



Неперсональные рекомендации

Андрей Зимовнов (Яндекс, ВШЭ)



Неперсональные

Для всех пользователей показываем
одни и те же рекомендации

Гарри Поттер и философский камень

Рейтинг фильма



8.118 162 617

IMDb: 7.50 (377 260)

Топ250: **199**

[об оценках и Топ-250](#)

Рейтинг кинокритиков

в мире 



153 + 38 = 191

★ 7.1

в России



2 + 0 = 2

[о рейтинге критиков](#)

Рейтинги кинокритиков



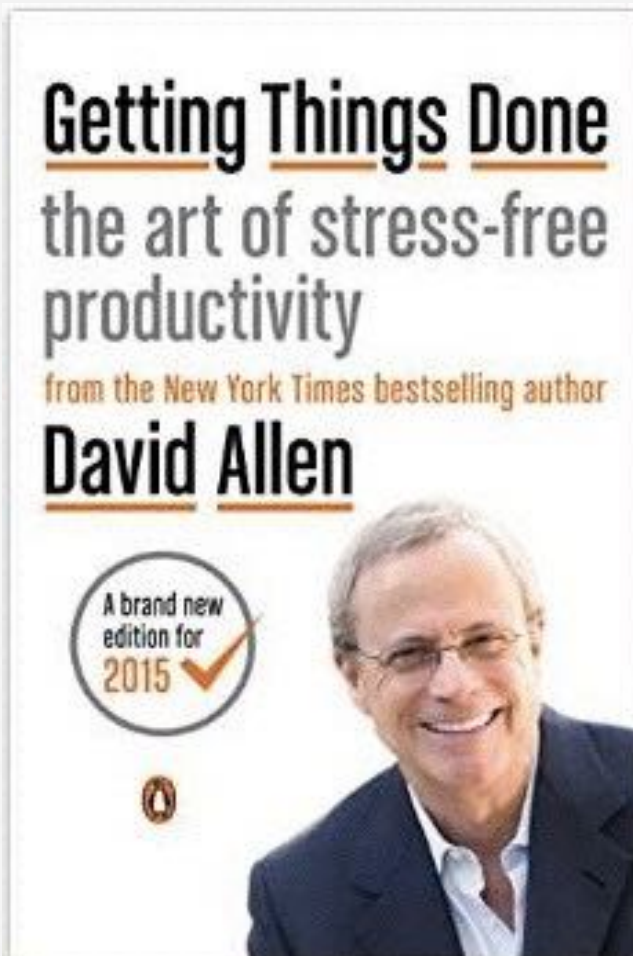
средняя оценка

всего рецензий

% положительных
рецензий

отрицательных

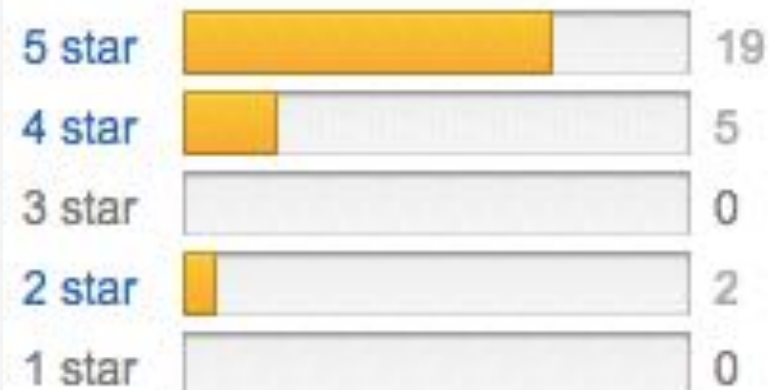
положительных



Customer Reviews

★★★★☆ 26

4.6 out of 5 stars





Проблемы с рейтингами

- **Явные рейтинги**
 - разная шкала (субъективная)
 - разброс рейтингов
- **Неявные рейтинги**
 - покупки (понравилось или нет?)
 - время на сайте (а если отвлекся?)
 - клик (является ли «не клик» сигналом, а что после клика?)
- **Накрутки**



Средний рейтинг



Явный рейтинг

It's ok

I love it

I like it



I hate it!

I don't like it

Средний рейтинг

$$P_i = \frac{\sum_{u=1}^n r_{ui}}{n}$$

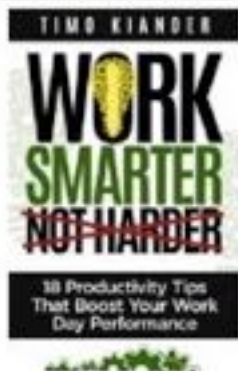
рейтинг пользователя u для товара i

item

количество пользователей оценивших i

Топ по среднему на Amazon

Все ОК?



Work Smarter Not Harder: 18 Productivity Tips That Boost Your Work Day Performance Mar 25, 2015

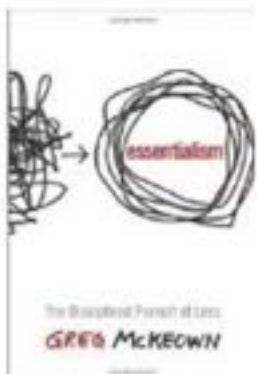
by Timo Kiander

Kindle Edition

\$0.00

Auto-delivered wirelessly

★★★★★ ▾ 2



Essentialism: The Disciplined Pursuit of Less Apr 15, 2014

by Greg McKeown

Hardcover

\$17.50 ~~\$23.00~~ ✓ Prime

Get it by **Wednesday, May 6**

More Buying Choices

\$10.61 used & new (62 offers)

Kindle Edition

\$10.99

Whispersync for Voice-ready

★★★★★ ▾ 508

Trade-in eligible for an Amazon gift card

FREE Shipping on orders over \$35

Excerpt

Page 5 : ... some new strategy in *time management*. It is about pausing ... [See a random page](#) in this book.



Проблема

Если мало рейтингов, то оценка неуверенная!

Регуляризация среднего

$$P_i = \frac{\sum_{u=1}^n r_{ui} + k\mu}{n + k}$$

← глобальное среднее

↑
контролирует
минимальное
количество
наблюдений





Лайки и дизлайки



Рекомендации на evanmiller.org

Все OK?

