Posudek oponenta bakalářské práce

Autor práce: **David Bubik**

Název práce: **Anotační aplikace pro úlohy z oblasti zpracování přirozeného jazyka**

Tématem práce Davida Bubika bylo vytvoření aplikace, která umožní anotaci dat pro vybrané úlohy z oblasti zpracování přirozeného jazyka (Natural Language Processing – NLP). Text práce je psán českým jazykem, nicméně občas se objeví překlepy a jednopísmenné předložky a spojky na konci řádků. Celkově je text na poměrně dobré úrovni a členění na jednotlivé kapitoly a sekce dává smysl.

V první teoretické části práce autor vysvětluje motivaci, představuje vybrané úlohy z oblasti NLP a také nabízí průzkum již existujících anotačních nástrojů. Na závěr je uvedena tabulka, která shrnuje jejich výhody a nedostatky. Ve druhé polovině své práce autor popisuje vyvíjenou aplikaci, včetně popisu implementace a testování. Aplikace je psána v jazyce Java za použití frameworku *Vaadin.* Jako databáze jsou použity MySQL a Elasticsearch.

Instalace a zprovoznění pomocí Dockeru byla bezproblémová a aplikaci se mi podařilo úspěšně spustit. V rámci administračního i anotátorského režimu jsem vyzkoušel všechny akce, které popisuje uživatelská příručka. Aplikace působí **intuitivním a přehledným dojmem**, nabízí celkem šest anotačních scénářů, které pokrývají hlavní úlohy v oblasti zpracování přirozeného jazyka (klasifikace textu, rozpoznávání pojmenovaných entit, sumarizace, analýza sentimentu a další).

Administrační mód

V rámci tohoto módu jsem narazil na **chybu v aplikaci při exportování dokončené anotované datové sady**. Stalo se, že v některých případech mi označená data nešla stáhnout (chyba: „žádný soubor“) a to jak v případě exportu do *JSON*, tak i do *TXT*. Zbývající akce (nahrání datové sady, vytvoření scénáře, pozvání anotátora) byly v pořádku.

Aplikace umožňuje porovnání anotací v případě, že na datové sadě pracuje více než jeden anotátor. Zpravidla jsou data při anotaci rozdělena mezi anotátory tak, že každý vzorek je anotován vždy více než jedním anotátorem. Pak lze spočítat anotátorskou shodu, která je jedním z důležitých ukazatelů kvality datové sady. Aplikace přehledně ukáže vzorky, ve kterých se anotátoři neshodli a navíc je k dispozici tzv. inter-annotator agreement Cohen’s Kappa – κ. Pro větší názornost a představu bych ale doplnil rovněž údaj o jednoduché anotátorské shodě uváděnou v procentech (kolik % vzorků označili anotátoři stejně).

Při vytváření scénáře u úlohy „Text classification“, je možnost zaškrtnutí multi-label a **force label**, přičemž force label není nikde vysvětlen. Nevšiml jsem si rozdílu se zapnutou volbou force label a bez ní.

Dále bych jen stručně zmínil dvě připomínky, které by mohly přispět k větší přehlednosti.

- Při větším množství scénářů bych je v jejich přehledu barevně označil a oddělil tak scénáře nové, rozpracované nebo již dokončené.
- Pro označenou datovou sadu (dokončený scénář) bych umožnil detailní pohled a statistiky (např. distribuce jednotlivých tříd, vyváženost datasetu apod.).

Anotátorský mód

V tomto jednoduchém módu jsou pouze dvě akce: přijmutí pozvání od administrátora do skupiny a samotná anotace. Vše bylo funkční, postrádal jsem jen tlačítko pro nahlášení problematického vzorku dat, např. když si anotátor není jistý co má označit. Oceňuji ukazatel progresu (kolikátý vzorek z kolika právě anotuji), možná by bylo lepší použít nějaký grafický prvek (např. progress bar ukazjící procenta).

Závěr

Aplikace je funkční a z mého pohledu i použitelná pro anotování. Zadání bylo splněno ve všech bodech. Text práce je na dobré úrovni a ocenil bych velké množství UML a jiných diagramů pro popis architektury programu. V seznamu literatury bych vytkl absenci datumu náhledu (datumu citace) u online zdrojů.

Celkově navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

Dotazy k práci

1) Dokázal byste vysvětlit výhody metriky Cohen' Kappa koeficientu oproti běžné percentuální anotátorské shodě?

2) Jak se budou řešit problematické situace, kdy se anotátoři neshodnou na nějakém vzorku dat?

V Plzni 20.7.2022 Ing. Jiří Martínek, Ph.D.