

Что нужно знать об уникальности текста

Глобализация массовых коммуникаций, предсказанная Мак-Люэном¹, стала реальностью, но его мрачные прогнозы по поводу будущего книги сбываются пока далеко не полностью. Все же, перефразируя один из самых известных афоризмов Мак-Люэна, можно сказать: «Гуттенберг дал всем возможность стать читателями, ксерокс дал всем возможность стать издателями, Интернет дал всем возможность стать писателями»[1].

В связи с этим хотим заметить, что вопрос об уникальности текстов, которые писатели предлагают читателям, не перестает быть актуальным. Информация об уникальности текстов достаточно широко представлена в сети Интернет, и «человек ищущий», приложив некоторые усилия и затратив небольшое количество времени, всегда сможет найти такие статьи.

Тем не менее, мы хотим сэкономить ваше время и предлагаем познакомиться с материалами специалистов, которые профессионально занимаются вопросами размещения информации в сети Интернет. Нами были взяты материалы Сайта для начинающих веб-мастеров (URL: <http://www.webadvisor.ru/unikalnost-teksta.html>) и сайта SEOpressa (URL: <http://seopressa.ru/stati/pochemu-tekst-dolzhen-byit-unikalnyim.html>). Первый материал нам показался интересным своими рассуждениями по общим вопросам уникальности авторских текстов, второй привлек внимание описанием работы конкретных программ и наличием полезных ссылок.

Заимствованные нами материалы печатаются в кавычках, снабжены ссылками, что означает цитирование авторского текста, размещенного в сети Интернет и разрешенного (не запрещенного) для копирования.

¹МАК-ЛЮЭН(McLuhan) Херберт Маршалл (1911 — 1980) — канадский философ и социолог, специалист по теории коммуникации и массовой культуре, автор оригинальной философии истории. Создатель культурно-исторической типологии, живописующей впечатляющую картину пожирания одного средства коммуникации другим.

«К слову «уникальный» можно подобрать массу синонимов – единственный, неповторимый, эксклюзивный, исключительный и др. Таким образом, становится понятным, уникальная статья – это авторское творение, не похожее на другие, единственное в своем роде. Наполнение сайта с помощью прямого копирования контента с других сайтов, вещь не только бесполезная, но и опасная. Копирование и публикация чужих статей, а также заимствование фрагментов чужих статей, без указания источника заимствования, является плагиатом. Плагиат (речевое воровство) подразумевает нарушение авторских прав и влечет за собой уголовную ответственность.

Поисковые системы определяют копированный контент и исключают его из поисковой выдачи. Это касается также легкого рерайта и копипаста. Рерайтинг, от английского *rewriting* – переписывание текста. Копипаст, от английского *cut, copy, paste* – вырезать, скопировать, вставить. При этом способе статья создается из фрагментов, скопированных из разных источников.

Что же делать, спросите вы, ведь Интернет заполнен всевозможной информацией, получается, что и писать то уже не о чем, обо всем уж написано. Здесь я однозначно могу сказать, сайты создают для того, чтобы делиться информацией. Если вы это можете сделать лучше других – очень хорошо. Совпадение отдельных идей не является плагиатом. Создание нового произведения невозможно без идеи, а идея авторов не имеет»[2].

«Не все наши заказчики знают, что такое уникальность текста, как ее проверять, какие при этом использовать параметры и какое значение она должна иметь. Это нас<сотрудников группы SEOpressa> и сподвигло написать развернутую статью.

1. Что такое уникальность?

Уникальность — это процентное соотношение материала, не имеющего совпадений с другими публикациями в Интернете, к заимствованному. К примеру, если это значение равно 60%, значит 40% статьи скопировано из других источников — это может быть как целый фрагмент (полное совпадение), так и отдельные слова или их сочетания (частичное совпадение). При уникальности 0 % можно говорить о полном плагиате, при 100% — о том, что в статье нет ни одного ранее используемого в Сети фрагмента. Существует даже общепринятая шкала оценки уникальности:

- очень высокая — 95-100%;
- высокая — 90-94%;
- средняя — 80-89%;
- низкая — 79% и ниже.

2. Зачем нужна высокая уникальность?

Значение, близкое к 100%, — это показатель того, что на сайте используется оригинальный контент, то есть не заимствованный с других ресурсов. «Яндекс» и Google это ценят, поэтому страницы с уникальными текстами имеют больше шансов получить хорошие позиции в поисковой выдаче, нежели «копипаст» — контент, скопированный с других источников. Разумеется, при равных условиях. Уникальный контент — это обязательное условие успешного продвижения любого сайта. Простое же копирование текстовой информации не только бесполезно, но и чревато санкциями поисковых систем. Перед тем как перейти к конкретным программам для проверки уникальности, давайте разберемся, как они это делают.