2. **normal**

209 up, 50 down  

A word made up by this corrupt society so they could single out and attack those who are different

Normal is nothing but a word made up by society

conformists

worker bees



in crowd

followers

mindless

by Bill Oct 6, 2005 [share this](#) [add comment](#)

3. **normal**

118 up, 25 down  

Сортировка по чистым лайкам («like» - «dislike»)



Проблема

Разности («like» - «dislike») для разных товаров несравнимы!



Вероятность лайка

Каждый рейтинг принимает только два значения 1 и 0 (like, dislike)

Каждый рейтинг – случайная величина Бернулли с вероятностью p

Распределение Бернулли

Вероятность лайка ($x = 1$): p (успех)

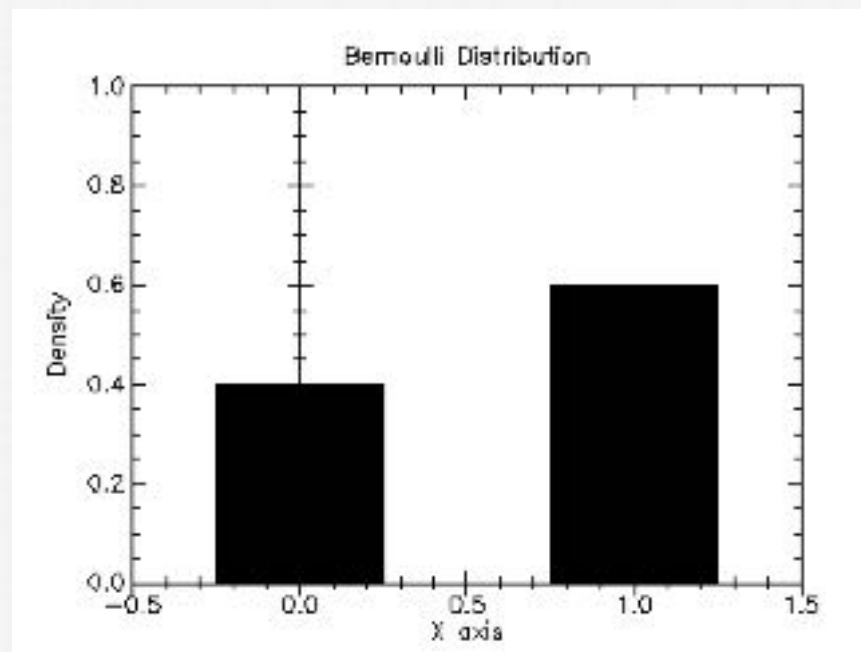
Вероятность дизлайка ($x = 0$): $1 - p$ (неуспех)



лайк

дизлайк

Частоты:



дизлайк

лайк

Сумма независимых величин

Проведем серию из n «подкидываний монеты»

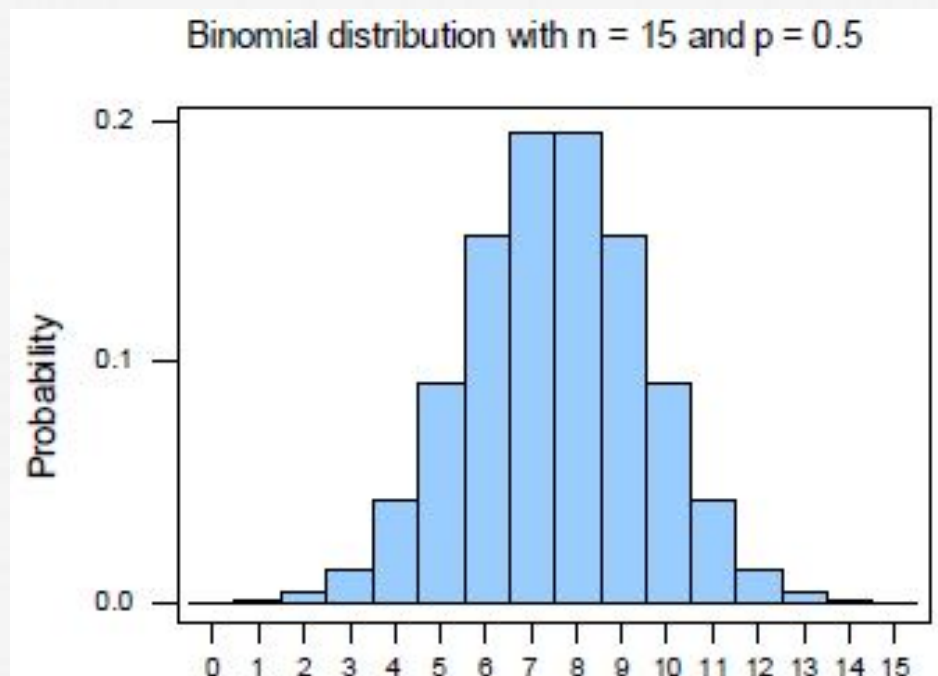
Какова вероятность получить k лайков?

