modelov do praxe narážajú na identický problém mnohé iné domény, kde je kľúčová precíznosť dát – napríklad **medicína** (kontrola extrakcie symptómov z lekárskych správ) alebo **právo** (validácia zmlúv a anonymizácia údajov). V týchto oblastiach, podobne ako v projekte MEMORISE, nemožno slepo dôverovať výstupom umelej inteligencie a prístup *Human-in-the-loop* sa stáva nevyhnutnosťou.

Vytvárať pre každú takúto doménu samostatnú, jednoúčelovú aplikáciu by však bolo neefektívne a dlhodobo neudržateľné. Existujúce riešenia často narážajú na limity flexibility – buď sú príliš všeobecné a nepokrývajú špecifické pracovné toky (ako je súbežná úprava segmentácie a prekladu), alebo sú naopak pevne naviazané na konkrétnu dátovú schému.

Motiváciou tejto práce je preto návrh a implementácia **univerzálnej, konfigurovateľnej platformy**. Cieľom je vytvoriť systém, ktorý dokáže dynamicky spracovávať výstupy z rôznych NLP služieb a prispôsobiť sa odlišným anotačným schémam iba prostredníctvom zmeny konfigurácie, bez nutnosti zásahov do zdrojového kódu.

Literatura

1. MEMORISE Consortium. *MEMORISE Project Methodology: Work Package 3 – Data Infrastructure*. Dostupné tiež z: `https://memorise.sdu.dk/about-memorise/`. [Cit. 2024-02-05].

Zoznam obrázkov

Zoznam tabuliek

Seznam použitých zkratek

A Přílohy

A.1 První příloha