Домашна задача 4

1. Насоки

За потребите на оваа домашна задача користете ги податочните множества movies_metadata.csv и ratings_small.csv достапни на дадените линкови.

2. Значење на колоните

Во рамки на податочното множество movies_metadata.csv достапни ви се следните колони:

- id - идентификатор на филмот
- adult - дали има ограничување за возрасни (TRUE, FALSE)
- budget - буџет за снимање на филмот
- genres - жанрови на филмот
- original_language - јазик на којшто оригинално е снимен филмот
- title - наслов на филмот
- overview - опис на филмот
- popularity - оцена за популарност на филмот (нерекината вредност)
- production_companies - компании одговорни за продукција на филмот
- production_countries - држави од кои потекнува продукицијата на филмот
- revenue - приход генериран од филмот
- runtime - времетраење на филмот (минути)
- vote_average - просечна оцена (од вкупно 10)
- vote_count - број на оцени

Во рамките на податочното множество ratings_small.csv достапни се следните колони:

- userId - идентификатор на корисникот кој го оценил филмот
- movieId - идентификатор на филмот
- rating - оцена која ја дал корисникот за филмот (од 1 до 5)

Двете множества може да ги споите преку movieId од ratings_small.csv и id колоната од movies_metadata.csv.

3. Задачи

1. Од колоната overview од податочното множество movies_metadata.csv генерирајте тројки со користење на некој јазичен модел (може да биде REBEL како во аудиториските

вежби, но може да пробате и друг модел по избор. Дозволено е да користите и некој чет модел како ChatGPT).

2. Креирајте хетероген граф којшто ќе содржи јазли за корисници и филмови.

2.1 За филмовите додадете атрибути од множеството movies_metadata.csv (adult, budget, genres, original_language, popularity, итн.). Некои од атрибутите може да ги претставите како посебен тип на јазол и да ги поврзете со соодветните филмови (може да изберете самите на кој начин ќе го направите тоа).

3. Креирајте врски помеѓу филмовите и корисниците на таков начин што врска ќе постои доколку корисникот го оценил филмот со оцена поголема или еднаква на 3.

Потребно ќе биде да изградите два модели за предвидување на врски (link prediction) врз хетерогениот граф.

4. Изградете го првиот модел. Потребно е да ги користи само филмовите (со нивните атрибути) и корисниците како јазли, како и врските меѓу нив (user-rating-movie). Целта е да се предвиди постоење на нови линкови од истиот тип (user-rating-movie). Направете поделба на линковите во следниот формат 80% за тренирање, 10% за валидација и 10% за тестирање.

4.1 Бидејќи нема атрибути за јазлите од типот корисници, креирајте ембединг слој во рамките на моделот со кој ќе се обидете да ги научите паралелно и карактеристики за корисниците.

5. Евалуирајте го моделот со користење на метриката ROC AUC.

6. Во графот додадете јазли и врски кои одговараат на тројките кои ги екстрахиравте од описот на филмот во првиот чекор (не мора да се додадат сите типови на тројки, само оние чии типови ги имате во доволно голем број во однос на големината на податочното множество). Поврзете ги овие тројки со соодветниот јазол - филм од чиј опис ги имате екстрахирано.

7. Изградете го вториот модел со користење и на дополнителните јазли и врски изгенерирани од описот на филмот (секако вклучете ги јазлите и врските кои беа вклучени во првиот модел). На крајот, целта повторно е да се предвиди постоење на линкови од типит user-rating-movie. Направете поделба на линковите во следниот формат 80% за тренирање, 10% за валидација и 10% за тестирање.

7.1 Обидете се да креирате ембединг за различните типови на јазли бидејќи освен за филмовите, за ниту еден друг тип на јазол нема атрибути. Креирајте посебен ембединг слој за секој тип на јазол во рамките на моделот и научете ги карактеристиките паралелно со тренирањето.

8. Евалуирајте го моделот со користење на метриката ROC AUC.

Кој модел е подобар?

РЕШЕНИЕТО НА ЗАДАЧАТА ТРЕБА ДА ГО ПОСТАВИТЕ КАКО COLAB ИЛИ JUPYTER NOTEBOOK ВО РАМКИТЕ НА КУРСОТ. НЕМА ПОТРЕБА ДА ПРИКАЧУВАТЕ БИЛО КАКОВ ДРУГ ДОКУМЕНТ.