



ТЕХНОСФЕРА

Поиск дубликатов. Часть 1

Сергукова Юлия,
программист отдела инфраструктуры проекта
Поиск@Mail.Ru

План лекции:

1. Дубликаты
 1. Терминология
 2. Примеры
 3. Шинглирование
2. Поиск дубликатов
 1. Улучшения
 2. Minshingle
 3. Алгоритм Бродера

Дубликаты



Капиbara, или водосвинка





Капиbara, или вodoсвинка

cyclowiki.org/wiki/Капиbara

Статья Обсуждение Читать Правка История Поиск

Капиbara

Капибáра, или водосвíнка (лат. *Hydrochoerus hydrochaeris*) — полуводное травоядное млекопитающее из семейства **водосвíнковых** (*Hydrochoeridae*), единственного представителя в семействе.

Капиbara — самый крупный среди современных грызунов.

Содержание [убрать]

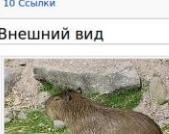
- 1 Внешний вид
- 2 Происхождение и разновидности
- 3 Ареал
- 4 Образ жизни
 - 4.1 Окружающая среда
 - 4.2 Сообщество
 - 4.3 Размножение
 - 4.4 Питание
 - 4.5 Болезни
 - 4.6 Содержание в неволе
 - 4.7 Продолжительность жизни
- 5 Охрана и статус вида
- 6 Популярность
- 7 Интересные факты
- 8 Источники
- 9 Литература
- 10 Ссылки

Внешний вид

Капиbara — это водосвинка, самый крупный современный грызун в мире. Длина тела капибары достигает полутора метров, вес — шестидесяти килограмм. Животное внешне напоминает **морскую свинку** с похожей симпатичной мордочкой, небольшими ушками и большим носом.

В переводе с языка индейцев **гуарани** «капиbara» — это «господин трав». В странах Южной и Центральной Америки это животное называют по-разному — корипи, капуги, калирино, почно.

Небольшие глазки находятся высоко на голове, несколько сзади. Рудиментарный хвост. Довольно короткие конечности. Толстая верхняя пасть, округлые, короткие уши, широко расставленные ноздри. Задние лапы капибары имеют по три пальца, передние — по четыре, причем между пальцами у нее, как у множества водоплавающих имются перепонки.



Капиbara, или водосвинка

1. <https://ru.wikipedia.org/wiki/%D0%9A%D0%B0%D0%BF%D0%B8%D0%B1%D0%B0%D1%80%D0%B0>
2. [http://www.ziganshin.ru/animals/k/Kapibara%20\(Hydrochoerus%20hydrochae ris\).html](http://www.ziganshin.ru/animals/k/Kapibara%20(Hydrochoerus%20hydrochae ris).html)
3. <http://cyclowiki.org/wiki/%D0%9A%D0%B0%D0%BF%D0%B8%D0%B1%D0%B0%D1%80%D0%B0>



Капиbara, или водосвинка

cyclowiki.org/wiki/Капиbara

Статья Обсуждение Читать Правка История П

Капиbara

Капибара, или **водосвинка** (лат. *Hydrochoerus hydrochaeris*) — полуводное травоядное млекопитающее из семейства **водосвинковых** (*Hydrochoeridae*), единствственный представитель в семействе.

Капиbara — самый крупный среди современных грызунов.

Содержание [вратаь]

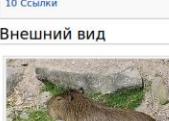
- 1 Внешний вид
- 2 Происхождение и разновидности
- 3 Ареал
- 4 Образ жизни
 - 4.1 Окружающая среда
 - 4.2 Сообщество
 - 4.3 Размножение
 - 4.4 Питание
 - 4.5 Болезни
 - 4.6 Содержание в неволе
 - 4.7 Продолжительность жизни
- 5 Охрана и статус вида
- 6 Популярность
- 7 Интересные факты
- 8 Источники
- 9 Литература
- 10 Ссылки

Внешний вид

Капиbara — это водосвинка, самый крупный современный грызун в мире. Длина тела капибary достигает полутора метров, вес — шестидесяти килограмм. Животное внешне напоминает **морскую свинку** с похожей симпатичной мордочкой, небольшими ушками и большим носом.

В переводе с языка индейцев **гуарани** «капиbara» — это «господин трав». В странах Южной и Центральной Америки это животное называют по-разному — корипинчо, калупи, каприничо, понко.

Небольшие глазки находятся высоко на голове, несколько сзади. Рудиментарный хвост. Довольно короткие конечности. Толстая верхняя губа, округлые, короткие уши, широко расставленные ноздри. Задние лапы капибара имеют по три пальца, передние — по четыре, причем между пальцами у нее, как у множества водоплавающих имются перепонки.



$$\text{КПД} = 1/3$$

[Участие](#)
[Создать об юнике](#)
[Портал сообщества](#)
[Форум](#)
[Свежие правки](#)
[Новые страницы](#)
[Страницы](#)
[Помощь](#)

[Инструменты](#)
[Ссылки сюда](#)
[Связанные правила](#)
[Спецстраницы](#)
[Постоянная ссылка](#)
[Сведения о странице](#)
[Цитировать страницу](#)

[Печатчик/экспорт](#)
[Создать книгу](#)
[Скачать как PDF](#)
[Версия для печати](#)

[В других проектах](#)
[Википедия](#)
[Википедида](#)
[Викиданные](#)

На других языках

- [Српски](#)
- [Deutsch](#)
- [English](#)
- [Soomaalzi](#)

Контент vs информация

Контент vs информация

1. Контент - текст + изображения + видео + другие данные на странице (в т.ч. стили)
2. Информация - семантический уровень данных(смысл)

Мы умеем работать только с контентом

Полезный контент - подмножество всего контента на странице.
Данные, полезные для индексации и поиска

Постановка проблемы (идеальный мир)

Полезный контент идёт в индекс

Больше **разнообразного** полезного контента - больше полнота индекса

Цель: качать больше разнообразного контента

Постановка проблемы (реальный мир)

Мы не можем заранее сказать, какой контент находится на странице

Только предполагаем: ранжирование, сад камней и т.д.

Цель 1: качать меньше потенциальных дубликатов

Цель 2: не допускать попадание дубликатов в индекс => поиск дубликатов *после* выкачки

План лекции:

1. Дубликаты
 1. Терминология
 2. Примеры
 3. Шинглирование
2. Практика
3. Поиск дубликатов
 1. Улучшения
 2. Minshingle
 3. Алгоритм Бродера
4. Домашняя работа

Какие бывают дубликаты?



Виды дубликатов. Зеркала

Совпадение 85-100% всего контента

http://lurkmore.to/%D0%A3%D0%BB%D1%8B%D0%B1%D0%B0%D0%B5%D0%BC%D1%81%D1%8F_%D0%B8_%D0%BC%D0%B0%D1%88%D0%B5%D0%BC

VS

https://lurklurk.com/%D0%A3%D0%BB%D1%8B%D0%B1%D0%B0%D0%B5%D0%BC%D1%81%D1%8F_%D0%B8_%D0%BC%D0%B0%D1%88%D0%B5%D0%BC



Виды дубликатов. Зеркала



ПОИСК

Перейти

Найти

навигация

- » Главная страница
- » Свежие правки
- » Нужные статьи
- » Истории
- » Случайная статья
- » Все статьи
- » Новые статьи
- » Участникам
- » Гедлайны (beta)
- » FAQ
- » Про цензуру

категории

- » Основы
- » Мены
- » Люди
- » Персонажи
- » Дач
- » ЖЖ
- » Фидо
- » ЛОР
- » Имена и ярлы
- » Занимательная география

счётчики

инструменты

- » Ссылки сюда
- » Связанные правки
- » Спецстраницы
- » Версия для печати
- » Постоянная ссылка

статья обсуждение просмотр непроверенные изменения править история

Представиться / зарегистрироваться

Улыбаемся и машем

(В этой версии ожидает проверки 1 изменение. Опубликованная версия была проверена 3 октября 2014.)

Улыбаемся и машем, парни, улыбаемся и машем! (англ. just smile and wave, boys. Smile and wave.) — фраза пингвинов (точнее, Шкипера) из мультифильма «Мадагаскар». В свою очередь, в мультифильме она появилась как аллюзия к фильму «Безумие короля Георга» (приуф®), в котором эту фразу любила говорить своей королевской семье королева Шарлотта — та, которая шарлатку изобрела.

В этой стране фраза пошла в народ, да так лихо пошла, что расположилась чуть менее, чем во всем **жежешекам, лирушечкам** и прочим быдлофорумам (гуль не даст сорвать). Вероятно,

разгадка скорости заражения заключается в том, что у поживших при ССР эта фраза ассоциируется с членами Политбюро ЦК КПСС и лично Леонидом Ильичом Брежневым, которые

стоят на **Мавзолее** во время парада на Красной площади 7 ноября или 1 мая. Стоят, улыбаются и машут.

Используется

- » Как аналог фразы «но мы-то с вами знаем...»
- » В качестве указателя на **тонкие обстоятельства** или **какую-либо гениальную идею**, посетившую светлую голову аффтара.
- » Как ироничная констатация собственной наивности или бездельяности.
- » В качестве ответа на **чё-то дурацкий совет** — как констатация собственного (возможно, минимого) превосходства над советчиком.
- » Как потребность привлечь чё-то внимание для исправления накосяченного (в роли обесцвеченного сплинигриза, то есть младенца).
- » Как инцидентовка благополучия, от которого хочется ридать. Популярна в стране Түркменистан и пионерлагерях в совке.
- » Как намек на **очередные прошки** очередной Кровавой Гбми или иных профильных организаций.
- » Когда сказать нехуй, а хочется.
- » Когда начальство пришло с проверкой (зашенти успехи пробывающихся работников), и вот оно уже тут, чём-то увлечено, и на самих работников особого внимания не обращает.

[править]



Отрывок из мульта



Королевский двор улыбается и машет

Улыбаемся и машем

(Опубликованная версия была досмотрена 3 октября 2014. В черновой версии имеются ожидающие проверки изменения файлов или шаблонов.)

Улыбаемся и машем, парни, улыбаемся и машем! (англ. just smile and wave, boys. Smile and wave.) — фраза пингвинов (точнее, Шкипера) из мультифильма «Мадагаскар». В свою очередь, в мультифильме она появилась как аллюзия к фильму «Безумие короля Георга» (приуф®), в котором эту фразу любила говорить своей королевской семье королева Шарлотта — та, которая шарлатку изобрела.

В этой стране фраза пошла в народ, да так лихо пошла, что расположилась чуть менее, чем во всем **жежешекам, лирушечкам** и прочим быдлофорумам (гуль не даст сорвать). Вероятно, разгадка скорости заражения заключается в том, что у поживших при ССР эта фраза ассоциируется с членами Политбюро ЦК КПСС и лично Леонидом Ильичом Брежневым, которые

стоят на **Мавзолее** во время парада на Красной площади 7 ноября или 1 мая. Стоят, улыбаются и машут.



Отрывок из мульта



Королевский двор улыбается и машет



Penguins.jpg

А вот так это обычно выглядит IRL.

А вот так это обычно выглядит IRL.

С новым годом, одноклассники!

С новым годом, одноклассники!

Используется

- » Как аналог фразы «но мы-то с вами знаем...»
- » В качестве указателя на **тонкие обстоятельства** или **какую-либо гениальную идею**, посетившую светлую голову аффтара.
- » Как ироничная констатация собственной наивности или бездельяности.
- » В качестве ответа на **чё-то дурацкий совет** — как констатация собственного (возможно, минимого) превосходства над советчиком.
- » Как потребность привлечь чё-то внимание для исправления накосяченного (в роли обесцвеченного сплинигриза, то есть младенца).
- » Как инцидентовка благополучия, от которого хочется ридать. Популярна в стране Түркменистан и пионерлагерях в совке.
- » Как намек на **очередные прошки** очередной Кровавой Гбми или иных профильных организаций.
- » Когда сказать нехуй, а хочется.
- » Когда начальство пришло с проверкой (зашенти успехи пробывающихся работников), и вот оно уже тут, чём-то увлечено, и на самих работников особого внимания не обращает.

В результате регулярного обострения СПГС у основных обитателей блогосферы, фраза лепится настолько, там где надо и не надо.

Официальный пакетом фраза используется при получении задания, заранее обреченнего на провал. Планктон как бы говорит нам, что он понимает бесперспективность порученного задания, но надо же чем-то заниматься, ибо в противоположном случае можно **огреть лопатой**.

Также фраза используется как намек, что пока на чё-то жопу сваливались неприятности (бэбс, тест на наркоту, поиск шпор на экзамене), и источник этих неприятностей занят, то этим надо воспользоваться (тихо пройти, списать, свалить).

Или полностью наоборот. Покажа фраза используется по отношению к планктону **крытны пацанами**. По типу — **расслабися и получай удовольствие**, пока мы тебя имеем. Например, в фильме Гая Ричи «Рок-н-ролщик»: «Да, это ограбление. Кладем сумки в машину, закрываем двери, улыбаемся, уходим».

Strongly related

- Foto 3560.gif
Один джентльмен загорал с сыном на пляже, а теща купалась. Вдруг она стала тонуть.
— Папа, смотри, наша бабушка что-то кричит и машет руками! — закричал сын.
— Что же ты сидишь, сынок? — сказал джентльмен. — Помаши и ты ей на помощь.

Бородатый анекдот

Ссылки

- » Песенка Пушного Ф

No u
turn
sign.png

«Улыбаемся и машем» имеет отношение к универсальным ответам.

[+]

42 • By design (Так надо) • Deal with it • GET OUT • НА НА НА, OH WOW • I dunno LOL • Null said • Not Your Personal Army • OK • One-liner • Sad but true (Это печально) • U MAD • А то! • Все

Виды дубликатов. Плагиат

Совпадение 85-100% **полезного контента**

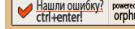
<http://annales.info/evrope/behaym/behaym18.htm>

vs

http://medieval_weapons.academic.ru/41



Виды дубликатов. Плагиат

 Сайт подключен к системе Orphus. Если Вы увидели ошибку и хотите, чтобы она была устранена, выделите соответствующий фрагмент текста и нажмите Ctrl+Enter!

[Назад К содержанию Дальше](#)

[Разновидности турниров]

I. «Механический» реннен

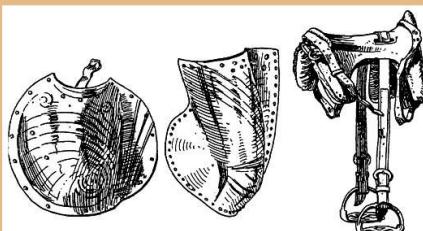
(нем. *Geschifttartschenrennen*)

Всадник одет в реннцойг, под доспехом — толстая ватная куртка — вамс с рукавами-буфами на упругой подкладке, заменяющими наручи. Ноги зачастую не имеют поножей. Защитой бедра служат ребристые набедренные щитки (нем. *Streifartschen*, рис. 621) или дильже (рис. 622) на ремнях, перекинутых или прорезанных через седло. Легкие реннен седла (ит. *silla rasa*) не имеют передних и задних лук (рис. 623). Лошадь покрыта кожаной попоной, голова защищена глухим налобником. В этом виде поединков было две разновидности. [405]

1. «Механический» реннен с тарчем

(нем. *Geschifttartschenrennen*)

При этом виде турнира удачный удар по тарчу противника позволял оторвать его от кирасы вместе со множеством металлических крепежных деталей. Эффект был вызван пружинным механизмом, установленным по центру нагрудника кирасы и соединенным с тарчем посредством штыря. Штырь проходил через отверстие в тарче и заклинивался снаружи металлической шайбой. Между тарчом и пружинным механизмом зажаты концентрические клинья таким образом, что они своим давлением на тарч удерживали пружину механизма, который своим усилием прижимал клинья.



Rис. 621. Набедренный щиток, для защиты бедра от удара о барьера. Кон. XVI в.
 Rис. 622. Дильже для правой ноги. Кон. XV в.
 Rис. 623. Легкое седло для турнира реннен. Кон. XV в.



★ Запомнить сайт Словарь на свой сайт RU ▾

Словари и энциклопедии на Академике

Ведите текст для поиска по словарям и энциклопедиям

Энциклопедия средневекового оружия

Найти!

АКАДЕМИК
dic.academic.ru

Энциклопедия средневекового оружия

Толкование

Разновидности турниров

I. «Механический» реннен

(нем. *Geschifttartschenrennen*)

Всадник одет в реннцойг, под доспехом — толстая ватная куртка — вамс с рукавами-буфами на упругой подкладке, заменяющими наручи. Ноги зачастую не имеют поножей. Защитой бедра служат ребристые набедренные щитки (нем. *Streifartschen*, рис. 621) или дильже (рис. 622) на ремнях, перекинутых или прорезанных через седло. Легкие реннен седла (ит. *silla rasa*) не имеют передних и задних лук (рис. 623). Лошадь покрыта кожаной попоной, голова защищена глухим налобником. В этом виде поединков было две разновидности.

1. «Механический» реннен с тарчем

(нем. *Geschifttartschenrennen*)

При этом виде турнира удачный удар по тарчу противника позволял оторвать его от кирасы вместе со множеством металлических крепежных деталей и выбросить тарч за голову всадника высоко в воздух. Этот эффект был вызван пружинным механизмом, установленным по центру нагрудника кирасы и соединенным с тарчем посредством штыря. Штырь проходил через отверстие в тарче и заклинивался снаружи металлической шайбой. Между тарчом и пружинным механизмом зажаты концентрические клинья таким образом, что они своим давлением на тарч удерживали пружину механизма, который своим усилием



Коды ответов

200 – успех! - их качает спайдер

30x – редирект

404 – страница не существует - нет контента для спайдера

50x – ошибка сервера

Страница не найдена. Примеры

404: <http://war-toys.ru/component/content/article/34/1-2012-01-28-09-03-06>

404 - Статья #34 не найдена!

Вы не можете посетить эту страницу из-за:

1. устаревшие закладки в избранном
2. поисковый сервер имеет устаревшие данные сайта
3. некорректный адрес
4. Вы не имеете доступа к этой странице
5. Запрашиваемый ресурс не найден.
6. Произошла ошибка при обработке вашего запроса!

Пожалуйста, выберите одну из следующих страниц:

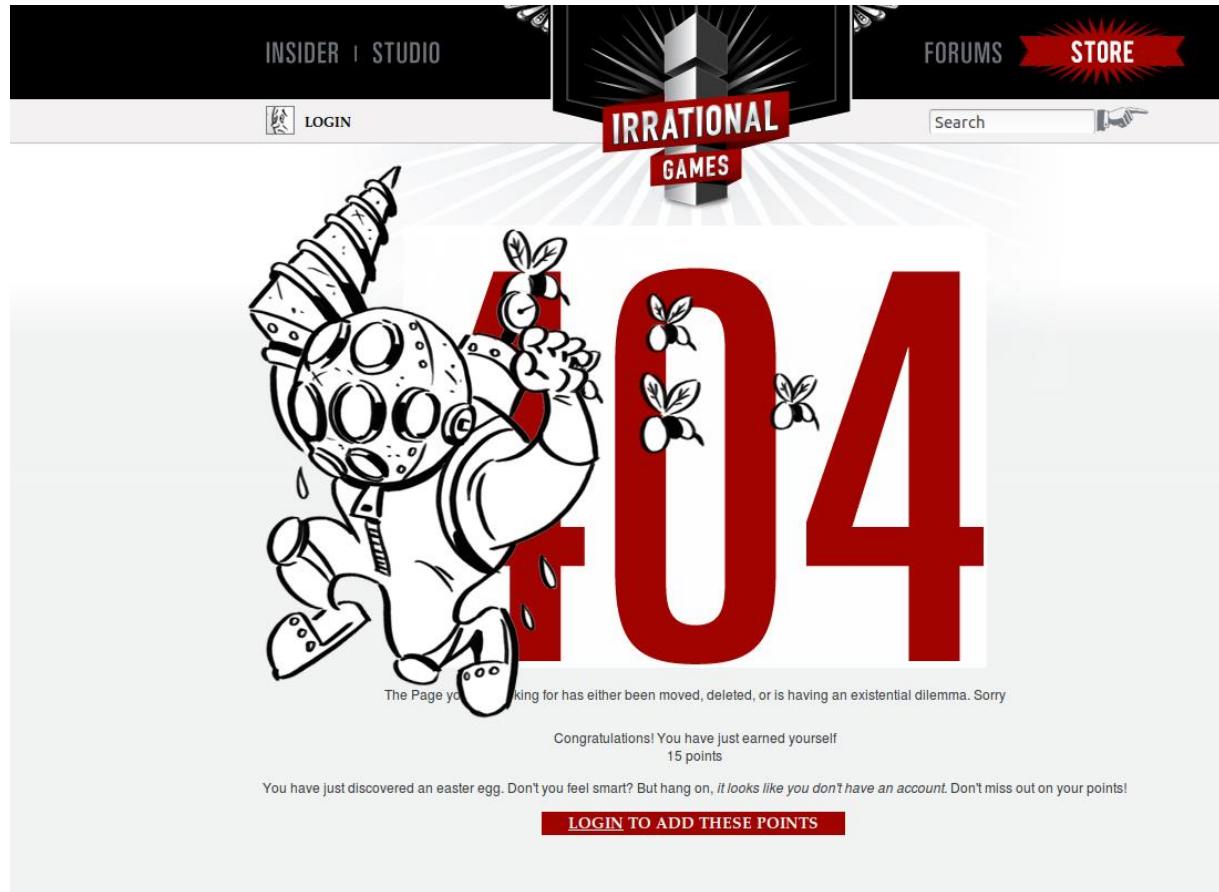
- [Главная страница](#)

Если у вас возникли сложности, пожалуйста, свяжитесь с администрацией этого сайта.

Статья #34 не найдена!

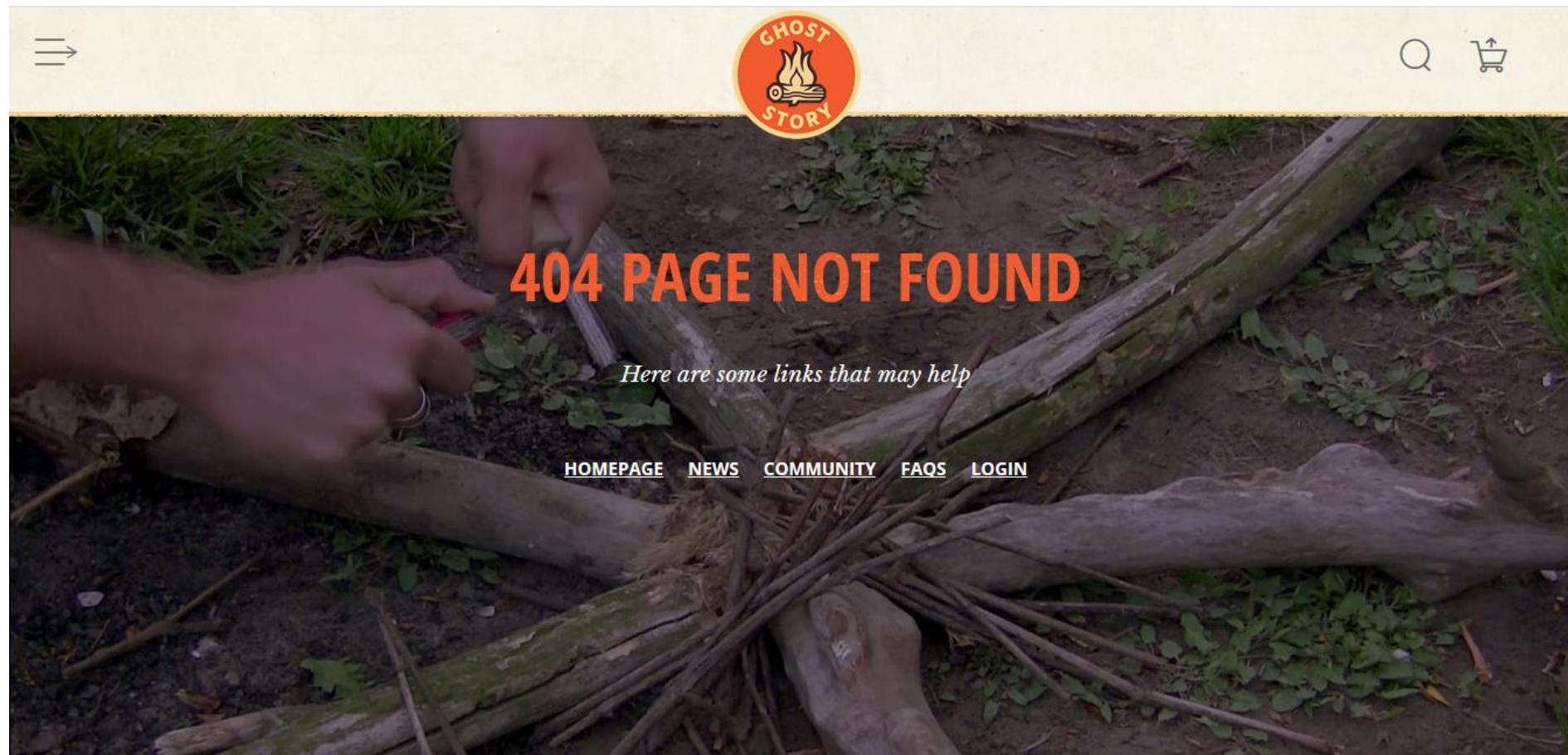
Страница не найдена. Примеры

404: <http://irrationalgames.com/asdfasdf>



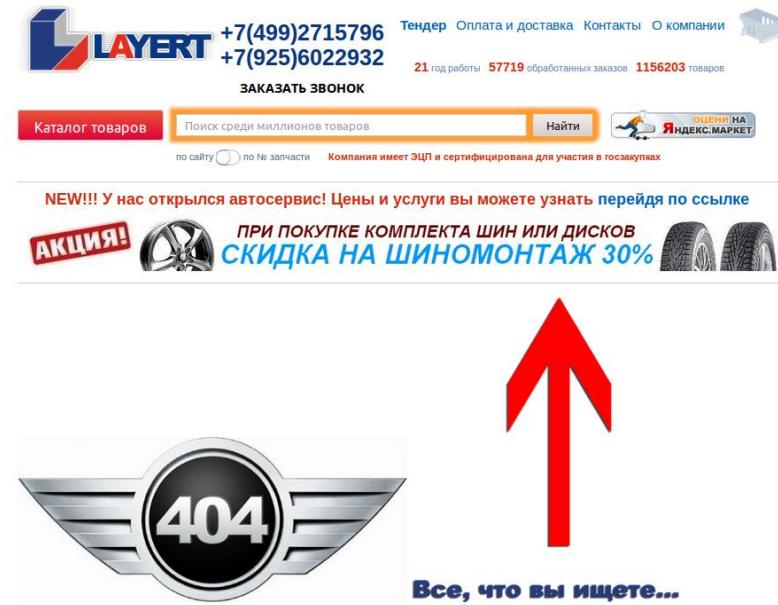
Страница не найдена. Примеры

404: <https://www.ghoststorygames.com/asdfasfsadfs>



Страница не найдена. Примеры

[http://layert.ru/site/menu/zpch vaz/dvig vaz.php](http://layert.ru/site/menu/zpch_vaz/dvig_vaz.php) 200(!)



Все, что вы ищете...

Запрошенной страницы не существует.

Это возможно при следующих обстоятельствах:

1. ссылка, по которой вы перешли, устарела
2. вы набрали в адресной строке неверный адрес

Если вы попали на эту страницу по ссылке на нашем сайте, напишите пожалуйста откуда и куда вы хотели попасть на copy@layert.ru

Отсюда вы можете:

1. Вернуться на главную страницу сайта

В чём проблема?

<http://layert.ru/site/menu/zapch vaz/dvig vaz.php> - честный 200

<http://layert.ru/site/menu/zpch vaz/dvig vaz.php>

<http://layert.ru/site/menu/zap vaz/dvig vaz.php>

<http://layert.ru/site/menu/zapch/dvig vaz.php>

404, которые говорят 200

Виды дубликатов. Soft 404

- 404
- “сайт заблокирован”
- “сайта больше нет”
- пользователя не существует
- и т.д.

1C-Bitrix – CMS-система
опция для настройки soft-404



Виды дубликатов. Похожие новости

Вечерние пригородные электрички №6095 и №6096 не будут курсировать по маршруту Тайга – Томск-1 – Тайга 7,9 и 15 октября в связи ремонтом на перегоне Богашево – Томск. Об этом сообщает пресс-служба ведомства.

Компания «Кузбасс-пригород» просит пассажиров быть внимательными и планировать свои поездки заранее с учетом изменений в расписании движения пригородных поездов.

Более подробную информацию о расписании движения электричек можно получить в кассах ОАО «Кузбасс-пригород», на сайте компании, а также с 8:00 до 20:00 по телефонам: (3842) 32-37-17, (38448) 7-20-54, 8(905) 968-90-70.

Ранее сообщалось, что РЖД отменит пригородных электричек из Томска и изменят частоту еще одного пригородного поезда из-за перехода на зимнее расписание.

Электропоезда № 6095 и № 6096 не будут совершать поездки по маршруту Тайга — Томск-1 — Тайга три дня в октябре из-за ремонтных работ, сообщает пресс-служба Западно-Сибирской железной дороги (филиал ОАО «РЖД»).

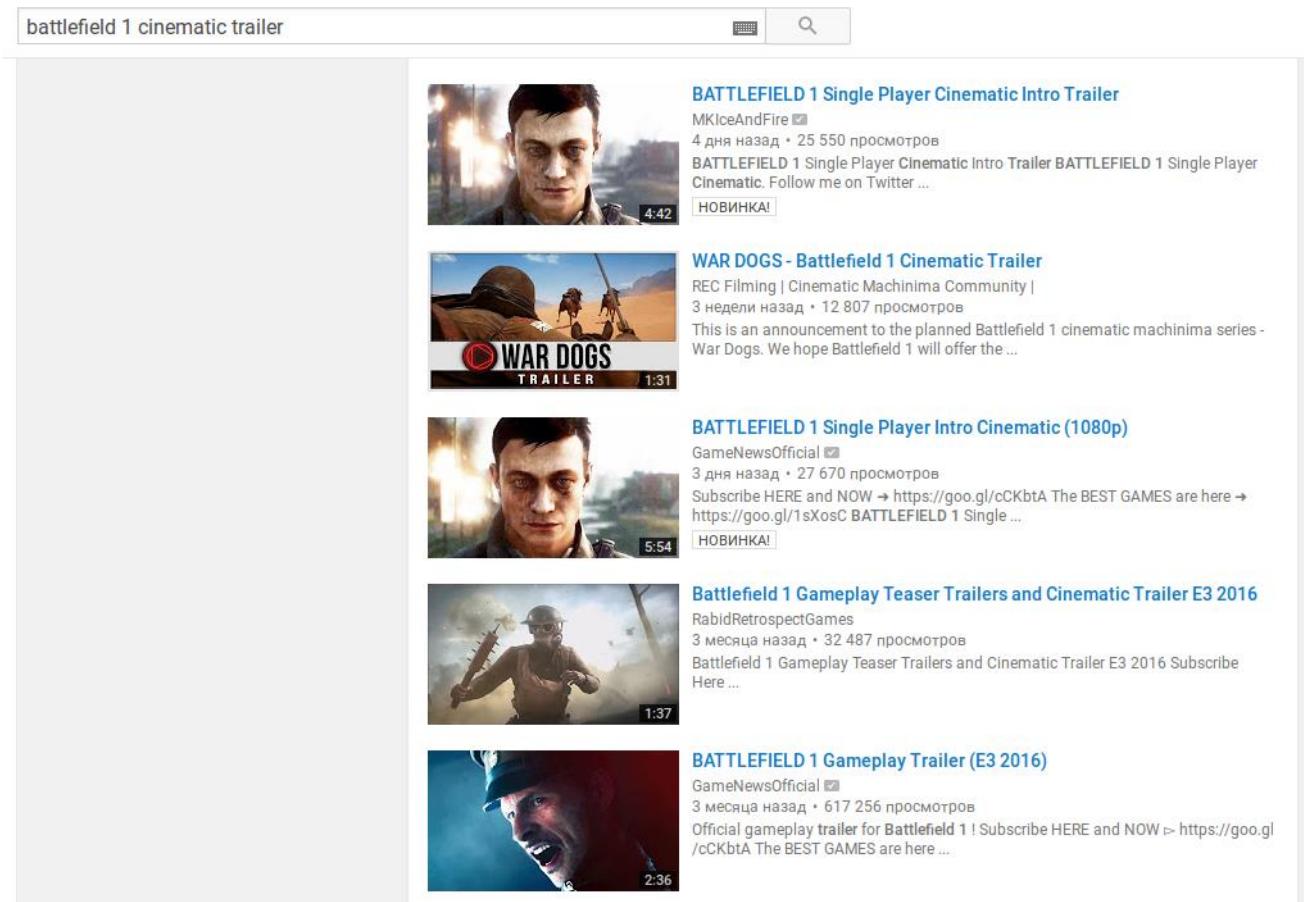
Вечерние пригородные электрички № 6095 и № 6096 не будут курсировать по маршруту Тайга — Томск-1 — Тайга 7, 9 и 15 октября в связи с проведением капитального ремонта на перегоне Богашево — Томск Кузбасского региона Западно-Сибирской железной дороги.

Компания «Кузбасс-пригород» просит пассажиров планировать свои поездки заранее с учетом изменений в расписании движения пригородных поездов.

Более подробную информацию о расписании движения электричек можно получить в кассах ОАО «Кузбасс-пригород», а также с 08:00 до 20:00 по телефонам 8 (3842) 32-37-17, 8 (3844) 87-20-54.

Виды дубликатов

Дубликатами могут быть не только текстовые документы



battlefield 1 cinematic trailer

BATTLEFIELD 1 Single Player Cinematic Intro Trailer
MKIceAndFire
4 дня назад • 25 550 просмотров
BATTLEFIELD 1 Single Player Cinematic Intro Trailer BATTLEFIELD 1 Single Player Cinematic. Follow me on Twitter ...
НОВИНКА! 4:42

WAR DOGS - Battlefield 1 Cinematic Trailer
REC Filming | Cinematic Machinima Community |
3 недели назад • 12 807 просмотров
This is an announcement to the planned Battlefield 1 cinematic machinima series - War Dogs. We hope Battlefield 1 will offer the ...
1:31

BATTLEFIELD 1 Single Player Intro Cinematic (1080p)
GameNewsOfficial
3 дня назад • 27 670 просмотров
Subscribe HERE and NOW → <https://goo.gl/cCKbtA> The BEST GAMES are here → <https://goo.gl/1sXosC> BATTLEFIELD 1 Single ...
НОВИНКА! 5:54

Battlefield 1 Gameplay Teaser Trailers and Cinematic Trailer E3 2016
RabidRetrospectGames
3 месяца назад • 32 487 просмотров
Battlefield 1 Gameplay Teaser Trailers and Cinematic Trailer E3 2016 Subscribe Here ...
1:37

BATTLEFIELD 1 Gameplay Trailer (E3 2016)
GameNewsOfficial
3 месяца назад • 617 256 просмотров
Official gameplay trailer for Battlefield 1 ! Subscribe HERE and NOW → <https://goo.gl/cCKbtA> The BEST GAMES are here ...
2:36

Поиск основанный на дубликатах

???

Поиск основанный на дубликатах

???



X: 1
T: Ev Chistr 'Ta, Laou
R: barndance
M: 4/4
L: 1/8
K: Gmin
f2 f2d2|e4c4|z2f2 f2e2|d2f2 f2d2|e4c3c|
d2B2 c4|z2f2 f2d2|e4c4|z2f2 f2e2|d2f2 f2d2|
e4c3c|d2B2 c4|z2c2 e2f2|g4g4|_a2f2 g4|
z2g2 f2f2|e4c4|d2B2 c4|z2c2 e2f2|
g4g4|_a2f2 g4|z2g2 f2f2|e4c3c|d2B2 c4|]
#Added by fiel 11 years ago.

 DOWNLOAD

 PRINT



План лекции:

1. Дубликаты
 1. Терминология
 2. Примеры
 3. Шинглирование
2. Поиск дубликатов
 1. Улучшения
 2. Minshingle
 3. Алгоритм Бродера

Поиск дубликатов

Дано: 2 документа

Задание: определить, являются ли они дубликатами

Поиск дубликатов. Подходы

1. Использовать весь текст
2. Использовать фрагмент текста
3. Использовать несколько фрагментов текста
4. Словари
5. Число/числа, вычисленные на основе особенностей текста
6. Др. сигнатура

Поиск дубликатов. Метрики

Характер сигнатуры определяет допустимое множество метрик

Метрика - функция(!), которая задает отношение между текстами

Поиск дубликатов. Простой пример

Мама мыла раму

VS

Мамма мыла раму

Поиск дубликатов. Шинглы

«Shingle» - «чешуйка», «черепица»

Шинглирование - получение множества фрагментов исходного текста

1 шингл - фрагмент текста длиной N

Поиск дубликатов. Шинглы.

Разбиение текста

Мама мыла раму

Как построим шинглы?

Поиск дубликатов. Шинглы.

Разбиение текста.

Последовательность шинглов

Мама_мыла_раму N = 3

{"Мам", "а_м", "ыла", "_ра", "му"}

Поиск дубликатов. Шинглы.

Разбиение текста.

Последовательность шинглов

Мама мыла раму $N = 3$

{"Мам", "а_м", "ыла", "_ра", "му"}

- Что делать с группой, меньше чем N ?
- Слишком чувствительно к неточным совпадениям:
"мамма мыла раму" -> {"мам", "ма_", "мыл", "а_р", "аму"}

Поиск дубликатов. Шинглы.

Разбиение текста.

Словарное разбиение

Мама мыла раму N = 1

{"Мама", "мыла", "раму"}

Поиск дубликатов. Шинглы.

Разбиение текста.

Словарное разбиение

Мама мыла раму N = 1

{"Мама", "мыла", "раму"}

- Достаточно большие тексты на похожую тематику основываются на практически одинаковых словарях
- Иногда порядок важен:
 - "Рыцаря нельзя было помиловать, и король решил его казнить"
 - "Рыцаря нельзя было казнить, и король решил его помиловать"

Поиск дубликатов. Шинглы.

Разбиение текста. Разбиение "внахлёт"

Мама мыла раму

N = 10

M	a	m	a			m	y	л	а			r	a	m	y	...
---	---	---	---	--	--	---	---	---	---	--	--	---	---	---	---	-----



shingle1

Поиск дубликатов. Шинглы.

Разбиение текста. Разбиение "внахлёт"

Мама мыла раму

N = 10



shingle2

Поиск дубликатов. Шинглы.

Разбиение текста. Разбиение "внахлёт"

Мама мыла раму

N = 10



shingle3



Шинглы. Сравнение документов

Построим матрицу смежности:

столбцы - множество документов

строки - всё возможное множество шинглов

	d1	d2	d3	...	dK
sh1	1	1	0		1
sh2	0	1	1		1
sh3	0	1	1		0
...					
shN	1	0	0		1

Шинглы. Сравнение документов

Все шинглы длины 8 для [a-zA-Z] -> $(26+26+1)^8$

Улучшение - нам не нужно всё множество шинглов. Достаточно множества шинглов из наших документов (т.е. удаляем строки из 0)



Сравнение документов.

У каждого документа – множество шинглов – вектор из 0 и 1

	d1	d2	d3	...	dK
sh1	1	1	0		1
sh2	0	1	1		1
sh3	0	1	1		0
...					
shN	1	0	0		1

Сравнение документов. Мера Жаккара

У каждого документа - множество шинглов

Мера Жаккара: $JC(A, B) = \frac{A \cap B}{A \cup B}$

Мера Жаккара. Пример

	d1	d2		
sh1	1	1		
sh2	0	1		
sh3	0	0		
sh4	1	0		
sh5	0	0		
sh6	0	1		

Мера Жаккара. Пример

	d1	d2		
sh1	1	1	*	
sh2	0	1		
sh3	0	0		
sh4	1	0		
sh5	0	0		
sh6	0	1		

Мера Жаккара. Пример

	d1	d2		
sh1	1	1	*	*
sh2	0	1		*
sh3	0	0		
sh4	1	0		*
sh5	0	0		
sh6	0	1		*

Мера Жаккара. Пример

	d1	d2			JC = 1/4
sh1	1	1	*	*	
sh2	0	1		*	
sh3	0	0			
sh4	1	0		*	
sh5	0	0			
sh6	0	1		*	

План лекции:

1. Дубликаты
 1. Терминология
 2. Примеры
 3. Шинглирование
2. Поиск дубликатов
 1. Улучшения
 2. Minshingle
 3. Алгоритм Бродера



Перерыв

План лекции:

1. Дубликаты
 1. Терминология
 2. Примеры
 3. Шинглирование
2. Поиск дубликатов
 1. Улучшения
 2. Minshingle
 3. Алгоритм Бродера

План лекции:

Чем больше документов, тем:

1. Больше множество всех шинглов этих документов
2. Больше сравнений пар документов

Кроме того - работать с текстом накладно

Переход к числам

Hash("Мама мыла ") = 172367463

Каждый шингл - в значение хэш-функции

Зачем? Экономим место.

1 char ~ 8bit

1 int ~ 32bit

Переход к числам

Hash("Мама мыла ") = 172367463

Каждый шингл - в значение хэш-функции

Зачем? Экономим место.

1 char ~ 8bit

1 int ~ 32bit

10 char ~ 80bit => "экономим" 48 бит на каждом шингле
PROFIT!

Сокращение множества шинглов

$sh(doc_1) = A, sh(doc_2) = B$

$A' \subseteq A : |A'| = N_s \ll |A|$

$B' \subseteq B : |B'| = N_s \ll |B|$

$P(\rho(A, B) \geq L \& \rho(A', B') \geq L) > 0.9$

Как сократить множество шинглов?

Вычеркиваем лишние строки

Этот метод **не** работает

	doc1	doc2	doc3	doc4	doc5
sh1	1	1	1	1	0
sh2	1	1	0	1	1
sh3	1	1	1	1	1
sh4	1	0	1	0	1
sh5	0	1	0	1	1
sh6	1	0	1	1	0

Какие строки лишние? Как формализовать их выбор?
Сколько можем вычеркивать?

Minshingle

		doc1	doc2	doc3	doc4	doc5
1	sh1	1	0	1	0	0
2	sh2	0	1	1	0	0
3	sh3	0	0	0	0	1
4	sh4	0	0	0	1	0
5	sh5	1	0	1	0	0
6	sh6	0	0	0	0	1



Msh(doc1) = 1...

Msh(doc2) = 2...

Msh(doc3) = 1...

Msh(doc4) = 4...

Msh(doc5) = 3...

Minshingle

		doc1	doc2	doc3	doc4	doc5
4	sh1	1	0	1	0	0
6	sh2	0	1	1	0	0
2	sh3	0	0	0	0	1
1	sh4	0	0	0	1	0
5	sh5	1	0	1	0	0
3	sh6	0	0	0	0	1

Пересортируем

Msh(doc1) = 1 ?...

Msh(doc2) = 2 ?...

Msh(doc3) = 1 ?...

Msh(doc4) = 4 ?...

Msh(doc5) = 3 ?...

Другой порядок шинглов!!!

Minshingle

		doc1	doc2	doc3	doc4	doc5	
1	sh4	0	0	0	1	0	
2	sh3	0	0	0	0	1	
3	sh6	0	0	0	0	1	
4	sh1	1	0	1	0	0	
5	sh5	1	0	1	0	0	
6	sh2	0	1	1	0	0	

Msh(doc1) = 1 ?...

Msh(doc2) = 2 ?...

Msh(doc3) = 1 ?...

Msh(doc4) = 4 ?...

Msh(doc5) = 3 ?...

Другой порядок шинглов!!!

Minshingle

		doc1	doc2	doc3	doc4	doc5
1	sh4	0	0	0	1	0
2	sh3	0	0	0	0	1
3	sh6	0	0	0	0	1
4	sh1	1	0	1	0	0
5	sh5	1	0	1	0	0
6	sh2	0	1	1	0	0

Msh(doc1) = 1 ?...

Msh(doc2) = 2 ?...

Msh(doc3) = 1 ?...

Msh(doc4) = 4 1...

Msh(doc5) = 3 ?...

Другой порядок шинглов!!!

Minshingle

		doc1	doc2	doc3	doc4	doc5
1	sh4	0	0	0	1	0
2	sh3	0	0	0	0	1
3	sh6	0	0	0	0	1
4	sh1	1	0	1	0	0
5	sh5	1	0	1	0	0
6	sh2	0	1	1	0	0

Msh(doc1) = 1 ?...

Msh(doc2) = 2 ?...

Msh(doc3) = 1 ?...

Msh(doc4) = 4 1...

Msh(doc5) = 3 2...

Другой порядок шинглов!!!

Minshingle

		doc1	doc2	doc3	doc4	doc5
1	sh4	0	0	0	1	0
2	sh3	0	0	0	0	1
3	sh6	0	0	0	0	1
4	sh1	1	0	1	0	0
5	sh5	1	0	1	0	0
6	sh2	0	1	1	0	0

Msh(doc1) = 1 ?...

Msh(doc2) = 2 ?...

Msh(doc3) = 1 ?...

Msh(doc4) = 4 1...

Msh(doc5) = 3 2...

Другой порядок шинглов!!!

Minshingle

		doc1	doc2	doc3	doc4	doc5
1	sh4	0	0	0	1	0
2	sh3	0	0	0	0	1
3	sh6	0	0	0	0	1
4	sh1	1	0	1	0	0
5	sh5	1	0	1	0	0
6	sh2	0	1	1	0	0

Msh(doc1) = 1 4...

Msh(doc2) = 2 ?...

Msh(doc3) = 1 4...

Msh(doc4) = 4 1...

Msh(doc5) = 3 2...

Другой порядок шинглов!!!

Minshingle

		doc1	doc2	doc3	doc4	doc5
1	sh4	0	0	0	1	0
2	sh3	0	0	0	0	1
3	sh6	0	0	0	0	1
4	sh1	1	0	1	0	0
5	sh5	1	0	1	0	0
6	sh2	0	1	1	0	0

Msh(doc1) = 1 4...

Msh(doc2) = 2 ?...

Msh(doc3) = 1 4...

Msh(doc4) = 4 1...

Msh(doc5) = 3 2...

Другой порядок шинглов!!!

Minshingle

		doc1	doc2	doc3	doc4	doc5
1	sh4	0	0	0	1	0
2	sh3	0	0	0	0	1
3	sh6	0	0	0	0	1
4	sh1	1	0	1	0	0
5	sh5	1	0	1	0	0
6	sh2	0	1	1	0	0

Msh(doc1) = 1 4...

Msh(doc2) = 2 6...

Msh(doc3) = 1 4...

Msh(doc4) = 4 1...

Msh(doc5) = 3 2...

Другой порядок шинглов!!!

Minshingle

		doc1	doc2	doc3	doc4	doc5
1	sh6	0	0	0	0	1
2	sh4	0	0	0	1	0
3	sh3	0	0	0	0	1
4	sh5	1	0	1	0	0
5	sh2	0	1	1	0	0
6	sh1	1	0	1	0	0



Msh(doc1) = 1 4 ?

Msh(doc2) = 2 6 ?

Msh(doc3) = 1 4 ?

Msh(doc4) = 4 1 ?

Msh(doc5) = 3 2 ?

Ещё одна перестановка

Minshingle

		doc1	doc2	doc3	doc4	doc5
1	sh6	0	0	0	0	1
2	sh4	0	0	0	1	0
3	sh3	0	0	0	0	1
4	sh5	1	0	1	0	0
5	sh2	0	1	1	0	0
6	sh1	1	0	1	0	0

$$\text{Msh}(\text{doc1}) = 1 \ 4 \ 4$$

$$\text{Msh}(\text{doc2}) = 2 \ 6 \ 5$$

$$\text{Msh}(\text{doc3}) = 1 \ 4 \ 4$$

$$\text{Msh}(\text{doc4}) = 4 \ 1 \ 2$$

$$\text{Msh}(\text{doc5}) = 3 \ 2 \ 1$$

Ещё одна перестановка

Minshingle

Почему это работает?

	doc1	doc2
A	1	1
B	0	1
C	1	0
D	0	0

А будет давать одинаковую свёртку при любой перестановке.

Minshingle

Почему это работает?

	doc1	doc2
A	1	1
B	0	1
C	1	0
D	0	0

А будет давать одинаковую свёртку при любой перестановке.

В и С будут давать расхождение в свёртках.

Д напрямую в свёртке не участвует.

Minshingle

Почему это работает?

	doc1	doc2		
A	1	1	*	*
B	0	1		*
C	1	0		*
D	0	0		

А будет давать одинаковую свёртку при любой перестановке
В и С будут давать расхождение в свёртках

$$JC = A/(A+B+C+D)$$

$$\text{Sim(minshingle)} \sim JC$$

Minshingle

	doc1	doc2	doc3	doc4	doc5
sh1	1	0	1	0	0
sh2	0	1	1	0	0
sh3	0	0	0	0	1
sh4	0	0	0	1	0
sh5	1	0	1	0	0
sh6	0	0	0	0	1

Msh(doc1) = 1 4 4

Msh(doc2) = 2 6 5

Msh(doc3) = 1 4 4

Msh(doc4) = 4 1 2

Msh(doc5) = 3 2 1

Minshingle

	doc1	doc2	doc3	doc4	doc5
sh1	1	0	1	0	0
sh2	0	1	1	0	0
sh3	0	0	0	0	1
sh4	0	0	0	1	0
sh5	1	0	1	0	0
sh6	0	0	0	0	1

$$\text{Msh}(\text{doc1}) = 1 \ 4 \ 4$$

$$\text{Msh}(\text{doc2}) = 2 \ 6 \ 5$$

$$\text{Msh}(\text{doc3}) = 1 \ 4 \ 4$$

$$\text{Msh}(\text{doc4}) = 4 \ 1 \ 2$$

$$\text{Msh}(\text{doc5}) = 3 \ 2 \ 1$$

пара	JC	Sim
d1-d2		
d1-d3		
d1-d4		
d1-d5		
d2-d3		
d2-d4		
d2-d5		
d3-d4		
d3-d5		
d4-d5		

Minshingle

	doc1	doc2	doc3	doc4	doc5
sh1	1	0	1	0	0
sh2	0	1	1	0	0
sh3	0	0	0	0	1
sh4	0	0	0	1	0
sh5	1	0	1	0	0
sh6	0	0	0	0	1

$$\text{Msh}(\text{doc1}) = 1 \ 4 \ 4$$

$$\text{Msh}(\text{doc2}) = 2 \ 6 \ 5$$

$$\text{Msh}(\text{doc3}) = 1 \ 4 \ 4$$

$$\text{Msh}(\text{doc4}) = 4 \ 1 \ 2$$

$$\text{Msh}(\text{doc5}) = 3 \ 2 \ 1$$

пара	JC	Sim
d1-d2	0/3	
d1-d3	2/3	
d1-d4	0/3	
d1-d5	0/4	
d2-d3	1/3	
d2-d4	0/2	
d2-d5	0/3	
d3-d4	0/4	
d3-d5	0/5	
d4-d5	0/3	

Minshingle

	doc1	doc2	doc3	doc4	doc5
sh1	1	0	1	0	0
sh2	0	1	1	0	0
sh3	0	0	0	0	1
sh4	0	0	0	1	0
sh5	1	0	1	0	0
sh6	0	0	0	0	1

Msh(doc1) = 1 4 4

Msh(doc2) = 2 6 5

Msh(doc3) = 1 4 4

Msh(doc4) = 4 1 2

Msh(doc5) = 3 2 1

пара	JC	Sim
d1-d2	0/3	0/3
d1-d3	2/3	3/3
d1-d4	0/3	0/3
d1-d5	0/4	0/3
d2-d3	1/3	0/3
d2-d4	0/2	0/3
d2-d5	0/3	0/3
d3-d4	0/4	0/3
d3-d5	0/5	0/3
d4-d5	0/3	0/3

Minshingle

	doc1	doc2	doc3	doc4	doc5
sh1	1	0	1	0	0
sh2	0	1	1	0	0
sh3	0	0	0	0	1
sh4	0	0	0	1	0
sh5	1	0	1	0	0
sh6	0	0	0	0	1

Msh(doc1) = 1 4 4

Msh(doc2) = 2 6 5

Msh(doc3) = 1 4 4

Msh(doc4) = 4 1 2

Msh(doc5) = 3 2 1

пара	JC	Sim
d1-d2	0/3	0/3
d1-d3	2/3	3/3
d1-d4	0/3	0/3
d1-d5	0/4	0/3
d2-d3	1/3	0/3
d2-d4	0/2	0/3
d2-d5	0/3	0/3
d3-d4	0/4	0/3
d3-d5	0/5	0/3
d4-d5	0/3	0/3

Minshingle. Как улучшить?

Качественный скачок: не обязательно знать номер шингла в перестановке. Шинглы уникальны => достаточно знать значение того шингла, что был первым в конкретной свертке

Minshingle

		doc1	doc2	doc3	doc4	doc5
1	sh1	1	0	1	0	0
2	sh2	0	1	1	0	0
3	sh3	0	0	0	0	1
4	sh4	0	0	0	1	0
5	sh5	1	0	1	0	0
6	sh6	0	0	0	0	1



Msh(doc1) = 1... | sh1...

Msh(doc2) = 2... | sh2...

Msh(doc3) = 1... | sh1...

Msh(doc4) = 4... | sh4...

Msh(doc5) = 3... | sh3...

Minshingle

		doc1	doc2	doc3	doc4	doc5	
1	sh4	0	0	0	1	0	
2	sh3	0	0	0	0	1	
3	sh6	0	0	0	0	1	
4	sh1	1	0	1	0	0	
5	sh5	1	0	1	0	0	
6	sh2	0	1	1	0	0	

Msh(doc1) = 1 4... | sh1 sh1...

Msh(doc2) = 2 6... | sh2 sh2...

Msh(doc3) = 1 4... | sh1 sh1...

Msh(doc4) = 4 1... | sh4 sh4...

Msh(doc5) = 3 2... | sh3 sh3...

Другой порядок шинглов!!!

Minshingle

		doc1	doc2	doc3	doc4	doc5	
1	sh6	0	0	0	0	1	
2	sh4	0	0	0	1	0	
3	sh3	0	0	0	0	1	
4	sh5	1	0	1	0	0	
5	sh2	0	1	1	0	0	
6	sh1	1	0	1	0	0	

Msh(doc1) = 1 4 4 | sh1 sh1 sh5

Msh(doc2) = 2 6 5 | sh2 sh2 sh2

Msh(doc3) = 1 4 4 | sh1 sh1 sh5

Msh(doc4) = 4 1 2 | sh4 sh4 sh4

Msh(doc5) = 3 2 1 | sh3 sh3 sh6

Ещё одна перестановка

Minshingle. Как улучшить?

Нужно хранить каждую перестановку

Случайный доступ к множеству шинглов - на больших объемах это случайный доступ к диску

Minshingle. Как улучшить?

Хэш-функция позволяет задать отношение порядка
 $H(A) < H(B) \Rightarrow A < B$

Задаем N разных хэш-функций, чтобы получить
minshingle размерностью N

Hash-функции для отношения порядка

	doc1
sh1	1
sh2	0
sh3	0
sh4	0
sh5	1
sh6	0

Hash-функции для отношения порядка

	sh	doc1
жить	12	1
самолет	35	0
зеленый	109	0
море	235	0
вулкан	265	1
изобразить	873	0

Hash-функции для отношения порядка

	sh	doc1
жить	12	1
самолет	35	0
зеленый	109	0
море	235	0
вулкан	265	1
изобразить	873	0

$H1(x)$:

$$H1(12) = 14$$

$$H1(35) = 18$$

$$H1(109) = 27$$

$$H1(235) = 32$$

$$H1(265) = 40$$

$$H1(873) = 52$$

$$H1(12) < H1(265) \Rightarrow msh_1 = 12$$

Hash-функции для отношения порядка

	sh	doc1
жить	12	1
самолет	35	0
зеленый	109	0
море	235	0
вулкан	265	1
изобразить	873	0

H1(x):

$$H1(12) = 14$$

$$H1(35) = 18$$

$$H1(109) = 27$$

$$H1(235) = 32$$

$$H1(265) = 40$$

$$H1(873) = 52$$

minshingle = 12

Hash-функции для отношения порядка

	sh	doc1
жить	12	1
самолет	35	0
зеленый	109	0
море	235	0
вулкан	265	1
изобразить	873	0

H2(x):
H2(12) = 143
H2(35) = 1982
H2(109) = 0
H2(235) = -15
H2(265) = 215
H2(873) = 102

$H2(12) < H2(265) \Rightarrow msh_2 = 12$

`minshingle = 12`



Hash-функции для отношения порядка

	sh	doc1
жить	12	1
самолет	35	0
зеленый	109	0
море	235	0
вулкан	265	1
изобразить	873	0

H2(x):

$H2(12) = 143$

$H2(35) = 1982$

$H2(109) = 0$

$H2(235) = -15$

$H2(265) = 215$

$H2(873) = 102$

minshingle = 12 12

Hash-функции для отношения порядка

	sh	doc1
жить	12	1
самолет	35	0
зеленый	109	0
море	235	0
вулкан	265	1
изобразить	873	0

H3(x):
H3(12) = 546
H3(35) = 14
H3(109) = -35
H3(235) = -100
H3(265) = 12
H3(873) = -102

$H3(265) < H3(12) \Rightarrow msh_3 = 265$

minshingle = 12 12

Hash-функции для отношения порядка

	sh	doc1
жить	12	1
самолет	35	0
зеленый	109	0
море	235	0
вулкан	265	1
изобразить	873	0

H3(x):
H3(12) = 546
H3(35) = 14
H3(109) = -35
H3(235) = -100
H3(265) = 12
H3(873) = -102

minshingle = 12 12 265

Hash-функции для отношения порядка

	sh		doc1
жить	12	sh1	1
самолет	35	sh2	0
зеленый	109	sh3	0
море	235	sh4	0
вулкан	265	sh5	1
изобразить	873	sh6	0

minshingle = 12 12 265

Msh = sh1 sh1 sh5

Minshingle. Реализация

```
for i in {0..99}; do  
    array[i] = null
```

shingleList – шинглы документа
h_i – i-ая хэш-функция (перестановка)
array[] – minshingle

```
for shingle in shingleList;  
do  
    for i in {0..99}; do  
        if array[i] == null or h_i(shingle) < h_i(array[i]); then  
            array[i] = shingle  
    done  
done
```

Minshingle. Реализация

```
for i in {0..99}; do  
    array[i] = null
```

```
for shingle in shingleList;  
do  
    for i in {0..99}; do  
        if array[i] == null or h_i(shingle) < h_i(array[i]); then  
            array[i] = shingle  
    done  
done
```

Minshingle. Реализация

```
for i in {0..99}; do
    array[i] = null

for shingle in shingleList;
do
    for i in {0..99}; do
        if array[i] == null or h_i(shingle) < h_i(array[i]); then
            array[i] = shingle
    done
done
```

Поиск дубликатов не для документов

Поиск дубликатов не для документов

- Похожие статьи и товары (облако тегов)
- Рекомендации для пользователей (похожие интересы)
- Неожиданное использование: детекция линкоферм

План лекции:

1. Дубликаты
 1. Терминология
 2. Примеры
 3. Шинглирование
2. Поиск дубликатов
 1. Улучшения
 2. Minshingle
 3. Алгоритм Бродера

Алгоритм Бродера

Алгоритм, который позволяет не использовать попарное сравнение.

Историческая справка:

Андрей Бродер - вице-президент AltaVista (исследовательский департамент), позже - вице-президент Yahoo! (исследования и реклама), сейчас - “выдающийся учёный” на службе у Google.

Автор многих алгоритмов в области биологии, генетических алгоритмов, поиска и оптимизации. Напр., алгоритм генерации лабиринта

Алгоритм Бродера. Шаг 1. пары шингл-документ

$Msh(doc1) = 1\ 4\ 4 \rightarrow \{ <1_1, doc1>, <2_4, doc1>, <3_4, doc1> \}$

<position_value, docId>

Алгоритм Бродера. Шаг 1. пары шингл-документ

$Msh(doc1) = 1\ 4\ 4 \rightarrow \{ <1_1, doc1>, <2_4, doc1>, <3_4, doc1> \}$

$Msh(doc2) = 2\ 6\ 5 \rightarrow \{ <1_2, doc2>, <2_6, doc2>, <3_5, doc2> \}$

$Msh(doc3) = 1\ 4\ 4 \rightarrow \{ <1_1, doc3>, <2_4, doc3>, <3_4, doc3> \}$

$Msh(doc4) = 4\ 1\ 2 \rightarrow \{ <1_4, doc4>, <2_1, doc4>, <3_2, doc4> \}$

$Msh(doc5) = 3\ 2\ 1 \rightarrow \{ <1_3, doc5>, <2_2, doc5>, <3_1, doc5> \}$

Алгоритм Бродера. Шаг 2. Группируем шинглы

<1_1, doc1>, <1_1, doc3>

<1_2, doc2>

<1_3, doc5>

...

<2_4, doc1>, <2_4, doc3>

...

Алгоритм Бродера. Шаг 3. Merge

$\langle 1_1, \text{doc1} \rangle, \langle 1_1, \text{doc3} \rangle \rightarrow \text{doc1 doc3}$

$\langle 1_2, \text{doc2} \rangle$

$\langle 1_3, \text{doc5} \rangle$

...

$\langle 2_4, \text{doc1} \rangle, \langle 2_4, \text{doc3} \rangle \rightarrow \text{doc1 doc3}$

...

Алгоритм Бродера. Шаг 3. Merge

$\langle 1_1, \text{doc1} \rangle, \langle 1_1, \text{doc3} \rangle \rightarrow \text{doc1 doc3}$

$\langle 1_2, \text{doc2} \rangle$

$\langle 1_3, \text{doc5} \rangle$

...

$\langle 2_4, \text{doc1} \rangle, \langle 2_4, \text{doc3} \rangle \rightarrow \text{doc1 doc3}$

...

$\langle 1_5, \text{doc7} \rangle, \langle 1_5, \text{doc8} \rangle, \langle 1_5, \text{doc9} \rangle \rightarrow$

doc7 doc8

doc7 doc9

doc8 doc9

Алгоритм Бродера. Шаг 3. Sum

$doc_i doc_j N$, где N - количество общих позиций в миншингле

$doc1 doc3 3$

Всего позиций в миншингле - 3

$$R(doc1, doc3) = 3/3 = 1$$

План лекции:

1. Дубликаты
 1. Терминология
 2. Примеры
 3. Шинглирование
2. Поиск дубликатов
 1. Улучшения
 2. Minshingle
 3. Алгоритм Бродера