3. Алгоритмы проверки уникальности текстов

Они бывают шингловыми или корреляционными. Программа для выявления плагиата может быть основана на одном из этих алгоритмов или сразу обоих.

Шингловые алгоритмы

Здесь за основу берется выявление совпадений текстовых фрагментов. Принцип работы следующий:

1. Сначала из текста удаляются все стоп-слова: знаки препинания, союзы, предлоги, местоимения, причастия, междометия, частицы, вводные слова и другие элементы, которые не несут смысловой нагрузки.

2. Очищенный от стоп-слов текст разбивается на фрагменты заданной в настройках длины, называемые шинглами.

3. Составляются фразы для поисковых систем, указанных в настройках. В результате отработанных запросов алгоритм получает множество страниц, с которыми в итоге и будет сравниваться исследуемый текст. О том, как именно это делается, разработчики умалчивают.

4. На этих страницах алгоритм ищет вхождения шинглов. Детального описания, как они это делают, вы тоже нигде не найдете.

В результате мы получаем общее значение уникальности текста, а также неуникальные фразы и ссылки на страницы с ними. Видя, какие именно фрагменты нужно уникализировать, копирайтер может улучшить этот показатель.

Корреляционные (нешингловые) алгоритмы

В этом случае тексты проверяются на схожесть по смыслу. Подробной информации о корреляционных алгоритмах мы не смогли найти. Известно

лишь, что они строже и лучше обнаруживают рерайты, поскольку сравнивают статьи целиком, а не по фрагментам.

4. Программы проверки уникальности

Их довольно много, поэтому рассмотрим только самые известные системы.

AdvegoPlagiat

<http://advego.ru/plagiat/top/>

Одна из самых распространенных бесплатных антиплагиат-программ, основанная на шингловом алгоритме. Несмотря на статус бета-версии, она полностью работоспособна, имеет основные функции и настройки, которых в большинстве случаев достаточно для проверки текстов. Особенности AdvegoPlagiat:

- Два режима проверки — быстрый и глубокий.
- Поддержка двух декапчеров — Antigat и RuCaptcha.
- Отображение в результатах не только уникальности, но и количества рерайта — результат работы корреляционного алгоритма. Его значение выводится в результатах только при 30% совпадений и выше, в противном случае отображается лишь уникальность.
- Функция «Игнорировать домены» — удобно, если нужно доработать текст на уже проиндексированной странице.

Внимание! Разработчик не рекомендует использовать показатель количества рерайта для оценки уникальности текста. Вероятно, это знак того, что корреляционный алгоритм Advego пока находится на стадии тестирования. Рекомендуемые разработчиком настройки AdvegoPlagiat:

- размер шингла и фразы — 4;
- таймаут — 15 с;

- максимальный размер документа — 256 килобайт;
- прерывание проверки — 0%;
- учет совпадений — от 1%.

Недостаток программы — совместимость только с операционными системами семейства Windows.

«eTXTАнтиплагиат»

<http://www.etxt.ru/antiplagiat>

Эта программа отличается впечатляющим набором возможностей, в том числе не имеющих никакого отношения к уникальности. Основана «eTXTАнтиплагиат» на двух алгоритмах: «Метод обнаружения копий» (шингловый) и «Метод обнаружения рерайта» (корреляционный). Ключевые особенности:

- Три основных режима проверки: экспресс, глубокая, на рерайт.
- Поддержка антикапчеров — в настройках можно ввести адрес любого сервиса и ключ.
- Проверка текстов в пакетном режиме — для этого достаточно указать папку, где находятся файлы.
- Проверка сайта — в этом случае программа самостоятельно скачивает текст по указанному адресу.
- Онлайн-проверка — возможность проверки текстов без установки каких-либо программ и без риска блокировки IP поисковыми системами (бесплатно можно проверять только несколько текстов в день).
- Сравнение текстов — удобная функция, позволяющая сравнить две статьи между собой. Удобно для проверки нескольких рерайтов одного и того же источника.

- Проверка уникальности изображения — новая функция, позволяющая сравнить два графических файла.
- SEO-сервис — определение более 10 параметров сайта (ТИЦ, PR, дата регистрации домена и т. д.). Адреса ресурсов можно вводить списком.

Удобно, что в настройках имеется несколько предустановленных профилей и кнопка «По умолчанию», позволяющая вернуть все установки в исходные значения. Программа «eTXTАнтиплагиат» существует в нескольких вариантах: для Windows, Linux и MacOS.

FindCopy

<http://miratools.ru/>или <http://findcopy.ru>

По этим адресам доступна программа проверки уникальности FindCopy от компании «Мирафокс», которой, кстати, принадлежат известные сервисы Miralinks и Gogetlinks. Нам сложно сказать, на каком алгоритме она основана, поскольку разработчик об этом не сообщает, а мы слишком редко ей пользуемся, чтобы делать предположения.

