



ТЕХНОСФЕРА

Поиск дубликатов. Часть 2

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

Что мы узнали в прошлый раз?

Что мы узнали в прошлый раз?

- Что такое дубликаты
- Сигнатуры и метрики для сравнения документов
- Шинглы и мера Жаккара
- Алгоритм Бродера

План занятия

1. Поиск дубликатов в больших коллекциях
2. Подготовка текста
3. Использование знаний о дубликатах

План занятия

1. Поиск дубликатов в больших коллекциях
 1. Обоснование
 2. Шинглы - LSH
 3. Алгоритмы с неделимой сигнатурой
 4. Сравнение алгоритмов

Почему нельзя пользоваться стандартными алгоритмами?

Пример: есть 80кМ документов

Почему нельзя пользоваться стандартными алгоритмами?

Пример: есть 80кМ документов

1. Попарное сравнение для меры Жаккара: $\sum_1^{N-1} = \frac{(N-1)(N-2)}{2}$,
т.е. порядка $32 \cdot 10^{20}$ сравнений

Почему нельзя пользоваться стандартными алгоритмами?

Пример: есть 80кМ документов

1. Попарное сравнение для меры Жаккара: $\sum_1^{N-1} = \frac{(N-1)(N-2)}{2}$,
т.е. порядка $32 \cdot 10^{20}$ сравнений
2. Алгоритм Бродера - если 1 элемент сигнатуры есть хотя бы у половины документов, то получаем $8 \cdot 10^{20}$ сравнений

Т.е. если 1 сравнение занимает хотя бы 1 мкс, то получаем 25М лет

Local Sensitive Hashing

Алгоритм поиска дубликатов в больших коллекциях с использованием шинглирования

Очевидно, что мы не должны перебирать все документы

Цель - сгруппировать потенциальные дубликаты. Это уменьшит количество сравнений

Local Sensitive Hashing

У нас есть:

Документы -> шинглы -> сигнатуры

Сигнатуры? Вспоминаем прошлую лекцию

Перестановки позволяют строить свертки векторов. В результате при сравнении получаем почти меру Жаккара

У каждого документа сигнатура - вектор из N чисел

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

LSH. Что сравниваем?

1. Для построения сигнатуры использовали N перестановок \Rightarrow размер сигнатуры - N элементов
2. Делим сигнатуру на b групп по r элементов в каждой ($N = b * r$)
3. Нумеруем группы (не забываем: смысл элементов с одним значением на разных позициях - разный)
4. В случае совпадения хотя бы одной группы считаем, что пара документов является **кандидатами в дубликаты**

LSH. Что сравниваем?



LSH. Что сравниваем?

S_1/1	S_1/2	S_1/3	...	S_1/n
-------	-------	-------	-----	-------

signature_1

S_2/1	S_2/2	S_2/3	...	S_2/n
-------	-------	-------	-----	-------

signature_2

S_3/1	S_3/2	S_3/3	...	S_3/n
-------	-------	-------	-----	-------

signature_3

LSH. Что сравниваем?

S_1/1	S_1/2	S_1/3	...	S_1/n
-------	-------	-------	-----	-------

signature_1

S_2/1	S_2/2	S_2/3	...	S_2/n
-------	-------	-------	-----	-------

signature_2

S_3/1	S_3/2	S_3/3	...	S_3/n
-------	-------	-------	-----	-------

signature_3

LSH. Что сравниваем?

S_1/1	S_1/2	S_1/3	...	S_1/n
-------	-------	-------	-----	-------

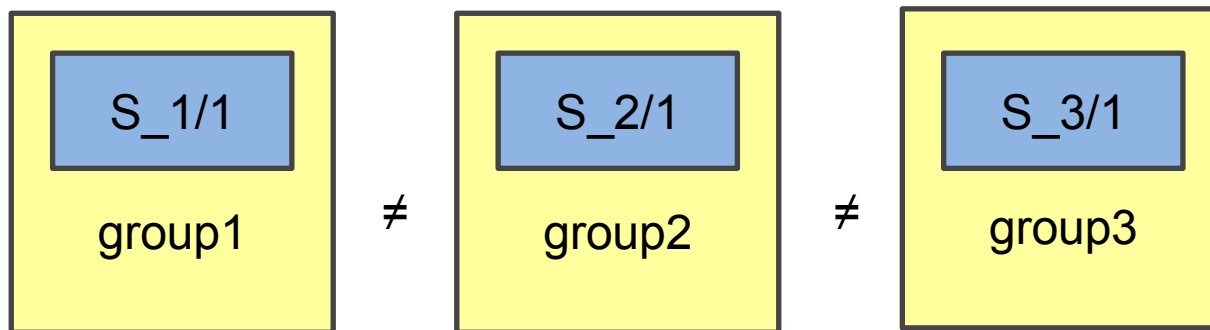
signature_1

S_2/1	S_2/2	S_2/3	...	S_2/n
-------	-------	-------	-----	-------

signature_2

S_3/1	S_3/2	S_3/3	...	S_3/n
-------	-------	-------	-----	-------

signature_3



LSH. Что сравниваем?

