Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Michaela Hubená
Název práce	Automatické generování Einsteinových hádanek v přirozeném jazyce
Rok odevzdání	2023
Studijní program	Informatika
Specializace	Programování a vývoj software

Autor posudku	Jan Hajič, Ph.D.	Oponent
Pracoviště	Ústav formální a aplikované lingvistiky	

K celé práci

	lepší	OK	horší	nevyhovuje
Obtížnost zadání		X		
Splnění zadání			X	
Rozsah práce *... textová i implementační část, zohlednění náročnosti*			X	

Práce implementuje generátor Einsteinových hádanek. Proces má dvě roviny: napřed se vygeneruje struktura hádanky na úrovni její vnitřní logiky, tedy pouze sady symbolů a sada nápověd různých druhů, a poté se tato struktura "přeloží"do přirozeného jazyka – češtiny – pomocí (téměř) state-of-the-art jazykového modelu ChatGPT 3.5. Kapitola 1 představuje Einsteinovy hádanky a předkládá dobrou analýzu a formalizaci nápověd. V kapitole 2 jsou stručně (bez vzorečků a bez diagramů) popsány Transformery jakožto jazykové modely. Kapitola 3 potom popisuje činnost aplikace: jak vygeneruje validní hádanku a jak poté konvertuje hádanku do přirozeného jazyka pomocí promptů pro ChatGPT. Kapitola 4 přináší dokumentaci – uživatelskou i vývojovou (neboť obě jsou poměrně stručné). Kapitola 5 poté přináší nejzajímavější část práce, a to analýzu vygenerovaných textových hádanek.

Silnou stránkou práce je především dobře navržená formalizace Einsteinových hádanek a skutečnost, že autorka věnovala prostor i analýze výsledků. Oceňuji i skutečnost, že jsou hádanky generovány v češtině. Další poměrně silnou stránkou je zpracování vývojové dokumentace a referenční dokumentace pomocí nástroje Sphinx, a také relativně čistý kód samotné implementace zachovávající zvyklosti jazyka Python.

Za slabou stránku práce považuji její rozsah. Bakalářská práce je zadaná jako softwarové dílo, avšak celý vlastní zdrojový kód se vejde do čtyř poměrně malých souborů o celkem ne více než 1500 řádkách. To by nebyl problém u experimentální práce, avšak na to zase text obsahuje příliš málo diskuse o alternativních metodách než ChatGPT (tj. žádnou). Už libovolné GUI (např. obalit aplikaci jako web pomocí knihovny Django) by stačilo, aby rozsah problém nebyl.

Software také neumožňuje uživateli ověřit řešení hádanky.

(pokračování na další straně)

Druhou slabou stránkou práce je minimum analýzy. Celý text působí jako technická zpráva o tom, co bylo provedeno, a jen málo se dotýká toho, *proč* bylo dílo provedeno právě tak. Zadání zmiňuje jako možnosti nástroje pro syntax a morfologii jako UDPipe či předtrénované embeddings jako FastText – vedle prokazatelných možností ChatGPT je jejich nepoužití zcela pochopitelné, nicméně v diskusi použité metody "přeložení"hádanky do přirozeného jazyka tyto metody zmíněny být mohly – se svými nedostatky, které volbu ChatGPT zdůvodňují. ChatGPT ale má i svoje nedostatky, např. nutnost mít API token pro OpenAI a připojení k internetu – v práci chybí diskuse toho, zda jsou tyto nedostatky podstatné při předpokládanem využívání aplikace. Je chvályhodné, že se v práci měří, kolik procent vygenerovaných hádanek je po konverzi do češtiny řešitelných, nicméně práce nejde za toto pozorování dál – považuje se dané procento neřešitelných hádanek za problém? Pokud ano, lze nepovedené hádanky nějak filtrovat? To je další aspekt tvorby softwarové aplikace, který v práci nebyl nijak zohledněn: kdo by takovou aplikaci mohl využít? Jaké má přesně výhody a nevýhody oproti aplikacím popisovaným na závěr kapitoly 1? Jaká uživatelská rozhraní by byla pro dané cílové skupiny vhodná?

Protože se tato analýza nevyskytuje v textu, mám jako dotazy pro obhajobu mám následující:

1. Kdo by mohli být uživatelé aplikace? Jak by mohly vypadat skutečné use-cases? Proč by Co by případně bylo třeba v aplikaci změnit, aby byla pro dané cílové skupiny a příklady užití jednoznačně vhodná?

2. Byla při návhru práce uvažována rizika spojená s používáním ChatGPT a OpenAI, jako (1) proprietárnost rozhraní a měnící se pricing, (2) potřeba síťového připojení pro cílovou skupinu?

3. Jaké alternativní metody než ChatGPT byste pro překlad hádanek do přirozeného jazyka navrhla, pokud by OpenAI API bylo nedostupné? Jaké budou mít oproti ChatGPT výhody a nevýhody?