Построим распределение для $x_1 + \dots + x_n$



лайк

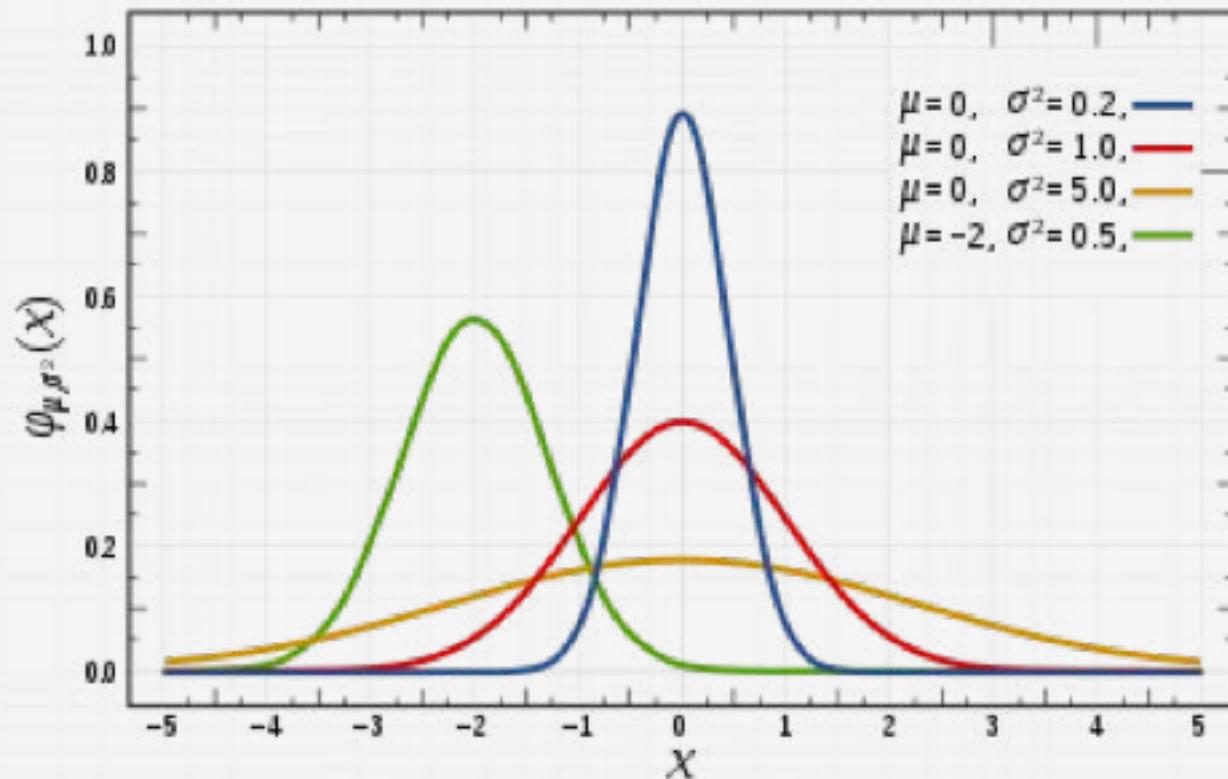
дизлайк



Биномиальное распределение

Нормальное распределение

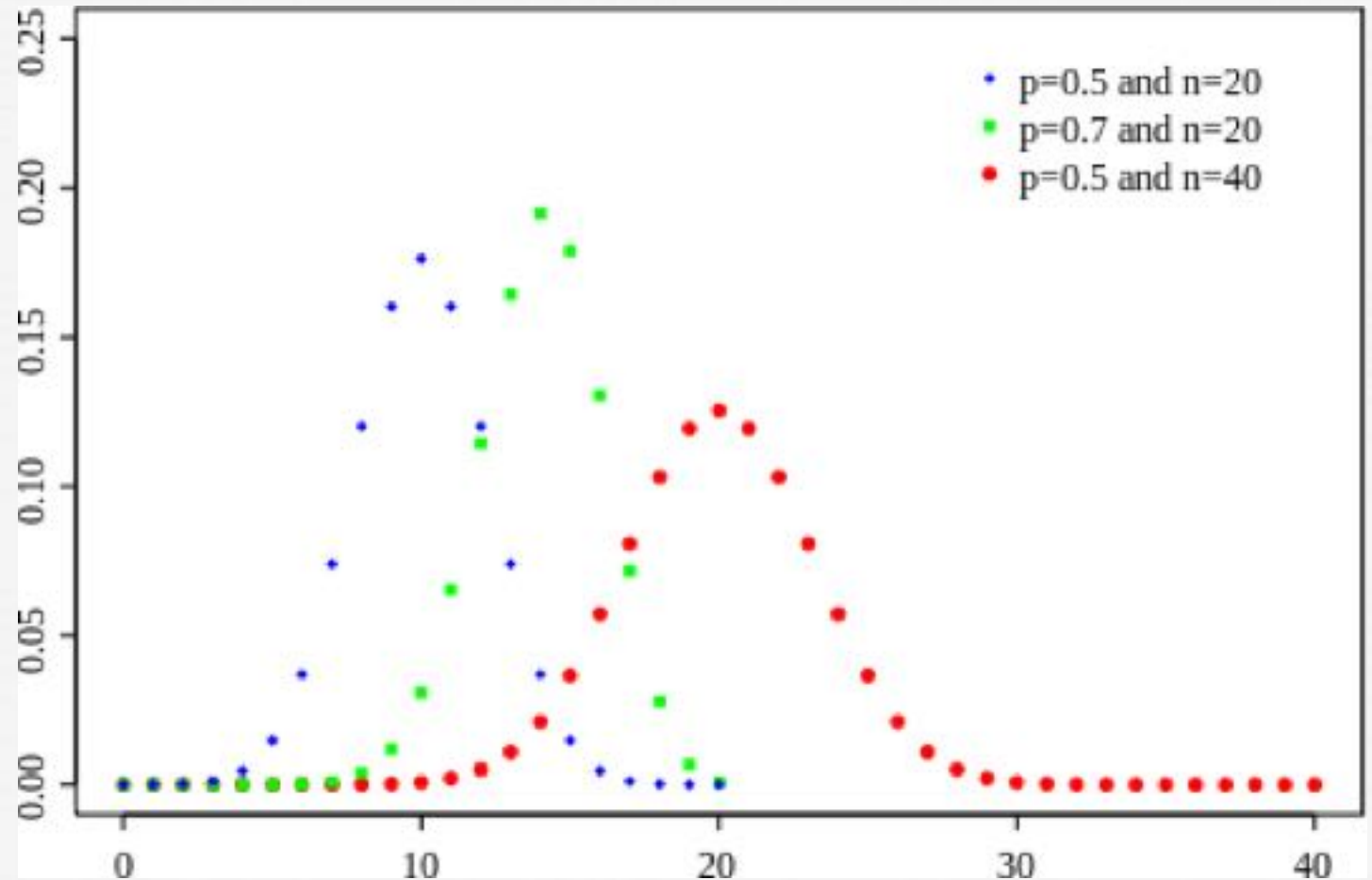
Величина объясняется суммой большого количества независимых компонент \rightarrow ее распределение близко к нормальному