S_1/1	S_1/2	S_1/3	...	S_1/n
-------	-------	-------	-----	-------

signature_1

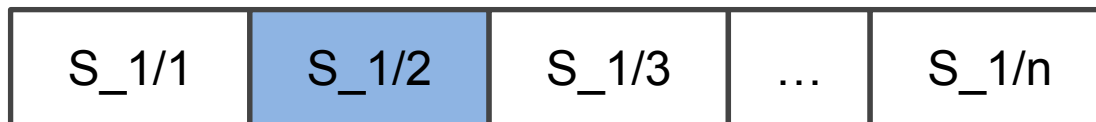
S_2/1	S_2/2	S_2/3	...	S_2/n
-------	-------	-------	-----	-------

signature_2

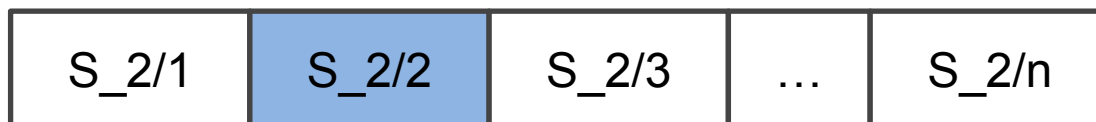
S_3/1	S_3/2	S_3/3	...	S_3/n
-------	-------	-------	-----	-------

signature_3

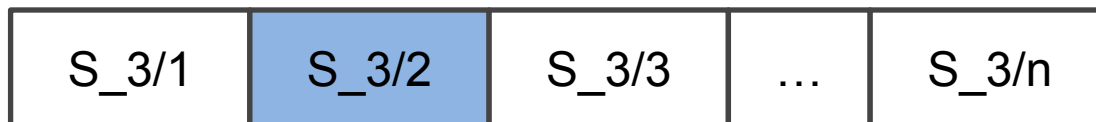
LSH. Что сравниваем?



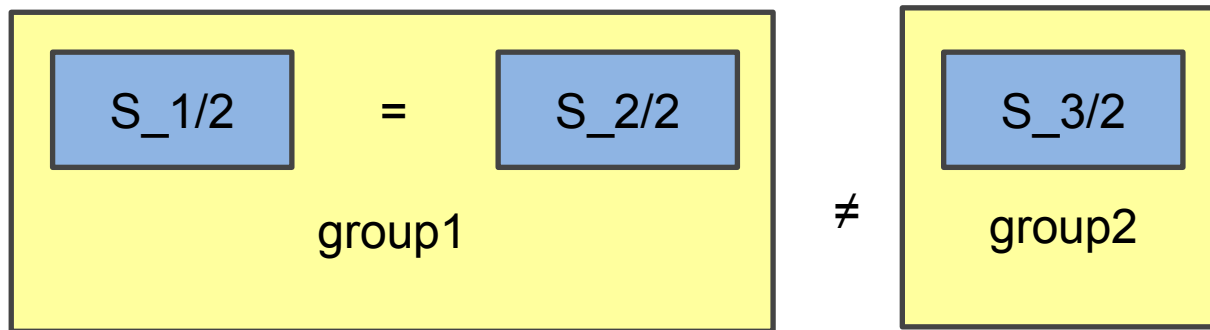
signature_1



signature_2



signature_3



LSH. Пример

Sign1 = 12 34 8654 2 3543 45654 234 298 12

Sign2 = 12 45 8654 34 123 39 234 298 12

LSH. Пример

Sign1 = [12 34 8654] [2 3543 45654] [234 298 12]

Sign2 = [12 45 8654] [34 123 39] [234 298 12]

3 группы по 3 шингла

LSH. Пример

Sign1 = [12 34 8654] [2 3543 45654] [234 298 12]

Sign2 = [12 45 8654] [34 123 39] [234 298 12]

Сравниваем:

1 из 3 групп совпадает

LSH. Пример

Sign1 = [12 34 8654] [2 3543 45654] [234 298 12]

Sign2 = [12 45 8654] [34 123 39] [234 298 12]

Улучшение:

Давайте захэшируем каждую группу, чтобы уменьшить объем тестовых данных

Sign1 = hash1 hash2 hash3

Sign2 = hash4 hash5 hash3

LSH. Пример

1. 100к документов
2. Размер сигнатуры - 100 элементов
3. Занимаемый объем - 305Мб (1 элемент - 32-битный int)
4. Порог подобия - 80%
5. Разбиваем сигнатуру на 20 групп по 5 элементов в каждой

Улучшение: можем сразу захэшировать каждую группу. Тогда сигнатура - не 100 int'ов, а 20 => 61Мб всего

LSH. Пример

Пусть сигнатуры $S1$ и $S2$ похожи на 80%
80% \Rightarrow из 100 элементов 20 различаются

LSH. Пример

Пусть сигнатуры $S1$ и $S2$ похожи на 80%

80% => из 100 элементов 20 различаются

Вероятность совпадения в отдельно взятой части

$$P = (0.8)^5 = 0.328$$

LSH. Пример

Пусть сигнатуры S1 и S2 похожи на 80%

80% => из 100 элементов 20 различаются

Вероятность совпадения в отдельно взятой части

$$P = (0.8)^5 = 0.328$$

Вероятность различия по всем корзинам (т.е. различные элементы равномерно распределены по 1 на группу)

$$P = (1-0.328)^{20} = 0.00035 \text{ (false-negative)}$$

LSH. Пример

Пусть сигнатуры $S1$ и $S2$ похожи на 90%
90% => из 100 элементов 10 различаются

LSH. Пример

Пусть сигнатуры $S1$ и $S2$ похожи на 90%

90% => из 100 элементов 10 различаются

Вероятность совпадения в отдельно взятой части

$$P = (0.9)^5 = 0.59049$$

LSH. Пример

Пусть сигнатуры $S1$ и $S2$ похожи на 90%

90% => из 100 элементов 10 различаются

Вероятность совпадения в отдельно взятой части

$$P = (0.9)^5 = 0.59049$$

Минимум 10 групп точно совпадут.

