Алгоритм генерации метатегов для поискового продвижения сайта

Апухтин Дмитрий Игоревич,
ООО «Инструменты генерации дохода»

Генерация метатегов — это задача, которой занимаются все SEO-специалисты для каждого проекта.

Написание текстов title и description для одного проекта отнимает несколько часов времени специалиста. Чтобы выделить время на более интеллектуальные задачи, решили автоматизировать процесс написания метатегов.

Нужно отметить, что метатеги подбираются на каждую страницу. При этом на одну страницу могут вести сразу несколько поисковых запросов, их все нужно учитывать при подборе метатегов.

Первичный анализ

На первом этапе был проведен анализ по ограничениям, которые накладываются на title и description. В первую очередь — это длина и допустимые символы.

Для этого провели консультацию с нашими экспертами, а также проверили метатеги, которые специалисты пишут вручную.

Для title выделили разрешенный список знаков препинания, а также наложили ограничения на переспам.

Для description установили границы от 100 до 300 символов. Знаки пунктуации допускаются.

Генерация title

Исследование нескольких тысяч title в выдаче по коммерческим запросам показало, что более 50% страниц в title содержат название компании и обычно оно указывается в конце title.

В связи с этим в алгоритм был встроен поиск названия компании/магазина на сайте по определенным шаблонам.

Например:

· В компании «Бодрый ростовщик»

· Интернет-магазин «Золотой плафон»

Анализ title в выдаче и title, которые пишут наши спецы показал, что title содержат запросы, по которым идет продвижение, но обычно это не полная форма. Также часто игнорируются коммерческие маркеры: цена, купить. Поэтому в основу алгоритма генерации title лег поиск кандидатов в title, а затем выбор кандидата с максимальным покрытием всех запросов, но с учетом фильтров.

В качестве кандидатов используются title страниц конкурентов, наш фактический title (если есть), наши заголовки на странице (по тегу h1). Каждый кандидат дополнительно разбивается на предложения, если это возможно. Если среди запросов есть 2 запроса, которые не пересекаются между собой по нормализованным словам, то в качестве кандидата используется их объединение через запятую. Также все запросы добавляются в кандидаты.

Все кандидаты проходят фильтрацию. В качестве одного из фильтров используется проверка, что все слова кандидата есть на продвигаемой странице. Сделано это, чтобы не выбрать

нерелевантного кандидата или кандидата, в котором есть, например, название нашего конкурента.

Другим фильтром является проверка на переспам. Для этого выработали правила, по которым определяли является ли title переспамленным или нет.

Считаем title переспамленным, если:

1. Одно слово повторяется 3 и более раз.

«Купить окна», «пластиковые окна недорого», «цены на окна»

2. Есть 3 и более слов, которые в title встречаются более 1 раза.

«Цена на узи почек москва», «узи мочевого пузыря и почек в москве»

3. Есть биграмма, которая встречается больше 1 раза.

«Запчасти Шевроле — каталог, цены», «купить запчасти Шевроле в Москве»

После фильтрации все оставшиеся кандидаты ранжируются по покрытию. Для расчета покрытия использовались триграммы. Триграмма — это набор 3 символов, которые мы получаем скользящим окном по запросу или кандидату с шагом один. Т.е. для запроса «купить детские сапоги» будут получены триграммы «куп», «упи», «пит» и т.д.

Покрытие одного запроса рассчитывалось следующим образом:

Определяем триграммы для запроса — T_q

Определяем триграммы для кандидата — T_c

Определяем триграммы запроса, которые есть также в триграммах кандидата: $T_{q \cap c} = T_q \cap T_c$

Покрытие определяем, как количество триграмм запроса, которые есть также в триграммах кандидата к количеству триграмм запроса:

$$Coverage_{qc} = \frac{|T_{q \cap c}|}{|T_q|}$$

Для определения покрытия кандидатом всех запросов на странице берется среднее значение покрытий:

$$Coverage_c = \frac{1}{N} \sum_{i=1}^{N} Coverage_{ic},$$

где N — число запросов на странице, c — индекс кандидата

В качестве другого решения рассматривалась нормализация запросов и кандидатов, а затем расчет tf-idf между запросом и кандидатом. Триграммы были выбраны, потому что они позволяют частично учитывать однокоренные слова. Например, «для дома» и «домашний» дадут ненулевое значение при покрытии триграммами, а при расчете tf-idf был бы 0.