Приблизим нормальным

$$\mu = P$$

$$\sigma = \sqrt{\frac{1}{n} P(1-P)}$$



Доверительный интервал

доля
положительных
рейтингов

квантиль стандартного
нормального
распределения

$$\frac{1}{1 + \frac{1}{n} z^2} \left[\hat{p} + \frac{1}{2n} z^2 \pm z \sqrt{\frac{1}{n} \hat{p}(1 - \hat{p}) + \frac{1}{4n^2} z^2} \right]$$

количество
наблюдений

Доверительный интервал

квантиль 95%

$$\frac{1}{1 + \frac{3.84}{n}} \left[\hat{p} + \frac{3.84}{2n} \pm 1.96 \sqrt{\frac{1}{n} \hat{p}(1 - \hat{p}) + \frac{3.84}{4n^2}} \right]$$

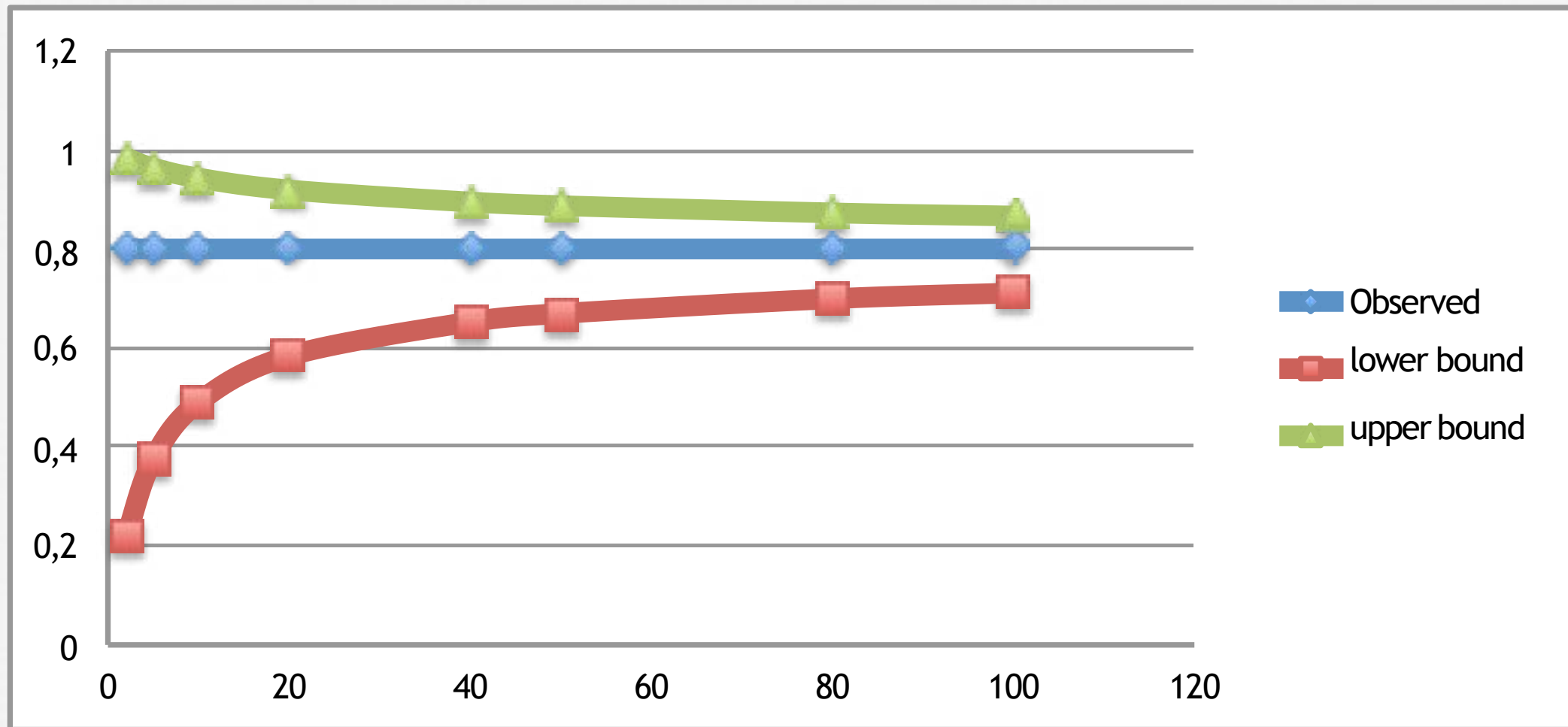


Доверительный интервал

n	2	5	10	20	40	50	80	100	1000
наблюдаемое p	80 %	80 %	80 %	80 %	80 %	80 %	80 %	80 %	80 %
нижняя граница	22 %	38 %	49 %	58 %	65 %	67 %	70 %	71 %	77 %
верхняя граница	98 %	96 %	94 %	92 %	90 %	89 %	87 %	87 %	82 %

Доверительный интервал

Зачем нам это все?





Учитываем уверенность

Ранжируем по нижней границе доверительного интервала!



Ранжирование на reddit.com



reddit

ALL

горячее

новое

набирающие популярность

спорное

рейтинговое

позолочено

хотите присоединиться? войдите



You DO have time to read! Download a free audiobook from Audible.com now. Choose from over 150k titles. Listen anytime, anywhere! (Audible.com)

рекламируется audiblerrreddit

поделиться

спонсируемая ссылка

что это?

1 4830



:l :l :P (i.imgur.com)

отправлено 4 hours ago автор DatOneBlindSloth в /r/funny

332 комментария поделиться

2 5239



I dont remember Mario being this hardcore (i.imgur.com)

отправлено 6 hours ago автор DonCairo в /r/gaming

1177 комментариев поделиться

3 4958

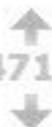


TIL that the writing staff of Futurama held three Ph.D.s, seven masters degrees, and cumulatively had more than 50 years at Harvard (en.wikipedia.org)

отправлено 5 hours ago автор juliokirk в /r/todayilearned

1036 комментариев поделиться

4 4712



My Neighbor is a Dick, no regret (livememe.com)

отправлено 5 hours ago автор factman5000 в /r/AdviceAnimals

1196 комментариев поделиться

5 4964



I just started getting spam in my primary email account... (i.imgur.com)

отправлено 6 hours ago автор BryanWake в /r/funny

563 комментария поделиться

6 4358



App idea... Hinder...tells you where Tinder matches are happening so you can show up to cock block.