Вероятность того, что другие 10 будут различаться (ошибка равномерно распределится):

$$P = (1 - 0.59049)^{10} = 0.00133$$

LSH. Пример

Пусть сигнатуры $S1$ и $S2$ похожи на 40%
40% \Rightarrow из 100 элементов различия в 60

LSH. Пример

Пусть сигнатуры $S1$ и $S2$ похожи на 40%

40% => из 100 элементов различия в 60

Вероятность совпадения в отдельно взятой части

$$P = (0.4)^5 \sim 0.1$$

LSH. Пример

Пусть сигнатуры S1 и S2 похожи на 40%

40% => из 100 элементов различия в 60

Вероятность совпадения в отдельно взятой части

$$P = (0.4)^5 = 0.01024$$

Вероятность различия по всем корзинам

$$P = (1-0.01024)^{20} = 0.81395$$

LSH. Пример

Пусть сигнатуры S1 и S2 похожи на 40%

40% => из 100 элементов различия в 60

Вероятность совпадения в отдельно взятой части

$$P = (0.4)^5 = 0.01024$$

Вероятность различия по всем корзинам

$$P = (1-0.01024)^{20} = 0.81395$$

Могут совпасть не более 8 групп из 20.

$$P = ((0.4)^5)^8 = 1.21 \cdot 10^{(-16)}$$

LSH. В чем профит?

Реальный пример: 250к документов рунета
73к дубликатов, сгруппированных в 12к групп
Худший случай: группа из 8к документов

LSH. В чем профит?

Реальный пример: 250к документов рунета
73к дубликатов, сгруппированных в 12к групп
Худший случай: группа из 8к документов

Без использования LSH: $3 \cdot 10^{10}$ сравнений (-> 5333 минуты)

LSH. В чем профит?

Реальный пример: 250к документов рунета
73к дубликатов, сгруппированных в 12к групп
Худший случай: группа из 8к документов

Без использования LSH: $3 \cdot 10^{10}$ сравнений (-> 5333 минуты)

С использованием LSH: максимальная группа из 8к документов
+ ~7к false positive => 10^8 сравнений (-> 1.6 минуты)

Как еще уменьшить количество сравнений?

Как еще уменьшить количество сравнений?

1. Группировать/фильтровать документы по языку
2. Брать только документы с "текстовым" контентом (pdf, txt, html)
3. Кластеризация по длине документа
4. Кластеризация по контенту

Алгоритмы для поиска в больших множествах

Документ \leftrightarrow сигнатура

$\text{Sig1} == \text{Sig2} \Leftrightarrow D(\text{doc1}, \text{doc2})$

Алгоритмы для поиска в больших множествах

Документ \leftrightarrow сигнатура

$\text{Sig1} == \text{Sig2} \Leftrightarrow D(\text{doc1}, \text{doc2})$

1. **Local-based** - работают с каждым документом в отдельности, сравнивают сигнатуры
2. **Global-based** - в построении сигнатуры используется знание о коллекции или ее части

Алгоритмы для поиска в больших множествах

1. MD5
2. TF
3. $TF*IDF$
4. Long Sent
5. Heavy Sent
6. Lex Rand
7. Descr Words

MD5

Приведен исключительно для примера и оценки качества алгоритмов

Baseline - очень точный, но чувствительный к любому изменению текста

TF

TF – term frequency

Частота, с которой тот или иной терм встречается в документе

$tf(t, d) = \frac{n_i}{\sum_k n_k}$, где n_i – количество вхождений искомого терма, сумма в знаменателе - размер документа в термах

Отбираем L самых частотных термов

Сигнатура => CRC32 от их конкатенации

Улучшения:

удаление стоп-слов, лемматизация, кластеризация документов

TF*IDF

TF – term frequency

IDF – inversed document frequency

$$\text{idf}(t, D) = \log \frac{|D|}{|(d_i \supset t_i)|}$$

TW – term weight

$TW = TF * IDF$

Работа с сигнатурой - аналогично TF:

Отбираем L термов с максимальным показателем

Сигнатура => CRC32 от их конкатенации

Long Sent

Разбиваем документ на предложения.

Отбираем 2-3 самых длинных предложения

Сигнатура => CRC32 от конкатенации отобранных предложений

Offtop: как детектировать предложения?

Offtop: как детектить предложения?

1. Знаки препинания (аккуратно: инициалы и сокращения)
2. Теги (иногда нет последней точки, но фрагмент выделен тегами)
3. Отдельно работать с таблицами
4. И многое другое

Более подробно - лекция №11

Heavy Sent

Аналогично Long Sent, но предложения сортируются не по длине, а по весу.

