УЧЕБНЫЙ ЦЕНТР «ТОПЭКСПЕРТ»

Модуль генерации контента

9 поток профессионального курса по SEO

Работу выполнила: Жанна Тепсуева

Дипломный руководитель: Дмитрий Иванов

ТОПЭКСПЕРТ
УЧЕБНЫЙ ЦЕНТР

Модуль генерации контента

Описание

Модуль позволяет задавать для конкретного типа страницы, формулу генерации контента. Формула включает в себя переменные из базы данных и словарные генерации. Модуль позволяет настраивать формулы для любого количества типов страниц. В конечном результате такой модуль позволяет автоматизировать наполнение страниц сайта. Типизация страниц может происходить либо по содержанию (категория товаров, подкатегория товаров, карточка товаров, обзор товара), либо по тематическому наполнению.

На основе переменных из базы данных, модуль должен сгенерировать уникальные тексты для страниц сайта, которые не будут расценены поисковыми системами, как спамные или сгенерируемые. Такой контент должен быть рассчитан исходя из метрик, которые мы определим в результате ручной выборки сайтов с неспамными текстами и на основе этих текстов, выведем параметры (входные данные), которые будем использовать для генерации собственного контента.

Для обучения нашего алгоритма, из интернета мы выгрузим 500 спамных текстов и 500 неспамных. На основе этих текстов составим показатели метрик спамных текстов и неспамных: количество вхождения ключевика, распределение по количеству слов в предложении, показатель BM25 для каждой леммы в текущем тексте, количество спамных слов в тексте не используя стоп-слова, средняя длина слова, сжимаемость, дисперсия длин слов и предложений и т.д. Все эти показатели будут сведены к лемме, т.е. в алгоритм так же требуется еще закодить лемматизатор, который будет перечислять все формы данного слова и указывать их морфологические признаки: род, число, падеж, время и др.

Задача модуля

1. Уникализация описания карточек товаров;
2. Автоматическое создание заголовков;
3. Автоматическое создание description.

Входные данные

1. Переменные из базы данных (наименование товара, цена, характеристики), на основе этих данных создаются шаблоны
2.Параметры генерации:

- количество вхождений ключевого запроса (тошностность);
- расчет распределения по количеству слов в предложении;
- расчет количества частей речи в тексте;
- расчет BM25 для каждой леммы в текущем тексте;

Автор: Жанна Тепсуева, 9-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов

- расчет дисперсии длин предложений;
- расчет количества спамных слов в тексте не используя стоп-слова.

3.В качестве входных данных также используем следующие факторы:

- длину документа;
- служебные теги ссылки;
- среднюю длину слова;
- сжимаемость;
- дисперсию длин слов и предложений;
- ЦИПФ;
- Доля вводных слов (безусловно, всем известно, что, без сомнения, бесспорно, как говорится, многие/все (знают, любят, выбирают итд), каждый/ не каждый/ всякий (знает, любит, выбирает итд), к счастью, между прочим, кстати, прямо скажем, иными словами, несомненно, сложно и т.д.)

Выходные данные

На выходе у нас получаются готовые тексты для страниц, которые выводятся в таблицу с параметрами, если параметры по которым был сгенерирован тот или иной текст оптимальные, то текст размещается на странице сайта, если его параметры сильно отклоняются от заданных, то текст выводится с пометкой СПАМ и в дальнейшем проверяется и редактируется вручную.

Формула

[Текст] [x] [Текст] [y] [z] [a] [b] [Текст] + параметры для определения спамности
Бывает, что в формуле полностью отсутствует статичный текст. Переменные могут быть выражены как элементом БД (наименование модели, название марки, цена), так и браться из словаря, который заранее готовит копирайтер.
Для эффективности работы модуля, включаем в формулу определение спамности.

На основе выборки спамных, неспамных текстов и 25 параметров, была составлена формула определения спамности, которая может сейчас коррелировать на 45%. С помощью этой формулы можно определить спамность уже имеющегося текста.

