

Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce Jan Černý
Název práce Měření lexikálního překvapení v právních textech
Rok odevzdání 2025
Studijní program Informatika
Specializace Umělá inteligence

Autor posudku Ivana Kvapilíková **Role** Vedoucí
Pracoviště Ústav aplikované a formální lingvistiky

Prosím vyplňte hodnocení křížkem u každého kritéria. Hodnocení *OK* označuje práci, která kritérium vhodným způsobem splňuje. Hodnocení *lepší* a *horší* označují splnění nad a pod rámec obvyklý pro bakalářskou práci, hodnocení *nevyhovuje* označuje práci, která by neměla být obhájena. Hodnocení v případě potřeby doplňte komentářem. Komentář prosím doplňte všude, kde je hodnocení jiné než *OK*.

K celé práci	lepší	OK	horší	nevyhovuje
Obtížnost zadání	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Splnění zadání	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Rozsah práce ... <i>textová i implementační část, zohlednění náročnosti</i>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<p>Komentář Bakalářská práce se věnuje tématu měření srozumitelnosti textu pomocí konceptu s názvem lexikální překvapení. Autor tento koncept nejpre formálně uvádí a následně popisuje implementaci nástroje Glitter, který ho dokáže měřit.</p> <p>Práce je motivovaná snahou o zlepšení srozumitelnosti úředních textů, které jsou cíleny na běžné občany, avšak svou formou jsou často těžko pochopitelné pro člověka bez právního vzdělání. Tato motivace je jasně uvedena na začátku práce a autor se k ní vrací i na konci, kdy ilustruje použití Glitteru právě na úryvcích z úředních textů.</p> <p>Práce úzce souvisí s komplexními tématy jazykového modelování a informační hodnoty textu. Pro měření lexikálního překvapení používá autor jazykové modely typu transformer. V začátku práce popisuje, jak takové modely fungují a jak jsou trénovány. Následně v implementační části uvádí, jakým způsobem z těchto modelů získává pravděpodobnosti pro vyhodnocení překvapivosti jednotlivých textových tokenů.</p> <p>Práce se věnuje důležitému problému a používá pokročilé technologie pro jeho řešení. Přístup je postaven na solidní teorii informační hodnoty textu. Zajímavé jsou výsledky analýzy skutečných úředních textů, kde se ukazuje, že jedním z indikátorů srozumitelnosti je délka neinformativních segmentů - tedy pasáží, které čtenáře zatěžují, aniž by přinášely podstatné informace.</p>				

Textová část práce	lepší	OK	horší	nevyhovuje
Formální úprava ... <i>jazyková úroveň, typografická úroveň, citace</i>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Struktura textu ... <i>kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu</i>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Analýza	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Vývojová dokumentace	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Uživatelská dokumentace	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<p>Komentář Teoretická část práce je poměrně stručná, ale věnuje se všem důležitým oblastem: motivace, cíle, popis problému, počáteční hypotéza, úvod do teorie lexikálního překvapení, informační entropie a jazykového modelování. Součástí teoretické části jsou odkazy na několik studií zabývajících se srozumitelností textu i na několik základních zdrojů pro jazykové modelování.</p> <p>Technická část je podrobná včetně detailního popisu zvolených abstrakcí a kvalitní uživatelské dokumentace. Autor popisuje, jak používat grafické i aplikační rozhraní, dále jak používat nástroj v příkazové řádce a jak ho spustit na serveru.</p> <p>V závěru práce autor ilustruje výsledky s použitím dvou různých jazykových modelů (RobeCzech a GPT2-XL-Czech) na několika vybraných textech a popisuje výsledky frekvenční analýzy na vybraných částech KUK korpusu obsahujících redigované úřední texty. Věnuje se i limitacím stávajícího řešení a možnostem budoucího rozvoje.</p> <p>Celkově je text přehledně strukturovaný a dobře čitelný, obsahuje pouze několik drobných chyb formálního charakteru.</p>				

Implementační část práce	lepší	OK	horší	nevyhovuje
Kvalita návrhu ... <i>architektura, struktury a algoritmy, použité technologie</i>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Kvalita zpracování ... <i>jmenné konvence, formátování, komentáře, testování</i>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Stabilita implementace	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<p>Komentář Implementace nástroje Glitter je z technického pohledu velmi kvalitní a promyšlená. Architektura je modulární a rozšiřitelná.</p> <p>Autor použil programovací jazyk Python. Základem pro práci s jazykovými modely je třída GlitterModel s jednotným rozhraním pro generativní i unmasking modely, což umožňuje do budoucna snadnou rozšiřitelnost pro nové modely obou typů. Autor pracoval s neuronovými modely pomocí knihovny PyTorch a během tvorby bakalářské práce prokázal dobré porozumění jejich vnitřnímu fungování.</p> <p>Glitter zpracovává text postupně pomocí abstrakce ContextWindow, které pluje textem a umožňuje i zpracování dlouhých textů přesahujících délku kontextového okna používaných modelů. Nástroj si ukládá zpracované segmenty pro úsporu času při opětovném generování.</p> <p>Součástí nástroje je webové rozhraní pro vizualizaci lexikálního překvapení přímo v textu. Pro analýzu většího množství dokumentů se dá spustit z příkazové řádky a výsledky exportovat.</p>				

Celkové hodnocení Výborně
Práci navrhuji na zvláštní ocenění Ne

Datum 10. června 2025

Podpis