### Обзор алгоритмов рекомендательных систем

К наиболее классическим относятся алгоритмы:

* Summary-based (неперсональные),
* Content-based,
* Collaborative Filtering,
* Matrix Factorization,
* и некоторые другие.

### Неперсонализированные рекомендации Они самые простые в реализации - них потенциальный интерес пользователя определяется просто средним рейтингом товара.

##### Актуальна проблема холодного старта, когда ещё не накоплено достаточное количество данных – Общий подход здесь — заменять то, что в данный момент не может быть посчитано, различными эвристиками (например, заменять средним рейтингом).

### В некоторых случаях также важно учитывать «свежесть» рекомендации - используются корректирующие коэффициенты (damping factors). Ниже пара формул для расчета рейтинга статей на медиа сайтах. (см. формулы)

### Content-based рекомендации

### Неструктурированные признаки описываются типичным для текста способом – векторами в пространстве слов ([Vector-Space model](https://en.wikipedia.org/wiki/Vector_space_model)).

### По мере взаимодействия пользователя с системой (скажем, он покупает фильмы), векторные описания приобретенных им товаров объединяются (суммируются и нормализуются) в единый вектор и, таким образом, формируется вектор его интересов. Далее достаточно найти товар, описание которого наиболее близко к вектору интересов, т.е. решить задачу поиска n ближайших соседей.

### Не все элементы одинаково значимы: например, союзные слова, очевидно, не несут никакой полезной нагрузки. Поэтому при определении числа совпадающих элементов в двух векторах все измерения нужно предварительно взвешивать по их значимости. Данную задачу решает хорошо известное в [Text Mining](https://en.wikipedia.org/wiki/Text_mining) преобразование [TF-IDF](https://en.wikipedia.org/wiki/Tf%E2%80%93idf), которое назначает больший вес более редким интересам. (см. формулу)

### В качестве меры близости двух векторов чаще всего используется косинусное расстояние. (см. формулу)

***Алгоритмы факторизации (Matrix Factorization)***

### Здесь мы можем говорить об обобщаемости моделей – уменьшение размерности данных, так как интересы будут описываться не вектором товаров, а существенно меньшим вектором предпочтений. Такие подходы еще называют спектральным разложением. Наиболее часто используемое разложение матриц называется SVD-разложением (singular value decomposition). (см. формулу)

### Из нашей матрице предпочтений мы получаем две матрицы факторов: U — компактное описание предпочтений пользователя, S — компактное описание характеристик продукта. Чтобы получить приближение матрицы предпочтений необходимо перемножить матрицы факторов.

#### **Collaborative Filtering (коллаборативная фильтрация)**

### В рамках данного подхода рекомендации генерируются на основании интересов других похожих пользователей (User-based вариант). Классическая реализация алгоритма основана на принципе k ближайших соседей, где для каждого пользователя ищем k наиболее похожих на него (в терминах предпочтений) и дополняем информацию о пользователе известными данными по его соседям. Пользовательская коллаборативная фильтрация (user-based CF) ищет для каждого клиента группу наиболее похожих на него (в терминах предыдущих покупок) клиентов и усредняет их предпочтения.

### (см. фото)

### На картинке выше проиллюстрирован принцип работы метода. В матрице предпочтений желтым цветом выделен пользователь, для которого мы хотим определить оценки по новым товарам (знаки вопроса). Синим цветом выделены три его ближайших соседа. «Похожесть» – в данном случае синоним «корреляции» интересов, достаточно часто используется косинусное расстояние как основная мера близости.

### У классической реализации алгоритма есть один явный минус – он плохо применим на практике из-за квадратичной сложности расчета матрицы расстояний O(n^2\*2m), где n — число пользователей, а m — число товаров.

В случае же с товарной коллаборативной фильтрацией (item-based CF) ближайшие соседи ищутся на множестве товаров — столбцов матрицы предпочтений. Если продукты содержательно похожи, то скорее всего они либо одновременно нравятся, либо одновременно не нравятся. Поэтому, когда мы видим, что у двух товаров оценки сильно коррелируют, это может говорить о том, что это товары-аналоги.

#### **Гибридные решения (hybrid)**

### Алгоритмы делятся по:

### типам комбинирования

### стратегиям объединения

### моделям

### подходам