Коллаборативная фильтрация

# Постановка задачи

В современном мире часто приходится сталкиваться с проблемой рекомендации товаров или услуг пользователям какой-либо информационной системы. В старые времена для формирования рекомендаций обходились сводкой наиболее популярных продуктов. Но со временем такие рекомендации стали вытесняться таргетированными (целевыми) предложениями: пользователям рекомендуются не просто популярные продукты, а те продукты, которые наверняка понравятся именно им.

Такая же ситуация обстоит и с книгами. Определить вкусы читателя подчас бывает довольно сложно, а вероятность того, что ему понравятся книги, которые читает n% населения равна n%. В то время как коллаборативная фильтрация, которую я и рассматриваю в данной работе, позволяет определять предпочтения читателей с гораздо большей вероятностью. И чем больше данных собраны в системе, чем больше параметров, по которым можно классифицировать книгу, тем точнее будет рекомендация.

# Теоретический раздел

**Совместная фильтрация** ( **CF** ) представляет собой метод , используемый рекомендательные системы . Совместная фильтрация имеет два чувства, узкие один и более общий.

В новом, более узком смысле, совместная фильтрация является методом создания автоматических прогнозов (фильтрация) об интересах пользователя, собирая предпочтения или вкус информации от многих пользователей (взаимодействующих). Лежит предположение о совместной фильтрации подхода заключается в том, что если человек имеет такое же мнение как человека *B* по вопросу, А, скорее всего, есть мнение B на другой проблеме , чем у случайно выбранного человека. Например, совместная система фильтрации рекомендации по телевизионным вкусам может делать прогнозы о том, какие телешоу пользователь хотел бы дан частичный список вкусов этого пользователя (нравится или не нравится). Обратите внимание , что эти предсказания являются специфическими для пользователя, но и использовать информацию , собранную из многих пользователей. Это отличается от более простого подхода дает среднее (неспецифической) оценку для каждого элемента , представляющего интереса, например , на основе его количества голосов.

В более общем смысле, совместная фильтрация представляет собой процесс фильтрации для получения информации или моделей с использованием методов с участием взаимодействия между несколькими агентами, точками зрения, источниками данных и т.д. Применение совместной фильтрации, как правило, связаны с очень большими наборами данных. Совместные методы фильтрации были применены ко многим различным видам данных, в том числе: зондирования и мониторинга данных, например, в разведке полезных ископаемых, зондирования окружающей среды на больших площадях или нескольких датчиков; финансовые данные, такие как институты финансовых услуг, которые объединяют множество финансовых источников; или в электронной коммерции и веб-приложений, где акцент делается на пользовательских данных и т.д. Остальная часть этого обсуждения основное внимание уделяется совместной фильтрации для пользовательских данных, хотя некоторые методы и подходы могут применяться к другим основным приложениям, а также.

Интуитивно понятно, что для рекомендации пользователю №1 какого-либо продукта, выбирать нужно из продуктов, которые нравятся каким-то пользователям 2-3-4-etc., которые наиболее похожи по своим оценкам на пользователя №1. Как же получить численное выражение этой «похожести» пользователей? Допустим, у нас есть M продуктов. Оценки, выставленные отдельно взятым пользователем, представляют собой вектор в M-мерном пространстве продуктов, а сравнивать вектора мы умеем. Среди возможных мер можно выделить следующие:

1. Косинусная мера
2. Коэффициент корреляции Пирсона
3. Евклидово расстояние
4. Коэффициент Танимото
5. Манхэттенское расстояние и т.д.

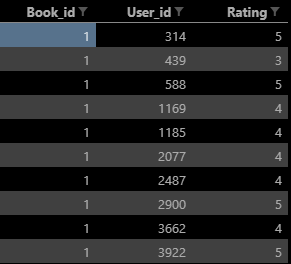
Основное допущение этого метода состоит в следующем: те, кто соглашался в прошлом, склонны согласиться и в будущем. Например, коллаборативная фильтрация или рекомендательная система по музыкальным вкусам способна прогнозировать, какая музыка понравится пользователю, имея неполный список его предпочтений (симпатий и антипатий).

# Dataset

В качестве входных данных мы будем рассматривать две таблицы:

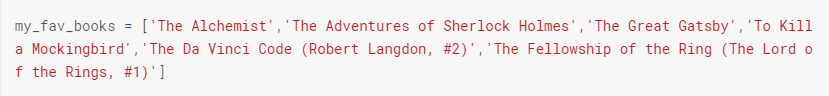
таблицу книг из 10000 записей

и таблицу с оценками этих книг другими читателями (981757 записей).

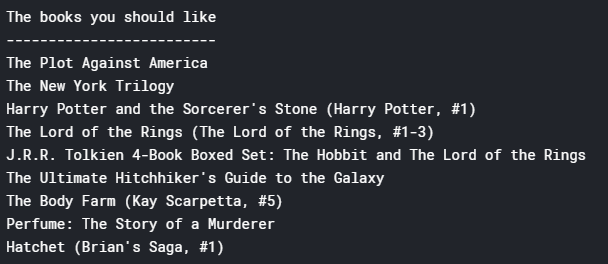


# Результаты

По входным данным с книгами, которые я уже прочёл и которые мне понравились



я получил рекомендации книг



Среди которых также были мои любимые, но которые я забыл указать в начале.

Исходя из полученных результатов могу сказать, что я получил верную рекомендацию.

# Выводы

Коллаборативная фильтрация позволяет очень точно определять предпочтения пользователей различных сервисов (в том числе и электронных библиотек), обеспечивая тем самым большой спрос на эти сервисы.