Pokud budou tyto dotazy při obhajobě zodpovězeny uspokojivě, práci doporučuji k obhajobě.

Textová část práce

	lepší	OK	horší	nevyhovuje
Formální úprava *... jazyková úroveň, typografická úroveň, citace*		X		
Struktura textu *... kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*		X		
Analýza			X	
Vývojová dokumentace		X		
Uživatelská dokumentace		X		

Text je srozumitelný, jasný a dobře strukturovaný. Citace a bibliografie jsou většinou v pořádku, byť internetové zdroje (především ty, které nemají charakter vědecké publikace) by měly být opatřeny v bibliografii URL a datem, kdy byly citovány. Bibliografie je minimální (pouze devět položek, z toho pouze pět publikací, což reflektuje ale aplikační spíše než výzkumný charakter práce. Výjimečně se v textu vyskytovaly hrubé chyby (např. sekce 5.1, první odstavec: „Vzory pro generování neobsahovali...").

(pokračování na další straně)

Uživatelská dokumentace standardně v souboru README.md je v pořádku - jedná se o konzolovou aplikaci, takže toho není třeba příliš. Nápověda funguje, jen bych ještě implementoval spuštění `einstein` bez parametrů jako ekvivalent `einstein -h`. Vývojová dokumentace je správně vygenerována s použitím nástroje Sphinx, ačkoliv docstrings jsou spíše minimální a předpokládají, že programátor čte jako součást dokumentace i kapitolu 1 textu práce.

Implementační část práce

	lepší	OK	horší	nevyhovuje
Kvalita návrhu *... architektura, struktury a algoritmy, použité technologie*		X		
Kvalita zpracování *... jmenné konvence, formátování, komentáře, testování*			X	
Stabilita implementace			X	

Samotná implementace nemá zásadní problémy. Jsou zvolené vhodné technologie (programovací jazyk, knihovny, dokumentační systém). Objektový návrh hádanky je rozumný. Jen bych asi nechtěl mít flag "used"u pravidel – udělal bych zvlášť třídu Solver, která je zodpovědná za běh řešení, protože myslím, že pravidla by neměla mít stav. Webová komunikace jako taková je správně oddělená (ve zvláštní funkci, ačkoliv by mohla možná mít spíše svou třídu), je správně použitý existující balíček v Pythonu implementující API pro OpenAI přímo.

V práci nejsou implementovány unit testy, ani jako doctesty. Očekával bych testy alespoň pro vše, co je součástí funkce, která umí Einsteinovu hádanku na základě množiny pravidel vyřešit (a ověřuje tak řešitelnost, případně jednoznačnost). Komentáře nad rámec docstringů nejsou, docstringy samotné jsou nicneříkající, avšak program málokdy vyžaduje více. Funkce `try_solve` by však lépe popsaná být měla.

Větší problémy jsou ve stabilitě, resp. při užívání programu. Používá se zasataralá metoda instalace přímo přes setup.py, což je sice nepříjemné, ale zároveň stále ještě ne neobvyklé. Horší však je, že se nikde v dokumentaci nespecifikuje vyžadovaná verze Pythonu (3.9), takže samotná instalace bez ručního čtení setup.py selhává. Závislost na balíčku `openai` je specifikována pouze v setup.py opět bez verze (ačkoliv zde naštěstí po instalaci správné verze pythonu se balíček již nainstaluje správně), standardem pro závislosti je při distribuci programů v Pythonu soubor `requirements.txt`.

Při běhu programu není uživatelsky přívětivě ošetřena situace, kdy chybí API klíč pro komunikaci se servery OpenAI: funkce `set_apikey()` by na chybějící hodnotu měla reagovat explicitní chybovou hláškou, a ne nastavením potenciálně prázdné hodnoty. Nikde v uživatelské dokumentaci se také nezmiňuje rate limiting OpenAI, který má na použitelnost programu zásadní vliv. V metodě, která se serverem OpenAI skutečně komunikuje, je sice pět pokusů se zvyšujícím se časovým odstupem, avšak toto by mohlo být konfigurovatelné.

Generování hádanek v symbolické reprezentaci však funguje spolehlivě a rychle.

Zároveň by se zde býval hodil způsob, jak vygenerovat textové zadání sice méně inteligentně než co umí ChatGPT, ale zato i v přítomnosti síťových potíží (viz výše). Na druhou stranu je pravda, že celá aplikace v podstatě nemá skutečné cílové publikum, a jako příklad obohacení výpočetní úlohy o rozhraní v přirozeném jazyce pomocí state-of-the-art metod a API to je pro autorku jistě dobrý začátek pro vývoj dalších aplikací využívajících nejnovější "velké"modely AI.

Celkové hodnocení Dobře

Práci navrhuji na zvláštní ocenění Ne

Datum Podpis