Автор: Жанна Тепсуева, 9-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов

ТОПЭКСПЕРТ
УЧЕБНЫЙ ЦЕНТР

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	URL	Наша оценка	Формула спамности	Количество слов	Количество слов в тайтле	Среднее количество букв в слове после очистки	Процент слов после леммати зации, находящихся в топ200	Процент слов после леммати зации, находящихся в топ500	Коэффициент сжатия GZip	Коэффициент сжатия BZip2	Процент существительных среди русских слов	Процент прилагательных среди русских слов	Процент глаголов среди русских слов	Процент союзов среди русских слов	Процент предлогов среди русских слов	Процент частиц среди русских слов	Средняя длина предложений	Дисперсия длины предложений
2			0,45123	0,16226358	-0,33162	0,069124	-0,18295	-0,11064	0,076335	-0,01398	-0,02598	-0,0434	0,093519	-0,04349	0,070652	0,090609	0,119971	0,146603
3	http://ru	1	33,94	1047	3	5,72	6,59	9,93	0,56	0,52	32,57	17,57	8,31	2,96	10,32	9,55	9,38	9,93
4	http://ru	1	30,34	3157	4	5,83	3,33	6,21	0,51	0,45	38,8	15,49	9,78	2,31	10,99	8,24	11,53	7,07
5	http://ru	1	31,48	3133	3	5,91	3,7	6,61	0,48	0,42	33	13,79	10,47	2,49	8,52	9,22	9,22	9,38
6	http://ru	1	25,17	146	3	6,85	14,38	19,86	0,61	0,66	37,67	15,07	8,9	4,79	5,48	10,96	8	5,95
7	http://ru	1	24,78	152	3	6,68	18,42	22,37	0,67	0,68	34,21	20,39	9,87	1,97	10,53	9,21	8,88	6
8	http://ru	1	33,77	398	4	5,84	11,81	15,33	0,58	0,59	36,18	10,05	8,04	2,76	13,07	8,04	13,66	14,78
9	http://re	1	25,32	1086	9	5,97	7,37	11,33	0,55	0,52	33,24	19,52	7,73	3,04	8,66	8,84	12,07	8,69
10	http://re	1	23,25	1591	11	5,82	6,35	10,62	0,53	0,48	34,82	15,08	13,52	4,46	10,31	8,93	17,12	8,8
11	http://re	1	18,26	518	10	5,97	9,85	15,83	0,53	0,56	36,1	12,74	11,2	2,7	7,92	10,23	9,7	6,31
12	http://re	1	18,24	445	9	5,79	9,66	14,38	0,51	0,56	40,9	10,34	9,44	2,7	9,66	8,99	6,65	6,3
13	http://re	1	22,11	197	10	6,3	13,2	20,81	0,66	0,7	42,13	16,24	12,69	1,52	8,63	6,09	9,8	5,71
14	http://re	1	18,03	862	11	6,06	6,26	10,32	0,44	0,43	41,3	15,08	9,86	2,09	13,34	7,77	13,68	8,89
15	http://re	1	22,49	881	11	5,96	9,08	15,32	0,57	0,53	32,01	16,35	12,15	2,5	10,33	8,74	15,19	8,37
16	http://re	1	18,66	596	10	5,57	13,42	18,79	0,58	0,57	28,02	18,12	10,57	3,19	10,23	9,06	10,64	7,28
17	http://re	1	9,21	283	13	6,09	12,01	17,31	0,47	0,53	43,11	13,07	10,24	3,18	6,71	9,54	8,03	6,02
18	http://re	1	8,12	245	15	6,09	11,02	15,51	0,49	0,58	43,27	10,61	10,62	4,08	4,9	10,2	7,61	4,74
19	http://re	1	10,76	276	13	6,04	13,77	18,84	0,51	0,58	42,75	11,23	10,5	3,62	6,16	9,06	8,29	6,93
20	http://re	1	8,24	273	15	5,96	12,45	16,85	0,51	0,59	40,29	12,82	9,89	3,66	6,23	10,99	7,97	5,09
21	http://wv	1	24	1290	10	5,56	8,45	13,64	0,53	0,49	24,03	13,57	14,34	4,57	7,98	9,84	11,53	10,65
22	http://wv	1	26,82	2481	8	5,34	5,68	9,75	0,53	0,48	25,15	11,45	14,87	4,07	9,88	11,16	9,64	8,7
23	http://wv	1	25,87	3745	7	5,65	3,82	7,4	0,48	0,41	24,25	12,6	13,72	4,94	8,92	11,59	7,14	7,89

