

Машинное обучение  
Лекция № 12, осень 2022

# Матчинг и ранжирование

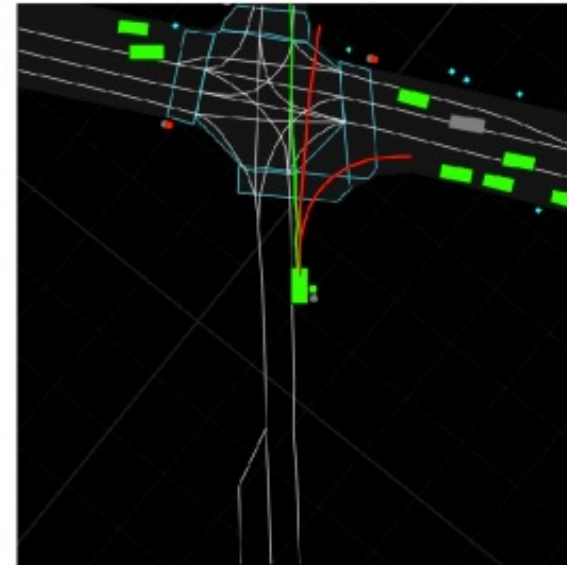
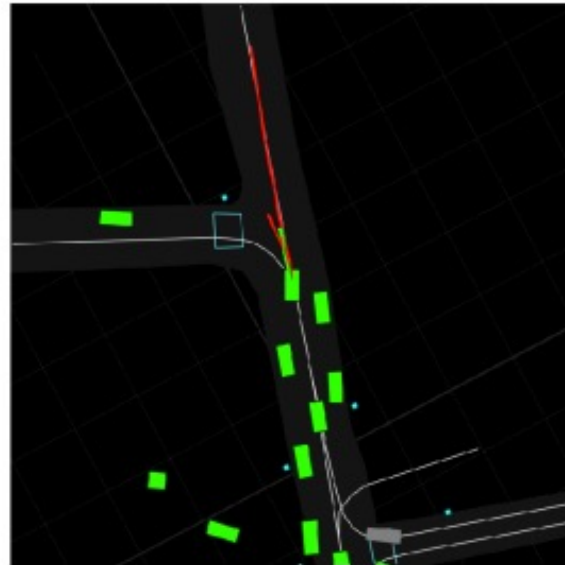
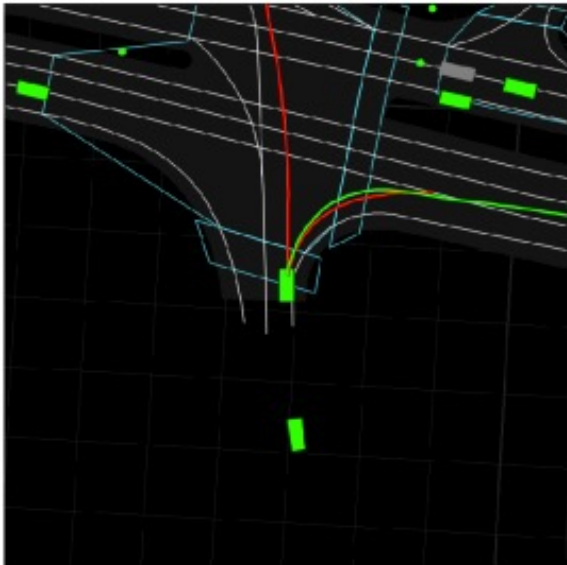


# План лекции

- Задача ранжирования
- Задача матчинга
- Подходы к решению задачи матчинга
- Метрики в задаче ранжирования
- Базовые методы решения задачи ранжирования

# Ранжирование

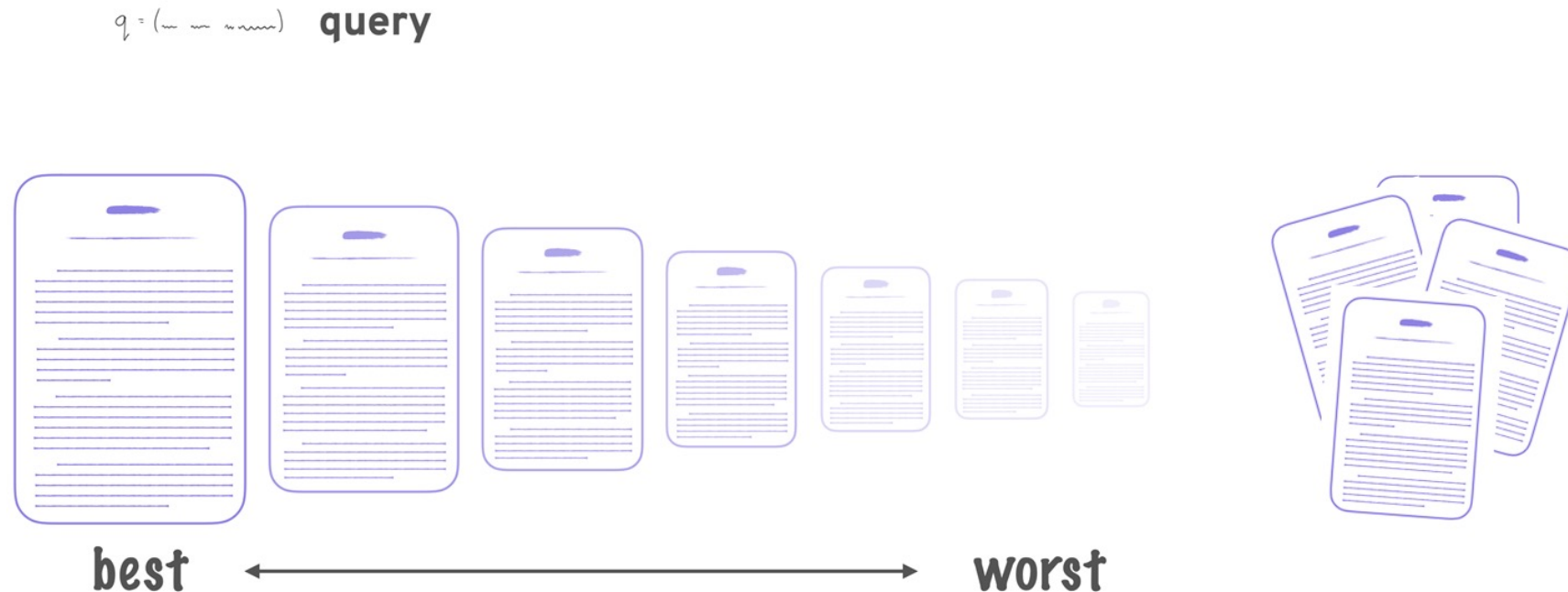
**Ранжирование** — процесс упорядочивания набора объектов в соответствии с некоторой мерой, то есть задание частично упорядоченного множества.



# Порядок имеет значение

Множество частично упорядочено, если указано, какие элементы следуют за какими (какие элементы больше каких).

В общем случае может оказаться так, что некоторые пары элементов не связаны отношением «следует за»

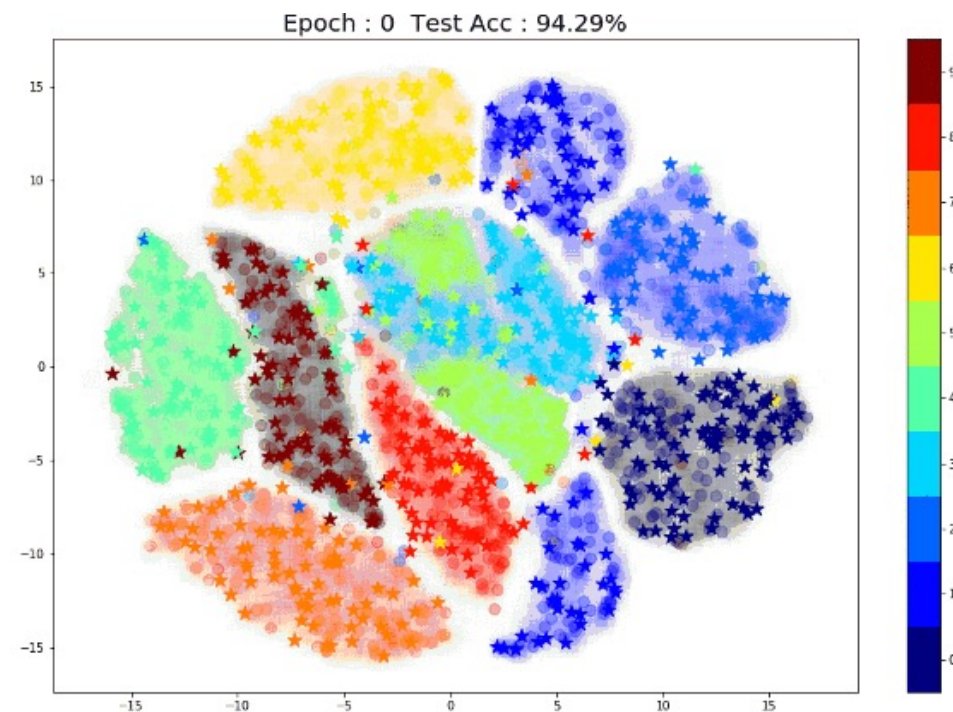


# Место задачи ранжирования

**Learning to rank** (обучение ранжированию) — класс задач машинного обучения с учителем (*supervised*) или с частичным привлечением учителя (*semi-supervised*), Цель — построить модель, которой является наилучшим приближением и обобщением способа ранжирования в обучающей выборке для новых данных.