Ранжирующая функция

$$y = \text{sign}(U - D)$$

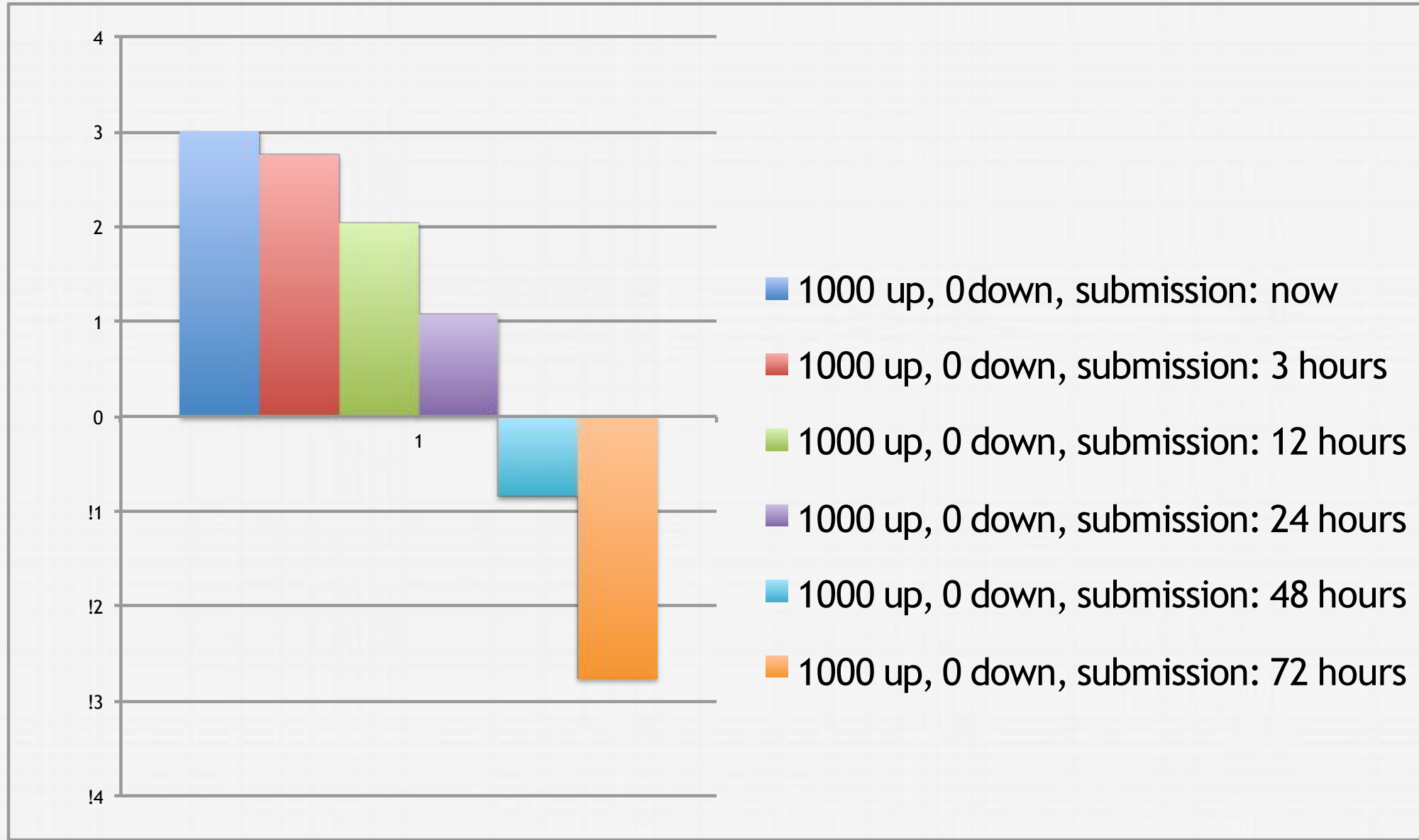
$$f(t_s, U, D) = y \log_{10} |U - D| - \frac{t_s}{45000}$$

