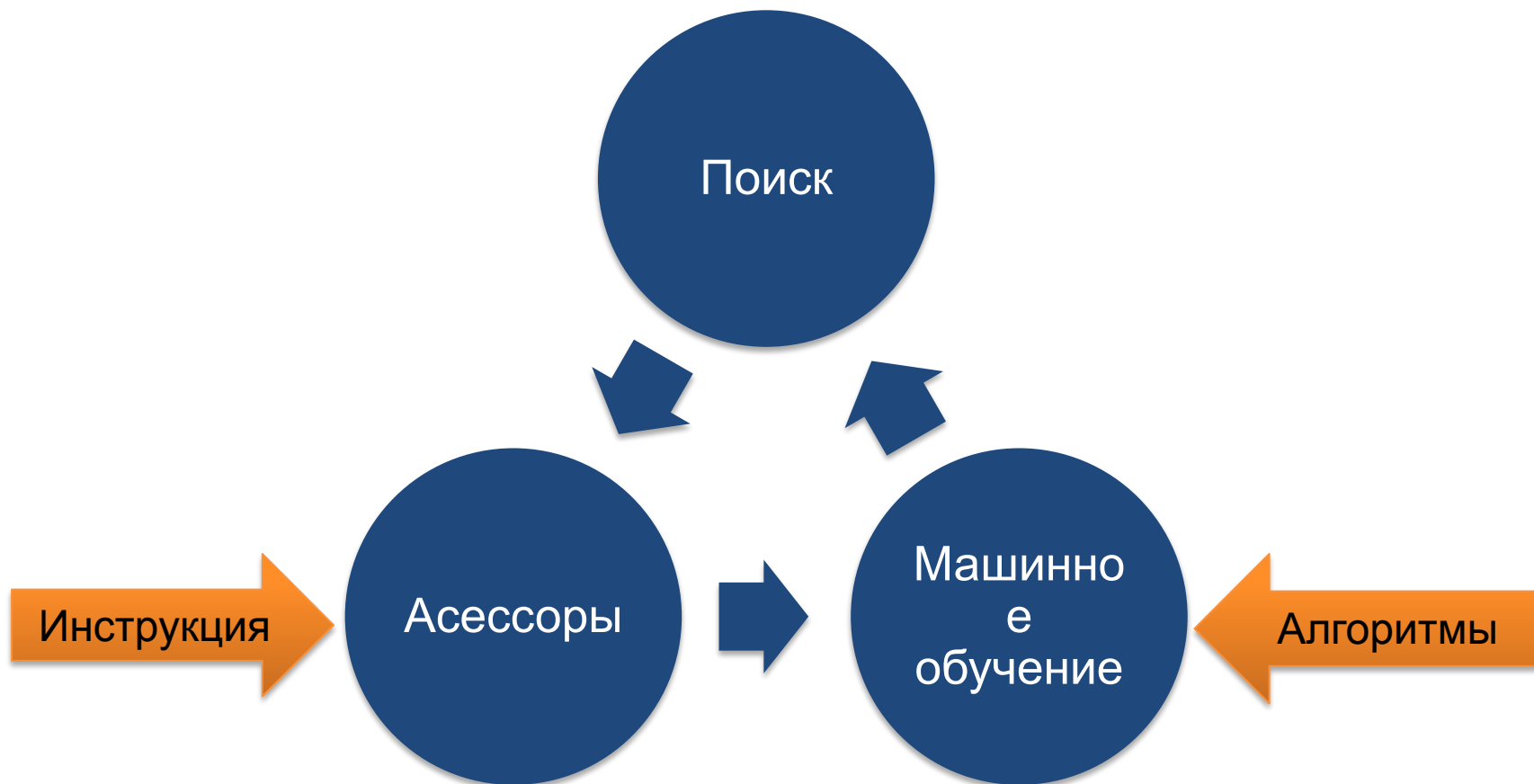


Ранжирование Спам и Антиспам

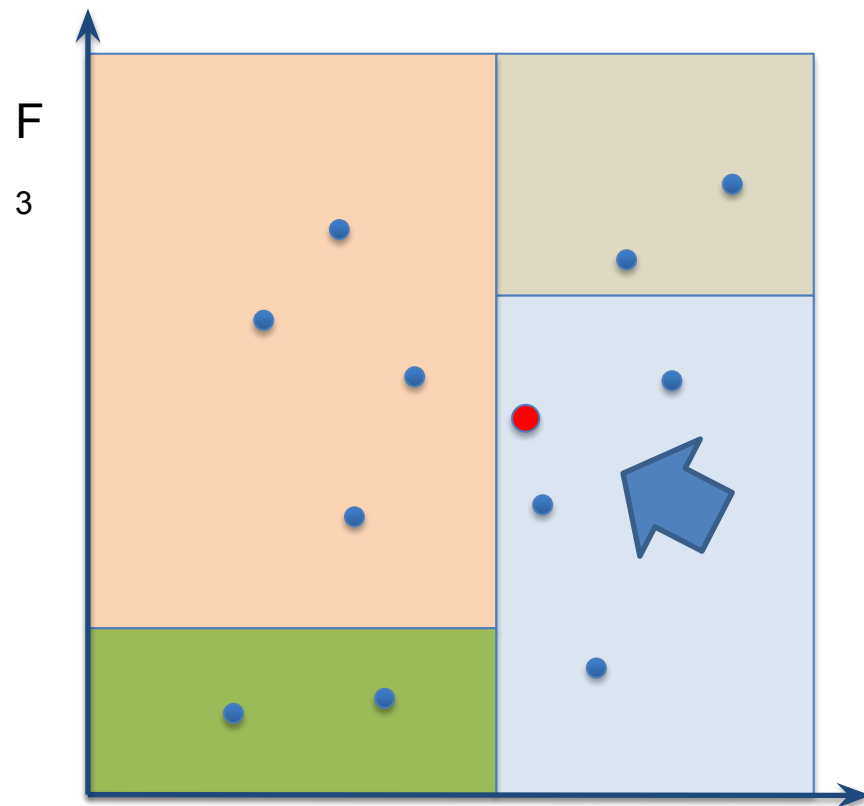
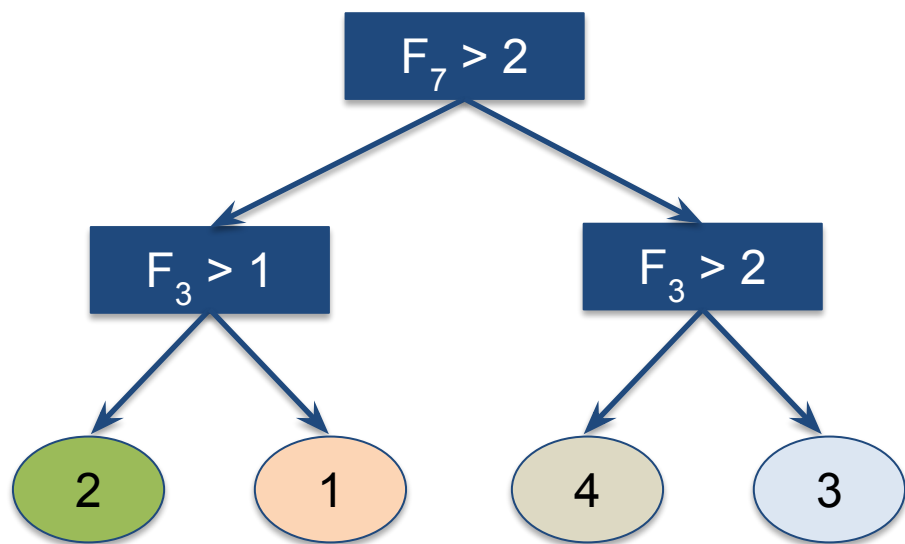
Что и для чего проходим?

- Апдейты ПС
- Текстовый спам
- Спам в коде страниц
- Дорвеи и редиректы
- Клоакинг и свопинг
- Заимствование контента
- Афффилиаты
- Ссылочный спам
- Поведенческий спам
- Спам в рекламе
- Парсинг

Машинное обучение



Деревья вариантов





Ассесоры

Система оценки качества Поиск@mail.ru

so Система оценки качества Поис... +

SQ @mail.ru бета

Вы вошли как **Андрей Калинин** ([сообщения](#))
[Администрирование](#) ▾ | [Список пакетов](#) ▾ | [Руководство асессора](#)

Пакет: [techforum](#)

Посмотреть выдачи: [mail.ru](#) | [yandex](#) | [google](#)
Искать в: [википедия](#) | [wordstat.yandex.ru](#) | [www.dic.academic.ru](#)

форум технологий mail.ru

Оценка запроса

Запрос из региона: Московская область, язык запроса: Русский

Класс запроса: **Навигационный** | [Информация](#) | [Цитата](#) | [Купить](#) | [Скачать](#) | [Картинки](#) | [Видео](#) | [Музыка](#) | [Игры](#) | [Книги](#) | [Люди](#)

Бессмысленный: ☐

Опечатка: ☐

Геоависимый: ☐ Да ☒ Нет ☐ Возможно

Порно-запрос: ☐ Да ☒ Нет ☐ Возможно

[Оценить](#)

© Mail.Ru
You are **Андрей Калинин**: [администрирование](#) | [список пакетов](#)

Система оценки качества Поиск@mail.ru

forum технологий mail.ru

[Скрыть/показать форму](#) Автоскрытие **отключено** [включить](#)

<http://techforum.mail.ru/>

Порно ☐ Да ☒ Нет ☐ Возможно
Спам ☐ Да ☒ Нет ☐ Возможно
Вредоносный ☐

☒ Обязательная страница
☐ Точный ответ
☐ Полезная страница
☐ Малополезная страница, но по теме
☐ Страница не по теме
☐ Нельзя оценить

Оценить страницу

Обязательно	4
Точный ответ	3
Полезно	2
Малополезно	1
Не по теме	0
Нельзя оценить	

Предыдущая [Следующая](#)

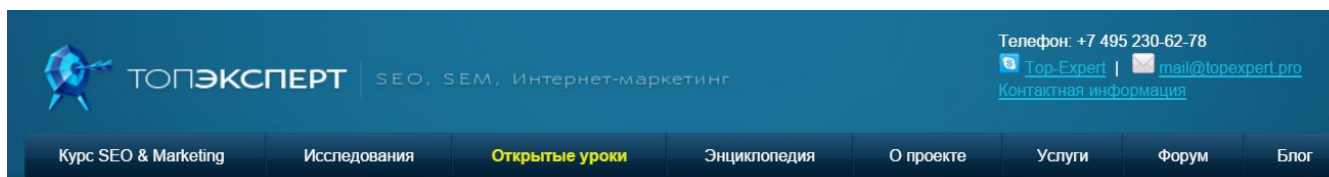
 > Конференция «ФОРУМ ТЕХНОЛОГИЙ» 2011 > [Хочу сюда!](#)

[О ФОРУМЕ](#) [РАСПИСАНИЕ](#) [ДОКЛАДЧИКИ](#) [ЗАДАТЬ ВОПРОС ДОКЛАДЧИКУ](#)



Классификация текстов и запросов

Тематическая классификация



Поток 18
19.02.16—28.08.16 ^{Уже идёт!}
[Прими участие сейчас!](#)

Поток 19
16.09.16—23.04.17
[Запишись заранее!](#)

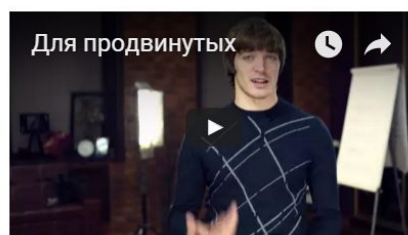
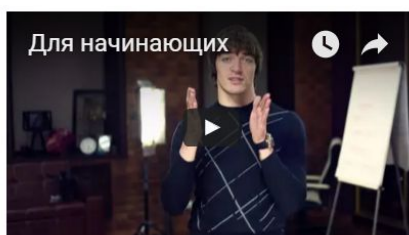
Профессиональный курс SEO&SEM от 3 до 6 мес.

Курс профессиональной подготовки специалистов по интернет-маркетингу, объединяющий знания и опыт от ведущих экспертов.

Курс проходит в формате онлайн  трансляций, после занятий слушателям предоставляются записи.

Обучение модульное. Вы можете выбрать одну из 5-ти комбинаций или пройти отдельные модули.

Для кого этот курс:



Продолжительность обучения
от 72 до 180 ч.

Старт ближайшей группы
19 февраля 2016 г.

До конца регистрации

25	20	30	50
дня	часов	минут	секунд

Расписание занятий

Поток 18
19.02.16—28.08.16

Вводное занятие

19.02.2016 19:00 — 22:00

Сайты

22.02.2016	19:00 — 22:00
24.02.2016	19:00 — 22:00
26.02.2016	19:00 — 22:00
29.02.2016	19:00 — 22:00
02.03.2016	19:00 — 22:00

Тематическая классификация

→ Плотность ключевых слов:

Ключ:	Вхожд/%	В Title	В mKeys	В mDesc	В <H>	В 	В
анализ	63(1.68%)						
курс	61(1.62%)	Да			Да		Да
сайт	50(1.33%)				Да		
работа	38(1.01%)				Да		
проверка	36(0.96%)						
ссылка	30(0.8%)						
поиск	24(0.64%)					Да	Да
страница	23(0.61%)						
поисковый	23(0.61%)					Да	Да
система	23(0.61%)						
Яндекс	20(0.53%)						Да
директор	17(0.45%)						
аналитик	17(0.45%)						
преподаватель	17(0.45%)				Да		
модуль	17(0.45%)						

Тематическая классификация

Тестирование классификации текста тематическими тегами

Ваш текст

Профессиональный курс SEO&SEM от 3 до 6 мес.

Курс профессиональной подготовки специалистов по интернет-маркетингу, объединяющий знания и опыт от ведущих экспертов.

Курс проходит в формате онлайн трансляций, после занятий слушателям предоставляются записи.

Обучение модульное. Вы можете выбрать одну из 5-ти комбинаций или пройти отдельные модули.

Для кого этот курс:

Выберите профессию:

Интернет-маркетолог SEO специалист Трафик-менеджер

SEO аналитик Директор по маркетингу

Вы можете выбрать одну из 5-ти специальностей. Пройдя курс, вы станете участником профессионального сообщества экспертов по интернет-маркетингу и сможете обмениваться опытом с другими участниками. Обучение по нашей системе уже прошли более 1000 интернет-маркетологов. Многие из них открыли свои агентства, стали

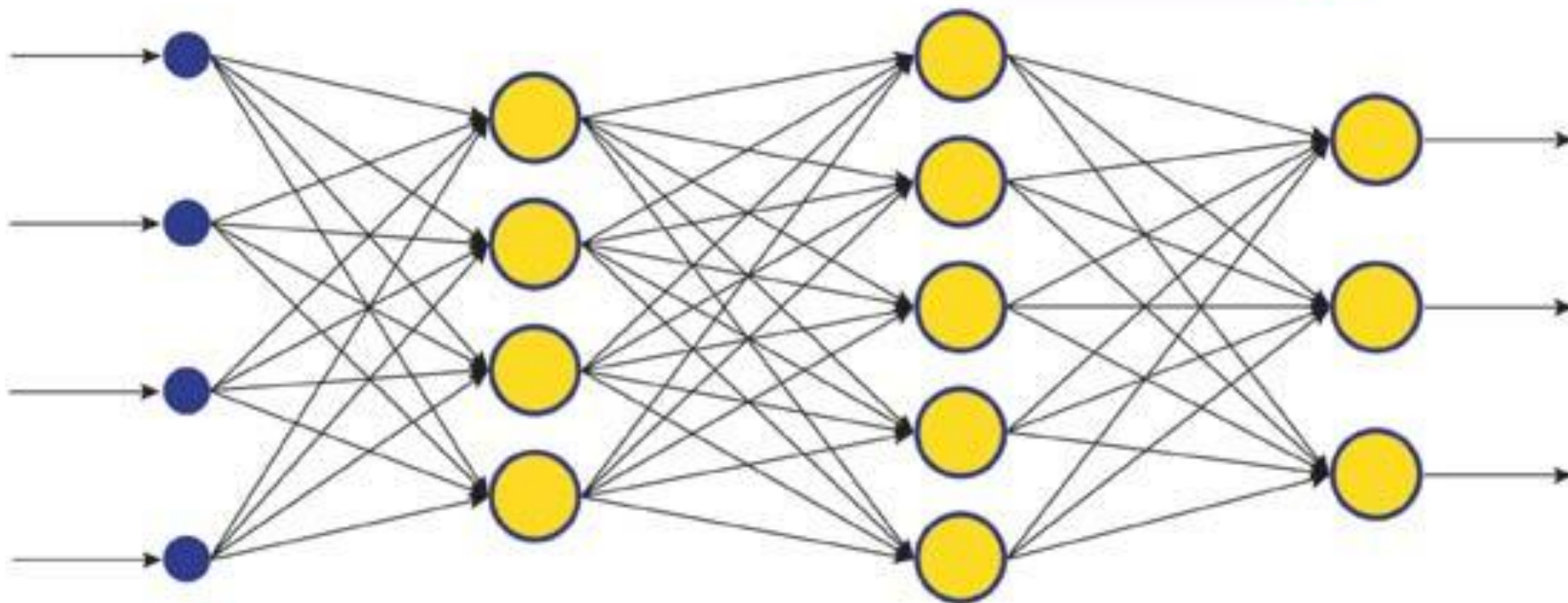
Тематические теги

бизнес интернет информация маркетинг наука образование обучение поиск проекты работа сайты финансы

<http://tag4site.ru/classify>

Тематическая классификация

Направление распространения данных



Направление распространения ошибки



ТОПЭКСПЕРТ

УЧЕБНЫЙ ЦЕНТР

BM25

Окарі ВМ25 — функція ранжирования

Пусть дан запрос Q , содержащий слова $q_1, \dots, q_n$. тогда функция ВМ25 даёт следующую оценку релевантности документа D запросу Q :

$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},$$

где $f(q_i, D)$ есть частота слова ([англ. term frequency, TF](#)) q_i в документе D , $|D|$ есть длина документа (количество слов в нём), а avgdl — средняя длина документа в коллекции. k_1 и b — свободные коэффициенты, обычно их выбирают как $k_1 = 2.0$ и $b = 0.75$.

$\text{IDF}(q_i)$ есть обратная документная частота ([англ. inverse document frequency, IDF](#)) слова q_i . Есть несколько толкований IDF и небольших вариации его формулы. Классически, она определяется как:

$$\log \frac{N}{n(q_i)},$$

где N есть общее количество документов в коллекции, а $n(q_i)$ — количество документов, содержащих q_i . Но чаще применяются «сглаженные» варианты этой формулы, например:

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

Окарі ВМ25 — функція ранжирования

$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},$$

оценка релевантности документа D запросу Q

Слово	Мотозэкипировка	Москва	Мотоцикл	Kawasaki	цена	шлем
Общее количество документов в коллекции	3581000000	3581000000	3581000000	3581000000	3581000000	3581000000
Количество документов содержащих слово вхождение	907000	269000000	12000000	8000000	650000000	9000000
Длина документа (слов), D	62	70	168	54	6	23
Частота слова в документе в документе D, f(q _i , D)	5153	5153	5153	5153	5153	5153
Средняя длина документа в коллекции, avgdl	0,012031826	0,01358432	0,03260237	0,01047933	0,00116437	0,00446342
k1 (стандартный коэф)	5151	5151	5151	5151	5151	5151
b (стандартный коэф)	2	2	2	2	2	2
формула	0,75	0,75	0,75	0,75	0,75	0,75
IDF, обратная документная частота	1,333160856	0,50009705	0,50009705	0,50009705	0,50009705	0,50009705
score (оценка релевантности)	3,596286782	1,090338047	2,47336528	2,649943	0,65410246	2,59866892
	4,794428766	0,545274841	1,23692268	1,32522868	0,32711471	1,29958666

- BM25F
- LinkBM25

- Посчитайте VM25 для анкор-листа



ТОПЭКСПЕРТ

УЧЕБНЫЙ ЦЕНТР

Апдейты ПС



Что и для чего проходим?

Проходит 107 дней до
индексации ссылки
(по данным
tools.promosite.ru)

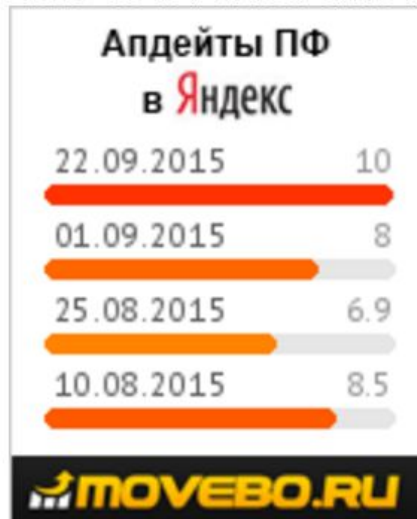
Может произойти
изменение
позиций без
изменений
индекса, т.е. вы
обновили тексты,
произошел апдейт,
но тексты остались
старыми

Что и для чего проходим?

23 сентября 2015 **Изменение выдачи** без выкладывания текстового индекса

7:43 Степень изменения выдачи: 85.2%

2015-09-24 17:00 отмечен Апдейт поведенческих факторов по Movebo



Виды апдейтов

18 сентября 2015 [2:40 RU+EN]

2:40 Текстовый апдейт: выложен индекс по 5 сентября 2015

1:45 Ссылочный апдейт: учтены ссылки, попавшие в индекс по 4 сентября 2015

7:49 Степень изменения выдачи: 87.2%

tools.promosite.ru

Как определять апдейты

пластиковые окна — 41 млн ответов

Найти

«Немецкие окна» — пластиковые окна Veka

[nem-okna.ru](#) ▼

Производство и монтаж **окон**, установка **пластиковых** откосов, остекление балконов. Онлайн-форма вызова замерщика. Ответы на вопросы. Фотогалерея выполненных работ.

+7 (495) 760-67-44

пн 10:00–19:00; вт 10:00–19:00; ср 10:00–19:00; чт 10:00–19:00; пт 10:00–17:00; сб 11:00...

м. Текстильщики

📍 Москва, ул. Юных Ленинцев, 25, оф. 9

Пластиковые окна - ...и размеры с установкой в Москве...

[MosOkna.ru](#) > [ceny-plastikovye-okna/](#) ▼

Пластиковые окна по низким ценам — не всегда выгодный вариант, поскольку в таком случае может страдать качество товара.

«Главокна» — пластиковые окна

[glavokna.ru](#) ▼

Изготовление и установка **окон**, остекление балконов и лоджий. Цены на продукцию. Информация о скидках. Технические описания профилей. Советы по эксплуатации. Контакты.

+7 (495) 665-35-65 · ежедневно, 9:00–21:00 · м. Савёловская

📍 Москва, ул. Бутырская, 46, стр. 1, оф. 612

пластиковые окна — 41 млн ответов

Найти

«Главокна» — пластиковые окна

[glavokna.ru](#) ▼

Изготовление и установка **окон**, остекление балконов и лоджий. Цены на продукцию. Информация о скидках. Технические описания профилей. Советы по эксплуатации. Контакты.

+7 (495) 665-35-65 · ежедневно, 9:00–21:00 · м. Савёловская

📍 Москва, ул. Бутырская, 46, стр. 1, оф. 612

...в Москве, цены на изготовление пластиковых окон ПВХ...

[OknaRosta.ru](#) > [plastikovye-okna/](#) ▼

Вы можете убедиться в низких ценах на продукцию, сравнив наше предложение с предложением других компаний, предлагающих вам **пластиковые окна** класса «А».

«Экоокна» — пластиковые окна и двери

[ecoookna.ru](#) ▼

Производство **пластиковых окон** и дверей, остекление балконов и лоджий, изготовление зимних садов и др. Онлайн-расчет стоимости **окна**. Информация о скидках. Адреса офисов продаж.

📍 Адреса на карте

По количеству изменений считается % шторма



Текстовый спам

Спам в тексте страниц

— избыток ключевых слов в тексте страницы, некачественные, бесполезные тексты, размещенные для манипулирования позициями.

Интернет-магазин холодильников в Екатеринбурге

Не один предмет бытовой техники не сравнится с холодильником. Он сразу решает целую гору ежедневных проблем, связанных с хранением продуктов. К приобретению холодильника необходимо основательно подготовиться, т.к. это предмет весьма продолжительного использования и стоит прилично. Теперь у жителей **Екатеринбурга** есть возможность приобрести **холодильник в интернет-магазине** по привлекательной цене

Представленный в магазинах ассортимент поражает. Модельные линейки популярных брендов отличаются не только внешними особенностями, огромное различие представляет собой функционал. Современные обычные и **встраиваемые холодильники** выполняют не только своё основное предназначение: сохранять пищу свежей. Они также могут морозить лёд для напитков, принимать и транслировать сигналы радиостанций, служить док-станцией для вашего телефона, транслировать ТВ-программы и многое другое. Выбор также осложняется различиями в размерной линейке.

Мы предлагаем вам **купить** абсолютно новый **холодильник** с незначительными внешними повреждениями, никак не влияющими на его работу. В нашем ассортименте представлены товары ведущих российских и зарубежных мировых производителей. **Холодильники** с дефектами во внешнем виде обычно убирают с витрин магазинов бытовой техники, но **интернет-магазин «Сток Маркет»** считает, это не причина для отказа от продажи техники, а прекрасный способ для покупателей значительно сэкономить.

В интернет-магазине «Сток Маркет» у вас есть возможность **купить** уцененный, но по-настоящему качественный **холодильник**, произведённый такими фирмами как: Ariston, Bosch, Electrolux, BEKO, Indesit, SHARP, Siemens и другими.

Для того чтобы **купить холодильник в интернет-магазине Екатеринбурга** достаточно позвонить нам по телефону +7 922 16 06 696 или оформить заказ через сайт.

Спам в тексте ссылок

— большое количество ссылок с одинаковыми анкерами, с коммерческими анкерами.

Облако анкоров



Анкор	Процент
купить медицинскую справку о болезни	33%
купить справку	33%

Спам в мета-тегах

- слишком длинные мета-теги,
- редкие n-граммы (купить детская каша, низкая цена),
- использование спецсимволов,
- сгенерированные по одной простой формуле,
- спам в содержимом микроразметки

Синонимайзинг, рерайт

— создание новых текстов по образцу старых без добавления к ним новой смысловой ценности путем изменения формы (переформулирования, замены слов на синонимы).

Распродажа скатертей от знаменитых брендов

Получить эксклюзивную вещь от прославленного европейского производителя можно, воспользовавшись таким выгодным предложением от интернет-магазина «Постель Бутик», как распродажа скатертей. В данном подразделе сайта предлагаются премиальные изделия, которые способны превратить обычный прием пищи в королевское застолье. Проводимая ресурсом распродажа скатертей включает только элитные товары с роскошным дизайном, из лучших, натуральных тканей. И все это можно приобрести теперь со скидками.

Распродажа скатертей ручной работы

Будучи невероятно выгодной для потребителей, распродажа скатертей включает множество моделей, подбирать которые можно в соответствии с особенностями помещения. Дополнительные преимущества, которые предлагает клиентам распродажа скатертей, можно выделить такие:

- Предлагается продукция ведущих торговых марок, с которыми сотрудничают прославленные, талантливые дизайнеры. Поэтому предметы интерьера создаются не только с учетом всех требований качества, но также в соответствии с новейшими тенденциями моды.
- Включает наша распродажа скатертей только дорогие вещи, скидка на которые позволяет приобрести за небольшие деньги действительно шикарную вещь.
- Особенности декора изделий, предлагаемых на нашем сайте со скидками, являются наличие ручной вышивки на пальцах. Чтобы создать всего одну часть узора (цветок, веточку, стрекозу) – мастерицам приходится тратить до девяти часов рабочего времени.
- Также распродажа скатертей включает изделия исключительно из природных, натуральных материалов. Весьма элегантно смотрится помещение, стол в котором украшен льняной скатертью. Материал является практичным, выглядит стильно, к тому же лен – полезен здоровью.

Всем, мечтавшим заполучить продукцию, разработанную маститыми дизайнерами и выполненную профессиональными мастерами – воплотить мечту поможет наша распродажа скатертей.

Выгодная распродажа скатертей в интернете

Осуществляемая на сайте «Постель Бутик» распродажа скатертей дает возможность желающим приобрести модную, уютную и шикарную деталь интерьера с приятной скидкой. Все изделия являются сертифицированными, поставляются напрямую от известных брендов. Присмотрев скатерть, также всегда можно к ней подобрать салфетки, столовые приборы и прочие, необходимые в любом доме, предметы.

Роскошные и уютные халаты со скидкой

Для клиентов, желающих стать обладателями роскошнейших брендовых изделий за доступную цену, сайтом «Постель Бутик» проводится распродажа халатов. Нежнейшие хлопковые модели от знаменитых мировых торговых марок, можно получить по действительно невысоким расценкам. Предлагаются халаты со скидкой для женщин классических светлых оттенков – все изделия отличаются изяществом, способностью подарить ощущение настоящей роскоши, изысканной отделкой и безукоризненной качественностью.

Выгодная распродажа халатов премиум-качества

Такое изделие как халат позволяет не только наслаждаться комфортом и ощущением уюта, находясь дома, но также придает облику особую элегантность. Реализуемые интернет-магазином халаты со скидкой предлагаются с такими характеристиками, как:

- Натуральный состав (100% хлопок). Предлагается распродажа халатов английского и турецкого производства из хлопковых волокон высочайшего качества.
- Многообразие моделей. Брендовые халаты со скидкой можно выбрать длинные или средней длины, с запахом или на пуговицах, дополненные поясом, капюшоном или необычными элементами отделки. Все товары предлагаются в жизнерадостных светлых оттенках – они могут быть однотонные или с нежным мелким рисунком.
- Сертифицированный товар ведущих ТМ. Осуществляемая распродажа халатов включает модели, пошитые в соответствии со всеми необходимыми нормами, а также мировыми модными тенденциями.
- Изящный, модный дизайн. Помимо разнообразных привлекательных моделей, халаты со скидкой могут быть дополнены такими изысканными деталями, как вышивка ручной работы или нежнейшее кружево.

Предполагающая реализацию множества моделей знаменитых производителей распродажа халатов, дает также возможность выбрать изделие подходящего размера. Предлагается таблица размеров в широком их диапазоне – подобрать можно халат, подходящий под любые индивидуальные особенности.

Элитные халаты со скидкой в интернет-магазине

Предлагаемая интернет-магазином «Постель Бутик» распродажа халатов дает возможность клиентам приобретать действительно роскошные модели из тончайшего хлопка по невысоким расценкам. Изысканный дизайн, который имеют халаты со скидкой, способен порадовать дам, обладающих тонким вкусом и чувством стиля, желающих даже дома всегда оставаться элегантными и привлекательными.

Автогенерация текстов

— автоматическое создание текстов по заранее заданным параметрам/формулам.

Панель инструментов

Сообщения (9)

▼ Вид в поиске ⓘ

- Структурированные данные
- Маркер
- Оптимизация HTML**
- Ссылки сайта

Оптимизация HTML

↩

Повторяющиеся заголовки (теги title)

Заголовок предоставляет пользователям и поисковым механизмам полезную информацию о вашем сайте. Текст, заключенный в теги заголовка, может отображаться на страницах результатов поиска, обеспечивая большую вероятность нажатия на релевантный описательный текст. Советуем просмотреть список и по возможности обновить теги заголовка.

Загрузить эту таблицу

Показать

Строк: 25 ▼

1-25 из 305

<

>



Спам в коде страниц

Манипулирование HTML-тегами

- Description
- Keywords
- H1, H2, H3, H4, H5, H6
- STRONG, B
- EM, I
- <A href=

Мы предлагаем вам **купить** абсолютно новый **холодильник** с незначительными внешними повреждениями, никак не влияющими на его работу. В нашем ассортименте представлены товары ведущих российских и зарубежных мировых производителей. **Холодильники** с дефектами во внешнем виде обычно убирают с витрин магазинов бытовой техники, но **интернет-магазин «Сток Маркет»** считает, это не причина для отказа от продажи техники, а прекрасный способ для покупателей значительно сэкономить.

Скрытие текста от посетителей

Белый текст на белом фоне

1 пиксельный текст

Скрытие через CSS (`display:none`)

Создание зон «текст в подвале»

ОКНА ОКНА

Белый текст на белом фоне

1 пиксельный текст

ОКНА

Скрытие через CSS (display:none)

Создание зон «текст в подвале»

ОКНА

ОКНА

Рерайт текстов

Разные методы «уникализации» текста

Генерация SEO-текстов

Множественное употребление слов в тексте

Задача: придумайте алгоритм определения спама в почте, на основе ТОЛЬКО ТЕКСТОВЫХ факторов



Дорвеи и редиректы

Дорвеи и редиреты

— сайты, которые создаются для манипулирования поисковыми системами, с целью получения трафика и его перепродажи, за счет создания низкокачественного контента и перенаправления пользователя, часто, путем обмана (автогенерация доменов, страниц, автоматическое заимствование контента с долгоиндексируемого сайта).



Клоакинг и СВОПИНГ

— создание страницы, контент на которой
меняется сразу после выхода в ТОП,
использовалось для продвижения
запрещенных тематик в Google

— показ разного контента роботу поисковой системы и пользователю (разные версии страниц, скрытые блоки, спамные страницы в скрытой части сайта).



ТОПЭКСПЕРТ

УЧЕБНЫЙ ЦЕНТР

Заимствование контента

Заимствование контента

— показ чужого контента на своем сайте
(автоматическое/ручное заимствование/iframe).



ТОПЭКСПЕРТ

УЧЕБНЫЙ ЦЕНТР

Аффилиаты

Разные сайты одной компании

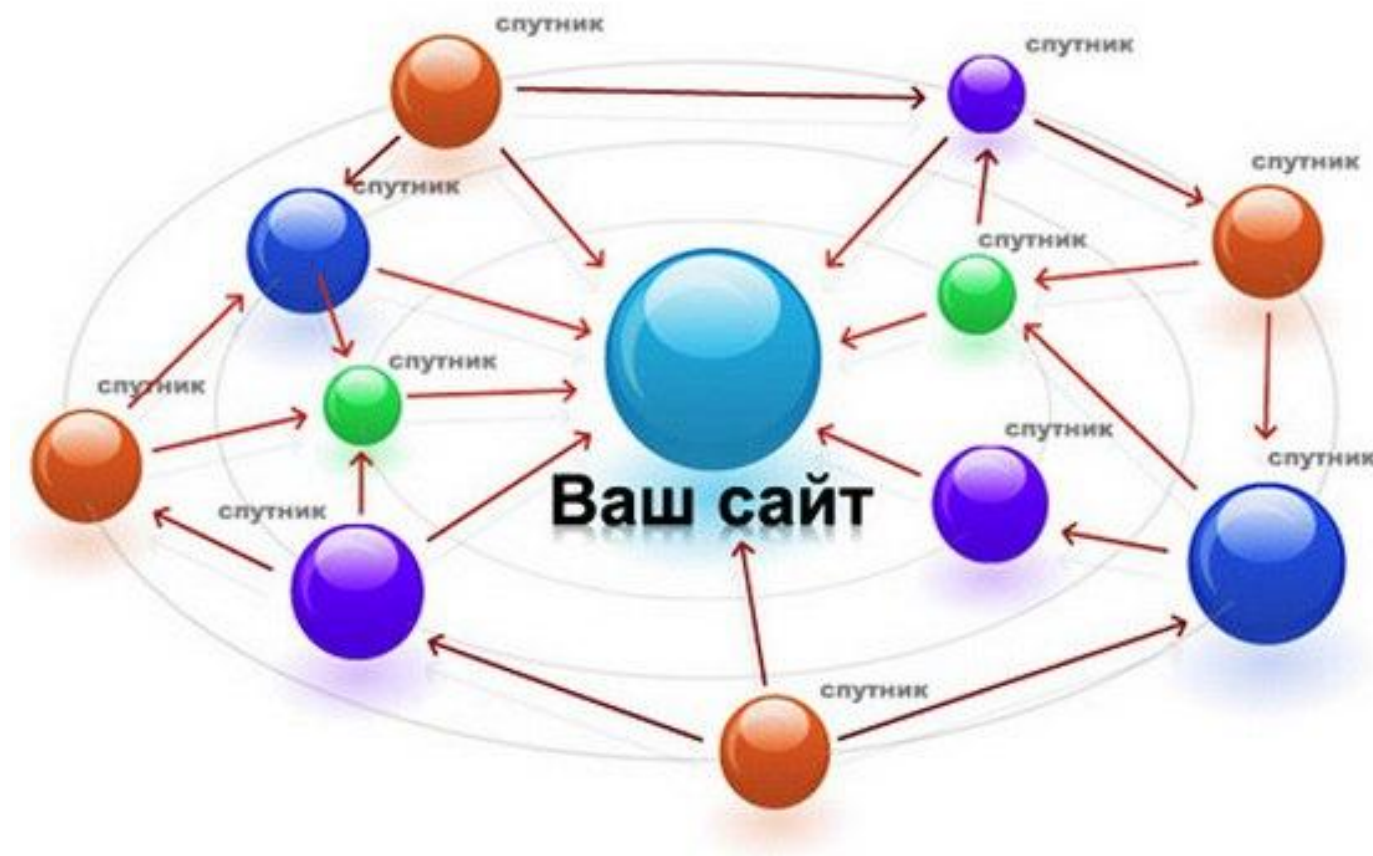
— попытка монополизировать рынок через манипулирование поисковой выдачей.

- совпадение контактных данных,
- один хостинг, один IP-адрес,
- одинаковый дизайн,
- одинаковый ассортимент, ценовое предложение,
- адрес самовывоза,
- владельца домена,
- один аккаунт Вебмастера или Метрики,
- оплата контекста с одного юр.лица



Сателлиты

Разные сайты одной компании





ТОПЭКСПЕРТ

УЧЕБНЫЙ ЦЕНТР

Ссылочный спам

- факт размещения ссылки для накрутки результатов ранжирования,
- создание сайтов для продажи ссылок (отдельные или сетки сателлитов):
 - наличие низкокачественного контента,
 - бесплатный дизайн и CMS,
 - размещение большого количество ссылок с главной страницы на внутренние,
 - прогон сайта по каталогам или низкокачественные ссылки,
 - молодой сайт без живого трафика,
 - появление ссылок на другие сайты в отсутствие трафика и позиций.

- взлом сайтов, использование вирусов (продажа ссылок с display:none),
- спам по форумам, гостевым книгам, доскам объявлений, каталогам (бесплатные CMS без обновления),
- покупка ссылок (ссылочный взрыв),
- обмен ссылками,
- публикация мусорных статей и ссылки из них.

Задача: придумайте
самообучающийся алгоритм
определения спама в
комментариях для движка
типа wordpress



Поведенческий спам

- имитация действий пользователей на выдаче,
- имитация переходов по ссылке,
- имитация поведения внутри сайта.



Спам в рекламе

- Pop-up реклама,
- Ссылки, ведущие на спамные сайты,
- Ссылки или реклама сайтов запрещенных тематик
- Злоупотребление партнерскими программами
- Отправь SMS, введи код, оплати доступ и т. п.

Проблемы сайта за сегодня (16-го Марта)

Другие дни

Нашлось **5 проблем** ● Фатальные 1 ● Критичные 0 ● Возможные проблемы 1 ● Рекомендации 3

● Фатальные

Обнаружены нарушения на сайте

На сайте обнаружены нарушения правил поисковой системы, в связи с чем к нему применены ограничения. Это может привести к ухудшению ранжирования сайта, а также к исключению из поиска его страниц.

[Подробнее](#)

Сигнал отправлен. Мы не гарантируем, что после снятия ограничений видимость сайта в поиске улучшится. Это зависит от многих факторов: насколько хорошо сайт проиндексирован, насколько удобно им пользоваться, нет ли на сайте других нарушений, ещё не представленных в этом разделе.

Кликджекинг

Использование на сайте технологии кликджекинга — механизма обмана пользователей, связанного с размещением на сайте невидимых элементов, взаимодействие с которыми производится без ведома пользователей.

Зачастую такой механизм применяется для получения персональных данных посетителя сайта без его согласия и выполнения действий от его имени в социальных сетях. Подробнее — в [Помощи](#).

Обычно ограничения снимаются в течение двух недель после устранения нарушения. При повторном обнаружении кликджекинга на сайте срок действия ограничений может быть увеличен.

Позиции сайта в результатах поиска понижены

Я всё исправил

Отправить повторный сигнал будет можно через 30 дней



ТОПЭКСПЕРТ
УЧЕБНЫЙ ЦЕНТР

Парсинг

- сервисы подбора запросов,
- поисковая выдача:
 - URL сайтов,
 - позиции,
 - сниппеты,
 - технические данные.
- сохраненные копии
- контекстная реклама
- Яндекс.Каталог



ТОПЭКСПЕРТ

УЧЕБНЫЙ ЦЕНТР

Уведомления

Настройка уведомлений



[Вебмастер](#) [Метрика](#) [Виджеты](#) [Рекламная сеть](#) [Директ](#) [Поиск для сайта](#) [API](#) [ещё](#)

вебмастер → настройки

[Мои сайты](#) [Добавить сайт](#)

Типы сообщений

 [Сообщения](#)

Типы сообщений

[Пересылка](#)

[Исключённые страницы](#)

[Выгрузка внешних ссылок](#)

[Подтверждение доставки](#)

[Настройки главной страницы](#)

Описание	Настройки
Глобальное Сообщения данного типа отправляются всем пользователям сервиса. Они содержат информацию о важных изменениях сервиса или поиска.	☐ Пересылать на электронную почту
Персональное Сообщения данного типа отправляются лично вам. Они используются службой поддержки.	☐ Пересылать на электронную почту
Обновление поисковой базы Сообщения данного типа отправляются после появления новой поисковой базы.	☐ Сохранять в разделе "Сообщения" ☐ Пересылать на электронную почту
Спам на страницах сайта Сообщения данного типа отправляются владельцу сайта, если на одной или нескольких страницах его сайта обнаружен спам.	☒ Сохранять в разделе "Сообщения" ☐ Пересылать на электронную почту
Значительное увеличение числа ошибок Сообщения данного типа отправляются владельцу сайта, если при очередном обходе робота увеличение числа ошибок обхода превысило заданный порог. Сообщать при увеличении числа ошибок более чем на 50 %	☒ Сохранять в разделе "Сообщения" ☐ Пересылать на электронную почту



Антиспам

Последствия спама для ПС

- мусорный контент,
- медленная работа серверов,
- ущерб качественным ресурсам,
- потеря аудитории

Что искать?

- КОНТЕНТ,
- ССЫЛКИ,
- ПОВЕДЕНИЕ

Что искать?

86% спама можно вычислить на основе анализа контента страниц.

Как искать?



Как искать?

Обнаружение спам страниц как задача бинарной классификации:

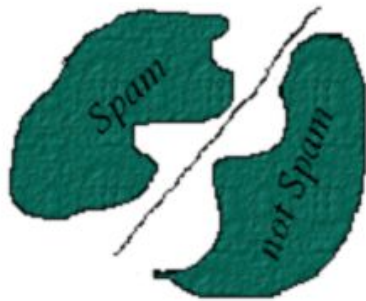
- 1 – спам
- 0 – не спам

Требуется:

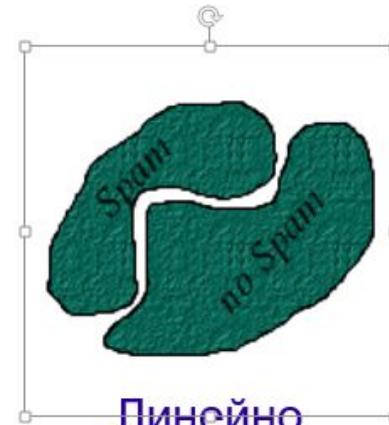
- Список признаков
- Метод классификации

Как искать?

Виды признаков



Линейно
разделимые
признаки



Линейно
неразделимые
признаки.

Выделяем небольшое количество хорошо разделимых признаков.

Как искать?

Попытка №1

Как искать?



ТОПЭКСПЕРТ
УЧЕБНЫЙ ЦЕНТР

Общее количество слов на странице в спамовых и неспамовых множествах



поиск@mail.ru

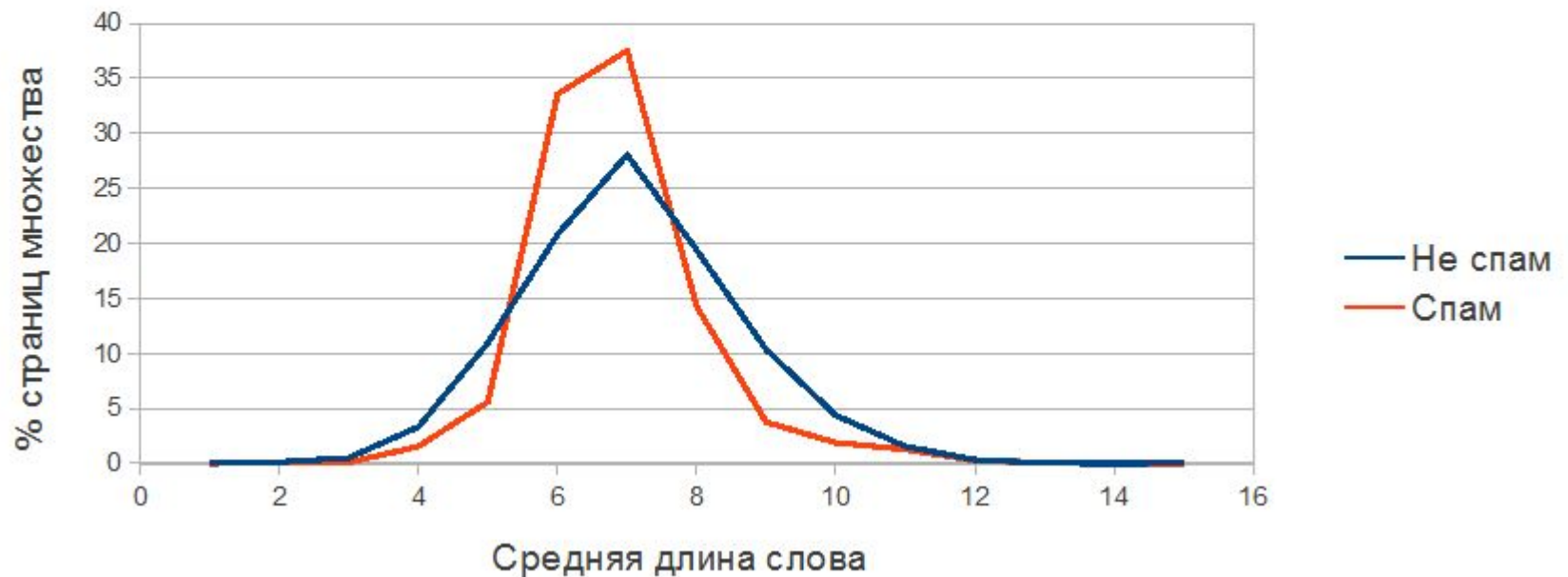
Как искать?

Количества слов в заголовке страниц
в спамовых и неспамовых множествах.



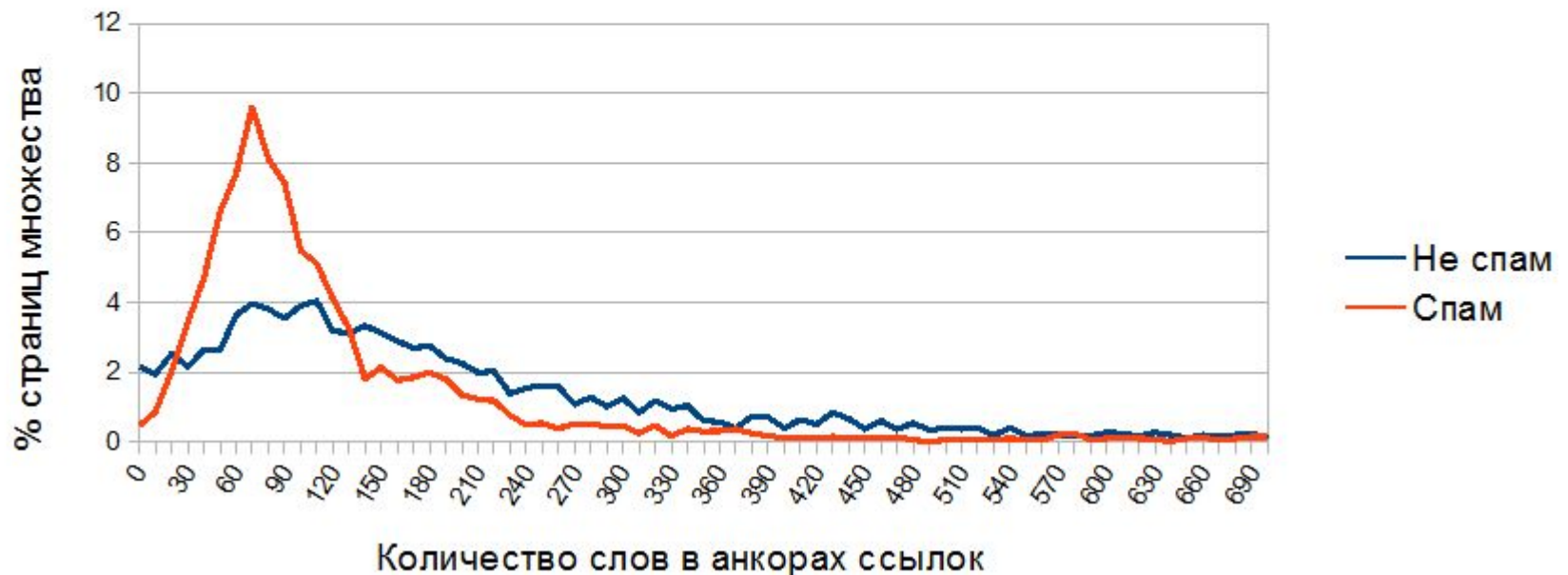
Как искать?

Распределение средней длины слова в спамовых и неспамовых множествах



Как искать?

Количество слов в анкерах ссылок для спамовых и неспамовых множеств



Как искать?

Степень сжатия документов в спамовых и неспамовых множествах



Как искать?

Попытка №2

Как искать?



ТОПЭКСПЕРТ
УЧЕБНЫЙ ЦЕНТР

Распределение усредненного веса ключевых слов для спам- и обычных страниц



Усредненное значение веса ключевых слов документа:

$$w_d^{kw} = \frac{\sum_{i=1}^N w_i}{N}$$

- w_i вес ключевого слова
 N количество ключевых слов

поиск@mail.ru

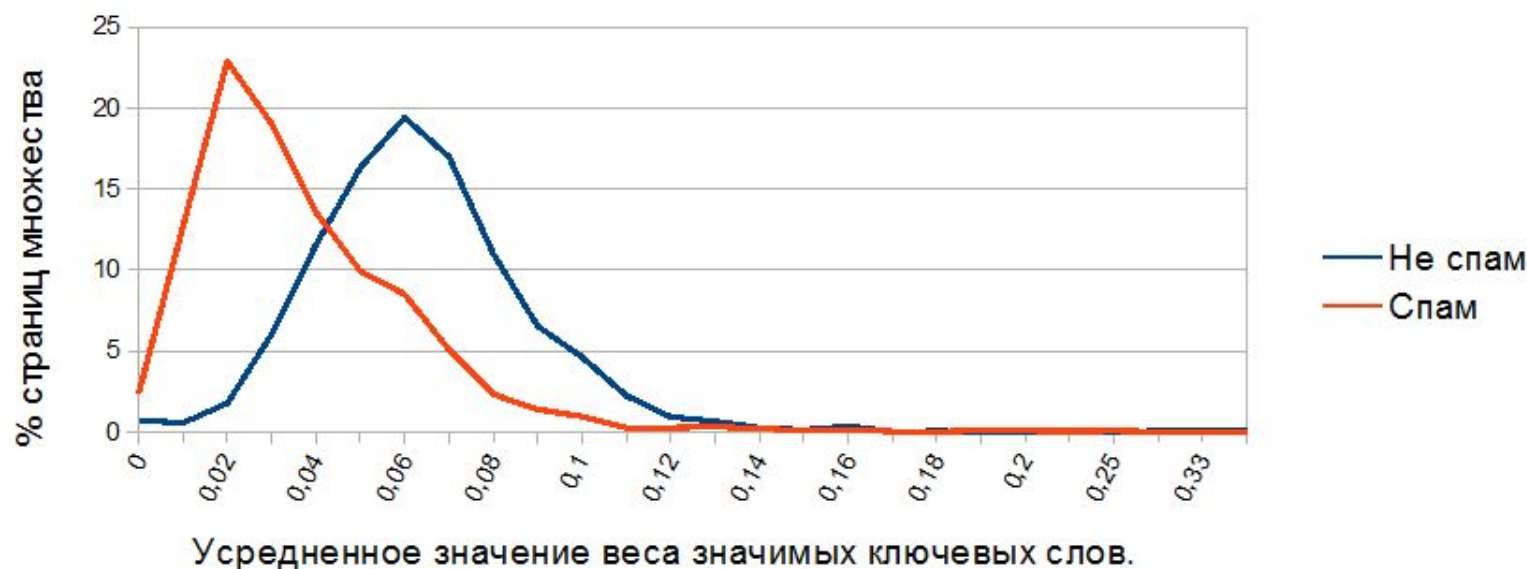
Как искать?



ТОПЭКСПЕРТ

УЧЕБНЫЙ ЦЕНТР

Распределение отношения веса значимых ключевых слов к общему количеству слов в спамовых и неспамовых множествах



Усредненное значение веса значимых ключевых слов документа:

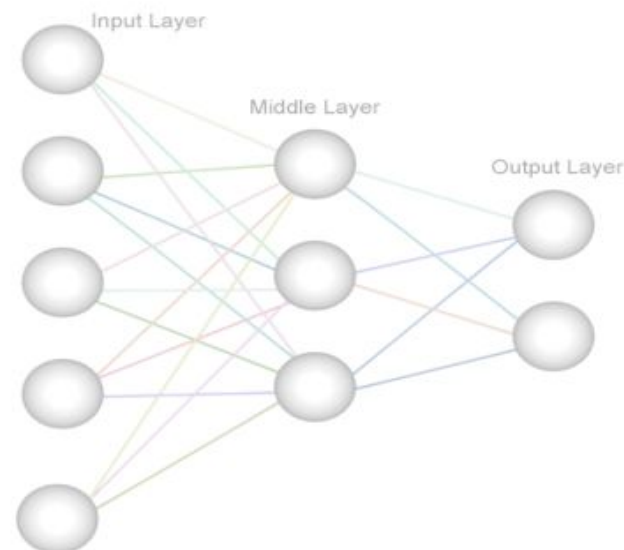
$$\bar{w}_d^{imp} = \frac{\sum_{i=1}^K w_i}{N}$$

- w_i вес ключевого слова
- N количество ключевых слов
- K количество значимых слов

поиск@mail.ru

Многослойный персептрон:

Входной слой — 80 нейронов ,
Скрытый слой — 96 нейронов
Выходной слой — 2 нейрона спам
и неспам
Функция активации — сигмоид



Снова помощь ассессоров

Точность	-	0,97
Полнота	-	0,94
F-мера	-	0,96

Вывод:

Языковые статистики

Грамматические признаки

Биграммные факторы

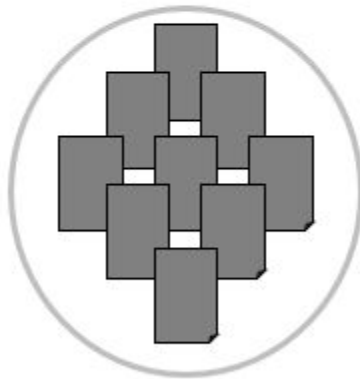
Синтаксис на уровне предложений

вскрывают 97%
ПОИСКОВОГО СПАМА

Как классифицировать сайт?

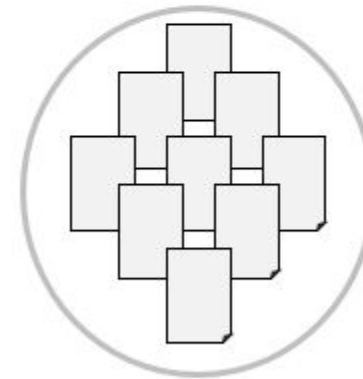
Спам или нет?

Спам сайт



100% = спам

Не спам сайт

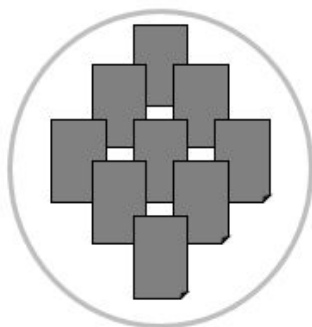


0% = не спам

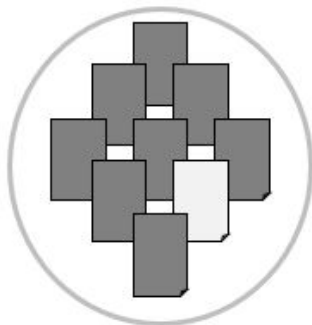
Как классифицировать сайт?

Спам или нет?

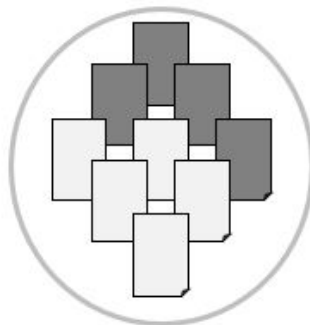
Спам сайт



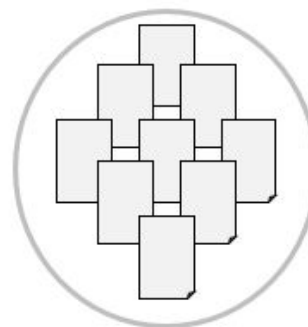
?



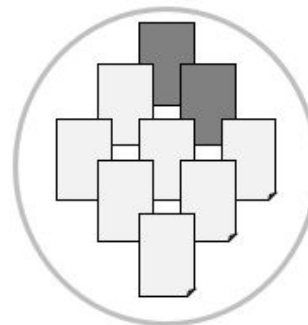
?



Не спам сайт



?



Как классифицировать сайт?

Как оно бывает:

- Хороший сайт со спам страницами:
 - Ошибка классификатора.
 - Взломанный сайт.
 - Переоптимизированный контент.
- Спам сайт с полезными страницами:
 - Ошибка классификатора.
 - Разбавление спама не спам страницами.

Как классифицировать сайт?

Характеристики сайта:

1. Доля спам страниц.
2. Расположение спам страниц.
3. Вероятность прихода/ухода на спам страницу с сайта.
4. На какие страницы ведут входящие/исходящие ссылки.
5. Вероятность участия в спам-ферме.

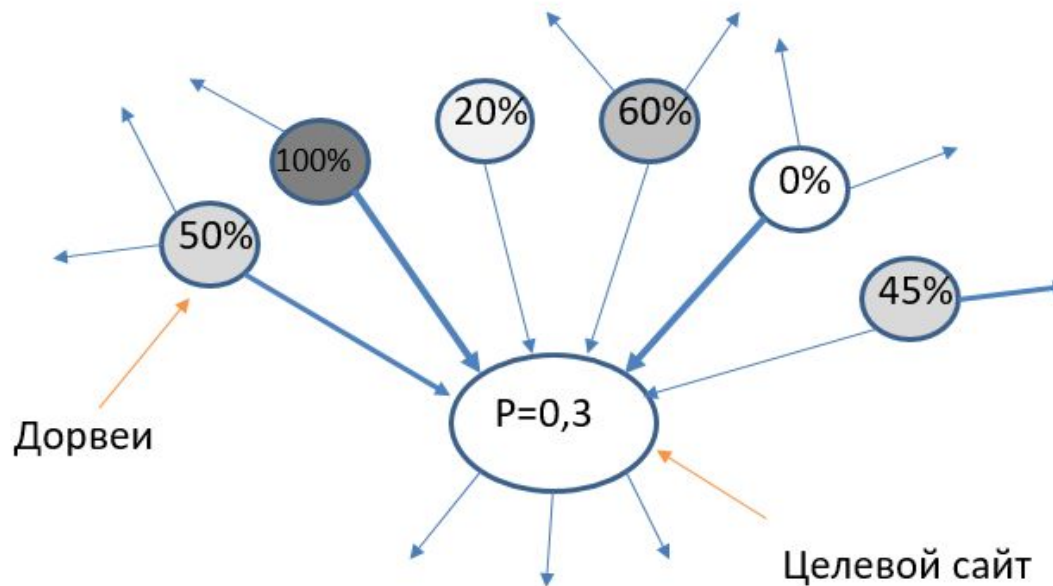
Как классифицировать?

Доля спам страниц



Как классифицировать сайт?

Участие в спам ферме



1. Вероятность того, что сайт раскручивается спам-сайтами.
2. Вероятность участия в спам-ферме.

Как классифицировать сайт?

Вероятность участия в спам-ферме





ТОПЭКСПЕРТ

УЧЕБНЫЙ ЦЕНТР

Оценка качества поиска

Показатели качества поиска

- Полнота - отношение числа найденных релевантных документов, к общему числу релевантных документов в базе.
- Точность - отношение числа релевантных документов, найденных ПС, к общему числу найденных документов.

Домашнее задание

1. Разработайте и опишите алгоритм генерации текста для интернет-магазина, который невозможно отследить автоматически. Почему это невозможно?
2. Опишите качества, по которым поисковая система может определить "спамность" ссылки. Предложите, какой должна быть ссылки, чтобы её невозможно было обнаружить по этим качествам.
3. Вы решили сделать сеть из 100 сателлитов. Придумайте такую схему ссылочных связей, которую невозможно определить в автоматическом режиме. Что может стать основой такого алгоритма? Как будет устроена такая система?
4. Опишите алгоритм накрутки поведенческих факторов, где поисковая система может 100% определить факт накрутки, но не может наложить санкции без риска забанить сайт, который накручивают конкуренты.
5. Определите, как работает сравнение поисковых систем по критерию "полнота и разнообразие" в сервисе <http://analyzethis.ru/>, разберитесь с алгоритмом и опишите его.

При выполнении домашнего задания пользуйтесь DataMining: ищите решения в интернете, изучайте статьи и публикации на эту тему, изучите решения, которые предлагаются для борьбы со спамом, чтобы лучше понимать механизмы. Публикации могут быть на таких сайтах, как РОМИП, HabraHabr, некоторые сообщества в ЖЖ, возможно видео-доклады сотрудников Яндекса.