*Pseudo-labeling* – одна из техник *semi-supervised* обучения

$\text{Loss per Batch} = \text{Labeled Loss} + \text{Weight} * \text{Unlabeled Loss}$

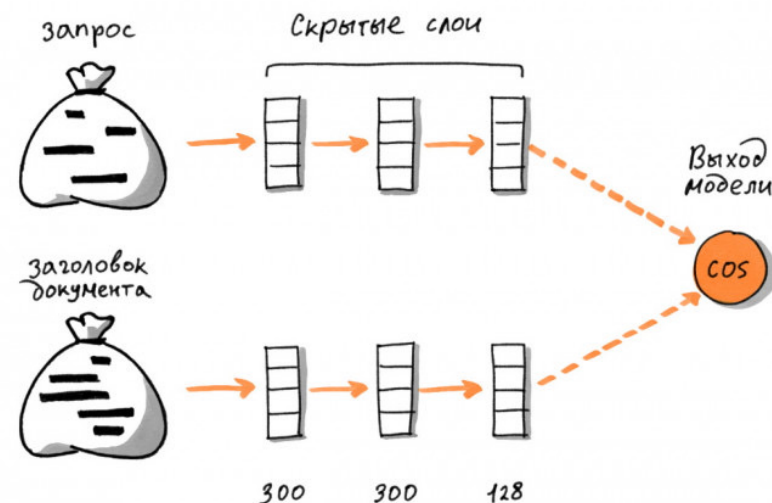


# Постановка задачи

**Мера релевантности** — степень соответствия между запросом и документом (или набором документов).

Чем выше это соответствие, тем выше в списке ранжирования должен находиться документ.

$$\text{Relevance}(\text{query}, \text{doc}) \sim \langle \text{Emb}^n(\text{query}), \text{Emb}^n(\text{doc}) \rangle$$



# Постановка задачи LTR

**Задача (learning to rank – LTR)** — создание глобальной модели  $F(q, D)$ , которая будет работать над всем корпусом документов.

$Q$  – набор запросов  $\{q_1, q_2, \dots, q_m\}$

$D$  – набор документов

$D_i = \{d_{i,1}, d_{i,2}, \dots, d_{i,n_i}\}$  – набор документов, релевантных  $i$  запросу  $q$

$d_{i,j}$  – элемент с индексом  $j$  в  $D_i$

$y_i = \{y_{i,1}, y_{i,2}, \dots, y_{i,n_i}\}$  – набор оценок релевантности для  $i$  запроса (размер тот же, что и у  $D_i$ )

$S = \{(q_i, D_i), y_i\}_{i=1}^m$  – тренировочный набор данных

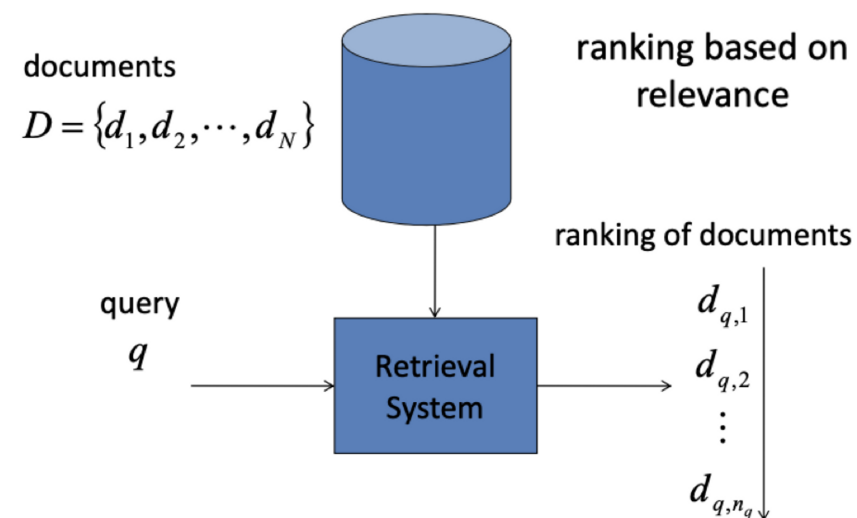
$x_{i,j} = \phi(q_i, d_{i,j})$  – вектор признаков для  $i$  запроса и  $j$  документа ( $i = 1, \dots, m, j = 1, \dots, n_i$ )

$\phi$  – функция для получения признаков (BM25, PageRank, мультимодальные модели)

$x_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,n_i}\}$  – признаки набора документов, релевантных  $i$  запросу  $q$

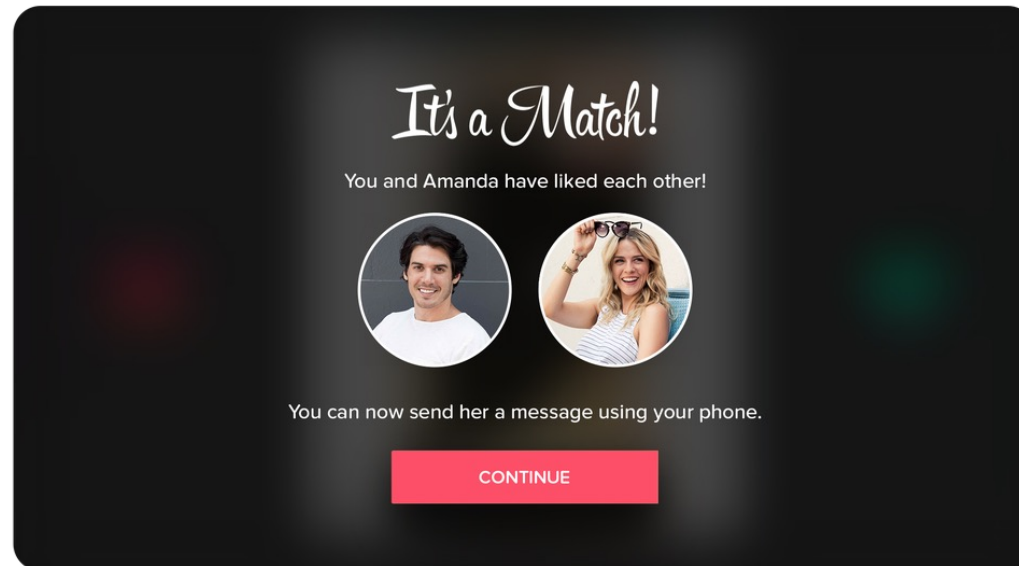
$f(q, d) = f(x)$  – ранжирующая модель, оценивающая релевантность для пары  $q, d$  на основе признаков  $x$

$F(q, D) = F(x)$  – глобальная ранжирующая модель



# Сопоставление == Матчинг

**Матчинг** — процесс сопоставления объектов на основе сравнения и расчета некоторой меры схожести, где объекты, с одной стороны, представляют собой «запросы», а с другой — «документы».





# Применимость матчинга

- Ценообразование;
- Мониторинг промоакций;
- Ассортиментное планирование;
- Мониторинг рынка (в том числе и со стороны производителя);
- Реализация функционала маркетплейса;
- Оперативный поиск.

СмакВил



- Товар «Чусовой (чёрный) хлеб»  
Цена - 11.2 рублей  
Масса - 220 грамм  
Поставщик – ООО «Моя оборона»
- ...

Алфавит Послевкусия



- Товар «Хлеб Чусовой Черный»  
Цена - 13.5 рублей  
Масса - 220 грамм  
Поставщик - ?
- ...