Кандидат с максимальным покрытием выбирается для этой страницы, после чего происходит расширение его названием компании или магазина, которое было получено ранее.

Примеры сгенерированных тайтлов:

Запросы	Title
мебель для телевизора мебель под телевизор мебель для тв	Мебель под телевизор в современном стиле от компании «ТумбаДом»
летний корпоративный отдых корпоративный отдых на природе корпоративный отдых активный корпоративный отдых организация корпоративного отдыха	Организация корпоративного выезда на природу | КорпКлаб
бассейн детский для дачи бассейн детский купить бассейн для детей купить детские бассейны для малышей детский бассейн купить в москве	Детские бассейны в Москве — купить бассейн для детей от компании «Акваполис»

Генерация Description

Генерация description представляется гораздо более сложной задачей, так как в description обычно используют длинные предложения, которые должны быть согласованы между собой. Даже использование нейронных сетей для генерации текстов не может дать приемлемый уровень текстов для русского языка. Поэтому идея писать тексты с нуля была отклонена практически с самого начала.

Если текст не получается сгенерировать, значит нужно использовать готовый. Но использовать текст с чужих страниц нельзя, так как у нас свои конкурентные преимущества и действительно релевантным будет текст только с нашего сайта. Поэтому задача генерации description превращается в задачу поиска подходящего текста на нашем сайте. Для того, чтобы не придумывать огромное количество правил, решили обучить классификатор, который будет принимать на вход некоторый текст и определять, является ли этот текст частью description или нет.

Для обучения необходимо большое количество данных. Для этого, аналогично title, выкачали страницы из выдачи по коммерческим запросам. В качестве положительных примеров использовались фактические description, разбитые на предложения. В качестве отрицательных примеров использовались остальные предложения на странице.

Все полученные предложения представляют собой корпус, предложения в этом корпусе предварительно нормализовали.

Для обучения необходимо представить предложения в виде вектора числе. Хорошо зарекомендовавший себя метод это TF-IDF. Библиотека sklearn предоставляет возможность такого преобразования с помощью метода TfidfVectorizer. В этом методе можно указать минимальное количество повторения терма, ниже которого терм не будет включен в словарь. Это позволяет очень просто удалить редкие слова. Часто это слова с орфографическими ошибками или специфичные термины. Удаление редких слов существенно сокращает объем словаря и снижает вероятность переобучения модели в дальнейшем.

Этот метод позволяет использовать не только одиночные слова, но и словосочетания из нескольких слов. Количество слов является входным параметром. Т.к. мы используем классификатор на длинных текстах, в которых много устойчивых словосочетаний (низкая цена, большой выбор и т.д.), интуитивно понятно, что использование биграмм (сочетаний из 2 слов) должно повысить качество модели.

Для обучения использовались несколько методов: LogisticRegression, RandomForest, SGD.

Самый лучший результат показал SGD (стохастический градиентный спуск).

В результате валидации на отложенной выборке из 92 тысяч был получен следующий результат:

Матрица ошибок классификации

	0	1
0	70061	2590
1	10257	9567

Отчет классификации

	precision	recall	f1-score	support
0	0.98	0.93	0.96	60996
1	0.24	0.50	0.32	2550
avg / total	0.95	0.92	0.93	63546

По результатам видно, что мы добились точности в 79% по положительному классу. Это очень хороший результат, если понимать, что при обучении, мы пометили все предложения на странице как отрицательные примеры и гарантировать что среди них действительно нет предложений похожих на description мы не могли.

Примеры ложноположительных предложений:

· У нас большой выбор медицинского оборудования, товары можно купить на выгодных условиях.

· Гарантия 5 лет!

· Преимущества покупки на сайте d-lock.

· Для записи на прием к ортодонту детской стоматологии «Мартинка», уточнения актуальной цены установки невидимых брекетов и стоимости других услуг звоните нашим консультантам по номеру, указанному в разделе Контакты.

· В магазине существует накопительная система скидок для постоянных клиентов.

· По лучшей цене быстро и удобно!

· Поэтому рассмотрим, по каким параметрам отличаются модели оперативной памяти.

· Выбирайте и заказывайте у нас — дешево, быстро и с надежной доставкой.

Полнота получилась очень низкой, но для работы алгоритма это не критично.

