|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ «Информатика и системы управления»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Лабораторная работа №4**

**по курсу «Проектирование рекомендательных систем»**

|  |  |
| --- | --- |
| **Тема: «Сравнение алгоритмов матричной факторизации»**  **Студент Горячев В. Г.**  **Группа ИУ7-33М**  **Преподаватель Быстрицкая А. Ю.** |  |

Москва

2023 г

# Описание алгоритмов

Матричная факторизация — разложение матрицы на произведение двух или более матриц.

***SVD в рекомендательных системах***

Разложение матрицы осуществляется на такие матрицы: *,* что элементы, лежащие на главной диагонали матрицы — это сингулярные числа, а все элементы остальные элементы являются нулевыми. Это позволяет получить приближённую матрицы к матрице М, с известным рангом. Это позволяет выделять главные факторы, способствующие выбору того или иного товара и осуществлять, на основе этого приближения, предсказания-рекомендации для пользователей.

***NMF***

Факторизация неотрицательных матриц (NMF) — это представление матрицы V в виде произведения матриц W и H, в котором все элементы трёх матриц неотрицательны.

# Библиотеки

В качестве языка программирования был выбран язык Python вместе с интерактивной средой Jupyter Notebook, поскольку они предоставляют удобный инструментарий для исследования, в частности, для выполнения лабораторных работ. Это определило выбор библиотек — нужно было найти совместимые с языком программирования.

В качестве источника алгоритмов использовались материалы из интернета для написания алгоритмов, а также сами библиотеки. Использованы стандартные библиотеки – numpy, pandas и sklearn. В этот раз к ним добавилась библиотека Surprise, разработанная специально для задач, связанных с рекомендательными системами.

# Данные

Для сравнения алгоритмов был повторно использован учебный вариант набора данных MovieLens (<https://grouplens.org/datasets/movielens/>). Он содержит оценки пользователей по фильмам и специально составлен для подобного рода учебных задач.

# Сравнение

Чтобы упростить сравнение, использовался стандартный подход из машинного обучения – разделение набора данных на тренировочную и обучающую части и применение метрики, сравнивающей предсказания модели и истинные ответы. В данном случае использовалась метрика RMSE (корень среднеквадратической ошибки) и MAE (средняя абсолютная ошибка, которая сравнивала прогнозы оценок пользователей для объектов, отсутствующих в обучающем наборе, с оценками из тестового. Также была возможность посмотреть время работы алгоритмов в зависимости от количества записей в наборе.

Таблица 1 — Качество работы алгоритмов

|  |  |  |
| --- | --- | --- |
|  | **MAE** | **RMSE** |
| **SVD** | 0,6612 | 0.8661 |
| **NMF** | 0,7105 | 0,9243 |

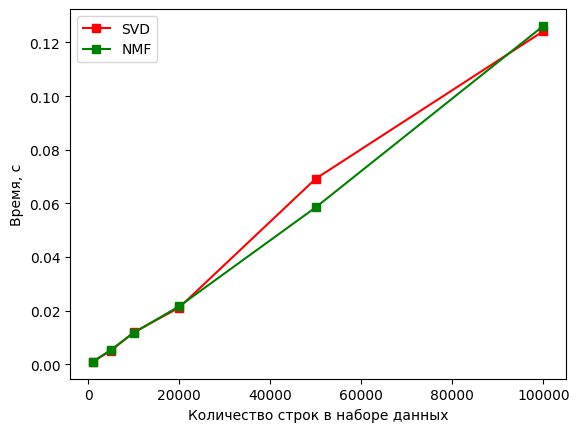


Рисунок 1 — Зависимость времени работы от размера тестовой выборки (по оси Х указан общий размер набора данных)

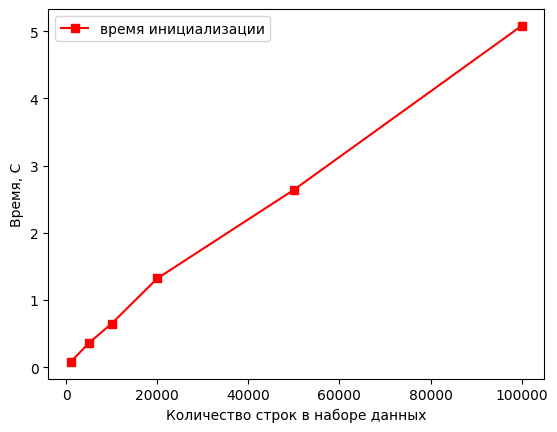


Рисунок 2 — Время инициализации алгоритмов в зависимости от размера набора данных (суммарно)

# Вывод

В ходе лабораторной работы были сравнены алгоритмы матричной факторизации. В данном случае алгоритмы повели себя схожим образом по времени: хотя алгоритму SVD на среднем количестве рекомендаций требуется больше времени, а на максимальном доступном размере – меньше, зависимость времени работы от количества данных, как и у NMF, имеет линейный характер. То же самое и с инициализацией – линейный характер графика, хоть и суммарного, позволяет сделать вывод, что и время настройки алгоритмов тоже имеет линейную зависимость. Что касается точности работы, MAE и RMSE алгоритма SVD на 7% и 6,3% соответственно лучше, чем результаты NMF. Это позволяет расценивать первый алгоритм как более удачный в рамках данной задачи с фильмами, получившими оценки от пользователей по 5-балльной шкале, хотя, конечно, результаты очень близки.