# Уникальный индекс – SKU

**SKU** (*Stock Keeping Unit*, складская учётная единица) — идентификатор товарной позиции (*артикул*), единица учёта запасов, складской номер, используемый в торговле для отслеживания статистики по реализованным товарам/услугам.

Каждой продаваемой позиции, будь то товар, вариант товара, комплект товаров (продаваемых вместе), услуга или некий взнос, назначается свой SKU.

**SKU** не всегда ассоциируется с физическим товаром, являясь скорее идентификатором сущности, представляемой к оплате.

Электроника › Смартфоны и гаджеты › Мобильные телефоны › Apple

## Смартфон Apple iPhone 14 128 ГБ, тёмная ночь

Выбор покупателей

4.9 111 отзывов Характеристики 148 вопросов 👍 экран, качество фотографий 663 500 человек интересовались за 2 месяца



Цвет товара: Тёмная ночь



Версия: Для других стран

Ростест (EAC) для других стран

Конфигурация памяти: 128 ГБ

128 ГБ 256 ГБ 512 ГБ

Количество SIM-карт: Dual SIM (nano-SIM + eSIM)

Dual SIM (nano-SIM + eSIM)

Dual SIM (nano-SIM) Dual SIM (eSIM)

Коротко о товаре

|                         |                                   |
|-------------------------|-----------------------------------|
| Экран                   | 6.1" (2532x1170)<br>OLED          |
| Память                  | встроенная 128 ГБ                 |
| Фото                    | двойная камера,<br>основная 12 МП |
| Процессор               | Apple A15 Bionic                  |
| SIM-карты               | Dual SIM (nano-SIM +<br>eSIM)     |
| Беспроводные интерфейсы | NFC, Bluetooth, Wi-Fi             |

# Предложение о продаже

Одна модель может содержать несколько SKU

Весь ассортимент – база документов

Все предложения – набор запросов

Мультимодальные данные – данные разной природы

Электроника > Смартфоны и гаджеты > Мобильные телефоны > iPhone 14 Результаты поиска во всех категориях

## Мобильные телефоны

Электроника  
Смартфоны и гаджеты  
Мобильные телефоны (800)

Все результаты поиска

Цена, ₽  
от 4 490 до 174 403

Срок доставки  
☐ Сегодня или завтра  
☐ До 5 дней  
☒ Любой

Чёрная пятница 70%

☐ Можно оплатить частями

Производитель  
☐ Apple  
☐ HONOR  
☐ OnePlus  
☐ Samsung  
☐ Ulefone  
☐ vivo  
☐ Xiaomi

Скидки и акции  
☐ кешбек баллами Плюса  
☐ скидки

☐ Со склада Яндекс  
☐ С доставкой Яндекс

Состояние товара  
☐ Новый  
☐ Ресейл

Внешний вид  
☐ Как новый  
☐ Отличный  
☐ Хороший

Линейка  
☐ iPhone 12  
☐ iPhone 14  
☐ iPhone 14 Plus  
☐ iPhone 14 Pro  
☐ iPhone 14 Pro Max

Еще 24

Операционная система  
☐ Android  
☐ iOS

Диагональ экрана  
☐ 3.5"-4.9"  
☐ 5.0"-5.4"  
☐ 5.5"-5.9"

Выбор покупателей

Смартфон Apple iPhone 14 128 ГБ, синий

Экран 6.1" (2532x1170) OLED  
Панель: встроенная 128 ГБ  
Фото: двойная камера, основная 12 МП  
Процессор: Apple A15 Bionic  
Сети: карты: Dual SIM (nano-SIM + eSIM)

4.9/110  
663 тыс. человек интересовались за 2 месяца

59870₽ • 599  
Доставка завтра [12]  
MOBILE-ROOMS.RU 4.9/5K  
В корзину  
88 предложений от 59499₽

Выбор покупателей

Смартфон Apple iPhone 14 128 ГБ, фиолетовый

Экран 6.1" (2532x1170) OLED  
Панель: встроенная 128 ГБ  
Фото: двойная камера, основная 12 МП  
Процессор: Apple A15 Bionic  
Сети: карты: Dual SIM (nano-SIM + eSIM)

4.9/110  
663 тыс. человек интересовались за 2 месяца

59670₽ • 597  
Доставка завтра [12]  
MOBILE-ROOMS.RU 4.9/5K  
В корзину  
92 предложения от 59660₽

Выбор покупателей

Смартфон Apple iPhone 14 128 ГБ, тёмная ночь

Экран 6.1" (2532x1170) OLED  
Панель: встроенная 128 ГБ  
Фото: двойная камера, основная 12 МП  
Процессор: Apple A15 Bionic  
Сети: карты: Dual SIM (nano-SIM + eSIM)

4.9/110  
663 тыс. человек интересовались за 2 месяца

58990₽ • 590  
Доставка завтра [12]  
MOBILE-ROOMS.RU 4.9/5K  
В корзину  
79 предложений от 58990₽

Выбор покупателей

Смартфон Apple iPhone 14 128 ГБ, сияющая звезда

Экран 6.1" (2532x1170) OLED  
Панель: встроенная 128 ГБ  
Фото: двойная камера, основная 12 МП  
Процессор: Apple A15 Bionic  
Сети: карты: Dual SIM (nano-SIM + eSIM)

4.9/110  
663 тыс. человек интересовались за 2 месяца

58980₽ • 590  
Доставка завтра [12]  
MOBILE-ROOMS.RU 4.9/5K  
В корзину  
79 предложений от 58970₽

Выбор покупателей

Смартфон Apple iPhone 14 128 ГБ, (PRODUCT)RED

Экран 6.1" (2532x1170) OLED  
Панель: встроенная 128 ГБ  
Фото: двойная камера, основная 12 МП  
Процессор: Apple A15 Bionic  
Сети: карты: Dual SIM (nano-SIM + eSIM)

4.9/110  
663 тыс. человек интересовались за 2 месяца

59820₽ • 599  
Доставка завтра [12]  
MOBILE-ROOMS.RU 4.9/5K  
В корзину  
59 предложений от 59800₽

Выбор покупателей

Смартфон Apple iPhone 14 256 ГБ, синий

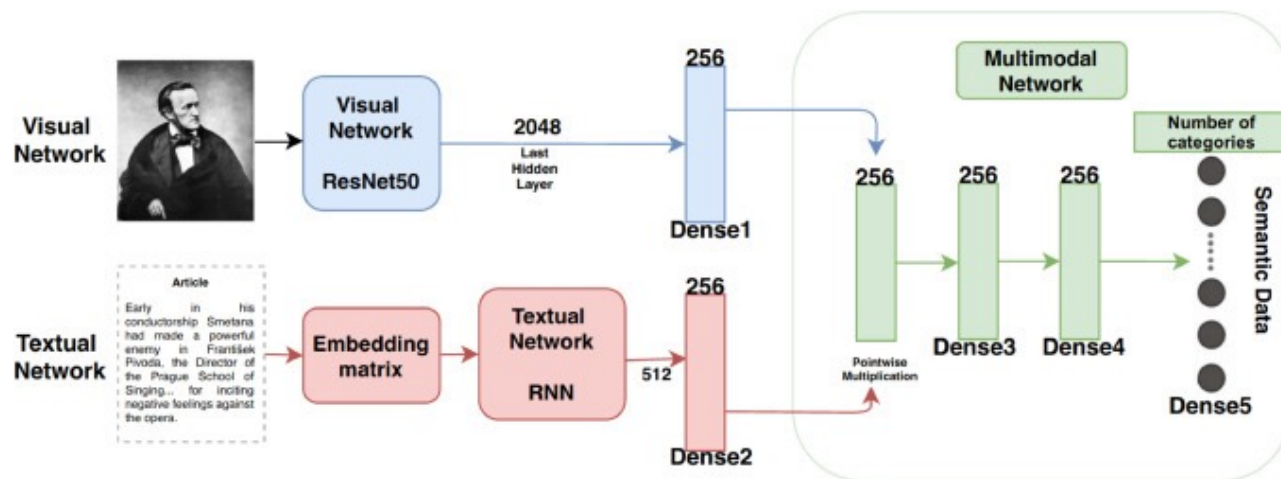
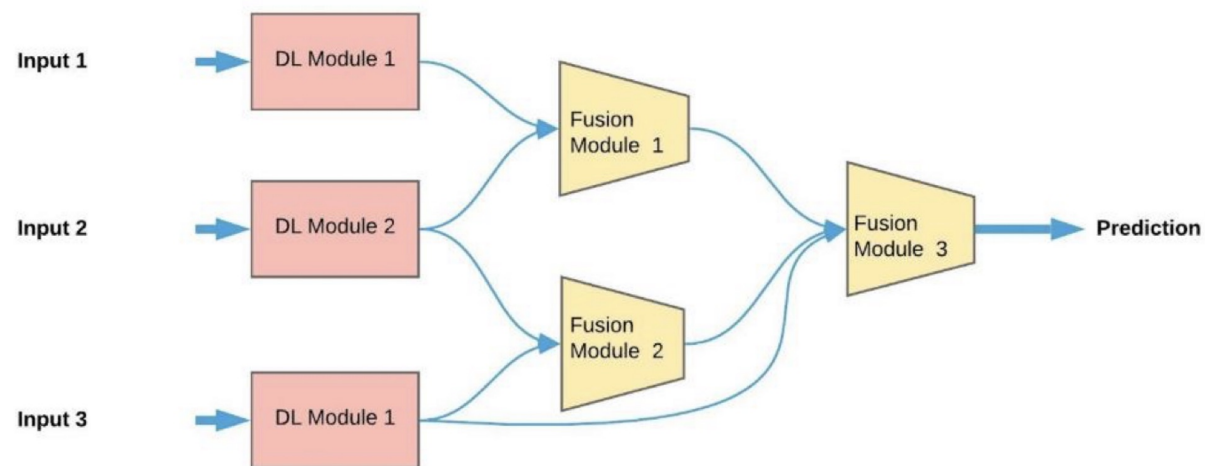
Экран 6.1" (2532x1170) OLED  
Панель: встроенная 256 ГБ  
Фото: двойная камера, основная 12 МП  
Процессор: Apple A15 Bionic  
Сети: карты: Dual SIM (nano-SIM + eSIM)

4.9/110  
663 тыс. человек интересовались за 2 месяца

67490₽ • 675  
Доставка завтра [12]  
MOBILE-ROOMS.RU 4.9/5K  
В корзину  
55 предложений от 67490₽

# Мультимодальные модели

**Мультимодальные модели** — модели, которые принимают и обрабатывают совместно данные разной природы; они оперируют совместными представлениями всех данных



# Проблемы в данных

- Неполнота информации (отсутствуют параметры товаров);
- Разные стандарты заполнения параметров;
- Англицизмы, сокращения, опечатки;
- Высокие требования к качеству;
- Большая схожесть между разными товарами.

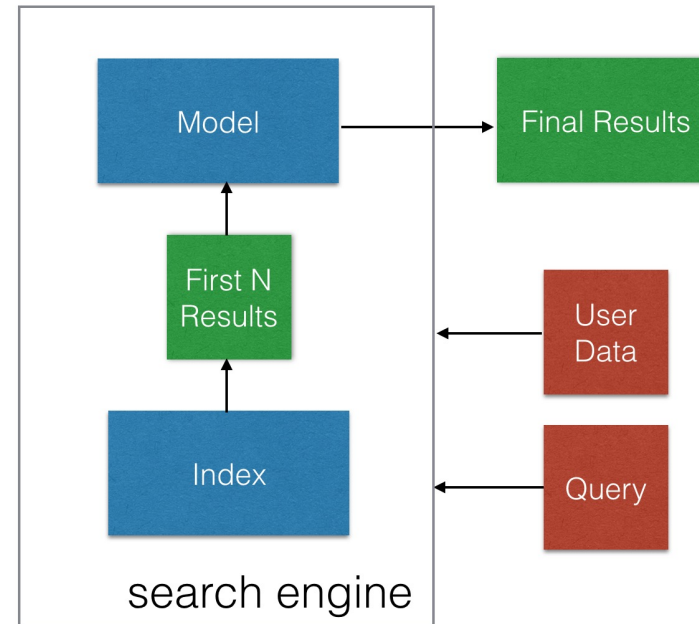
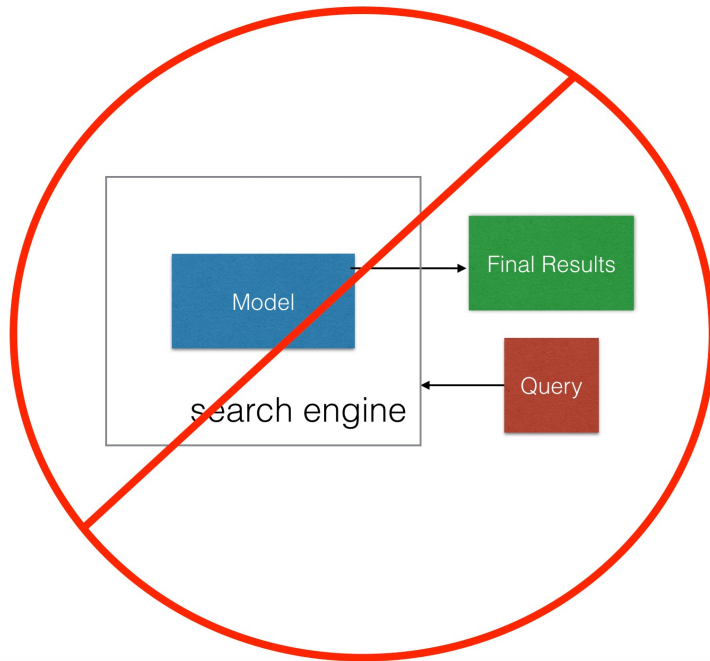
*Сложные, непривычные для обычных пользователей товары с сильно выраженной спецификой. Например, так выглядят названия товаров: «ВВГнг(А)-LS 4х6,0+1х4,0 1Кв, ТУ 16.К71-310-2001».*

На сайте производителя: «Посудомоечная машина ВЮ серия Economy qwerty123456 »

На сайте ритейлера: «Посудомоечная машина Economy (QWERTY123456) »

# Пайплайн матчинга

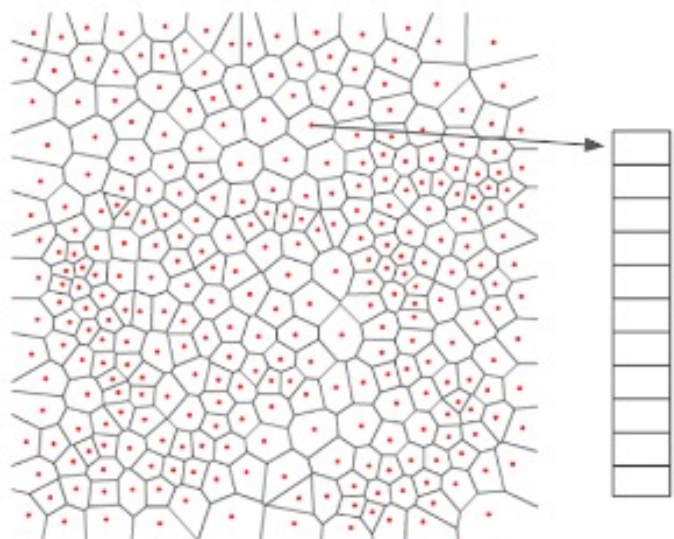
- Поиск и удаление дубликатов в базе документов;
- Кластеризация входного потока запросов;
- Замена товаров из корзины пользователя.





# Поиск кандидатов

Разбиение пространства поиска



FAISS, Annoy, NMSLib  
(ANN)

العربية  
العربية - Поиск Google  
العربية بين يديك  
العربية بين يديك **скачать**  
العربية بين يدي أولادنا **pdf**  
العربية **перевод русский**

Эвристики

# Метрики в ранжировании

**Качество/точность** — *насколько аккуратна система ранжирования?*

Измеряем возможности системы ранжировать релевантные документы выше нерелевантных.

**Эффективность** — *насколько быстро выдается ответ? Какое количество ресурсов необходимо для формирования ответа?*

Измеряем затраты на память и время формирования ответа.

**Удобство использования** — *насколько полезна система для решения задач?*

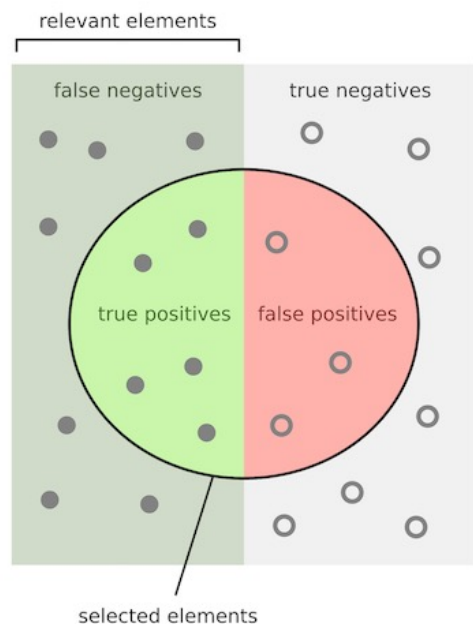
Пользовательские ощущения, UX.

## **Оценка качества ранжирования**

- Зафиксированный набор документов;
- Зафиксированный набор запросов;
- Оценки релевантности пар (в идеале оценки даются пользователями системы);
- Наборы должны быть репрезентативными.



# Оценка по топу – metric@k



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

## Precision

**Precision** — доля объектов, отнесённых классификатором к положительным и действительно являющимися положительными.

**relevant documents** — релевантные документы

**retrieved documents** — выданные документы

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap |\{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

## Recall

**Recall** показывает, какую долю объектов положительного класса из всех объектов положительного класса нашёл алгоритм.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap |\{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

## Fb-мера

$F_\beta$ -мера — агрегированный критерий качества **precision** и **recall**, где  $\beta$  показывает вес точности в метрике.

$F_1$  — среднее гармоническое **precision** и **recall** при  $\beta = 1$ .

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

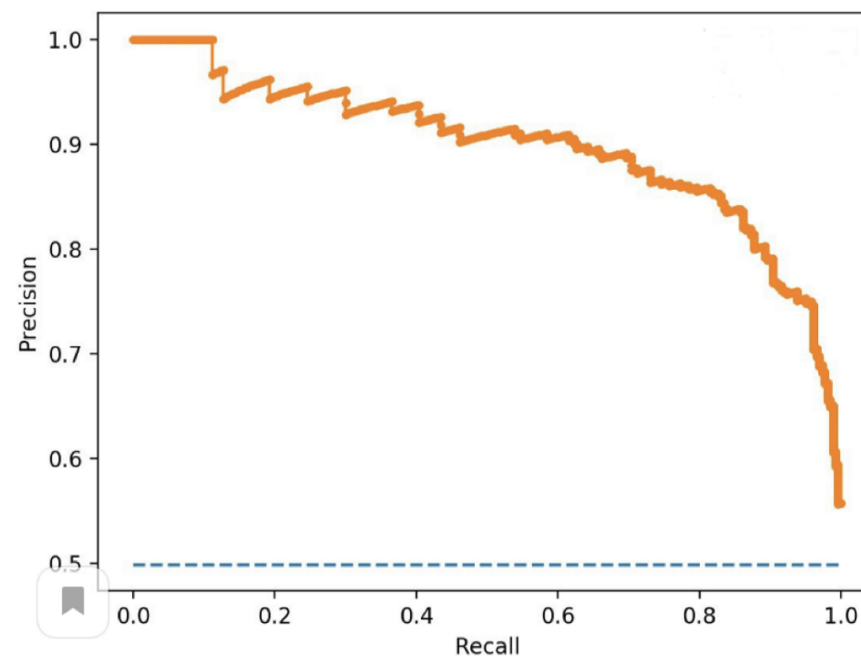
# PR-кривая

Алгоритм построения кривой

1. Сортируем предсказания по убыванию релевантности.
2. Считаем значение точности и полноты по первой паре.
3. Понижаем значение порога, чтобы выше порога было две пары.
4. Повторяем до тех пор, пока не добавим все элементы.
5. Опционально применить отсечение (Recall@Precision=N).

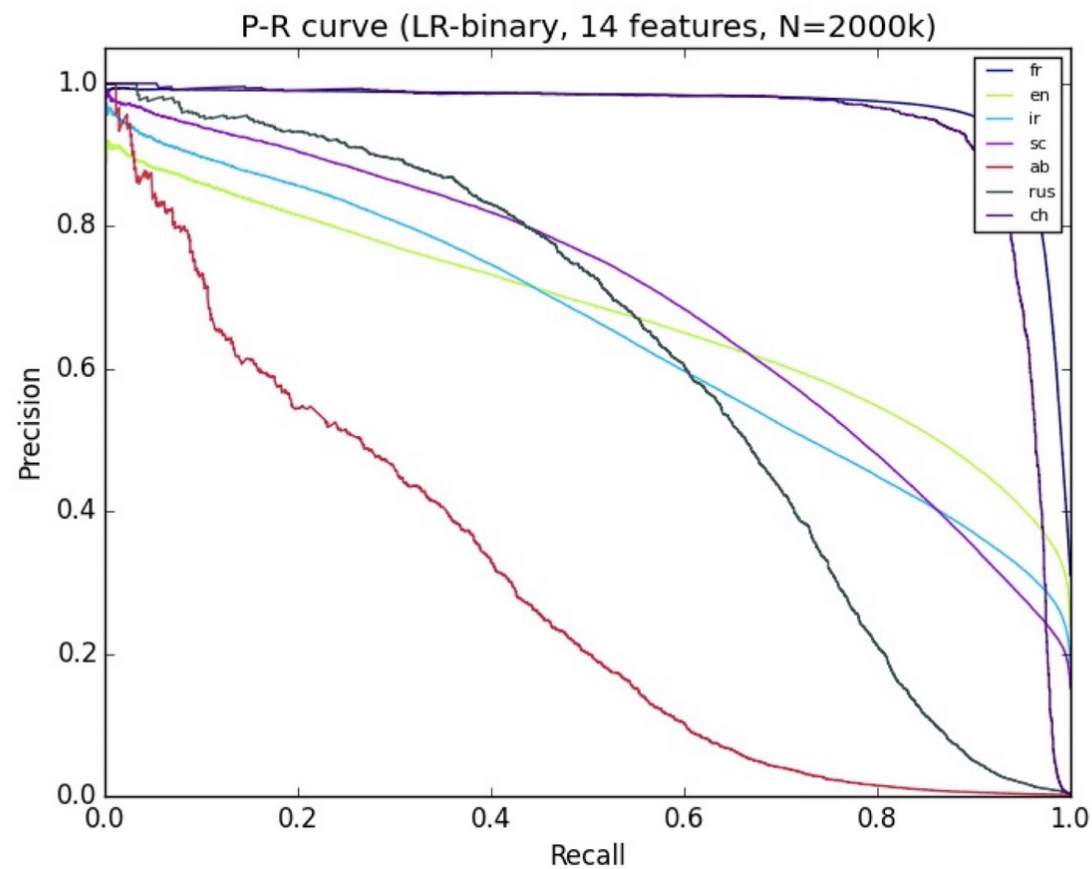
Метрикой будет площадь под PR-кривой (*PR-AUC*)

| ID оффера | ID модели | Предсказание формулы | Правильный ответ |
|-----------|-----------|----------------------|------------------|
| a01       | 1         | 6.4                  | 1                |
| a01       | 3         | 0.7                  | 0                |
| b02       | 2         | 0.6                  | 1                |
| c03       | 2         | -0.8                 | 0                |



# Оценка качества по PR

- PR-auc
- PR-auc@N



# Оценка качества ранжирования

**Average Precision (AP)** показывает, насколько много релевантных объектов сконцентрировано среди самых высоко оценённых. Чувствительна к порядку ранжирования в топе.

$$AP = \sum_K (Recall@k - Recall@[k - 1]) \cdot Precision@k$$

*MAP* — среднее *AP* по всем запросам *Q*

$$MAP = \frac{\sum_{q=1}^Q AP(q)}{Q}$$

| <i>k</i> | Document ID | Predicted Relevance | Actual Relevance   |
|----------|-------------|---------------------|--------------------|
| 1        | 06          | 0.90                | Relevant (1.0)     |
| 2        | 03          | 0.85                | Not Relevant (0.0) |
| 3        | 05          | 0.71                | Relevant (1.0)     |
| 4        | 00          | 0.63                | Relevant (1.0)     |
| 5        | 04          | 0.47                | Not Relevant (0.0) |
| 6        | 02          | 0.36                | Relevant (1.0)     |
| 7        | 01          | 0.24                | Not Relevant (0.0) |
| 8        | 07          | 0.16                | Not Relevant (0.0) |

Релевантных

Сумма

1

$0 + 1/1 = 1$

1

1

2

$1 + 2/3 = 1,67$

3

$1,67 + 3/4 = 2,42$

3

2,42

4

$2,42 + 4/6 = 3,08$

4

3,08

4

3,08

# Многоуровневый пример

Уровни релевантности:

1. Нерелевантно;
2. В целом релевантно;
3. Очень релевантно, точное соответствие.

