

## Posudek oponenta diplomové práce

**Student:** Holubec Michael, Bc.  
**Téma:** Uspořádání fragmentů textu s pomocí jazykového modelu (id 23379)  
**Oponent:** Kocour Martin, Ing., UPGM FIT VUT

- 1. Náročnost zadání** **obtížnější zadání**  
Obtížností zadania spočíva hlavne v nedostatku dát. Autor si v rámci práce musel sám označiť správne poradie čítania textových regiónov v dokumente. Takto vytvorený dataset využil pri vyhodnocovaní svojho riešenia. Nový dataset obsahuje články z českých novin a sám o sebe je tiež určitým prínosom.
- 2. Splnění požadavků zadání** **zadání splněno s podstatným rozšířením**  
Autor rozšířil prácu o tzv. priestorovú analýzu, ktorá využíva priestorové informácie o regióne na správne určenie poradia. Tento prístup dosahuje značne lepších výsledkov než pôvodné riešenie, založené len na jazykovej analýze. Autor navyše navrhol riešenie, v ktorom kombinuje oba prístupy.
- 3. Rozsah technické zprávy** **je v obvyklém rozmezí**
- 4. Prezentací úroveň předložené práce** **80 b. (B)**  
Prvá polovica správy s teoretickým úvodom je prehľadná a dobre sa číta. Problematická je druhá polovica, kde sa autor najprv venuje implementácii (Kapitola 4) a až následne opisuje navrhované riešenie (Kapitola 5). Niektoré časti navrhovaného riešenia by bolo dobré lepšie vysvetliť (napr. 5.2.2).
- 5. Formální úprava technické zprávy** **75 b. (C)**  
Správa je písaná v češtine. V texte sa objavuje zopár preklepov. Formálnu úpravu hodnotím ako priemernú. V opise priestorovej analýzy (5.3) sa vyskytujú fragmenty s kódom, ktoré pôsobia skôr rušivo a nemajú žiadnu výpovednú hodnotu. Na druhú stranu práca obsahuje niekoľko užitočných schém a obrázkov, ktoré čitateľovi pomáhajú lepšie pochopiť zámer autora.
- 6. Práce s literaturou** **95 b. (A)**  
Práca sa odkazuje na relevantné zdroje.
- 7. Realizační výstup** **95 b. (A)**  
Realizačný výstup je prehľadný, dobre štrukturovaný kód v jazyku Python. Vytvorený nástroj bol použitý k získaniu experimentálnych výsledkov. Nástroj mi bol taktiež osobne odprezentovaný. V zdrojových kódoch chýba záhlavie s menom autora.
- 8. Využitelnost výsledků**  
Táto práca nadväzuje na projekt PERO, ktorý si kladie za cieľ digitalizovať historické dokumenty. Autor celkom jasne ukázal, že na určenie poradia textových regiónov nestačí analyzovať ich nadväznosť len pomocou jazykového modelu, ale je potrebné použiť aj ich priestorové údaje. Výsledný systém pracuje celkom spoľahlivo na dokumentoch, kde sú textové regióny z jedného celku zreteľne oddelené medzerou od ostatných.
- 9. Otázky k obhajobě**  
V technickej správe spomínate jazykový model založený na LSTM neurónovej sieti, experimentovali ste aj s inými typmi jazykových modelov?
- 10. Souhrnné hodnocení** **90 b. výborně (A)**  
Aj napriek drobným nedostatkom hodnotím prácu veľmi kladne. Oceňujem hlavne šírku naštudovania danej problematiky. Po praktickej stránke autor odviedol kus poctivej roboty. Neuspokojil sa s nepresnou jazykovou analýzou. Namiesto toho navrhol nové riešenie založené na priestorových informáciách, ktoré taktiež zkomboval spolu s jazykovým modelovaním. Práca bola publikovaná na študentskej konferencii Excel@FIT.

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 4. června 2022

Kocour Martin, Ing.  
oponent