После обучения модели, которая сможет выделять предложения с нашей страницы, необходимо разработать алгоритм, как эти предложения соединять между собой. После анализа существующих description пришли к выводу, что большое количество description формируются по шаблону:

«Основная часть» + «Преимущества»

Основная часть — это предложение, которое отражает, что именно пользователь может найти на этой странице.

· Детская ортопедическая подушка TRELAX Optima Baby П03 создана для полноценного сна и отдыха детей в возрасте от трех лет.

· Новогодние платья для девочек, эффектные и оригинальные, можно купить в интернет-магазине Evikris.

· Высокоскоростной вибромассажер ERGO HAND — устройство для выполнения массажа в домашних условиях.

Преимущества — это предложения про выгоду, которую пользователь найдет на сайте, часто это информация про доставку, гарантию, качество, выгодные цены и т.д.

· Цена от 550 тыс.рублей

· Лучшее качество и прекрасный сервис

· Удобный каталог, а также бесплатная услуга доставки — наши главные преимущества

Для определения, к какому же типу принадлежит предложение, а также для их последующего ранжирования используется TF-IDF. Для этого считаем TF-IDF между всеми запросами и предложением, усредняем его. Если TF-IDF меньше некоторого порога, то говорим, что это предложения типа «Преимущества», иначе это «Основная часть» и предложения этого типа по среднему значению TF-IDF ранжируются между собой.

В качестве порога для минимального TF-IDF был выбран не ноль, т.к. есть ряд популярных слов, которые часто есть в запросах, но при этом часто встречаются и в предложениях типа «Преимущества». Это предлоги, слова купить, заказать, цена и т.д.

Далее формируем уже финальный description:

Среди предложений типа «Основная часть» выбираем предложение с максимальным средним TF-IDF. Если длина этого предложения больше, чем минимальная длина description, то возвращаем результат. Если нет, то среди оставшихся предложений типа «Основная часть» пытаемся найти предложение, у которого нет пересечений с уже выбранным предложением, по нормализованным словам, при этом игнорируются служебные части речи (предлоги, союзы).

Если такое предложение удалось найти, то добавляем его к первому и снова проверяем, достаточно ли длины.

Если не удалось, то аналогичным образом пытаемся набрать предложения типа «Преимущества». При это при выборе очередного предложения проверяем, что нет пересекающихся слов со словами, которые уже точно будут в финальном description. Сделано это для того, чтобы результирующий description выглядел естественно и не содержал переспама. Также есть ограничение на добавление преимуществ, потому что если их больше 3, то description выглядит, как несколько несвязных между собой коротких предложений. В этом случае и в некоторых других случаях (например, на странице нет текста, который мы можем использовать и поэтому нам не удалось подобрать description), эту работы придется выполнить специалисту.

Примеры сгенерированных description:

Запросы	Description

служба уничтожения клопов в квартире москваслужба уничтожения клопов в квартире ценауничтожение клоповуничтожение клопов в квартиреуничтожение клопов в квартире москвауничтожение клопов в квартире цена	Получить подробную консультацию, а также заказать выезд службы по уничтожению клопов в квартирах Москвы или области и обработку квартиры методом холодного или горячего тумана по доступной стоимости можно по телефонам: 7 495 975-70-69, 8 800 200-36-77.
выпечка оптом от производителявыпечка оптом от производителя в москвекондитерская фабрика в москвепроизводитель выпечкипроизводитель выпечки в москвефабрика кондитерских изделийфабрика кондитерских изделий москва	Наша фабрика производит более 250 наименований выпечки и кондитерских изделий по оригинальным рецептам. Мы предлагаем клиентам огромный выбор и выгодные цены.
мойка окон Бескудниковскиймойка окон Бибиревомойка окон Бирюлёво Восточноемойка окон Бирюлёво Западноемойка окон Богородское	Мы оказываем услуги по Мойке окон в Богородском и других районах Москвы. Доступные цены, Высокое качество, оперативное исполнение – вот основные преимущества сотрудничества с компанией «Клинта».

Сейчас алгоритм генерации метатегов имеет серьезный недостаток: в обоих случаях нам нужен контент страницы, а для description нужно еще и наличие подходящего текста на этой странице. Если в работу берется проект, для которого требуется создать новые страницы, то эти алгоритмы не будут работать. Дальнейшие улучшения будут направлены на то, чтобы довести процент автоматически генерируемых description и title до 100%.