Особенности сервиса:

- отсутствие каких-либо настроек и режимов проверки;
- не нужно устанавливать никаких программ;
- проверка занимает от 20 до 40 секунд;
- размер проверяемого текста — до 12 000 знаков;
- отсутствие ограничений для зарегистрированных пользователей;
- генерация публичных ссылок на результаты проверки.

С помощью FindCopy мы проверяем только тексты, которые предстоит размещать на площадках Miralinks.

Text.ru

<http://text.ru>

Онлайновый сервис, основанный на корреляционном алгоритме. Мнения интернет-общественности об этой системе отличаются — одни хвалят ее за способность выявлять рерайтинг, другие жалуются на то, что ей крайне сложно угодить. От лица нашей группы могу сказать, что нешингловый алгоритм text.ru часто мешает писать хорошие тексты. Выбирая эту систему как основное мерило уникальности, будьте готовы попрощаться с красивыми устойчивыми выражениями. В погоне за непохожестью вы можете получить безликие конструкции и странные сочетания слов. Особенности сервиса:

- отсутствие каких-либо настроек и режимов проверки;
- проверка выполняется в режиме онлайн, то есть прямо в браузере;
- возможность добавить проверку текста в очередь, не ожидая ее окончания;
- размер текста для незарегистрированных пользователей — 2000, после регистрации — 15 000 знаков;
- формирование ссылок на результаты проверок, доступ к ним;
- возможность установить на своем сайте счетчик уникальности — своего рода защита против воров контента.

Content-watch.ru

<http://www.content-watch.ru>

Как и text.ru, этот сервис основан на корреляционном алгоритме. По результатам он нам показался очень похожим на вышеупомянутый инструмент. Особенности:

- Без регистрации можно проверить до 5 текстов в день размером не более 3000 знаков. Зарегистрированным пользователям доступно 20 проверок по 20 000 знаков.
- Платные тарифы, с которыми можно выполнять от 100 до 500 проверок в день и активировать функцию регулярной проверки страниц вашего сайта на уникальность.
- Возможность проверять сайты, в том числе в пакетном режиме.

5. Настройка программ по проверке уникальности

Кроме размера шингла и поисковой фразы, существует еще ряд параметров, доступных в программах AdvengoPlagiatus и «eTXT Антиплагиат». Разберем самые важные из них:

- Поисковые системы — помимо стандартного набора («Яндекс» и Google) программа может также поддерживать Yahoo, Bing и другие сервисы. Подключать их имеет смысл для текстов на иностранных языках, в противном случае вы только потеряете время.
- Прокси — используя прокси-сервер, вы можете минимизировать риск бана вашего IP при проверке текстов. Чтобы активировать эту функцию, введите в соответствующих полях адрес сервера, порт, логин и пароль.
- Антигейт — функция, позволяющая не вводить капчу самостоятельно. За вас это будут делать специально обученные люди за вполне разумные деньги. Чтобы воспользоваться этой возможностью, зарегистрируйтесь на сайте <http://antigate.com> или <https://rucaptcha.com>, затем пополните счет, включите соответствующую службу в программе и введите свой персональный ключ.

- Таймаут — время, в течение которого программа будет ждать ответа от запрашиваемого сайта. Увеличивать это значение следует при частом превышении интервала ожидания, уменьшать особого смысла нет — проверка будет проходить быстрее, но некоторые сайты могут не успеть ответить. По умолчанию таймаут в AdvegoPlagiatus составляет 15 с, в «eTXTАнтиплагиат» — 30 с.

6. Какая антиплагиат-система самая лучшая?

Любой опытный копирайтер вам скажет, что такой системы не существует — каждая из них имеет свои недостатки. Для большей части текстов мы рекомендуем (и пользуемся сами) «eTXTАнтиплагиат» и AdvegoPlagiatus, которые давно развиваются и регулярно обновляются. Проверенные ими страницы хорошо ранжируются, поэтому у нас нет сомнений в их эффективности. К плюсам этих систем можно отнести отсутствие ограничений по размеру и количеству проверок, к минусам — некоторые технические трудности в виде блокировки вашего IP поисковыми системами, необходимости вводить капчу и подобным проблемам. Их решение — тема отдельной статьи. Если у вас нет желания этим заниматься, воспользуйтесь любым онлайн-сервисом. Я рекомендую «eTXTАнтиплагиатOn-line» в бесплатном варианте или за деньги, если проверять нужно часто и много»[3].

Список использованных источников:

1. Предпринимательская деятельность на современном книжном рынке России [Электронный ресурс] URL: <http://www.bestreferat.ru/referat-132711.html>
2. Сайт для начинающих веб-мастеров. [Электронный ресурс] URL: <http://www.webadvisor.ru/unikalnost-teksta.html>
3. SEOprensa [Электронный ресурс] URL: <http://seopressa.ru/stati/pochemu-tekst-dolzhen-byit-unikalnyim.html>