# POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

## I. IDENTIFIKAČNÍ ÚDAJE

| | |
|---|---|
| **Název práce:** | **Creating a Knowledge Base from Websites** |
| **Jméno autora:** | **Štěřovský Josef** |
| **Typ práce:** | diplomová |
| **Fakulta/ústav:** | Fakulta elektrotechnická (FEL) |
| **Katedra/ústav:** | Katedra počítačů |
| **Oponent práce:** | Ing. Jan Drchal, PhD. |
| **Pracoviště oponenta práce:** | AIC |

## II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

**Zadání** **průměrně náročné**

*Hodnocení náročnosti zadání závěrečné práce.*

Zadání je velmi obecné. Při rozumném způsobu řešení náročnost odhaduji jako průměrnou: spojuje scrapping a použití metod strojového učení.

**Splnění zadání** **splněno**

*Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.*

Zadání bylo splněno bez výhrad. Rozsahem práce odpovídá průměru.

**Zvolený postup řešení** **správný**

*Posuďte, zda student zvolil správný postup nebo metody řešení.*

Se zvoleným způsobem řešení souhlasím až na menší výjimku: v podsekci 5.4.2 („Text classification by prompting“) student popisuje použití TF-IDF pro výběr nejdůležitějších vět. Důvodem je snížení objemu textu pro následné zpracování velkým jazykovým modelem. Tato operace v podstatě odpovídá jednoduché verzi extraktivní sumarizace, která, jak student sám píše, snižuje souvislost textu. Častou metodou v podobném příkladě je vstupní text nekrátit, ale rozdělit jej (mnohdy s překryvem) na více bloků, které jsou klasifikovány zvlášť (výstupy je pak třeba agregovat).

Rovněž by bylo vhodné ručně analyzovat menší množinu špatně klasifikovaných záznamů a přidat jejich ukázky.

**Odborná úroveň** **B - velmi dobře**

*Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.*

Práce má velmi dobrou odbornou úroveň. Výrazněji postrádám jen statistiky jednotlivých získaných datových sad (počty stránek, tokenů, apod.).

**Formální a jazyková úroveň, rozsah práce** **A - výborně**

*Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.*

Text je psán kvalitní angličtinou s minimem překlepů. Práce je dobře strukturována a dobře se čte. Kapitola 2 (Theoretical foundation) zabíhá v některých případech až do přílišných detailů obecně známé problematiky (např. úvod sekce 2.3 Classification).

**Výběr zdrojů, korektnost citací** **A - výborně**

*Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.*

Zdroje jsou citovány správně.

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

# POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

**Další komentáře a hodnocení**
*Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.*

**III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE**

*Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.*

K práci mám následující dotaz:
Zkoušel jste, vzhledem k problémům s katalánštinou zamíchanou do španělských dat, nahradit USE v Top2Vec jiným multilinguálním modelem, který by tento jazyk podporoval? Pokud ne, jak by to bylo náročné?

Celkově se jedná o velmi pěknou diplomovou práci. Hodnotím ji klasifikačním stupněm **B - velmi dobře.**

Datum: 11.6.2024 Podpis: