



ТЕХНОСФЕРА

Поиск дубликатов. Часть 1

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

Практика и домашка

Практика:

- середина занятия
- тема: "Шинглирование" + "Коэффициент подобия Жаккара"

Домашка:

- выдаётся сегодня(17.03.2018)
- дедлайн ~ через 2 недели(02.04.2018)
- тема: "Поиск дубликатов" + "Алгоритм Бродера"

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

Дубликаты



Капибара, или водосвинка



Капибара, или водосвинка

cyclowiki.org/wiki/Капибара



cyclowiki.org

Навигация
Помощь Циклопедии
Сообщить об ошибке
Форум
FAQ
Формат статей
Качественные статьи
Случайная статья
Новые страницы
Свежие правки

Инструменты
Ссылки сюда
Связанные правки
Спецстраницы
Версия для печати
Постоянная ссылка

Статья Обсуждение

Читайте Правка История

Капибара

Капиба́ра, или **водосви́нка** (лат. *Hydrochoerus hydrochaeris*) — полуводное травоядное млекопитающее из семейства **водосвинковых** (Hydrochoeridae), единственный представитель в семействе.

Капибара — самый крупный среди современных грызунов.

Содержание [убрать]

- Внешний вид
- Происхождение и разновидности
- Ареал
- Образ жизни
 - Окружающая среда
 - Сообщество
 - Размножение
 - Питание
 - Болезни
 - Содержание в неволе
 - Продолжительность жизни
- Охрана и статус вида
- Популярность
- Интересные факты
- Источники
- Литература
- Ссылки

Внешний вид



Капибара — 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** «капибара» — это «господин трав». В странах Южной и Центральной Америки это животное называют по-разному — корпинчо, калугита, калпинчо, пончо.

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

Участие
Сообщить об ошибке
Портал сообщества
Форум
Свежие правки
Новые страницы
Справка
Помощь
Инструменты
Ссылки сюда
Связанные правки
Спецстраницы
Версия для печати
Постоянная ссылка
Печать/ксерокопировать статью
Создать книгу
Скачать как PDF
Версия для печати
В других проектах
Викисказ
Викивиды
Викиновости
На других языках
• Слэш
• Deutsch
• English
• Suomi

Капиба́ра^[1], или **водосви́нка**^[2] (лат. *Hydrochoerus hydrochaeris*) — полуводное травоядное млекопитающее из семейства водосвинковых (Hydrochoeridae), единственный представитель в семействе. Капибара — самый крупный среди современных грызунов. На языке индейцев гуарани слово капибара означает «господин трав»^[3].

Содержание [убрать]

- Внешний вид
- Распространение
- Образ жизни и питание
- Социальная структура и размножение
- Капибара в истории
- Статус популяции
- Примечание
- Источники
- Ссылки

Внешний вид [править | править вики-текст]

Длина тела взрослой капибары достигает 1—1,35 м, высота в холке — 50—60 см. Самцы весят 34—63 кг, а самки — 36—65,5 кг (измерения произведены в венесуэльских лыносах)^[4]. Самки, как правило, крупнее самцов.

Телосложение тяжёлое. Внешне капибара напоминает гигантскую большеголовую морскую свинку. Голова крупная, массивная с широкой, тупой мордой. Верхняя губа толстая. Уши короткие, округлые. Ноздри широко расставлены. Глаза маленькие, расположены высоко на голове и отнесены несколько назад. Хвост рудиментарный. Конечности довольно короткие, передние — 4-палые (пальцы было шесть), задние — 3-палые. Пальцы соединены небольшими плавающими перепонками и снабжены короткими сильными когтями. Тело покрыто длинными (30—120 мм) и жёсткими волосами; подшёрсток отсутствует. Окрас верхней стороны тела от рыжеватого до сероватого, броушного, как правило, желтовато-бурый. Молодик окрашен светлее. У половозрелых самцов на верхней части морды расположен участок кожи с многочисленными крупными салынными железами. У самок имеется 6 пар броушных сосков.

Череп массивный, с широкими и сильными скуловыми дугами. Зубов 20. Щёчные зубы без корней, растут в течение всей жизни животного. Резцы широкие, имеют продольную бороздку на наружной поверхности^[4]. Малая и большая берцовые кости частично срастаются между собой. Ключицы нет. Хромосом в диплоидном наборе 66.

Вот как описывает капибара Джеральд Даррелл в «Трёх билетах до Звёнденер»:

«Этот гигантский грызун представляет собой жирного зверька с продолговатым телом, покрытым жёсткой лохматой шерстью густой коричневой расцветки. Передние лапы у капибары длиннее задних, массивный оступок не имеет хвоста, и поэтому у неё всегда такой вид, будто она вот-вот собирается сесть. У неё крупные лапы с широкими перепончатыми пальцами, а когти на передних лапах, короткие и тупые, удивительно напоминают миниатюрные копыта. Вид у неё

← www.ziganshin.ru/animals/k/Kapibara (Hydrochoerus hydrochaeris).html

[Домашнее животное / Капибара, или водосвинка (Hydrochoerus hydrochaeris) /

Алфавитный указатель

А Б В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Э Ю Я

Капибара, или водосвинка

исключать | удалять | с минимальным риском

EX EW CR EN VU NT LC

Капибара, или **водосвинка** (**Hydrochoerus hydrochaeris**) — полуводное травоядное млекопитающее из семейства водосвинковых (Hydrochoeridae), единственный представитель в семействе. Капибара — самый крупный среди современных грызунов. На языке индейцев гуарани слово капибара означает «господин трав».



Внешний вид

Длина тела взрослой капибары достигает 1-1.35 м, высота в холке - 50-60 см. Самцы весят 34-63 кг, а самки - 36-65.5 кг. Самки, как правило, крупнее самцов.

Телосложение тяжёлое. Внешне капибара напоминает гигантскую большеголовую морскую свинку. Голова крупная, массивная с широкой, тупой мордой. Верхняя губа толстая. Уши короткие, округлые. Ноздри широко расставлены. Глаза маленькие, расположены высоко на голове и отнесены несколько назад. Хвост рудиментарный. Конечности довольно короткие; передние - 4-палые (пальцы было шесть), задние - 3-палые. Пальцы соединены небольшими плавающими перепонками и снабжены короткими сильными когтями. Тело покрыто длинными (30-120мм) и жёсткими волосами; подшёрсток отсутствует. Окрас верхней стороны тела от рыжеватого-бурого до сероватого, броушного, как правило, желтовато-бурый. Молодик окрашен светлее. У половозрелых самцов на верхней части морды расположен участок кожи с многочисленными крупными салынными железами. У самок имеется 6 пар броушных сосков.

Череп массивный, с широкими и сильными скуловыми дугами. Зубов 20. Щёчные зубы без корней, растут в течение всей жизни животного. Резцы широкие, имеют продольную бороздку на наружной поверхности. Малая и большая берцовые кости частично срастаются между собой. Ключицы нет. Хромосом в диплоидном наборе 66.

Вот как описывает капибара Джеральд Даррелл в «Трёх билетах до Звёнденер»: Этот гигантский грызун представляет собой жирного зверька с продолговатым телом, покрытым жёсткой лохматой шерстью густой коричневой расцветки. Передние лапы у капибары длиннее задних, массивный оступок не имеет хвоста, и поэтому у неё всегда такой вид, будто она вот-вот собирается сесть. У неё крупные лапы с широкими перепончатыми пальцами, а когти на передних лапах, короткие и тупые, удивительно напоминают миниатюрные копыта. Вид у неё весьма аристократический: её плоская широкая голова и туловище, почти квадратная морда имеют благодушно-порочительственное выражение, придающее ей сходство с воздушными лысами. По земле капибара передвигается характерной шаркающей походкой или скачет вразвалку галопом, в воде же плавает и ныряет с поразительной лёгкостью и проворством. Капибара - флегматичный добродушный вегетарианец, лишённый ярких индивидуальных черт, присущих некоторым его сородичам, но этот недостаток возмещается у неё спокойным и дружелюбным нравом.

Распространение и среда обитания

Капибара встречается по берегам разнообразных водоёмов в тропических и умеренных частях Центральной и Южной Америки, восточное Анд - от Панамы до Уругвая и северо-востока Аргентины (до 38°17' ю. ш., провинция Буэнос-Айрес).

Семейство

Водосвинковые (Hydrochoeridae)

[править | править вики-текст]

Капибара



Научная классификация

Царство: Животные

Тип: хордовые

Класс: Млекопитающие

Отряд: Грызуны

Семейство: Водосвинковые

Род: Водосвинки

Вид: Капибара

Латинское название

Hydrochoerus hydrochaeris

Linnaeus, 1766

Синонимы на Викискладе

Синонимы на Викискладе

Охранный статус

исключать | удалять | с минимальным риском

EX EW CR EN VU NT LC

Вызывающие наименьшие опасения

IUCN 3.1 Least Concern: 10300-φ

Капибара, 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\(Hydrochoerus%20hydrochaeris\).html](http://www.ziganshin.ru/animals/k/Kapibara%20(Hydrochoerus%20hydrochaeris).html)
3. <http://cyclowiki.org/wiki/%D0%9A%D0%B0%D0%BF%D0%B8%D0%B1%D0%B0%D1%80%D0%B0>

Капибара, или водосвинка

cyclowiki.org/wiki/Капибара



cyclowiki.org

Навигация
Помощь
Циклопедия
Сообщить об ошибке
Форум
FAQ
Формат статей
Качественные статьи
Случайная статья
Новые страницы
Свежие правки

Инструменты
Ссылки сюда
Связанные правки
Спецстраницы
Версия для печати
Постоянная ссылка

Статья Обсуждение

Капибара

Капиба́ра, или **водосви́нка** (лат. *Hydrochoerus hydrochaeris*) — полуводное травоядное млекопитающее из семейства **водосвинковых** (Hydrochoeridae), единственный представитель в семействе.

Капибара — самый крупный среди современных грызунов.

Содержание [убрать]

- Внешний вид
- Происхождение и разновидности
- Ареал
- Образ жизни
 - Окружающая среда
 - Сообщество
 - Размножение
 - Питание
 - Болезни
 - Содержание в неволе
 - Продолжительность жизни
- Охрана и статус вида
- Популярность
- Интересные факты
- Источники
- Литература
- Ссылки

Внешний вид



Капибара — 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** «капибара» — это «господин трав». В странах Южной и Центральной Америки это животное называют по-разному — корпинчо, калугита, калпинчо, пончо.

Небольшие глаза находятся высоко на голове, несколько сзади. Рудиментарный хвост. Довольно короткие конечности. Толстая верхняя губа, 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 имеются перепонки.

www.ziganshin.ru/animals/k/Kapibara (Hydrochoerus hydrochaeris).html

[Домашние животные / Капибара, или водосвинка (Hydrochoerus hydrochaeris) /

Алфавитный указатель

А Б В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Э Ю Я

Капибара, или водосвинка

исключая утконосы с млекопитающих рылец
EX EW CR EN VU NT LC

Капибара, или **водосвинка** (**Hydrochoerus hydrochaeris**) — полуводное травоядное млекопитающее из семейства водосвинковых (Hydrochoeridae), единственный представитель в семействе. Капибара — самый крупный среди современных грызунов. На языке индейцев гуарани слово капибара означает «господин трав».



Внешний вид

Длина тела взрослой капибары достигает 1-1,35 м, высота в холке — 50-60 см. Самцы весят 34-63 кг, а самки — 36-65,5 кг. Самки, как правило, крупнее самцов.

Телосложение тяжёлое. Внешне капибара напоминает гигантскую большеголовую морскую свинку. Голова крупная, массивная с широкой, тупой мордой. Верхняя губа толстая. Уши короткие, округлые. Ноздри широко расставлены. Глаза маленькие, расположены высоко на голове и отнесены несколько назад. Хвост рудиментарный. Конечности довольно короткие; передние — 4-палые (палец было шесть), задние — 3-палые. Пальцы соединены небольшими плавательными перепонками и снабжены короткими сильными когтями. Тело покрыто длинными (30-120мм) и жесткими волосами; подшерсток отсутствует. Окрас верхней стороны тела от рычавато-бурого до сероватого, броушного, как правило, желтовато-бурый. Молодчик окрашен светлее. У половозрелых самцов на верхней части морды расположен участок кожи с многочисленными крупными сильными железами. У самок имеется 6 пар борошных сосков.

Череп массивный, с широкими и сильными скуловыми дугами. Зубов 20. Щечные зубы без корней, растут в течение всей жизни животного. Резцы широкие, имеют продольную бороздку на наружной поверхности. Малая и большая берцовые кости частично срастаются между собой. Ключицы нет. Хромосом в диплоидном наборе 66.

Вот как описывает капибару Джеральд Даррелл в «Трёх билетах до Звёнденер»: Этот гигантский грызун представляет собой жирного зверька с продолговатым телом, покрытым мягкой лохматой шерстью густой коричневой расцветки. Передние лапы у капибары длиннее задних, массивный оступ не имеет хвоста, и поэтому у неё всегда такой вид, будто она вот-вот собирается сесть. У неё крупные лапы с широкими перепончатыми пальцами, а когти на передних лапах, короткие и тупые, удивительно напоминают миниатюрные копыта. Вид у неё весьма аристократический: её плоская широкая голова и тулая, почти квадратная морда имеют благодушно-порочительственное выражение, придающее ей сходство с задумчивым лисом. По земле капибара передвигается характерной шаркающей походкой или скачет вразвалку галопом, в воде же плавает и ныряет с поразительной лёгкостью и проворством. Капибара — флегматичный добродушный вегетарианец, лишённый ярких индивидуальных черт, присущих некоторым его сородичам, но этот недостаток восполняется у неё спокойным и дружелюбным нравом.

Распространение и среда обитания

Капибара встречается по берегам разнообразных водоёмов в тропических и умеренных частях Центральной и Южной Америки, восточное Анд — от Панамы до Уругвая и северо-востока Аргентины (до 38°17' ю. ш., провинция Буэнос-Айрес).

Семейство

Водосвинковые (Hydrochoeridae)

Капиба́ра^[…], или **водосви́нка**^[…] (лат. *Hydrochoerus hydrochaeris*) — полуводное травоядное млекопитающее из семейства водосвинковых (Hydrochoeridae), единственный представитель в семействе. Капибара — самый крупный среди современных грызунов. На языке индейцев гуарани слово капибара означает «господин трав»^[…].

Содержание [убрать]

- Внешний вид
- Распространение
- Образ жизни и питание
- Социальная структура и размножение
- Капибара в истории
- Статус популяции
- Примечание
- Источники
- Ссылки

Внешний вид [править · править вики-текст]

Длина тела взрослой капибары достигает 1—1,35 м, высота в холке — 50—60 см. Самцы весят 34—63 кг, а самки — 36—65,5 кг (измерения произведены в венесуэльских лыносах)^[…]. Самки, как правило, крупнее самцов.

Телосложение тяжёлое. Внешне капибара напоминает гигантскую большеголовую морскую свинку. Голова крупная, массивная с широкой, тупой мордой. Верхняя губа толстая. Уши короткие, округлые. Ноздри широко расставлены. Глаза маленькие, расположены высоко на голове и отнесены несколько назад. Хвост рудиментарный. Конечности довольно короткие; передние — 4-палые (пальцы было шесть)^[…], задние — 3-палые. Пальцы соединены небольшими плавательными перепонками и снабжены короткими сильными когтями. Тело покрыто длинными (30—120 мм) и жесткими волосами; подшерсток отсутствует. Окрас верхней стороны тела от рычавато-бурого до сероватого, броушного, как правило, желтовато-бурый. Молодчик окрашен светлее. У половозрелых самцов на верхней части морды расположен участок кожи с многочисленными крупными сильными железами. У самок имеется 6 пар борошных сосков.

Череп массивный, с широкими и сильными скуловыми дугами. Зубов 20. Щечные зубы без корней, растут в течение всей жизни животного. Резцы широкие, имеют продольную бороздку на наружной поверхности^[…]. Малая и большая берцовые кости частично срастаются между собой. Ключицы нет. Хромосом в диплоидном наборе 66.

Вот как описывает капибару Джеральд Даррелл в «Трёх билетах до Звёнденер»:

«Этот гигантский грызун представляет собой жирного зверька с продолговатым телом, покрытым жёсткой лохматой шерстью густой коричневой расцветки. Передние лапы у капибары длиннее задних, массивный оступ не имеет хвоста, и поэтому у неё всегда такой вид, будто она вот-вот собирается сесть. У неё крупные лапы с широкими перепончатыми пальцами, а когти на передних лапах, короткие и тупые, удивительно напоминают миниатюрные копыта. Вид у неё

КПД=1/3

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. Плагиат

Совпадение 85-100% **полезного контента**

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

VS

http://medieval_weapons.academic.ru/41

Виды дубликатов. Плагиат

Нашли ошибку? Сообщите об этом

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

[Назад К содержанию](#) [Дальше](#)

[Разновидности турниров]

I. «Механический» реннен

(нем. Geschiftrennen)

Всадник одет в реннойг, под доспехом — толстая ватная куртка — вамс с рукавами-буфами на упругой подкладке, заменяющими наручи. Ноги за *набедренные щитки* (нем. Streiftartschen, рис. 621) или *дильже* (рис. 622) на ремнях, перекинутых или продернутых через седло. *Легкие реннен седла*. Лошадь покрыта кожаной попоной, голова защищена глухим налобником. В этом виде поединков было две разновидности. [405]

1. «Механический» реннен с тарчем

(нем. Geschiftartschenrennen)

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 при ударе проходил через отверстие в тарче и заклинивался снаружи металлической шайбой. Между тарчем и пружинным механизмом зажаты концентрические клинья таким образом, что они своим давлением на тарч прижимали клинья.

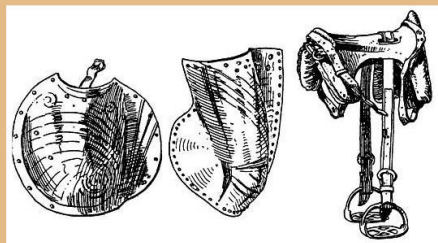
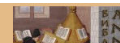


Рис. 621. Набедренный щиток, для защиты бедра от удара о барьер. Кон. XVI в.

Рис. 622. Дильже для правой ноги. Кон. XV в.

Рис. 623. Легкое седло для турнира реннен. Кон. XV в.



АКАДЕМИК
dic.academic.ru

Словари и энциклопедии на Академике

Введите текст для поиска по словарям и энциклопедиям

Найти!

Энциклопедия средн...

Толкования

Переводы

Книги

Энциклопедия средневекового оружия

Разновидности турниров это:

Толкование

Разновидности турниров

I. «Механический» реннен

(нем. Geschiftrennen)

Всадник одет в реннойг, 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. Streiftartschen, рис. 621) или дильже (рис. 622) на ремнях, перекинутых или продернутых через седло. *Легкие реннен седла* (ит. silla rasa) не имеют передних и задних луков (рис. 623). Лошадь покрыта кожаной попоной, голова защищена глухим налобником. В этом виде поединков было две разновидности.

1. «Механический» реннен с тарчем

(нем. Geschiftartschenrennen)

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 тарчем и пружинным механизмом зажаты концентрические клинья таким образом, что они своим давлением на тарч удерживали пружину механизма, который своим усилием прижимал клинья.



Коды ответов

200 – успех! - их качает спайдер

404 - страница не существует - нет контента для спайдера

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:

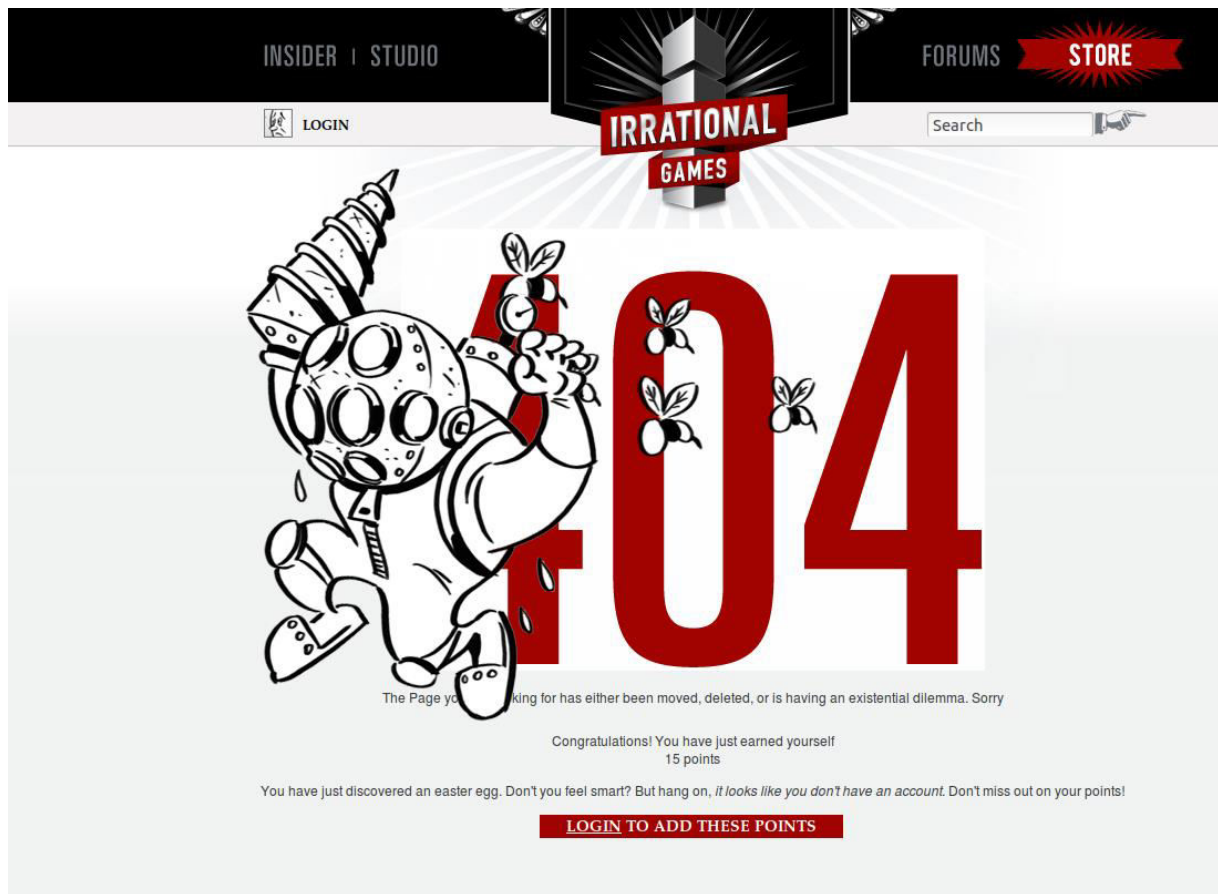
- [Главная страница](#)

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

Статья #34 не найдена!

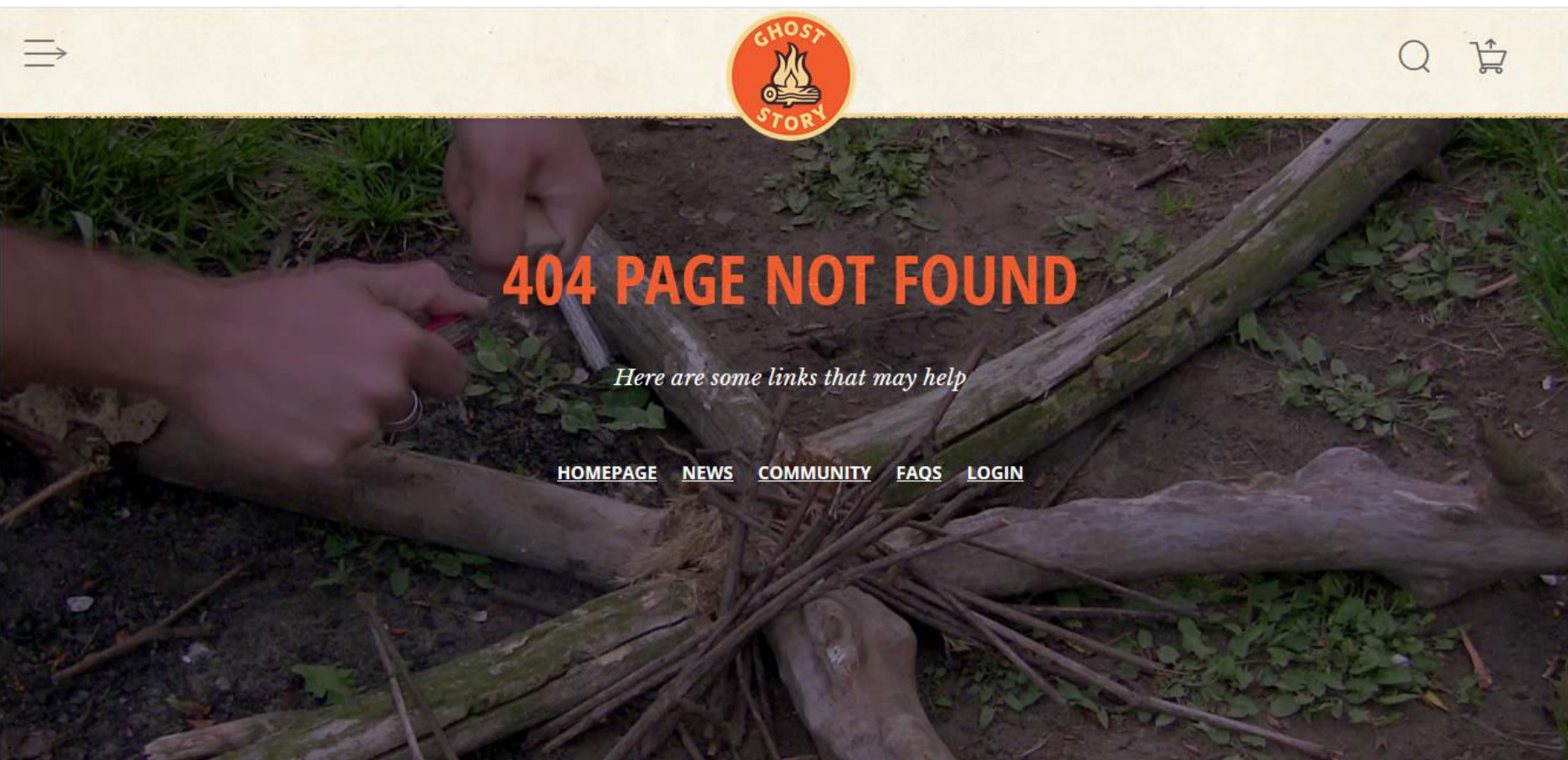
Страница не найдена. Примеры

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



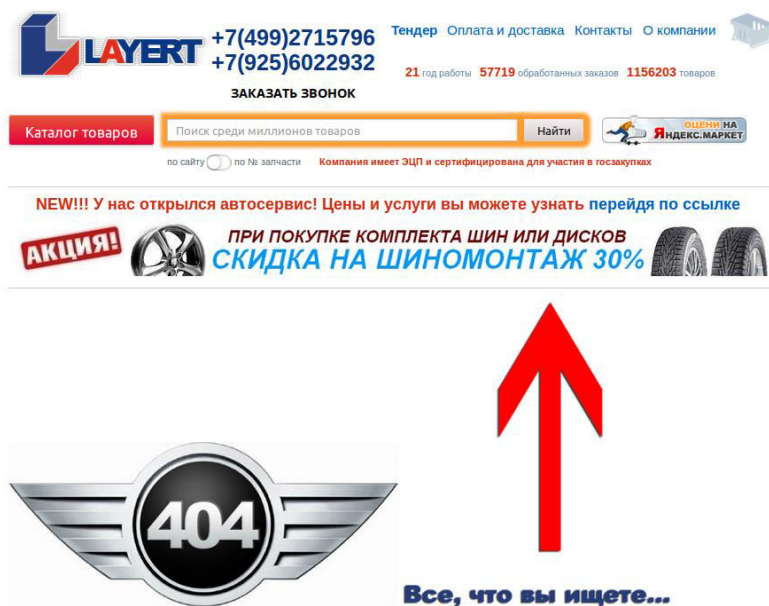
Страница не найдена. Примеры

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



Страница не найдена. Примеры

http://layert.ru/site/menu/zpch_vaz/dvig_vaz.php 200(!)



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.

1С-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.

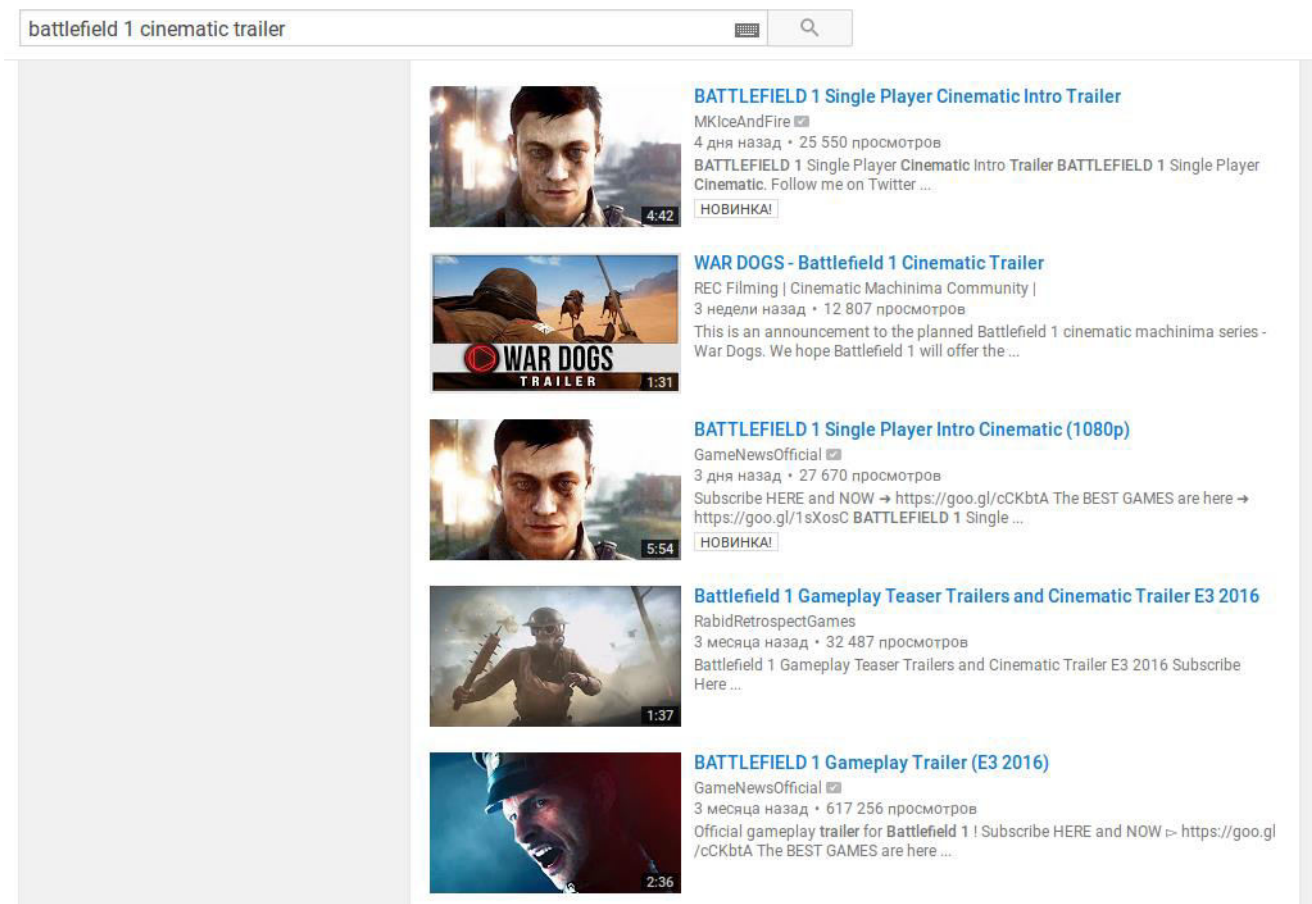
Компания «Кузбасс-пригород» просит пассажиров планировать свои поездки заранее с учетом изменений в расписании движения пригородных поездов.

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!

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 ...

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!

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 ...

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 ...

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

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"}

Поиск дубликатов. Шинглы.

Разбиение текста.

Последовательность шинглов

Мама_мыла_раму $N = 3$

{"Мам", "а_м"}

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 "внахлёст"

Мама мыла раму

$N = 10$



shingle 1

Поиск дубликатов. Шинглы.

Разбиение текста. Разбиение "внахлёст"

Мама мыла раму

$N = 10$



shingle2

Поиск дубликатов. Шинглы.

Разбиение текста. Разбиение "внахлёст"

Мама мыла раму

$N = 10$

М	а	м	а		м	ы	л	а		р	а	м	у	...
---	---	---	---	--	---	---	---	---	--	---	---	---	---	-----



shingle3

Поиск дубликатов. Шинглы.

Разбиение текста. Разбиение "внахлёст"

Что делать с конечными шинглами?

1. Оставить как есть. Т.о. первая буква будет только в 1 шингле
2. Зациклить текст. Тогда получаем

“Мама мыла раму” -> { ..., “ мыла раму”, “мыла рамуМ”, “ыла рамуМа”, “ла рамуМам”, ..., “уМама мыла”}

Шинглы. Сравнение документов

Построим матрицу:

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] $\rightarrow (26+26+1)^8$

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

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

Мера Жаккара. Пример

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

Мера Жаккара. Пример

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

Мера Жаккара. Пример

	d1	d1			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. Практика
3. Поиск дубликатов
 1. Улучшения
 2. Minshingle
 3. Алгоритм Бродера
4. Домашняя работа

Практика

1 - скачиваем архив с данными

<https://cloud.mail.ru/public/Ahcc/ReQ7k8NXM>

Внутри – 6 документов.

Задание: проверить, есть ли среди них дубликаты

Практика. План действий

2 - извлечь из html текст

Python: html2text, BeautifulSoup

console:

```
lynx --dump ./1.html > file.txt
```

OSX:

```
textutil -convert txt *.html
```

Часть для сильных духом. Остальные могут сразу взять .txt

Практика. План действий

3 - шинглирование

А) Размер шингла = 8 символов

Шаг = 1

В) Размер шингла = 3 слова

Шаг = 1

Практика. План действий

4 - поиск дубликатов

Практика

15 минут



Перерыв

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

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

$\text{Hash}(\text{"Мама мыла"}) = 172367463$

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

1. Одинаковая размерность для всех документов
2. Новое множество достаточно мало, чтобы оперировать им в памяти
3. Новое множество достаточно велико, чтобы не потерять основную часть информации о подобии документов

$$P(sim(doc_1, doc_2) | sim(Sig(doc_1), Sig(doc_2))) > 0.9$$

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

Msh(doc1) = 1 ?...

Msh(doc2) = 2 ?...

Msh(doc3) = 1 ?...

Msh(doc4) = 4 ?...

Msh(doc5) = 3 ?...

Другой порядок шинглов!!!

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

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
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

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
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

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
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

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
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

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
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

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
6	sh1	1	0	1	0	0
5	sh2	0	1	1	0	0
3	sh3	0	0	0	0	1
2	sh4	0	0	0	1	0
4	sh5	1	0	1	0	0
1	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

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

Minshingle

Почему это работает?

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

A будет давать одинаковую свёртку при любой перестановке

B и C будут давать расхождение в свёртках

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

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

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

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		
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

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	
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. Как улучшить?

p1	p2	p3		doc 1	doc 2	doc 3	doc 4	doc 5
1	4	6	sh1	1	0	1	0	0
2	6	5	sh2	0	1	1	0	0
3	2	3	sh3	0	0	0	0	1
4	1	2	sh4	0	0	0	1	0
5	5	4	sh5	1	0	1	0	0
6	3	1	sh6	0	0	0	0	1

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

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

$H1(x)$:

$$H1(12) = 14$$

$$H1(35) = 18$$

$$H1(109) = 27$$

$$H1(235) = 32$$

$$H1(265) = 40$$

$$H1(873) = 52$$

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

$H_2(x)$:

$$H_2(12) = 143$$

$$H_2(35) = 1982$$

$$H_2(109) = 0$$

$$H_2(235) = -15$$

$$H_2(265) = 215$$

$$H_2(873) = 102$$

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

$H_3(x)$:

$$H_3(12) = 546$$

$$H_3(35) = 14$$

$$H_3(109) = -35$$

$$H_3(235) = -100$$

$$H_3(265) = 12$$

$$H_3(873) = -102$$

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

$H_3(x)$:

$H_3(12) = 546$

$H_3(35) = 14$

$H_3(109) = -35$

$H_3(235) = -100$

$H_3(265) = 12$

$H_3(873) = -102$

minshingle = 12 12 265

Hash-функции для отношения порядка

	sh		doc 1
жить	12	sh1	1
самолет	35	sh2	0
зеленый	109	sh3	0
море	235	sh4	0
вулкан	265	sh5	1
изобразит ь	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 hi(shingle) < hi(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  $h_i(\text{shingle}) < h_i(\text{array}[i])$ ; then
            array[i] = shingle
        done
    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. Практика
3. Поиск дубликатов
 1. Улучшения
 2. Minshingle
 3. Алгоритм Бродера
4. Домашняя работа

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

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

$\langle position_value, docId \rangle$

Алгоритм Бродера. Шаг 1. пары шингл-документ

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

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

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

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

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

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

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

<1_2, doc2>

<1_3, doc5>

...

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

...

<1_5, doc7>, <1_5, doc8>, <1_5, doc9> ->

doc7 doc8

doc7 doc9

doc8 doc9

Алгоритм Бродера. Шаг 3. Sum

$\text{doc}_i \text{ doc}_j N$, где N - количество общих позиций в миншингле

doc1 doc3 3

Всего позиций в миншингле - 3

$$R(\text{doc1}, \text{doc3}) = 3/3 = 1$$

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

Домашняя работа

To: hw3duplicates@mail.ru

Subj: [HW3] BD-21, Иванов Иван

номер группы

ваши фамилия и имя

Срок сдачи - до 3 ноября (включительно)

<https://cloud.mail.ru/public/FmmS/anz96rDas> - данные для тренировки

чтение — zcat ... | src/docreader.py

Тестироваться будет на этой выборке + "секретная часть"

Домашняя работа

Environment:

Ubuntu 14.04 (x86_64)

Python 2.7

numpy (1.11.2)

google protobuf (3.1.0)

Restrictions:

20 min

4GB

Домашняя работа

Выполнение:

В архивах с данными - протобуфы (поля: url, body, text)

Работаем с .text - строим шинглы пословно (размер шингла – 5 слов), мощность миншингла - 20, коэффициент подоби́я 0.75 (для Бродера и Жаккара)

Выходные данные:

url1 url2 0.9

Только те, у кого коэффициент ≥ 0.75 !!!

Порядок урлов в паре и порядок пар не важен

Домашняя работа

К письму нужно приложить архив (только
`\.tar\.gz | \.tgz | \.tar | \.tar.bz2`)

В архиве:

- `./preinstall.sh` - если надо что-то доустанавливать; считаем, что есть `root` и `sudo` без пароля. Но лучше без него
- `./run.sh` - должен запускать код; выхлоп - в стандартный `output`.
Примерно так: `python ./broader_shingles.py ${@}`
- Исходный код :)

Спасибо за внимание

Вопросы?