Используем TW из алгоритма TF*IDF

Вес предложения (SW) - это сумма весов составляющих его термов

Сигнатура => CRC32 от топ-3 предложений

Lex Rand

1. Строим словарь по всем документам. Выкидываем максимальные и минимальные по весу (TW)
2. Сокращаем полученный словарь на 30%. 10 разных способов - 10 новых словарей
3. Пересекаем каждый новый словарь с терминами документа. Если пересечение достаточно большое - хэш от него становится сигнатурой документа. Т.о. у документа до 10 сигнатур
4. Считаем, что совпадение хотя бы одной сигнатуры у пары документов позволяет считать их дубликатами

Descr Words

В "ослабленном" варианте можно считать алгоритмом кластеризации

1. Из общего словаря выбираем 2к опорных слов. Покрытие: 35-65% документов
2. Для каждого опорного слова подбираем tf_i - пороговое значение для TF.
3. Fingerprint=[11100010...11], где 1 на i -ой позиции, если $TF_i \geq tf_i$
4. Выбрасываем из проверки документы с менее, чем 3 единицами в отпечатке (недостаточно информации)
5. Одинаковые отпечатки - документы в одном кластере (считаем дубликатами)

Как сравнивают алгоритмы?

Берётся тестовое множество. Из множества получаем "идеальный" ответ - то, с чем будем сравнивать.

Как сравнивают алгоритмы?

Берётся тестовое множество. Из множества получаем "идеальный" ответ - то, с чем будем сравнивать.

Что сравнивается:

1. Полнота - сколько урлов из правильного ответа было найдено

Как сравнивают алгоритмы?

Берётся тестовое множество. Из множества получаем "идеальный" ответ - то, с чем будем сравнивать.

Что сравнивается:

1. Полнота - сколько урлов из правильного ответа было найдено
2. Точность - сколько урлов из общего количества найденных - правильные

Как сравнивают алгоритмы?

Берётся тестовое множество. Из множества получаем "идеальный" ответ - то, с чем будем сравнивать.

Что сравнивается:

1. Полнота - сколько урлов из правильного ответа было найдено
2. Точность - сколько урлов из общего количества найденных - правильные
3. F-мера (мера ван Ризбергена) - считается по полноте (recall) и точности (precision)

$$F = 2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$$

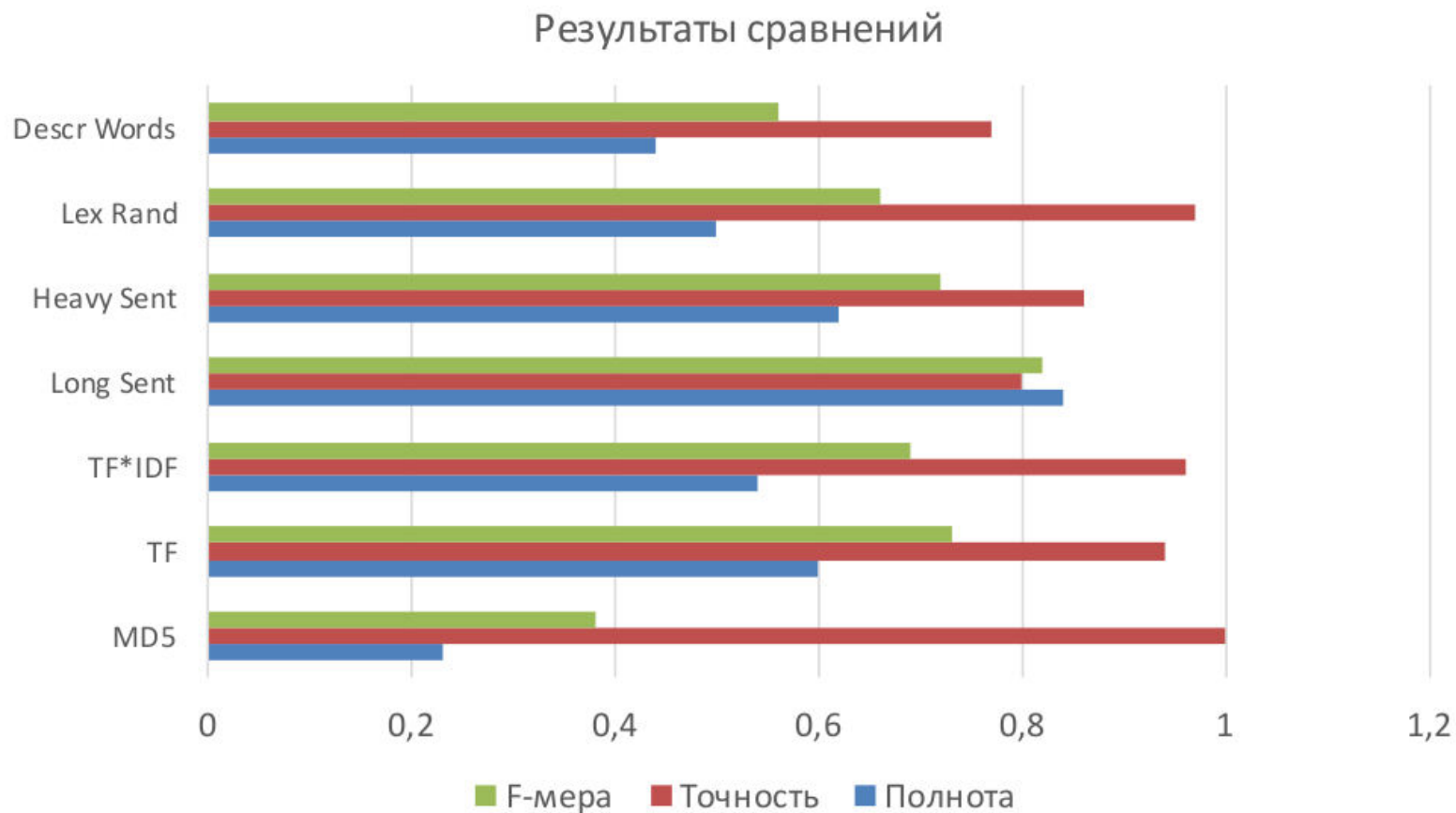
Сравнение алгоритмов

Взяли множество РОМИП 2007 (500к документов)

Сравнительный анализ методов определения нечетких дубликатов для Web-документов

© Зеленков Ю.Г, Сегалович И.В

Сравнение алгоритмов



Перерыв



План занятия

1. Поиск дубликатов в больших коллекциях
2. Подготовка текста
3. Использование знаний о дубликатах

План занятия

1. Поиск дубликатов в больших коллекциях
2. Подготовка текста
 1. Постановка проблемы
 2. Нормализация
 3. Удаление обвязки

Что не так с данными?

Что не так с данными?

1. Обязка
2. Пробелы
3. Разбиение на абзацы
4. Ошибки
5. Прочие странности

история пророка мухам... — ابراهيم بن شعبان
13 participants in the group

<https://ok.ru/group/57002394255396/topic/62669420576804>



Daxshat StatuSlar — prince Jaju

[join](#)[illegible]

Что делать?

1. Нормализация текста

Нормализация текста

1. Lower-case
2. Отсутствие пробелов
3. Отсутствие знаков препинания
4. Отсутствие псевдографики
5. Что делать с числами?

Нормализация текста

1. Lower-case
2. Отсутствие пробелов
3. Отсутствие знаков препинания
4. Отсутствие псевдографики
5. Что делать с числами?
 1. Можно оставлять "как есть" (проблема - автогенеренные время и дата, курсы валют и т.д.)
 2. Можно удалять (проблема - некоторые торговые названия пострадают + проблемы с таблицами с данными)

Что делать?

1. Нормализация текста
2. Удаление обвязки

Что такое обвязка?

АВТО@mail.ru Безопасность за рулем Секреты опытных водителей

Поиск по сайту

Новости Тест-драйвы Водителям Форум Каталог Отзывы Покупка и продажа

Новая Toyota Supra может стать гибридом

Несмотря на сотрудничество с BMW, японцы планируют оснастить спорткар собственной силовой установкой

4



Технология для неё будет позаимствована у гоночных автомобилей Toyota, выступающих в кольцевых гонках «24 часа Ле-Мана». Об этом со ссылкой на президента европейского подразделения Toyota Йохана ван Зила [сообщает](#) Auto Express.



ЛЮБИШЬ СВОЙ АВТО?
Скорее расскажи
нам о нём!

Что такое обвязка?



4

TOYOTA SUPRA

Toyota Supra с пробегом 2

Характеристики Toyota Supra
Отзывы о Toyota Supra 7

[Все новости](#)
[Все тест-драйвы](#)

Купи и продай автомобиль на cars.mail.ru

Лучшие предложения на кроссоверы и внедорожники

Самые надежные седаны за разумные деньги

Автомобили с передним приводом на автомате: Недорого

Рекомендуем

Honda CR-V – представлено новое поколение

Девушки автосалона в Париже: прекрасны все до единой!

Видео дня: агрессивные собаки против автомобиля

Синоттики рассказали, когда в Россию придет «день жестянц»...

«Скорая» против командира ГИБДД: разгорается новый скандал

Фото дня: водитель спас самое дорогое

Дилер Fiat Chrysler: актер Ельчин сам виноват в своей смерти

Госдума начала рассматривать законопроект об отмене...

☐ Подписаться на рассылку

Отзывы об авто

Toyota Chaser 10 декабря

Даже не знаю с чего начать. До этого у меня были советские автопромы))) даже вспоминать нет желания. Летом 2008 купил Тойоту Лезви 1.6L 180 бешеных японских кобыл. ДВС 4A-GE 5МКПП. Не машина а сказка...

★★★★★ 4.7

Toyota Prius 7 июля

Вторая машина в семье. Езжу 1 месяц. Все познается в сравнении - приходится оценивать после Хайлендера. Негатив. Зимой по нечищеным дорогам ездить плохо. Очень низкая подвеска для реальных российских...

★★★★☆ 4.2

Другие отзывы о Toyota Supra

Все отзывы 1

Расскажите о своем авто

Что такое обвязка?

1. Навигационные блоки
2. Рекомендационные блоки
3. Рекламные блоки
4. Header / footer

Т.е. та информация, что не несет непосредственной полезной семантической нагрузки

Ключевое у них всех - блоки

Удаление обвязки

Удаление обвязки

1. **Local-based** - работает только с текущим документом
2. **Global-based** - работает с группой документов, имеет доступ к дополнительной информации о коллекции

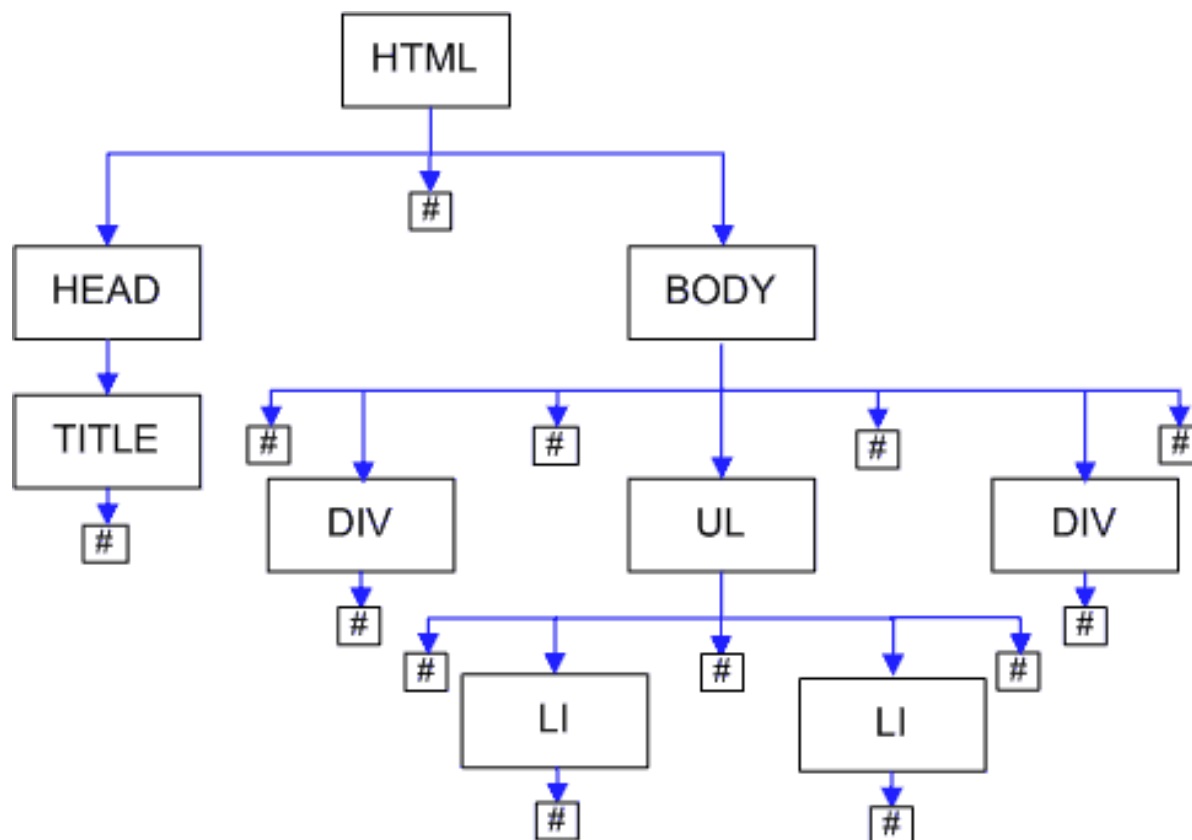
Local-based

Принцип:

имеем набор предположений о том, где обычно располагается полезный контент и как он выглядит (например, мы считаем, что это должен быть текст из более, чем 3 слов, это не должен быть список ссылок, и т.д.)

Local-based

Основная мысль: html – это дерево!!



Local-based. Детекция визуализации



Selenium:

отрисовываем страницу, определяем взаимное расположение блоков. Считаем, что полезный контент где-то относительно в центре.

Local-based. Детекция визуализацией



Плюсы:

- Можем учитывать js/css
- Подход с точки зрения итогового вида страницы

Минусы:

- Проблемы с таблицами
- Долго
- Не всякая вёрстка так очевидна:

<http://animanga.ru/manga.aspx?id=1422>



АНИМАНГА
Крупнейший каталог переводов манги

каталог » манга » дорохедоро

Дорохедоро

林田球
Dorohedoro

Номер в топе: 48
Голосов: 334
Добавили в список: 923

Мой статус: отсутствует
Мой рейтинг: ?
Обсудить (95)

Названия

Транскрипция: Dorohedoro
Оригинал: ドロヘドロ

Информация

Тип: манга
Авторы: ХАЯСИДА Кю:
Жанры: приключения, фэнтези, комедия, сэйнэн (найти похожие)
Год выхода: 2001
Томов: 19
Глав: —
Статус: выход продолжается
Комментарий:

Издания

стандартное, 2002 г., 19 т., издательство Сёгакукан

Все обложки

коллекционное, 2011 г., 1 т., издательство Сёгакукан

Человек с головой каймана, это странно? А если внутри этого человека живёт кто-то другой? Ещё более странно? Это мир магов и мир "дыры" — места, которое маги используют, как полигон для своих испытаний. Места, где небо пропитано магическими обрывками, а раз в год мертвецы вылезают из своих могил... Но, тем не менее, манга эта не мрачная, и очень даже интересная. А что ждёт нас дальше? Да всё то же... и ещё немного больше безумия. Ведь это Dorohedoro!

Переводы

показывать по релизам

KusoSekai

заморожен

Перевод: 9,56 (152)
Мой рейтинг: ?
Источники перевода:
Скачать с сайта

Local-based. Специальная вёрстка

1. Html5 - тег <article>
 1. <http://schema.org/Article>
 2. Используется на > 1М доменов
 3. Пример: <https://lenta.ru/articles/2016/03/16/deathstar/>
2. Доп.разметка (в основном для social graph)
 1. Пример: http://www.rbc.ru/technology_and_media/15/03/2016/56e6c7aa9a7947d4e6fb362d
 2. og:title, og:type, og:description

Local-based. Анализ контента

1. Content-based
 1. Словарь ключевых слов (которые удалять / оставлять)
 2. Анализ предложений
 3. Анализ размера и плотности текста
2. Context-based
 1. Какие теги в тексте
 2. Дерево тегов: родительские и дочерние блоки

Local-based. Анализ контента

Самый популярный: boilerpipe

Аналоги: Safari Reader, Evernote

Есть реализация под Python:

<https://github.com/ptwobrussell/python-boilerpipe/>

Но: jpyre + jni + chardet

Онлайн-реализация: <https://boilerpipe-web.appspot.com/>

коллекционное, 2011 г., 1 т., издательство Сёгакукан

Человек с головой каймана, это странно? А если внутри этого человека живёт кто-то другой? Ещё более странно? Это мир магов и мир "дыры" — места, которое маги используют, как полигон для своих испытаний. Места, где небо пропитано магическими отбросами, а раз в год мертвецы вылезают из своих могил... Но, тем не менее, манга эта не мрачная, и очень даже интересная. А что ждёт нас дальше? Да всё то же... и ещё немного больше безумия. Ведь это Dorohedoro!

Как работает boilerpipe?

1. Находим крупные фрагменты текста (опорные)
2. Постепенно добавляем соседей
 1. Считаем вес текста
 2. Считаем вес ссылок
 3. Если подходит - присоединяем, если нет - отсекаем всё дерево
3. Допущение, которое позволяет так делать: мы считаем, что внутри дерева обвязка и нормальный текст не перемешаны, а структурированы

Иногда всё не очень хорошо

Википедия: Страстоцвет

<https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D1%80%D0%B0%D1%81%D1%82%D0%BE%D1%86%D0%B2%D0%B5%D1%82>

ЭТИМОЛОГИЯ [править | править вики-текст]

Пассифлоры были среди первых цветов Нового Света, попавших в сады Европы. Первое известное описание пассифлоры дал в 1553 году Педро Сьеса де Леон, описав «гранадиллы», росшие в Колумбии. *Granadilla* в переводе с испанского означает «маленький гранат». В 1610 году изображение цветка пассифлоры попало в руки итальянского историка и религиозного деятеля Джакомо Босио. Босио начал собирать и другие описания и изображения цветка, привозимые мексиканскими иезуитами, и в том же году опубликовал доклад «Della Trionfante e Gloriosa Croce», где описал цветок пассифлоры как наглядное воплощение страданий Христа. Три рыльца пестика символизировали гвозди, которыми были прибиты к кресту ступни и руки Христа. Внешняя корона олицетворяла терновый венец, тычинки — пять ран. Семьдесят две венечные нити внутренней короны были приняты за количество шипов тернового венца. Копьевидные листья обозначили копье, пронзившее Христа. Желёзки, найденные на обратной стороне листа, должны были означать тридцать сребреников, полученных Иудой за предательство. Эти сравнения дали повод к названию растения *Passiflora*, от латинского «passio» — страдание и «flos» — цветок, то есть страстоцвет. Позже предпринимались и другие попытки найти религиозные символы в различных частях растения. Но были и люди, осуждавшие их как суеверие.

Тут в головах заметил я цветок;
~~загадочный — лиловый с золотистым,~~
Он странен был, но каждый лепесток
Проникнут был очарованием чистым.

В ту ночь, когда лилася кровь Христа
(В народе есть предание об этом) —
Впервые он расцвёл в тени креста
И потому зовётся страстоцветом,

Как будто бы в застенке палача,
На нём видны орудья мук Христовых:
Всё, от креста, верёвок и бича,
До молота — с венцом из игл терновых.



Passiflora caerulea



Boilerpipe. Немного практики

<https://cloud.mail.ru/public/B3xJ/SmYZBv5KX>

```
> java -jar boilerpipe.jar ./test.html
```

Что делаем:

1. документы $m^*.html$; $m^*.text$ - уже извлеченный текст
2. Определяем степень подобия документов ($m^*.text$)
3. Очищаем от обвязки ($m^*.html$)
4. Повторяем п.2

Как определить степень подобия? TF-IDF, Long Sent, Heavy Sent.
Можно использовать шинглы

Readability

<https://www.readability.com/>

Основное допущение: характер тегов соответствует характеру заключенного в них текста

Есть теги, более подходящие для полезного контента, есть - более подходящие для обвязки

Тегам присвоены веса

Readability

Тегам присвоены веса

Положительные (id | class):

`.*post.*, .*hentry.*, .*content.*, .*text.*, .*body.*`

Отрицательные:

`.*comment.*, .*meta.*, .*footer.*, .*cloud.*`

Readability

1. Берем все параграфы (`<p>`)
2. Смотрим на родительский тег:
 1. Относится к "текстовым" - увеличиваем суммарный вес
 2. Относится к "обвязке" - уменьшаем суммарный вес
3. Идём выше по дереву и повторяем п.2. Ключевой момент: информация снизу дерева переносится не вся, а с понижающим коэффициентом
4. По итогам: у каких поддеревьев достаточно большой вес - те и оставляем

