

# Big Data 10.0



Team-One Project 2  
Content Based Recommender System



# Рекомендации на карточке товара

## Бандлы



## Аксессуары (сопутствующие товары)



## Связанные товары



Что рекомендуют на карточке товара в Ozon.ru:

- Бандлы
- Аксессуары
- Связанные товары

## Бандлы

Бандлы могут подбираться экспертно и статистически.  
Статистические строятся на основе частоты совместных заказов.

Тройки подбираются из пар, для каждой пары ищем общий третий товар как пару к каждому из товаров. Например, если для пары товаров А и В существует товар С, который покупается и с товаром А и с товаром В, то образуется тройка ABC



ozon.ru  
выбирайте

© OZON.ru 12

## Связанные товары (Рекомендуем также)

- В данной полке показываются как товары заменители, так и дополнители
- Покрытие полкой более 98% товаров
- Примерно каждый 10й товар добавленный в корзину был найден клиентом OZON.ru именно в этой полке

- Строится как комбинация алгоритмов, основанных на:
  - заказах (товары в одном заказе, товары заказанные одним клиентом)
  - логах (статистика совместных просмотров и т.п.)
  - контенте (товарные характеристики)

ozon.ru  
выбирайте

© OZON.ru 14

## Аксессуары (сопутствующие товары)

Ручная полка, генерируется на основе ручных связок типов и фильтров (ограничений), которые должны быть применены к этим связкам.



ozon.ru  
выбирайте

© OZON.ru 13

Нет  
инфо

Есть  
инфо

На проекте дана инфо:

- Статистика совместных просмотров
- Контент
  - Текстовый
  - Логический (каталоги)

# Идея 1: Текстовый контент

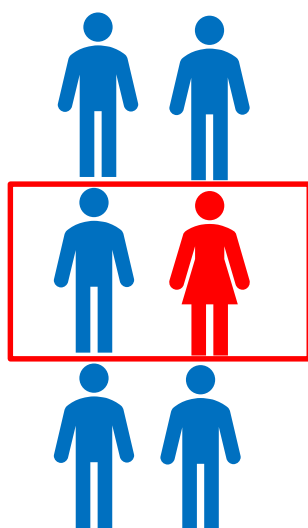


- Можно попробовать решить задачу «в лоб» используя текстовый контент в описании товара
- Word2Vec 😊
- Tfidf 😊
- Cosines Similarity 😞
  - crossJoin даже матриц test x train => 8 млрд вариантов
  - Наш Spark «умирает»
- Prod2Vec - новая система на основе нейронных сетей, которая сейчас реализуется в Ozon - слишком сложно для проекта  
<https://habr.com/ru/company/ozontech/blog/432760/>

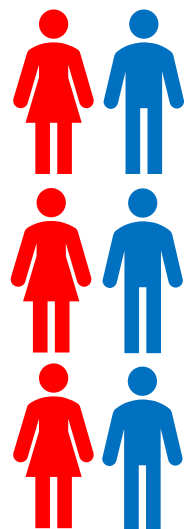
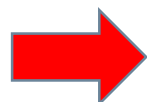
# Идея 2: Алгоритм поиска друзей



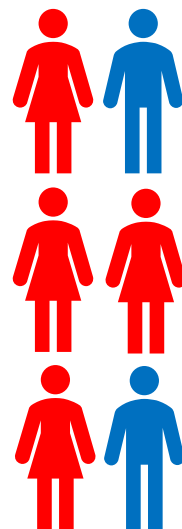
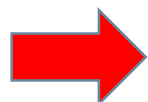
- На основе статистики совместных просмотров товаров (TrueRecoms = FRIENDS)
- TRAIN join TEST ( `left\_semi` ) join TRAIN ( `left\_outer` )
- BOYS join GIRLS ( `left\_semi` ) join BOYS ( `left\_outer` ) =>



FRIENDS



BOY-FRIENDS



FRIENDS OF FRIENDS



PROBLEM: ~22% of GIRLS  
have no BOY-FRIEND!

# Идея 3: Проживание в одном каталоге



- Логическая структура каталогов охватывает большинство товаров
- Поиск пар с одинаковым catalogid ( `соседи по двору` )
  - TEST join CATALOG (itemid, left\_outer) join CATALOG (catalogid, left\_outer)




- Поиск пар с одинаковым pathid ( `соседняя улица, район` )
  - TEST join CATALOG (itemid, left\_outer) join PATH1/2/3 (catalogid, left\_outer)
- Поиск пар с одинаковым parent\_id (e.g. `учились/работали вместе` )
  - TEST join DETAILS (itemid, left\_outer) join DETAILS (parent\_id, left\_outer)
- **Filter: rating is not null** (не будем рекомендовать непонятных маргинальных личностей даже если они тусуют в нашем дворе) !!!
- Friends + Catalogs => 99%   , остальным 1% предлагаем MOST POPULAR

# Идея главная: приоритет рекомендаций

- Наши связи с другими людьми можно отранжировать по приоритету - то же самое с рекомендациями
- Находим подходящие пары для TEST  и присваиваем им вес для ранжирования:

Pairs	Weight (alpha=32)
Friends	Rating + alpha * Views
Friends of Friends	Rating + (alpha/4) * Views
Common Catalog	Rating
Common Path 1/2/3	Rating /2/4/8
Common Parent_id	Rating

22.85%

24% of test items  have  
<100 recoms => space for  
improvement!!!

- alpha - scaling factor for implicit feedback
- **Note: `Friends` are important => we keep them even if they have no Rating!!!**



# **BONUS: Harvard CS50 Final Project**

## **2 min video**

- See video presentation of the project by the link below

<https://youtu.be/ein2VnyxFTQ>