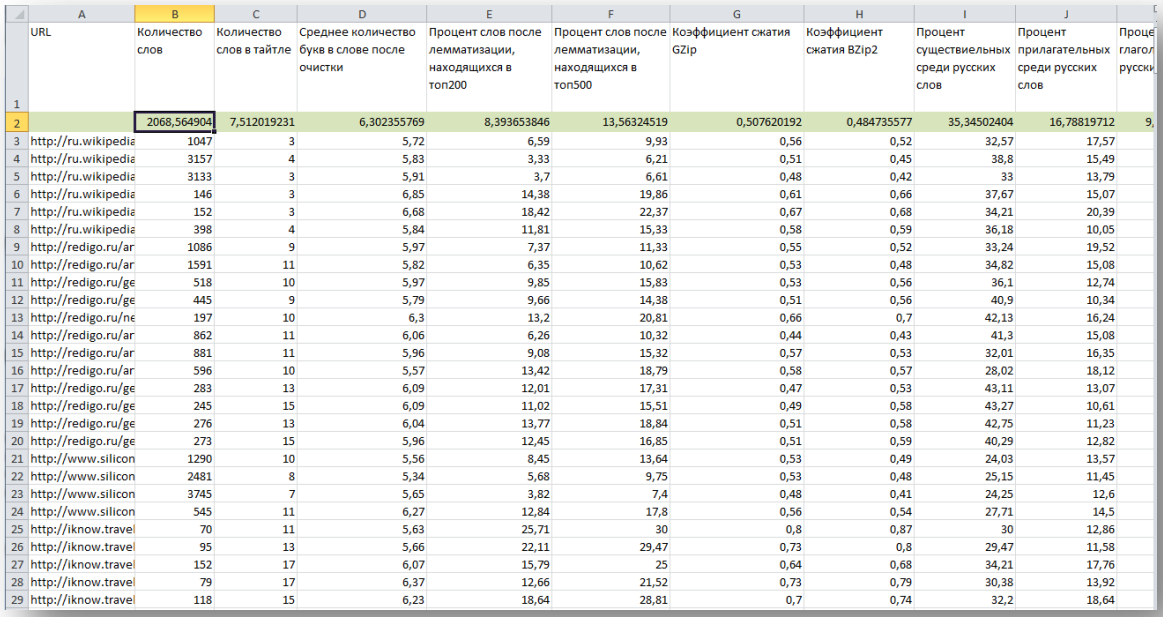
В качестве выборки были взяты по 5 документов с каждого сайта (неспам/СПАМ) для быстрого переобучения модуля

Для того чтобы модуль мог генерировать контент, нам требуется классификатор тематик. Составить классификатор можно на основе словаря русского языка, т.е. мы берем слова из словаря и смотрим, сколько каждого слова присутствует в каждом документе. И все что относится, к какому-то слову разделяем на кластеры, поэтапно. Например: кофеварка - бытовая техника - электроника.

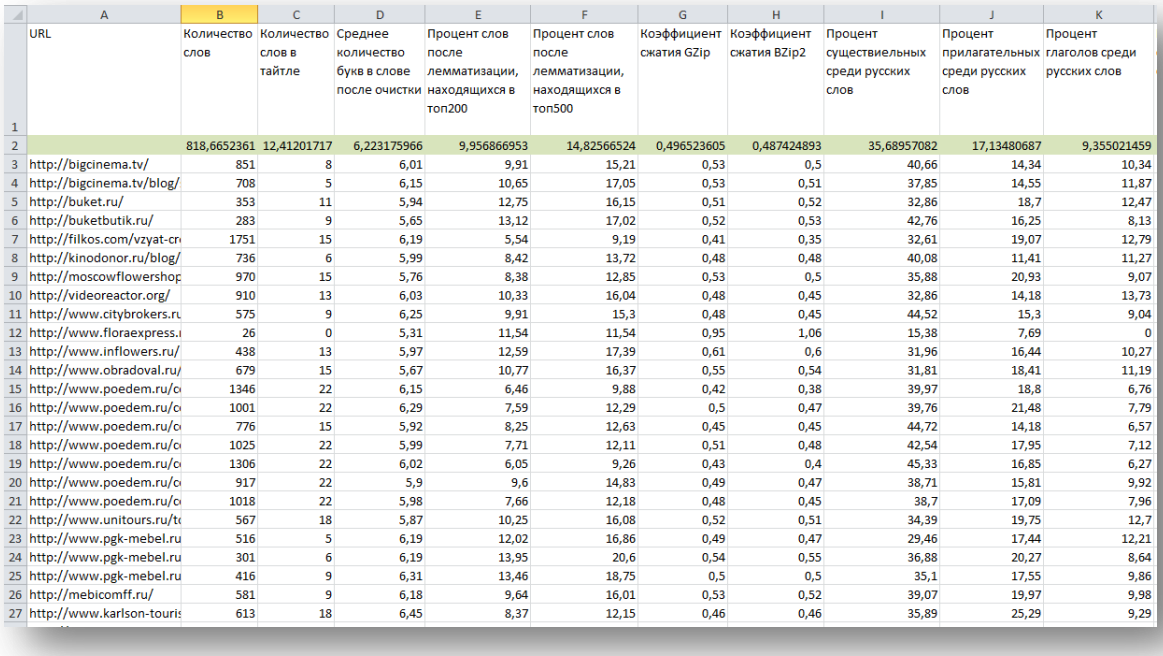
Среднее значение для неспамных и спамных текстов по каждому параметру. Данные средних значений заносятся в алгоритм модуля, для того чтобы он учитывал эти показатели при составлении контента.

Автор: Жанна Тепсуева, 9-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов

ТОПЭКСПЕРТ
УЧЕБНЫЙ ЦЕНТР



	A	B	C	D	E	F	G	H	I	J
1	URL	Количество слов	Количество слов в тайтле	Среднее количество букв в слове после очистки	Процент слов после лемматизации, находящихся в топ200	Процент слов после лемматизации, находящихся в топ500	Коэффициент сжатия GZip	Коэффициент сжатия BZip2	Процент существительных среди русских слов	Процент прилагательных среди русских слов
2		2068,564904	7,512019231	6,302355769	8,393653846	13,56324519	0,507620192	0,484735577	35,34502404	16,78819712
3	http://ru.wikipedia	1047	3	5,72	6,59	9,93	0,56	0,52	32,57	17,57
4	http://ru.wikipedia	3157	4	5,83	3,33	6,21	0,51	0,45	38,8	15,49
5	http://ru.wikipedia	3133	3	5,91	3,7	6,61	0,48	0,42	33	13,79
6	http://ru.wikipedia	146	3	6,85	14,38	19,86	0,61	0,66	37,67	15,07
7	http://ru.wikipedia	152	3	6,68	18,42	22,37	0,67	0,68	34,21	20,39
8	http://ru.wikipedia	398	4	5,84	11,81	15,33	0,58	0,59	36,18	10,05
9	http://redigo.ru/ar	1086	9	5,97	7,37	11,33	0,55	0,52	33,24	19,52
10	http://redigo.ru/ar	1591	11	5,82	6,35	10,62	0,53	0,48	34,82	15,08
11	http://redigo.ru/ge	518	10	5,97	9,85	15,83	0,53	0,56	36,1	12,74
12	http://redigo.ru/ge	445	9	5,79	9,66	14,38	0,51	0,56	40,9	10,34
13	http://redigo.ru/ne	197	10	6,3	13,2	20,81	0,66	0,7	42,13	16,24
14	http://redigo.ru/ar	862	11	6,06	6,26	10,32	0,44	0,43	41,3	15,08
15	http://redigo.ru/ar	881	11	5,96	9,08	15,32	0,57	0,53	32,01	16,35
16	http://redigo.ru/ar	596	10	5,57	13,42	18,79	0,58	0,57	28,02	18,12
17	http://redigo.ru/ge	283	13	6,09	12,01	17,31	0,47	0,53	43,11	13,07
18	http://redigo.ru/ge	245	15	6,09	11,02	15,51	0,49	0,58	43,27	10,61
19	http://redigo.ru/ge	276	13	6,04	13,77	18,84	0,51	0,58	42,75	11,23
20	http://redigo.ru/ge	273	15	5,96	12,45	16,85	0,51	0,59	40,29	12,82
21	http://www.silicon	1290	10	5,56	8,45	13,64	0,53	0,49	24,03	13,57
22	http://www.silicon	2481	8	5,34	5,68	9,75	0,53	0,48	25,15	11,45
23	http://www.silicon	3745	7	5,65	3,82	7,4	0,48	0,41	24,25	12,6
24	http://www.silicon	545	11	6,27	12,84	17,8	0,56	0,54	27,71	14,5
25	http://iknow.travel	70	11	5,63	25,71	30	0,8	0,87	30	12,86
26	http://iknow.travel	95	13	5,66	22,11	29,47	0,73	0,8	29,47	11,58
27	http://iknow.travel	152	17	6,07	15,79	25	0,64	0,68	34,21	17,76
28	http://iknow.travel	79	17	6,37	12,66	21,52	0,73	0,79	30,38	13,92
29	http://iknow.travel	118	15	6,23	18,64	28,81	0,7	0,74	32,2	18,64