|    | Gain |
|----|------|
| D1 | 3    |
| D2 | 2    |
| D3 | 1    |
| D4 | 1    |
| D5 | 3    |
| D6 | 1    |
| D7 | 2    |

# Многоуровневый пример

Уровни релевантности:

1. Нерелевантно;
2. В целом релевантно;
3. Очень релевантно, точное соответствие.

|    | Gain | Cumulative Gain |
|----|------|-----------------|
| D1 | 3    | 3               |
| D2 | 2    | 3+2             |
| D3 | 1    | 3+2+1           |
| D4 | 1    | 3+2+1+1         |
| D5 | 3    | 3+2+1+1+3       |
| D6 | 1    | 3+2+1+1+3+1     |
| D7 | 2    | 3+2+1+1+3+1+2   |

# Многоуровневый пример

Уровни релевантности:

1. Нерелевантно;
2. В целом релевантно;
3. Очень релевантно, точное соответствие.

|    | Gain | Cumulative Gain | Discounted Cumulative Gain                         |
|----|------|-----------------|--|
| D1 | 3    | 3               | $3$  |
| D2 | 2    | $3+2$           | $3 + 2/\log(3)$                                    |
| D3 | 1    | $3+2+1$         | $3 + 2/\log(3) + 1/\log(4)$                        |
| D4 | 1    | $3+2+1+1$       | $3 + 2/\log(3) + 1/\log(4) + 1/\log(5)$            |
| D5 | 3    | $3+2+1+1+3$     | ...  |
| D6 | 1    | $3+2+1+1+3+1$   | ...  |
| D7 | 2    | $3+2+1+1+3+1+2$ | $\text{DCD@7} = 3 + 2/\log(3) + \dots + 2/\log(8)$ |

# Многоуровневый пример

Уровни релевантности:

1. Нерелевантно;
2. В целом релевантно;
3. Очень релевантно, точное соответствие.

|    | Gain | Discounted Cumulative Gain                                   |
|----|------|--|
| D1 | 3    | 3  |
| D2 | 2    | $3 + 2/\log(3)$  |
| D3 | 1    | $3 + 2/\log(3) + 1/\log(4)$                                  |
| D4 | 1    | $3 + 2/\log(3) + 1/\log(4) + 1/\log(5)$                      |
| D5 | 3    | ...  |
| D6 | 1    | ...  |
| D7 | 2    | $\text{DCD@7} = 3 + 2/\log(3) + \dots + 2/\log(8) \sim 7.38$ |



# Многоуровневый пример

Уровни релевантности:

1. Нерелевантно;
2. В целом релевантно;
3. Очень релевантно, точное соответствие.

## Normalized DCG

$$nDCG@k = \frac{DCG@k}{IdealDCG@k}$$
$$nDCG \in [0, 1]$$

|    | Gain | Discounted Cumulative Gain                            |
|----|------|---|
| D1 | 3    | 3   |
| D2 | 2    | $3 + 2/\log(3)$                                       |
| D3 | 1    | $3 + 2/\log(3) + 1/\log(4)$                           |
| D4 | 1    | $3 + 2/\log(3) + 1/\log(4) + 1/\log(5)$               |
| D5 | 3    | ...   |
| D6 | 1    | ...   |
| D7 | 2    | $DCD@7 = 3 + 2/\log(3) + \dots + 2/\log(8) \sim 7.38$ |

$$IdealDCD@7 = 3 + 3/\log(3) + \dots + 1/\log(8) \sim 7.83$$

$$nDCD@7 = 0.942$$

# PFound

Значение метрики будет **оценкой вероятности** найти релевантный результат в выдаче модели:

$$p_{found} = \sum_{i=1}^n pLook[i] \cdot pRel[i]$$

$pLook[i]$  — вероятность просмотреть  $i$ -й документ из списка

$pRel[i]$  — вероятность того, что  $i$ -й документ окажется релевантным (например, 0%, 50%, 100% для трёхуровневой шкалы).

Для расчёта  $pLook[i]$  используются два предположения:

- результаты ранжирования просматриваются сверху вниз
- процесс прекращается в случае нахождения релевантного результата либо без каких-то определённых причин (например, если "надоело")

$$pLook[i] = pLook[i - 1] \cdot (1 - pRel[i - 1]) \cdot (1 - pBreak)$$

$pBreak$  — вероятность прекращения просмотра выдачи

# Базовые метрики

Среднеобратный ранг (Mean reciprocal rank, **MRR**) — среднее гармоническое между рангами.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

| Запрос  | Ответы                            | Правильный ответ | Ранг | Обратный ранг |
|---------|-----------------------------------|------------------|------|---------------|
| кочерга | кочерг, кочергей, <b>кочерёг</b>  | кочерёг          | 3    | 1/3           |
| попадья | попадь, <b>попадей</b> , попадьёв | попадей          | 2    | 1/2           |
| турок   | <b>турок</b> , турков, турчан     | турок            | 1    | 1             |

$$MRR = (1/3 + 1/2 + 1) / 3 = 11/18 \sim 0.61$$

# Базовые метрики

Среднеобратный ранг (Mean reciprocal rank, **MRR**) — среднее гармоническое между рангами.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

| Запрос  | Ответы                            | Правильный ответ | Ранг | Обратный ранг |
|---------|-----------------------------------|------------------|------|---------------|
| кочерга | кочерг, кочергей, <b>кочерёг</b>  | кочерёг          | 3    | 1/3           |
| попадья | попадь, <b>попадей</b> , попадьёв | попадей          | 2    | 1/2           |
| турок   | <b>турок</b> , турков, турчан     | турок            | 1    | 1             |

$$MRR = (1/3 + 1/2 + 1) / 3 = 11/18 \sim 0.61$$

## Kendall rank correlation coefficient

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}$$

$$\tau \in [-1, 1]$$

**number of concordant pairs** — количество согласованных пар (верно упорядоченных)

**number of discordant pairs** — количество несогласованных пар (неверно упорядоченных)

$$\binom{n}{2} = \frac{n(n-1)}{2} \text{ — биномиальный коэффициент}$$

Часто используется в статистике для оценки **ранговых корреляций**.

# Промежуточные выводы

- Имеем привилегию отказаться от выдачи;
- Важны только самые-самые первые результаты (1-3);
- Огромный дисбаланс (от нуля до тысяч матчей);
- Финальное решение можно предоставить классификатору;
- Отдельные метрики для разных этапов пайплайна;
- Метрики могут агрегироваться на уровне одного SKU;

# Подходы к решению задачи ранжирования

## 1. Pointwise (поточечный)

- Функция ошибки по конкретному объекту (в пару к запросу).

$$\sum_{q,j} l(f(\mathbf{x}_j^q), r_j^q) \rightarrow \min$$

## 2. Pairwise (попарный)

- Функция ошибки по паре объектов (в пару к запросу).

$$\sum_q \sum_{i,j: r_i^q > r_j^q} l(f(\mathbf{x}_i^q) - f(\mathbf{x}_j^q)) \rightarrow \min$$

## 3. Listwise (списочный)

- Функция ошибки на всём списке документов (для конкретного запроса).

$$l(\{f(\mathbf{x}_j^q)\}_{j=1}^{m_q}, \{r_j^q\}_{j=1}^{m_q}) \rightarrow \min$$

# Pointwise - BM25

$$score(D, Q) = \sum_{i=1}^n IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$$

$f(q_i, D)$  — частота слова  $Q_i$  в документе  $D$  (запрос  $Q_i$  содержит в себе слова  $Q_1, \dots, Q_n$ )

$$IDF = \log \frac{N}{n(q_i)}$$

$$IDF(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right) \quad \text{Сглаженный вариант IDF}$$

## Недостатки BM25

- Значение отрицательно, если производится расчёт для слова, входящего более чем в половину документов (частотные слова, stop-слова);
- Функция сконструирована вручную

# Pairwise

$A > B$  — документ A должен быть отранжирован выше документа B

$P(A > B)$  — вероятность того, что документ A должен быть отранжирован выше, чем B

$f : R^d \mapsto R$  — функция отображения документа в меру релевантности

$f(x_1) > f(x_2)$  — модель оценила релевантность одного документа выше другого

Функция потерь

$$C_{ij} \equiv C(o_{ij}) = -\bar{P}_{ij} \log P_{ij} - (1 - \bar{P}_{ij}) \log(1 - P_{ij})$$

$P_{ij}$  — предсказание модели

$\bar{P}_{ij}$  — целевая метка (*target*)



# Pairwise

Функция потерь

$$C_{ij} \equiv C(o_{ij}) = -\bar{P}_{ij} \log P_{ij} - (1 - \bar{P}_{ij}) \log(1 - P_{ij})$$

$P_{ij}$  — предсказание модели

$\bar{P}_{ij}$  — целевая метка (*target*)

$o_i \equiv f(x_i)$  — предсказание нашего алгоритма для одного объекта (*логит* или *скор*)

$o_{ij} \equiv f(x_i) - f(x_j)$  - сходство порядка в ранжировании

$P_{ij} = \frac{e^{o_{ij}}}{1 + e^{o_{ij}}}$  — функция отображения предсказания (логита) в вероятность.

Тогда функция потерь

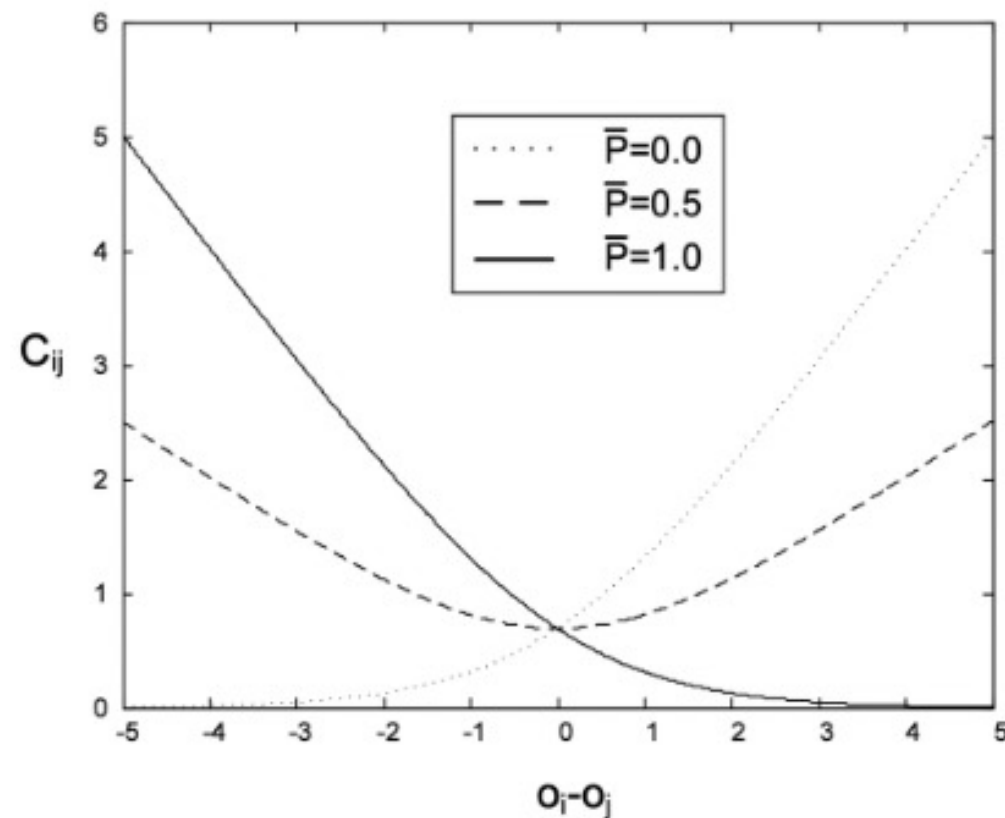
$$C_{ij} = -\bar{P}_{ij} o_{ij} + \log(1 + e^{o_{ij}})$$

# Pairwise

Тогда функция потерь

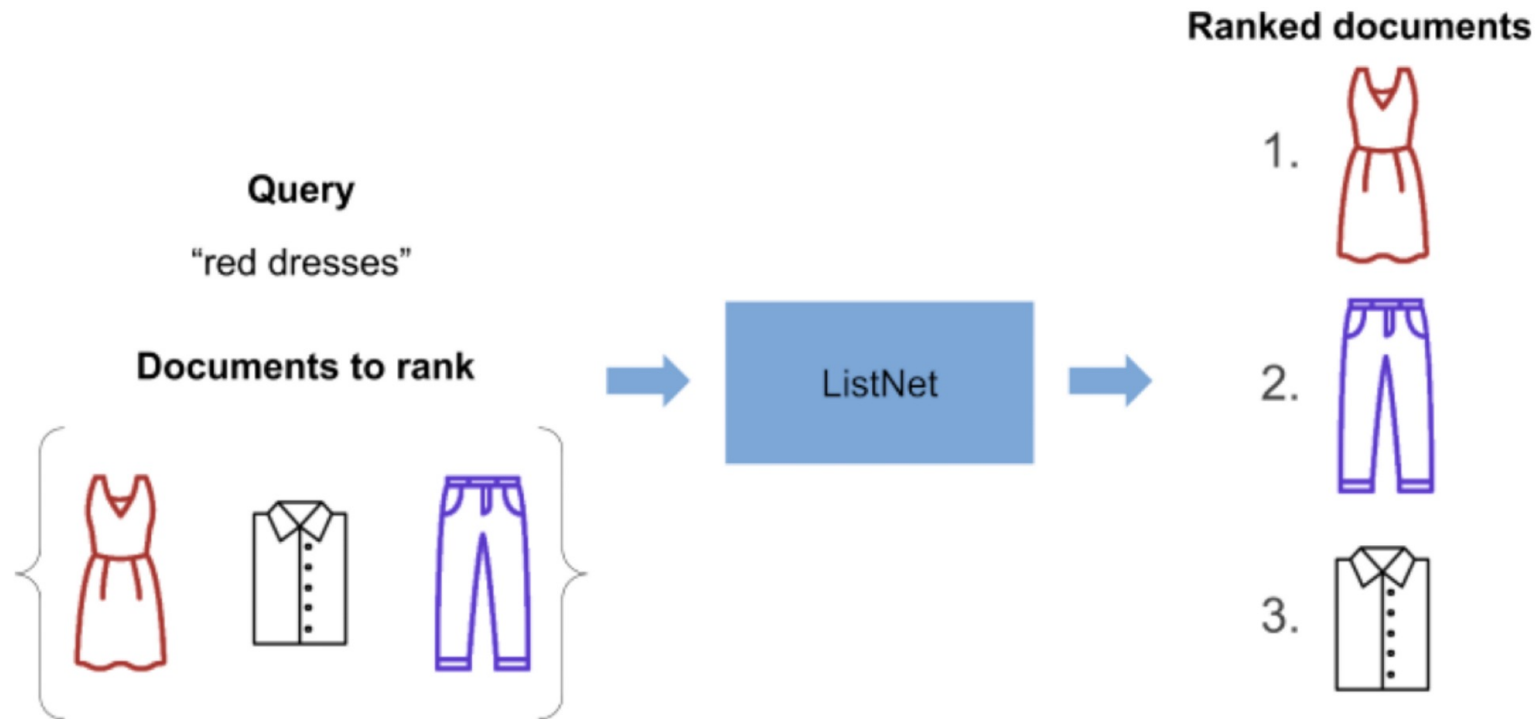
$$C_{ij} = -\bar{P}_{ij}o_{ij} + \log(1 + e^{o_{ij}})$$

- Линейная асимптотика более робастна при шумных метках.
- При таргете 0.5 симметрична, позволяет тренироваться на объектах одного ранга.



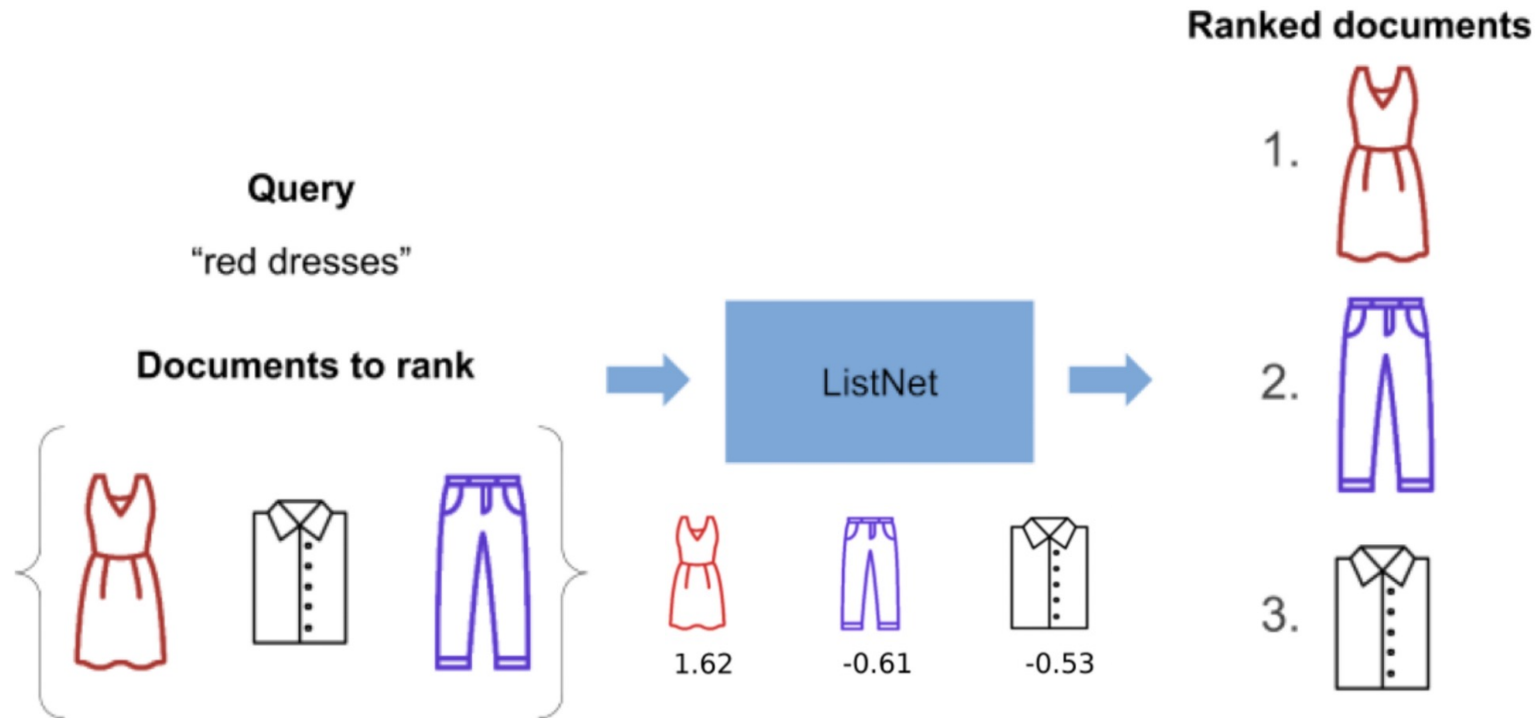
# Listwise

Хотим, чтобы алгоритм учился на всем множестве релевантных документах для запроса



# Listwise

Хотим получить распределение вероятностей похожее на нашу разметку



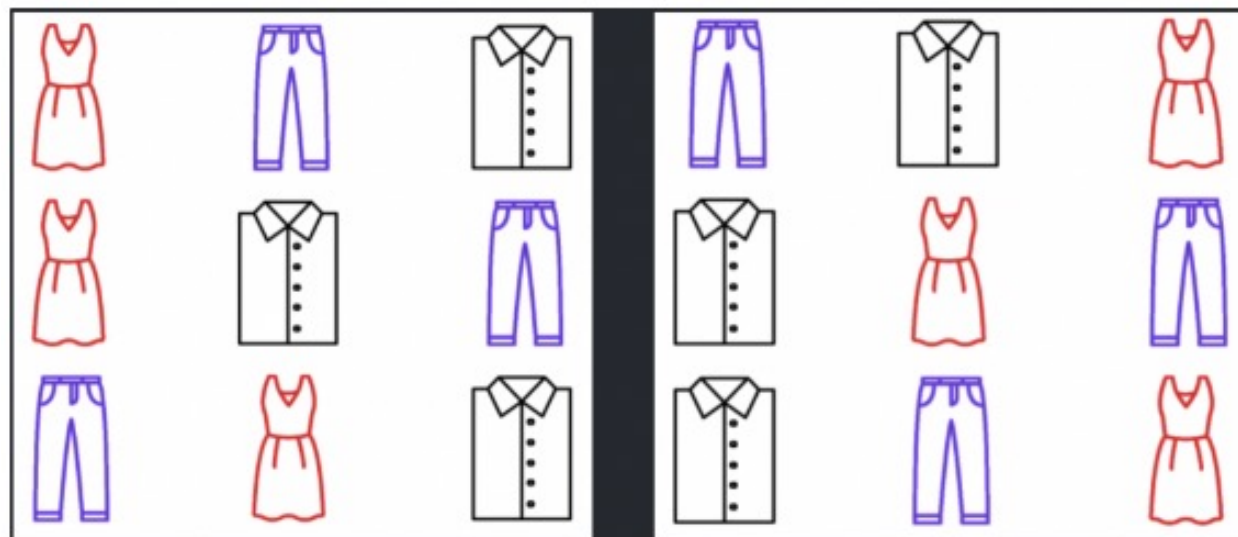
# Listwise

Сделаем предположение, что **любая перестановка документов возможна**, но при этом разные перестановки могут иметь **разную вероятность**

$\pi = \langle \pi(1), \pi(2), \dots, \pi(n) \rangle$  перестановка

$$P_s(\pi) = \prod_{j=1}^n \frac{\phi(s_{\pi(j)})}{\sum_{k=j}^n \phi(s_{\pi(k)})}$$

- вероятность возникновения такой перестановки



# Listwise

$$P_s(\pi) = \prod_{j=1}^n \frac{\phi(s_{\pi(j)})}{\sum_{k=j}^n \phi(s_{\pi(k)})}$$



1.62



-0.61



-0.53

$$P_s(\pi) = \prod_{j=1}^n \frac{\phi(s_{\pi(j)})}{\sum_{k=j}^n \phi(s_{\pi(k)})} = \frac{\phi(s_1)}{\phi(s_1) + \phi(s_2) + \phi(s_3)} \cdot \frac{\phi(s_2)}{\phi(s_2) + \phi(s_3)} \cdot \frac{\phi(s_3)}{\phi(s_3)} = \frac{e^{s_{dress}}}{e^{s_{dress}} + e^{s_{pants}} + e^{s_{shirt}}} \cdot \frac{e^{s_{pants}}}{e^{s_{pants}} + e^{s_{shirt}}} \cdot \frac{e^{s_{shirt}}}{e^{s_{shirt}}} = 0.3917 \text{ (39.17\%)}$$

Выводы для метода:



















- Наибольшая вероятность у перестановки, в которой объекты отсортированы в порядке убывания.
- Наименьшая вероятность у перестановки, в которой объекты отсортированы в порядке возрастания.
- Количество перестановок равно  $n!$ .

| Rank 1 | Rank 2 | Rank 3 |        |
|--------|--------|--------|--------|
|        |        |        | 42.59% |
|        |        |        | 39.17% |
|        |        |        | 8.58%  |
|        |        |        | 0.92%  |
|        |        |        | 7.83%  |
|        |        |        | 0.91%  |

# Listwise

Выводы для метода:

- Наибольшая вероятность у перестановки, в которой объекты отсортированы в порядке убывания.
- Наименьшая вероятность у перестановки, в которой объекты отсортированы в порядке возрастания.
- Количество перестановок равно  $n!$  – медленно будем обучать...

| Rank 1  | Rank 2  | Rank 3  |        |
|---|---|---|--------|
|    |    |    | 42.59% |
|    |    |    | 39.17% |
|    |    |    | 8.58%  |
|    |    |    | 0.92%  |
|   |   |   | 7.83%  |
|  |  |  | 0.91%  |

# Listwise

Выводы для метода:



















- Наибольшая вероятность у перестановки, в которой объекты отсортированы в порядке убывания.
- Наименьшая вероятность у перестановки, в которой объекты отсортированы в порядке возрастания.
- Количество перестановок равно  $n!$
- TopOne Probability – вероятность того, что объект  $j$  находится на первом месте в отранжированном списке

$$P_s(j) = \sum_{\pi(1)=j, \pi \in \Omega_n} P_s(\pi) \cdot \text{все перестановки, где первый } j$$

$$P_s(j) = \frac{\phi(s_j)}{\sum_{k=1}^n \phi(s_k)}$$

Сводим задачу к оценке распределений, что  $j$  объект в разметке должен иметь заданную вероятность

$$L(y^{(i)}, z^{(i)}) = - \sum_{j=1}^n P_{y^{(i)}}(j) \log(P_{z^{(i)}}(j)),$$

| Rank 1  | Rank 2  | Rank 3  |        |
|---|---|---|--------|
|    |    |    | 42.59% |
|    |    |    | 39.17% |
|    |    |    | 8.58%  |
|    |    |    | 0.92%  |
|   |   |   | 7.83%  |
|  |  |  | 0.91%  |