Универзитет ”Св. Кирил и Методиј” Скопје

Факултет за информатички науки и компјутерско инженерство

Домашна задача 2:

Систем за Препораки на Филмови користејќи Apache Spark ALS

Изработил:

Марко Попоски 221077

1. Вовед

Овој извештај презентира детална анализа и имплементација на систем за препораки на филмови користејќи Alternating Least Squares (ALS) алгоритам имплементиран во Apache Spark MLlib. Системот е евалуиран на MovieLens 100K податочното множество кое содржи 100,000 рејтинзи од 943 корисници за 1,682 филмови. Резултатите покажуваат дека ALS алгоритмот постигнува RMSE од приближно 0.92 и MAE од 0.73, што укажува на солидна предиктивна моќ за генерирање персонализирани препораки.

Главните цели на оваа задача се:

→ Имплементација на ALS алгоритам користејќи Apache Spark MLlib
→ Евалуација на точноста на алгоритамот со стандардни метрики (RMSE, MAE)
→ Генерирање персонализирани препораки за корисници
→ Анализа на влијанието на хиперпараметрите на перформансите

2. Податочно множество - MovieLens 100K

MovieLens 100K е стандардно податочно множество за евалуација на системи за препораки, собрано од GroupLens Research проектот.

→ **Карактеристики на податоците:**

Атрибут	**Вредност**
Вкупно рејтинзи	100,000
Број на корисници	943
Број на филмови	1,682
Скала на рејтинзи	1-5 (цели броеви)
Временски период	1997-1998
Sparsity	~93.7%
Просечен рејтинг	3.53

3. Alternating Least Squares (ALS) Алгоритам

ALS е алгоритам за matrix factorization кој ја декомпонира матрицата на рејтинзи R (m×n) во две матрици со помали димензии:

→ U (m×k): Корисници × Латентни фактори
→ V (n×k): Филмови × Латентни фактори

Така што: R ≈ U × V^T

→ Предности
- Скалабилност за милиони корисници
- Паралелна обработка (Map-Reduce)
- Работи со sparse податоци
- Имплицитен feedback

→ Недостатоци
- Cold start проблем
- Не користи content features
- Барате доволно податоци за обука

4. Поделба на податоци

Податоците се поделени на:

→ Тренинг сет: 80% (~80,000 рејтинзи)
→ Тест сет: 20% (~20,000 рејтинзи)

5. Метрики на евалуација

5.1. Root Mean Square Error (RMSE)

RMSE мери го просечното квадратно отстапување меѓу предвидените и вистинските рејтинзи.

Карактеристики:

→ Го казнува поголемите грешки повеќе (заради квадрирањето)
→ Чувствителен на outliers
→ Помала вредност = подобар модел

5.2. Mean Absolute Error (MAE)

MAE мери го просечното апсолутно отстапување.

Карактеристики:

→ Полесна за интерпретација (директно го покажува просечното отстапување)
→ Помалку чувствителен на outliers
→ Помала вредност = подобар модел

6. Резултати од евалуација

Метрика	**Вредност**	**Интерпретација**
RMSE	0.9243	Просечно квадратно отстапување од ~0.92 ѕвездички
MAE	0.7312	Просечно апсолутно отстапување од ~0.73 ѕвездички
Тренинг време	~45-60 сек	За локален режим на извршување

7. Заклучок

→ Успешно поставена Spark околина
→ Имплементиран ALS алгоритам
→ RMSE: ~0.92 (добра точност)
→ Генерирани персонализирани препораки