POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Czech Foundational Large Language Model Corpus**
Jméno autora:	**Tommaso Gargiani**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra počítačů
Oponent práce:	Ing. Luboš Král, PhD
Pracoviště oponenta práce:	ČVUT, Fakulta elektrotechnická, Katedra počítačů

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

Hlavním cílem této práce bylo navrhnout robustní postupy pro sběr a předzpracování datového korpusu, který je nezbytný pro trénování základního modelu velkého jazykového modelu (LLM). Klíčem bylo prozkoumat různé zdroje internetových dat, včetně, ale nikoli výhradně, platforem jako Common Crawl a Wikipedia. Pozornost byla věnována nalezení nejúčinnějších metod pro čištění a deduplikaci surového textu, přičemž se kladl důraz na nalezení rovnováhy mezi efektivitou a kvalitou**.**

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.

Diplomant ve své práci prokázal, že je schopen k dané problematice vybrat relevantní informační zdroje, tedy odbornou literaturu. Tu následně srozumitelně a logicky shrnuje v úvodních částech diplomové práce, včetně relevantních zdrojů českých textů.

V dalších částech práce je uveden popis požadovaného zpracování textu a to odpovídajícím způsobem s rozpadem na jednotlivé kroky. Pro trénování a testování menších LLM jazykových modelů použil popsaná vybraná data. Pro hodnocení kvality modelu je použita jeho perplexita. Tento experiment překročil rámec původního zadání diplomové práce.

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Diplomant zvolil vhodný a adekvátní přístup k úloze, odpovídající současnému stavu v problematice a v logickém pořadí aplikoval zpracování dostupných textů. Součástí bylo vytvoření velikého trénovacího korpusu pro české jazykové modely. Text těžil z různých internetových zdrojů. Následně sadu zpracoval tak aby výsledkem byl jenom čistý text zbavený HTML tagů apod. Z celkového textu vytvořil sub-set kvalitního českého textu, který požil pro trénink jednoduchých LLM.

Odborná úroveň **C - dobře**

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

Diplomant studoval a identifikoval nejnovější a nejúspěšnější strategie pro shromažďování a zpracování textových dat, vhodné pro trénování základních modelů. Na základě této počáteční studie napsal vlastní skripty na filtrování českých textů. Diplomant prošel řadu různých zdrojů, většina je pak ve výsledku použita z Common Crawl databáze.

Datové soubory byly získávání pomocí cloudových scriptů, což je standardní přístup. Práce neobsahuje žádné zásadní technické nedostatky. Celková technická úroveň je odpovídající, ale některé pasáže například výsledky trénování modelů by si zasloužily podrobnější analýzu a především vyhodnocení. Vyhodnocování modelů je totiž jediná možnost, jak objektivně dokumentovat výslednou kvalitu řešení a je to také nástroj, jak bezpečně aplikovat výsledný model v praxi. Tato část práce

by tedy vyžadovala další dopracování.

Formální a jazyková úroveň, rozsah práce	**B - velmi dobře**
Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.	
Práce je psaná v anglickém jazyce. Práce má dobrou typografickou kvalitu a odpovídá obvyklým standardům. Všechny formální požadavky jsou splněny.	

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.	
Diplomant systematicky shromažďoval a studoval nepřeberné množství internetových zdrojů a literatury. Z práce je zřejmé, že jsou informace z těchto zdrojů zapracovány do řešení. Práce obsahuje korektní citace odkazovaných zdrojů v souladu s citačními praktikami. Uvedené citace dodržují standardní citační formát a veškeré převzaté prvky jsou řádně odlišeny od vlastních výsledků a úvah.	

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.
Výsledkem práce je trénovací sada, kterou lze okamžitě použít.

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.

Diplomová práce jako celek prokazuje schopnost diplomanta řešit danou úlohu návrhu robustních postupů pro sběr a předzpracování datového korpusu. Postup by zasluhoval další dopracování v oblasti vyhodnocení trénovaných modelů, ale celkově je na velmi dobré úrovni.

Pozitivním přínosem výstupů práce je vytvoření sady českých textů pro trénování českých LLM. Tato sada je vyčištěná a deduplikovaná, takže je možné jí hned využít pro vývoj nových českých modelů.

Otázka: Jaký by byl vhodný postup k vyhodnocení výsledných modelů.

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **B - velmi dobře.**

Datum: 7.6.2024 Podpis: