Описание предметной области

В мире существует множество поисковых систем: Bing, Google, Яндекс. Рынок поисковых систем очень конкурентный: любая поисковая система стремится привлечь к себе как можно больше пользователей, чтобы иметь большую прибыль с показа рекламы. Именно поэтому важными аспектами любой поисковой системы являются релевантность и качество поисковых страниц на выдаче. Как правило, учитываются первые топ 10 страниц, ведь именно их увидит пользователь. Яндекс каждый год вкладывает огромные средства на улучшения алгоритмов ранжирования. Запросы пользователя бывают разной тематики, например: медицина, спорт, новости. Такие тематики называются поисковыми срезами. Например, одна команда может улучшать выдачу на спортивном срезе, другая на медицинском. В этой работе речь пойдет об мониторинге качества поисковой выдачи на новостном срезе. Это задача крайне важна, т.к. доля новостных запросов очень велика и Яндексу крайне важно иметь качественную выдачу, которая сможет в полной мере удовлетворить интересы пользователей. Например, если человек идет по улице и видит, что на улице пожар, он может захотеть узнать, что случилось. Для этого он напишет запрос в поисковую систему: «пожар в Москве что случилось». На такой запрос хорошая поисковая система должна правильно обработать поисковый запрос и документы-кандидаты, которые пойдут пользователю на выдачу. В данном случае поиску важно показать новые, актуальные, «свежие», как мы говорим, документы. В жизни часто происходят незапланированные внезапные новостные события: аварии, случаи смерти, стихийные бедствия. Как правило, такие внезапные новостные событие сильно интересуют людей в первые секунды их происшествия. В таком случае задача поисковой системы как можно раньше актуализировать выдачу по тому или иному новостному событию, от этого будет зависеть счастье пользователей и прибыль компании в целом. В данной работе представлена разработка метрики качества поисковой выдачи на таких новостных событиях. Чтобы иметь возможность построить такую метрику было решено множество задач внутри компании. Для предсказания новостных запросов о внезапных новостных событиях на которые Яндексу важно иметь актуальную выдачу с первых секунд были использованы алгоритмы машинного обучения, catboost (градиентный бустинг), различные нейросетевые факторы, а также нейросеть LSTM для анализа временных рядов динамики запроса. На различных этапах подсчета метрики были применены нейронные сети, а именно BERT, реализовано множество Map Reduce вычислений на языке запросов YQL (разработка Яндекса, аналог SQL). Не обошлось и без регулярной ручной, асессорской разметки. Затраты на разметку были снижены во много раз благодаря машинному обучению.