	A	B	C	D	E	F	G	H	I	J	K
1	URL	Количество слов	Количество слов в тайтле	Среднее количество букв в слове после очистки	Процент слов после лемматизации, находящихся в топ200	Процент слов после лемматизации, находящихся в топ500	Коэффициент сжатия GZip	Коэффициент сжатия BZip2	Процент существительных среди русских слов	Процент прилагательных среди русских слов	Процент глаголов среди русских слов
2		818,6652361	12,41201717	6,223175966	9,956866953	14,82566524	0,496523605	0,487424893	35,68957082	17,13480687	9,355021459
3	http://bigcinema.tv/	851	8	6,01	9,91	15,21	0,53	0,5	40,66	14,34	10,34
4	http://bigcinema.tv/blog/	708	5	6,15	10,65	17,05	0,53	0,51	37,85	14,55	11,87
5	http://buket.ru/	353	11	5,94	12,75	16,15	0,51	0,52	32,86	18,7	12,47
6	http://buketbutik.ru/	283	9	5,65	13,12	17,02	0,52	0,53	42,76	16,25	8,13
7	http://filkos.com/vzyat-cr	1751	15	6,19	5,54	9,19	0,41	0,35	32,61	19,07	12,79
8	http://kinodonor.ru/blog/	736	6	5,99	8,42	13,72	0,48	0,48	40,08	11,41	11,27
9	http://moscowflowershop	970	15	5,76	8,38	12,85	0,53	0,5	35,88	20,93	9,07
10	http://videoreactor.org/	910	13	6,03	10,33	16,04	0,48	0,45	32,86	14,18	13,73
11	http://www.citybrokers.ru	575	9	6,25	9,91	15,3	0,48	0,45	44,52	15,3	9,04
12	http://www.floraexpress.r	26	0	5,31	11,54	11,54	0,95	1,06	15,38	7,69	0
13	http://www.inflowers.ru/	438	13	5,97	12,59	17,39	0,61	0,6	31,96	16,44	10,27
14	http://www.obradoval.ru/	679	15	5,67	10,77	16,37	0,55	0,54	31,81	18,41	11,19
15	http://www.poedem.ru/c	1346	22	6,15	6,46	9,88	0,42	0,38	39,97	18,8	6,76
16	http://www.poedem.ru/c	1001	22	6,29	7,59	12,29	0,5	0,47	39,76	21,48	7,79
17	http://www.poedem.ru/c	776	15	5,92	8,25	12,63	0,45	0,45	44,72	14,18	6,57
18	http://www.poedem.ru/c	1025	22	5,99	7,71	12,11	0,51	0,48	42,54	17,95	7,12
19	http://www.poedem.ru/c	1306	22	6,02	6,05	9,26	0,43	0,4	45,33	16,85	6,27
20	http://www.poedem.ru/c	917	22	5,9	9,6	14,83	0,49	0,47	38,71	15,81	9,92
21	http://www.poedem.ru/c	1018	22	5,98	7,66	12,18	0,48	0,45	38,7	17,09	7,96
22	http://www.unitours.ru/t	567	18	5,87	10,25	16,08	0,52	0,51	34,39	19,75	12,7
23	http://www.pgk-mebel.ru	516	5	6,19	12,02	16,86	0,49	0,47	29,46	17,44	12,21
24	http://www.pgk-mebel.ru	301	6	6,19	13,95	20,6	0,54	0,55	36,88	20,27	8,64
25	http://www.pgk-mebel.ru	416	9	6,31	13,46	18,75	0,5	0,5	35,1	17,55	9,86
26	http://mebicomff.ru/	581	9	6,18	9,64	16,01	0,53	0,52	39,07	19,97	9,98
27	http://www.karlson-touris	613	18	6,45	8,37	12,15	0,46	0,46	35,89	25,29	9,29

Метрики	Ср. значение
Количество слов	2000
Количество слов в тайтле	8
Среднее количество букв в слове после очистки	6
Процент слов после лемматизации, находящихся в топ200	8
Процент слов после лемматизации, находящихся в топ500	13
Процент существиельных среди русских слов	35

Автор: Жанна Тепсуева, 9-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов

Процент прилагательных среди русских слов	17
Процент глаголов среди русских слов	10
Процент союзов среди русских слов	2
Процент предлогов среди русских слов	10
Процент частиц среди русских слов	7
Средняя длина предложений	12
Дисперсия длины предложений	8
Максимальное количество слов в предложении	45
Доля предложений с несколькими глаголами	33
Среднее количество существительных в предложении	4

Взаимодействие с другими модулями

Модуль проверки уникальности контента

Модуль проверки орфографии

Модуль выгрузки контента

Модуль защиты от копирования

Логическая схема работы модуля

Особые требования

Возможность ручного редактирования для каждой страницы

Описание процессов взаимодействия

После того как текст будет готов для размещения на сайте, его предварительно пропускаем через выше описанные модули, т.е. проверяем орфографию, уникальность текста и т.д.

Автор: Жанна Тепсуева, 9-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов

Предполагаемая нагрузка

Нагрузка предполагается небольшая

Процесс остановки модуля

Модуль должен останавливаться автоматически, либо вручную по требованию или при возникновении ошибки. Ошибки: если модуль генерирует все тексты как СПАМ; если модуль пропускает спамные тексты в этом случае требуется остановка модуля.

Процесс запуска модуля

Запуск модуля также происходит либо в автоматическом режиме, либо в ручном.

Автор: Жанна Тепсуева, 9-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов