Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Prokop Divín
Název práce	Srovnání sekvenční a strukturních metod strojového učení pro predikci protein-ligand vazebných reziduí
Rok odevzdání	2023
Studijní program	Informatika
Studijní obor	Umělá inteligence
Autor posudku	Škoda Petr Oponent
Pracoviště	Katedra softwarového inženýrství

K celé práci

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

Cílem práce bylo porovnat metody predikce vazebných ze sekvence a P2Rank. Autor v práci vybere několik metod strojového učení, které zkombinuje s vybranými způsoby reprezentace reziduí. Chybí mi přehled existujících / publikovaných metod, které by bylo možné využít.

Složitost práce vidím zejména v doménové specificitě a využití výpočetního clustru, zejména pokud si musel řešitel sám připravit běhové prostředí.

Při porovnání s P2Rankem pak není jasné, zda-li byl použitý výstup P2Ranku po filtraci (bod 4 stránka 14). Pokud ano, bylo by zajímavé podobnou filtraci provést i na výstupech sekvenčních metod.

Autor v práci dále tvrdí, cituji: "modely používající embeddingy se dokázali značně přiblížit k výsledkům P2rank."Hodnota MCC/Recall je přitom 0.548/0.7 pro P2Rank a 0.394/0.275 pro T5 embedding. Zde mi přijde, že mohla být provedena hlubší analýza a širší porovnání.

Textová část práce

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

Textová, i implementační, část práce obsahuje gramatické chyby v podobě chybějících, či prohozených znaků. Překlepy jsou zajímavé zejména v kódu, kde by je mělo rozumné IDE dnešní doby samo detekovat.

Vývojová dokumentace se soustředí na popis skriptů. Uživatelská je pak pokryta README soubory v GitHub repositáři. Vzhledem k povaze práce to považuji za dostatečné.

Celkově tedy práce obsahuje všechny potřebné části v přijatelném rozsahu. Čitelnost místy kazí opakování sdělení, například strana 8 kolem začátku sekce 1.2 strana 8. Nicméně, pokud práci nečteme příliš pozorně, tak je to naopak nápomocné

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	X	

Chybí popis instalace a požadavků na běhové prostředí, například použitelná verze Pythonu. Postup spuštění je pak nutné vyčíst z repozitáře, případně README souborů.

V souboru requirements.txt chybí knihovna bio. Bez její instalace není možné spustit některé skripty.

Spuštění příkazu

```
python checkData.py \
  trainData/bindings_labeled \
  trainData/fasta-raw \
  trainData/residues \
  datasets
```

Vyprodukuje velké množství chybových logů. Neb skript samotný doběhne, není jasné, jestli se jedná o chyby, které je třeba řešit. Uživatel také není informován o tom, jak by mohl chyby řešit.

Python skripty mohou využívat knihovny, nebo modulu pro zpracování argumentů z příkazové řádky. Některé skripty toto zpracování provádí ručně a neposkytují žádnou nápovědu ke svému spuštění.

Obecně je poměrně značný rozdíl mezi skripty ve složkách pbsPrediction a trainData. Kód je nekonzistentně formátován. Je tak možné najít třeba:

```
path=""
...
chain = name[-1]
...
chain =line[0]
```

A to v jednom souboru. Kód by vypadal lépe, pokud by byl použití linter/prettier. Repozitář projektu obsahuje prázdné soubory, například "./pbsPrediction/zk.py".

Přinejmenším překvapivé je využití CSV formátu ve složce ./pbsPrediction/finalResults. Autor využívá kombinace prázdných řádků a řádku s jednou hodnotou pro oddělení skupin hodnot. Tento přístup mi přijde přinejmenším nevhodnou praxí.

Výše uvedené jsou jen některé příklady neobvyklosti či nedotažeností v softwarové části práce.

Celkové hodnocení Velmi dobře (spíše horší)

Práci navrhuji na zvláštní ocenění Ne

Datum 25.8.2023 Podpis