U - лайки
D - дизлайки

секунд с
момента
загрузки

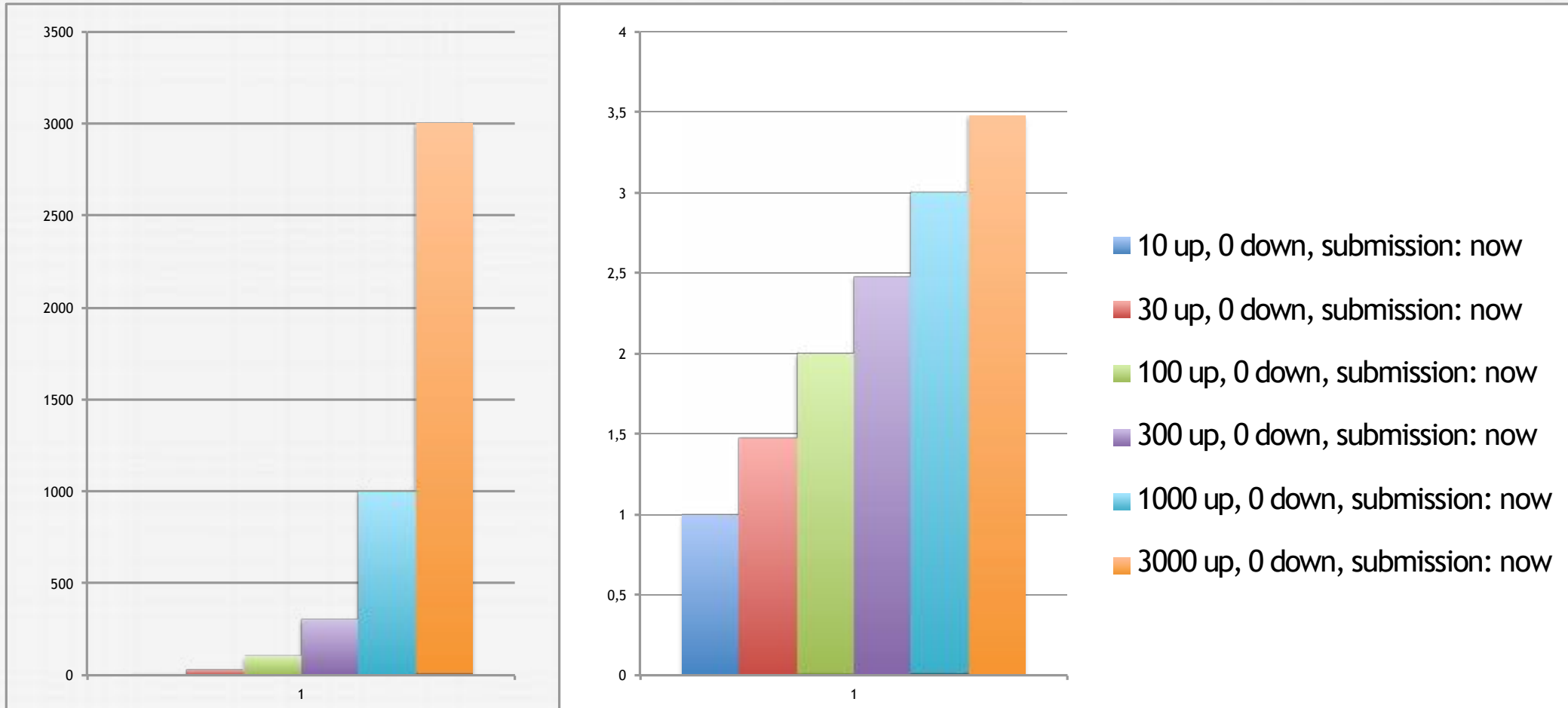
ad hoc
константа

Естественное устаревание



Логарифмы линеаризуют

Оба слагаемых теперь линейно меняются





Неперсональное ранжирование

Плюсы:

- Легко сделать
- Хорошо работает для новых пользователей

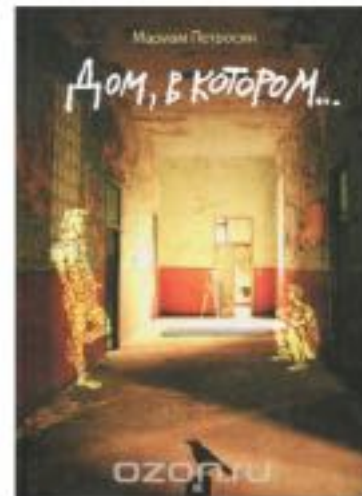
Минусы:

- Нет персонализации
- Смещенные оценки (люди больше жалуются, чем хвалят)



Неперсональные рекомендации в ozon.ru

Рекомендации к товару



Дом, в котором...

ID 24277965

Новинка Бестселлер

★★★★★ (155 отзывов) 566 189 У меня это есть

Автор: Мариам Петросян

Издательство: Гаятри/Livebook

ISBN 978-5-904584-69-6; 2015 г.

Язык: Русский

[Дополнительные характеристики](#)

Рекомендуем также



Дом, в котором... В
3 томах (комплект)
509,60 Р

В корзину



Тринадцатая
сказка
332 Р

В корзину



Дом странных
детей
326,40 Р

В корзину



Дом, в котором...
164,90 Р

Скачать



Убить
пересмешника...
287,20 Р

В корзину

Откуда их взять?

Не зависит от
пользователя





Наивный подход

Можно посчитать, как часто два товара покупают вместе.



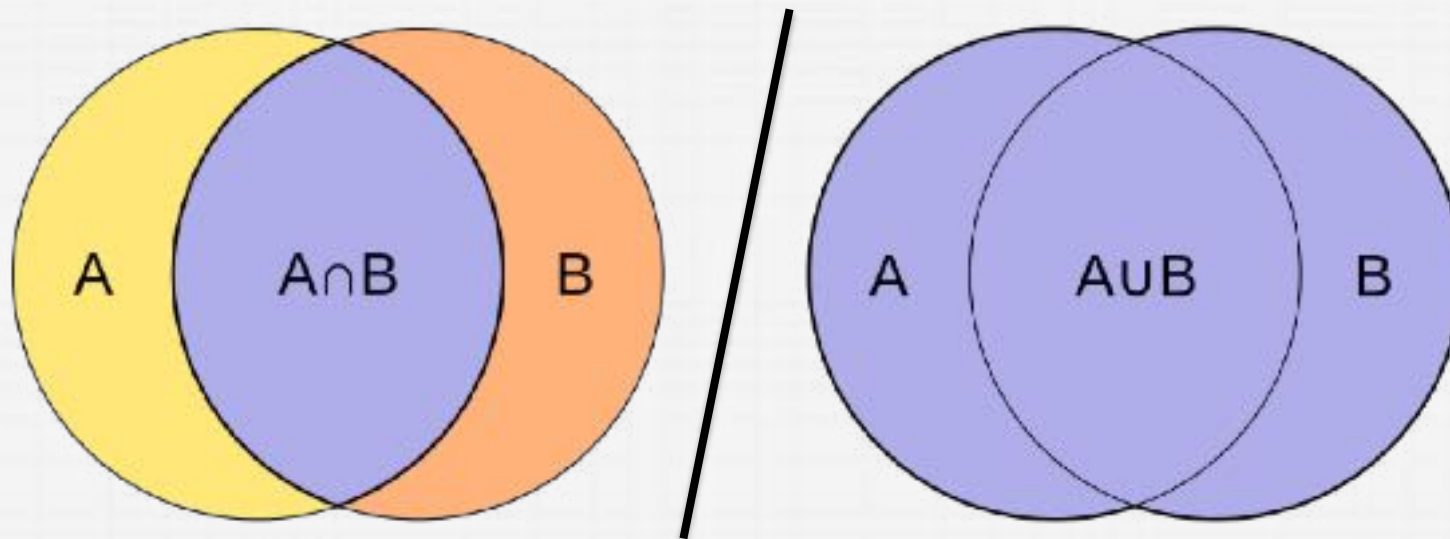
Наивный подход

Можно посчитать, как часто два товара покупают вместе.

Окажется, что туалетную бумагу покупают ко всему 😊

Мера Жаккара для множеств

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



Что в нашем случае множество?



В нашем случае

Множество для товара – это все пользователи, которые его купили.

Тогда мы будем измерять похожесть двух товаров с точки зрения купивших их пользователей.

Чем чаще покупают вместе редкие товары, тем лучше.

В нашем случае

Рассмотрим матрицу Item-User, где в ячейке записана 1, если пользователь u покупал товар i .

	u1	u2	u3	u4
i1	1		1	1
i2		1		
i3	1	1		

Одним из признаков рекомендательной системы может быть мера Жаккара между строчками матрицы (товарами).

Пример расчета

	u1	u2	u3	u4
i1	1		1	1
i2		1		
i3	1	1		

$$J(i_1, i_2) = \frac{0}{4} = 0$$

Пример расчета

	u1	u2	u3	u4
i1	1		1	1
i2		1		
i3	1	1		

$$J(i_1, i_2) = \frac{0}{4} = 0$$

$$J(i_1, i_3) = \frac{1}{4} = 0.25$$



Алгоритмическая сложность

В реальной задаче:

- Миллионы пользователей (N)
- Миллионы товаров (M)

Как посчитать меру Жаккара для всех товаров?



Алгоритмическая сложность

В реальной задаче:

- Миллионы пользователей (N)
- Миллионы товаров (M)

Как посчитать меру Жаккара для всех товаров?

- Наивный подход: $O(M * M * N)$

Оптимальный алгоритм

Вклад в числитель только от совместных покупок каждого пользователя!

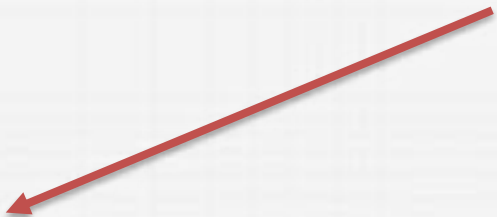
	u1	u2	u3	u4
i1	1		1	1
i2		1		
i3	1	1		





Считаем на MapReduce

Инвертированный индекс



Map: $(u, (i_1, i_2, \dots)) \rightarrow ((i_k, i_l), 1)$ для всех пар товаров в истории

Reduce: $((i_k, i_l), (1, 1, \dots)) \rightarrow ((i_k, i_l), \text{мощность пересечения})$

Как посчитать мощность объединения?

Инвертированный индекс

Map: $(u, (i_1, i_2, \dots)) \rightarrow ((i_k, i_l), 1)$ для всех пар товаров в истории

Reduce: $((i_k, i_l), (1, 1, \dots)) \rightarrow ((i_k, i_l), \text{мощность пересечения})$

Как посчитать мощность объединения?

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



Считаем на MapReduce

В реальности терабайт промежуточных данных
И пол дня счета

SELECT

```
ic1.itemId,  
ic2.itemId AS jointItemId,  
SUM(ic1.val * ic2.val)
```

FROM ic **AS** ic1

```
JOIN ic AS ic2 ON ic1.clientId = ic2.clientId
```

WHERE ic1.itemId < ic2.itemId**GROUP BY** ic1.itemId, ic2.itemId;

Архитектура платформы





Формула успеха

- **40%** Apache Spark (Python) + **50%** Hive on TEZ + **10%** Hive UDF (Java).
- Парсить данные удобно в Spark на Python, дальше их можно сложить в Hive таблицу и продолжить обработку SQL запросом.
- ~ **70%** code reuse между прототипом и продакшеном: как правило на UDF переписываются только критичные по производительности и не очень сложные функции, которые достаточно универсальны.
- Математики могут улучшать алгоритмы (нужно знать Python и SQL)!

Визуализируйте рекомендации!

```
In [50]: from NextGen import SparkUtils as SU
from pyspark import HiveContext
import ujson as json

sc = SU.createSparkContext("sb_test", 33, "math")
hivec = HiveContext(sc)
j = hivec.sql("select related from sb_final where search='ipad'").collect()[0]
items = map(lambda x: x["ItemId"], json.loads(j.related))
SU.itemIdsToHTML(items[0:100])
```

Out[50]:

30481115



[Apple iPad Air 2 Wi-Fi + Cellular 16GB, Space Gray](#)

30481116



[Apple iPad Air 2 Wi-Fi + Cellular 16GB, Silver](#)

30481117



[Apple iPad Air 2 Wi-Fi + Cellular 64GB, Space Gray](#)

30481118



[Apple iPad Air 2 Wi-Fi + Cellular 64GB, Silver](#)

30481119



[Apple iPad Air 2 Wi-Fi + Cellular 128GB, Space Gray](#)

Визуализируйте рекомендации!

In [3]: SU.htmlAccs(accs)

Out[3]: **31682325**



Футболка-поло
женская Vaop,
цвет: зеленый.
B205204. Размер
XS (42)

Рекомендуем

Брюки

Джинсы

Шорты

Туалетная вода

Юбка

Парфюмерная вода

31785599



Брюки женские
Broadway, цвет:

31785601



Брюки жен.
Broadway, цвет:

31643287



Брюки женские
Sela, цвет:

31678474



Джинсы
женские Calvin

31678475



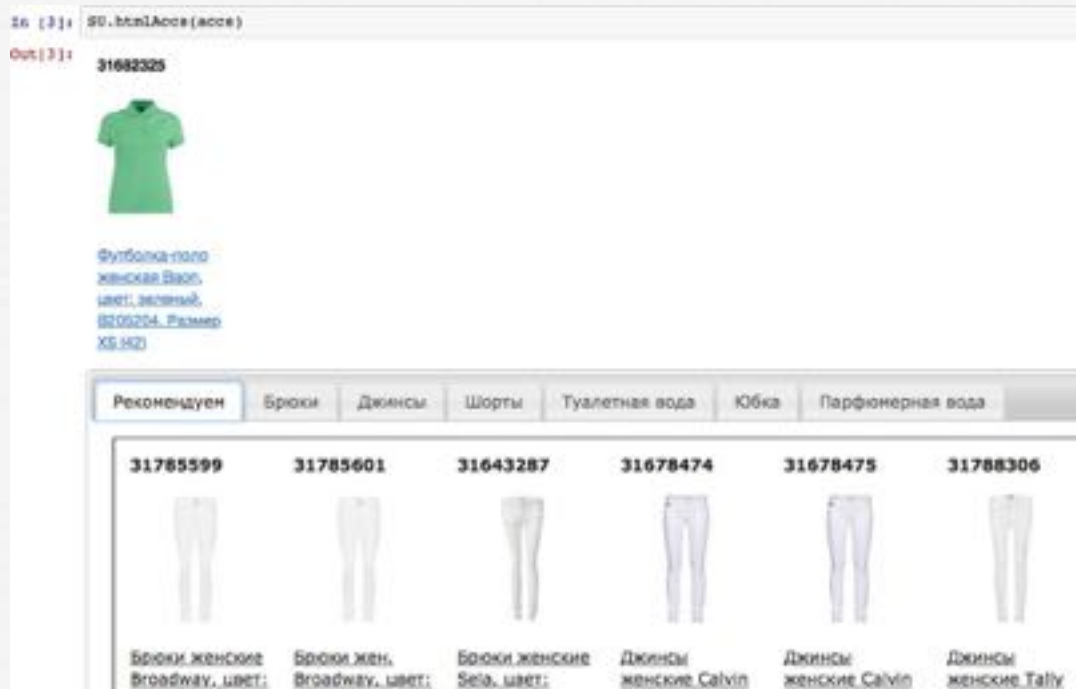
Джинсы
женские Calvin

31788306



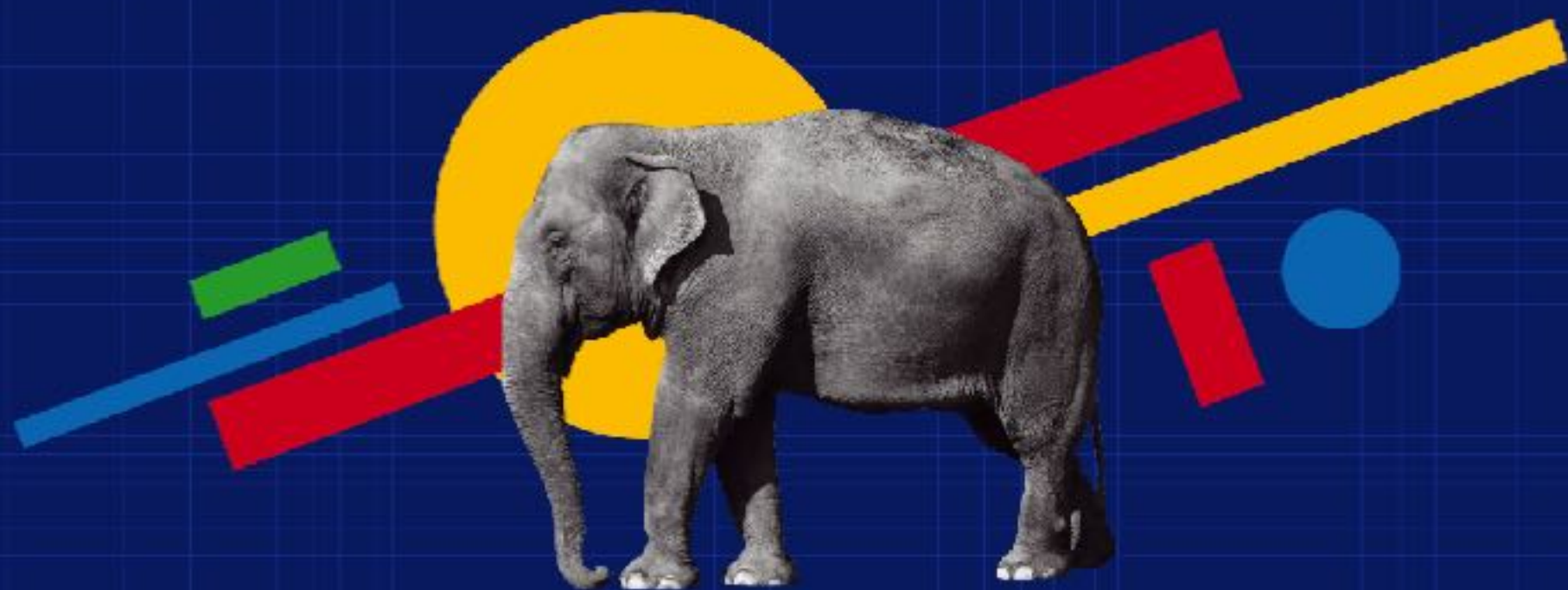
Джинсы
женские Tally

Это не сложно



И помогает дебажить в офлайне...

```
def htmlAccs(accs):
    uuid_str = str(uuid.uuid1())
    hostId, accs = accs
    tabNames = [tab[0] for tab in accs]
    tabsCode = u"<ul>"
    for idx, tabName in enumerate(tabNames):
        tabsCode += u""""<li><a href="#tabs{uuid}-{idx}">{name}</a></li>""", format(
            idx=idx+1, name=tabName, uuid=uuid_str)
    tabsCode += u"</ul>"
    divsCode = u""
    tabItems = [tab[1] for tab in accs]
    for idx, itemList in enumerate(tabItems):
        divsCode += u""""<div id="tabs{uuid}-{idx}"><table><tr><td>""", format(
            idx=idx+1, uuid=uuid_str)
        for item in itemList:
            divsCode += itemIdToDiv(item)
        divsCode += u""""</td></tr></table></div>""""
    from IPython.display import HTML
    output = itemIdToDiv(hostId) + "<br>"*15 + """"<div id="tabs{uuid}">""", format(
        uuid=uuid_str) + tabsCode + divsCode + \
        """"</div>
        <script>
        ${function() {
            $("#tabs"" + uuid_str + """).tabs();
        }};
        </script>""""
    return HTML(output)
```

BIG DATA IS LOVE

NEWPROLAB.COM