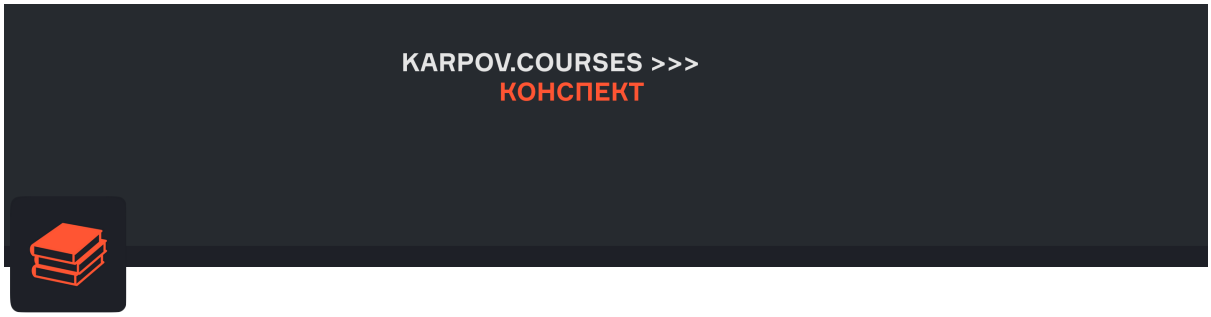


> Конспект > 10 урок > Толока как инструмент для оценки моделей и обновления датасетов

> Оглавление

> Краудсорсинг

Краудсорсинг — это привлечение для выполнения некоторой работы большого числа добровольцев и вместе с тем передача некоторых функций и задач неопределённому кругу исполнителей. Происходит от английского слова crowd (толпа).

Смысл сервиса краудсорсинга заключается в организации некоторой площадки, где с одной стороны предлагается некая деятельность от заказчика, а с другой — рабочие руки, которые умнее и быстрее компьютеров в разных непонятных и субъективных задачах.

Примеры таких площадок:

1. Яндекс.Толока́
2. Amazon Mechanical Turk

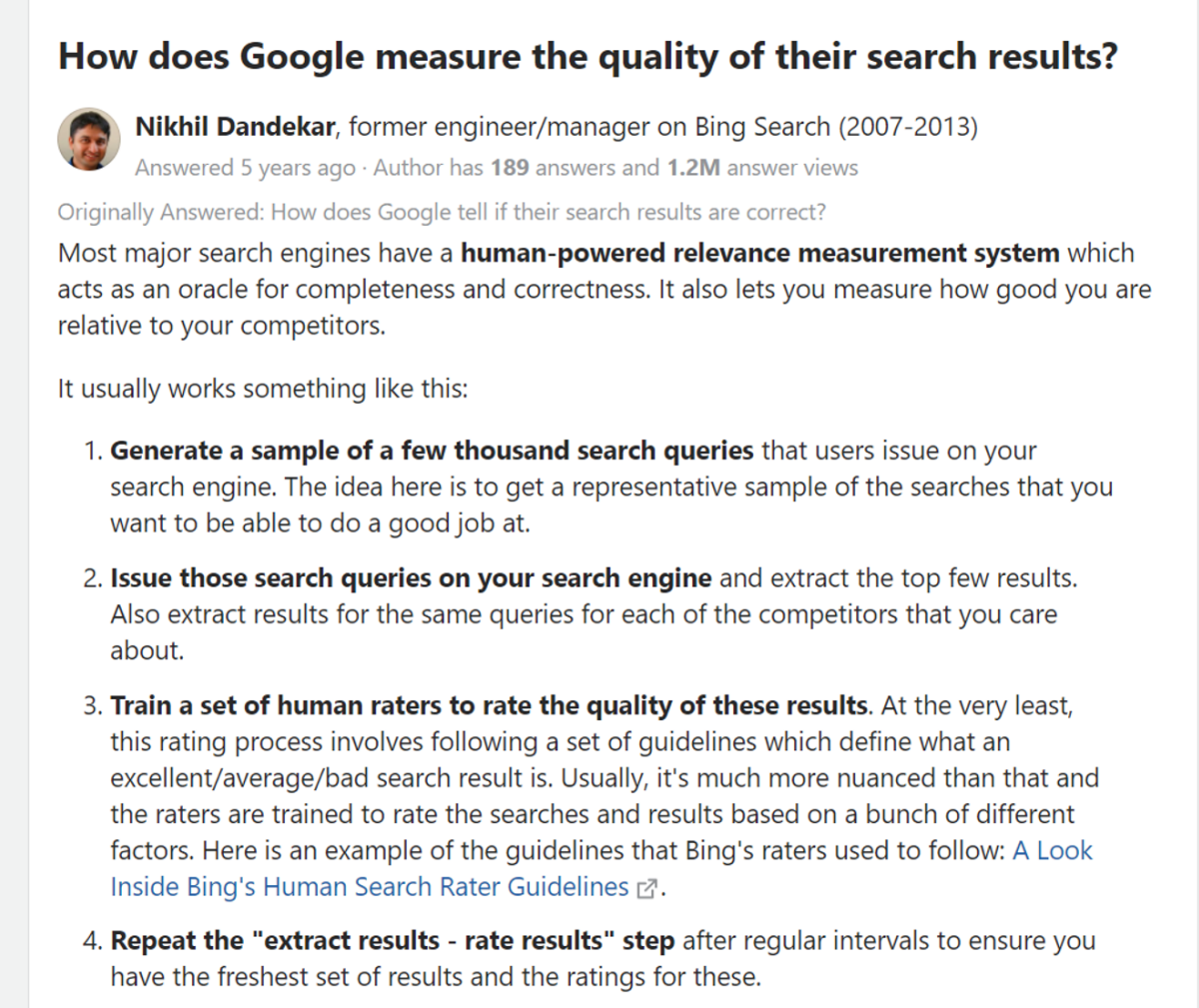
Главный принцип задач — это простота каждого отдельного задания. Для его достижения необходимо одну большую и сложную задачу разбивать на подзадачи, каждая из которых будет решаться сотнями или тысячами людей индивидуально.

> Построение пайплайна для ранжирования или поиска

Представим следующую ситуацию: вы внедрили новый мощный алгоритм или метод (например, BERT в задачу ранжирования), и необходимо продемонстрировать начальнику, что он работает намного лучше, чем предыдущий метод. С эти может помочь подобный сервис.

Можно создать такой валидационный набор данных, который не будет собран с логов с какими-то прокси-метриками, а будет оцениваться естественным внутренним чувством релевантности множества живых людей, перед которыми поставили задачу оценить по шкале от 1 до 5, насколько им нравится в выдаче для конкретного запроса данный сайт вместо старого.

Обычно именно так крупные корпорации замеряют качество модели. Приведём здесь в качестве примера ответа менеджера компании Google, который на рубеже 2010-х годов работал в поисковой системе Bing.



How does Google measure the quality of their search results?

Nikhil Dandekar, former engineer/manager on Bing Search (2007-2013)
Answered 5 years ago · Author has **189** answers and **1.2M** answer views

Originally Answered: How does Google tell if their search results are correct?

Most major search engines have a **human-powered relevance measurement system** which acts as an oracle for completeness and correctness. It also lets you measure how good you are relative to your competitors.

It usually works something like this:

1. **Generate a sample of a few thousand search queries** that users issue on your search engine. The idea here is to get a representative sample of the searches that you want to be able to do a good job at.
2. **Issue those search queries on your search engine** and extract the top few results. Also extract results for the same queries for each of the competitors that you care about.
3. **Train a set of human raters to rate the quality of these results**. At the very least, this rating process involves following a set of guidelines which define what an excellent/average/bad search result is. Usually, it's much more nuanced than that and the raters are trained to rate the searches and results based on a bunch of different factors. Here is an example of the guidelines that Bing's raters used to follow: A Look Inside Bing's Human Search Rater Guidelines.
4. **Repeat the "extract results - rate results" step** after regular intervals to ensure you have the freshest set of results and the ratings for these.

Ответ на соответствующий вопрос менеджера компании Google

Процесс проверки будет выглядеть следующим образом:

1. Подготавливаем выборку из генеральной совокупности, чтобы она отображала "естественное распределение запросов";
2. Обученная модель генерирует предполагаемую выдачу;
3. Формулируем чёткие критерии (основные правила) системы оценивания для "толпы";
4. Получаем результаты разметки "толпой";
5. Проделываем цикл заново (переходим к пункту 1).

Особенность такого подхода в том, что можно сравнивать не только две выдачи вашей внутренней разработки, но и смотреть на конкурента, у которого лучше топ-3 ссылки в ответ на некоторый запрос. Хоть вы и не знаете, какой PR-AUC у Google, но можете понять, как часто люди предпочитают вашу выдачу калифорнийской.

Рассмотрим теперь достаточно простой вопрос: для чего именно логично применять краудсорсинг? Про 2-й и 3-й пункт мы поговорили — это расчёт реальных метрик и сравнение моделей. Очевидно, что можно собирать датасет и с помощью исполнителей, и при этом даже не использовать базовую кандидатную модель!

Можно поставить задачу, к примеру, следующим образом: "сходи на сайт X и посмотри, есть ли там товар Y, и если он есть, то укажи на него ссылку." Таким образом, во-первых, вам не нужно пробегать по ссылкам всех магазинов, а во-вторых, для конкретного товара становится ясно, какая ссылка магазина-конкурента или магазина-партнёра относится к нему. Поставив подобный процесс на поток, можно значительно расширить скромную исходную выборку.

Помимо этого, можно выстраивать двухступенчатый пайплайн с верификацией ответов. То есть:

1. Один человек в рамках одного задания нашёл ссылку на товар;
2. Трое других в рамках другого задания проверяют, что по ссылке действительно тот товар.

Таким образом, у вас и данные чище, и деньги за неправильные ссылки можно не платить, то есть можно отменять выплату. Этот механизм поддерживается толокóй и описан в документации.

Можно пойти ещё дальше и представить более продвинутый пайплайн (скажем, что необходимо разметить светофоры на картинках в рамках разработки автопилота). Вопросы для разметчика тогда будут следующие:

1. Содержит ли фотография светофор?
2. Выделите светофор рамочкой (bounding box).
3. Правильно ли указана рамка для объекта? (при валидации чужой разметки)

Последний пример использования, который хочется отметить отдельно — это разметка данных для матчинга вблизи пороговых значений. Это те матчи, которые оценены высоко, однако модель не уверена, и, дообучив модель именно на них, можно более явно очертить разграничивающую линию между корректным и неправильным матчем. По сути это самые сложные примеры для обучения модели — и в то же время самые интересные. Это применимо и вне матчинга: допустим, в рамках модерации сообщений, где текст, получивший оценку выше порога 0.95, означает бан на форуме, а посты с оценкой от 0.85 до 0.95 отправляются на рассмотрение человеком. Абсолютно так же, как и в

примерах выше, после получения вердикта можно переиспользовать данные, получив таким образом active learning пайплайн.

Резюмируя, можно выделить следующие функции "толпы" разметчиков в пайплайне:

1. Аннотация тренировочных данных, сбор из интернета;
2. Оценка качества модели (разметка датасета под валидацию);
3. Сравнение выдачи разных алгоритмов;
4. Проверка размеченных или вновь собранных данных;
5. Разметка данных около порогового значения, т.е. на границе "уверенности" модели.

> Советы по настройке проектов

Теперь дадим несколько ценных советов и расскажем о лучших практиках:

- Инструкция не должна быть сложной: она не должна превышать по объёму три четверти листа А4, в ней должно быть не более 4-5 аспектов, на которые стоит обратить внимание.
- В идеале работа должна быть формализована в виде решающего дерева таким образом, что на каждое новое задание исполнитель идёт по инструкции и принимает решение (матч или не матч, 1 релевантность или 5, и т.д.).
- Действие в задании также должно быть простым.

Применительно к матчингу это означает, что не лишено смысла разбиение на какие-то крупные категории: техника, товары для дома, еда и т.д. В каждой из них выделен список характеристик, на которые точно нужно обращать внимание, приведены конкретные примеры, покрывающие 95% самых частых кейсов, будь то цвет, размер или вес.

- Задание должно быть удобным, им должно быть приятно пользоваться.

Если вы предлагаете пользователю, например, найти товар на сайте — сгенерируйте ссылку на поиск именно на том сайте и добавьте кнопку “перейти в поиск”, чтобы одним кликом экономить время исполнителя, увеличивая тем самым производительность труда.

Поскольку речь идёт о заработке денег, то площадка непременно привлечёт внимание жуликов. Либо это будут люди, которые кликают просто наугад, либо это будут боты, выполняющие ту же функцию, но куда быстрее.

- Используйте CAPTCHA, либо ограничение на максимальное количество отправок в минуту, а также проверку того, что все ответы разные, а текстовые поля для заполнения не пустые.
- Для понимания минусов инструкции, а также оценки временных затрат и выставления фильтров необходимо самостоятельно прорешать задание, выступив в роли разметчика. Так вам сразу станут очевидны недостатки описания действий пользователя, и у вас появится идея о том, на что нужно обращать внимание.
- Разметить часть данных самостоятельно полезно и потому, что необходимо иметь не только один датасет для разметки, но и два дополнительных. Первый — это тренировочный набор, для которого будет расписано, что, где, как и почему. Второй вспомогательный датасет — это экзамен, где уже нет подсказок и где вы вольны выставить порог знаний, долю ошибок или правильных ответов.

Если пользователь не проходит порог, то не допускается до выполнения задания. Тут вы отсеиваете и самых примитивных ботов, и немотивированных и ленивых людей, которые даже не могут прочитать инструкцию и подсказки.

Вам может показаться сложным настройка совокупности всех фильтров и факторов, но это нормально — определение цены разметки та ещё наука. Самое главное в работе с толокой — это итерации изменения и улучшения задания. За создание проекта нужно садиться как можно раньше: пока вы соберёте витрину данных, обучите модель и пофиксите все баги, появятся уже новые данные. Этот процесс в общем-то и не должен прекращаться. Проверять итерации можно на малом количестве данных и занедорого. Методично, итеративно, пошагово — вот ключевые слова.

- Но есть ещё один нерассмотренный вопрос: как же валидировать и подытожить работу разметчиков? Для этого на самом деле нужен третий датасет, для которого также известна разметка. Из всех трёх он должен быть самым большим. Это датасет так называемых "ханнипоинтов" (honey-points), о которых исполнитель не знает. Они подмешиваются в общую выдачу, по 2–4 штуки на страницу. Если вы видите, что человек за 5 страниц ни разу не ответил правильно — можно его забанить. Поскольку таких секретных вопросов будет показываться пользователю много, то и

датасет должен быть немаленький, иначе исполнитель просто поймёт, что что-то не так, и может начать отвечать на них правильно, при этом портя остальную разметку. Также это позволяет устанавливать пользователю значение навыка разметки для вашей задачи, равный, условно, доле правильных ответов за последние 20 "ханнипоинтов". Есть система добавочных поощрений за высокий уровень для стимуляции исполнителей. Эти коэффициенты опять же требуют настройки и итераций.

- Но что более важно — такая система навыков, на лету оценивающая исполнителей, помогает и в агрегации их ответов.

Под агрегацией понимается проставление финального ответа по данным нескольких пользователей, ведь обычно каждое задание, каждый матч или поисковая выдача показываются не одному, а 3, 5 или даже 7 людям. Это увеличивает кратно время и цену задания, однако без разметки с перекрытием никогда нельзя быть уверенным в том, что вообще отмечают люди. Примитивный метод — выбор ответа, за который голосует большинство, однако можно брать голоса взвешенно. Есть также и алгоритм умной агрегации по методу Дэвида-Скина (Dawid-Skene), который предлагает та же толока. Он учитывает индивидуальные особенности разметчиков, их склонность отвечать так же, как и большинство. Детальнее с его описанием можно ознакомиться, вбив в поисковую строку имена авторов метода.

> Резюме

- Доразметка данных толпой сторонних людей за деньги может пригодиться как для оценки обучаемых моделей, так и для многоуровневых пайплайнов подготовки данных.
- Задание для разметчиков должно быть максимально простым и понятным, а инструкция — короткой и последовательной.
- Платформы краудсорсинга имеют большое количество параметров для настройки, и особенно внимательно нужно следить за составляемой инструкцией к выполнению и параметрами бана нерадивых пользователей.
- Для достижения успеха нужно повторить процесс несколько раз, пока большая часть неочевидных проблем не будет решена. Потратить на это два-три месяца вполне естественно. В процессе подготовки важно не забыть собрать 2–3 датасета, которые будут использоваться в качестве

контроля исполнителей, в том числе и скрытого, на основе которого можно рассчитать вес, или среднюю корректность пользователя, и использовать её для усреднения ответов.