Проблемы: те же, что и у boilerpipe

Global-based

1. Страницы на сайте никогда не существуют сами по себе
2. Допускаем, что страницы, объединенные в кластер, не только имеют схожий по некоторым свойствам контент, но и похожую структуру => обвязку



XPath

http://www.w3schools.com/xsl/xpath_intro.asp

Язык регулярных выражений для указания конкретного тега / группы тегов с использованием информации о его родительских тегах и их атрибутах

XPath. Пример

```
1 from lxml import html
2 import requests
3 page =
requests.get("https://ru.wikipedia.org/wiki/%D0%A0%D1%8B%D
1%81%D0%B8")
4 tree = html.fromstring(page.content)
5 info_block =
tree.xpath('//div[@class="NavContent"]/descendant::* /td /* /text
()')
6 for str in info_block: print str.encode('utf-8')
```


XPath. Пример

Подцарство

Эуметазои

Без ранга

Двусторонне-симметричные

Без ранга

Вторичноротые

Подтип

Позвоночные

Инфратип

Челюстноротые

Надкласс

Четвероногие

Подкласс

Звери

Инфракласс

Плацентарные

Надотряд

Подотряд

Кошкообразные

Подсемейство

Малые кошки

Global-based. Детекция человеком

Для кластера задаётся XPath (или несколько) для удаления
обязки/извлечения контента

Так работают рекомендации в <http://go.mail.ru/>

Плюсы: очень точное выделение контента

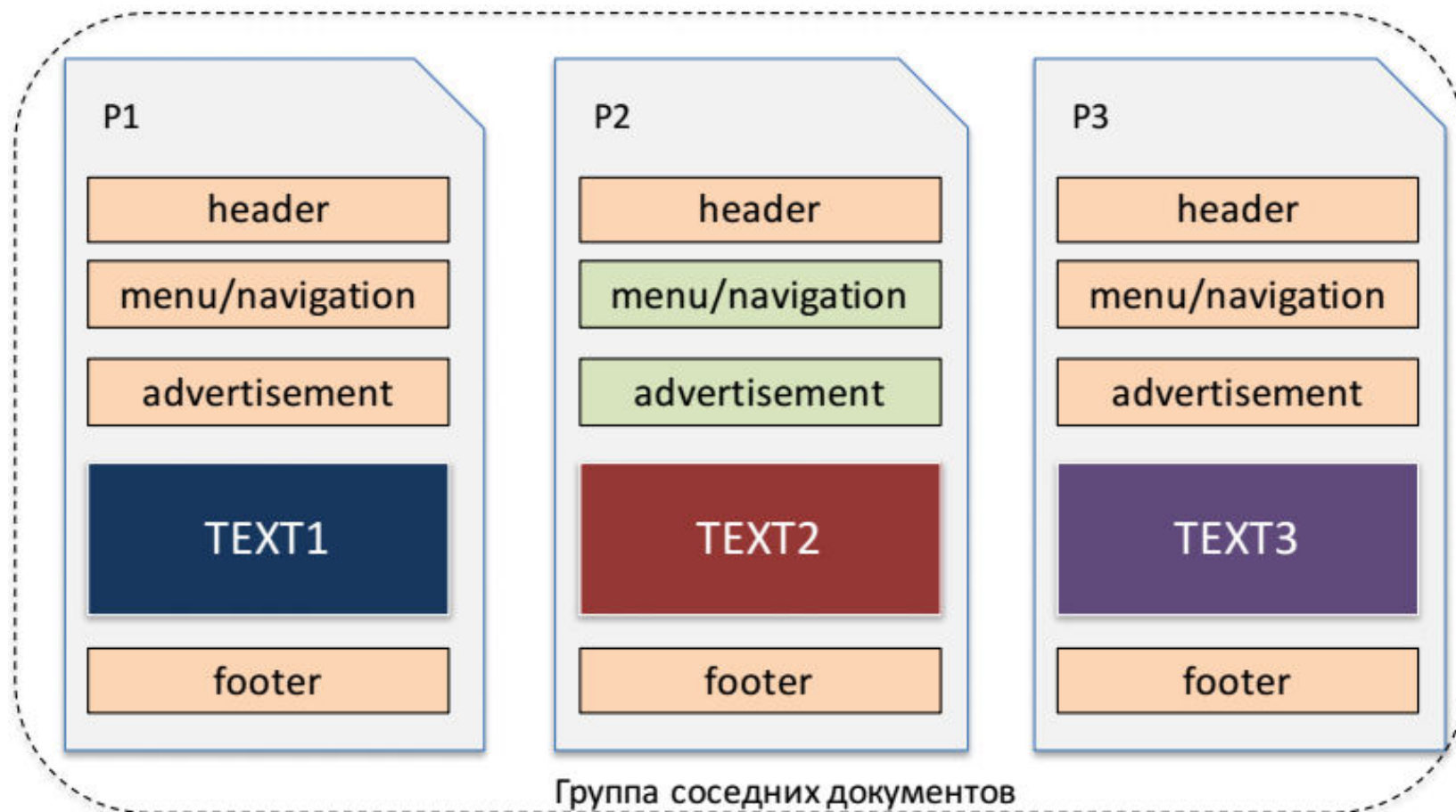
Минусы: система разваливается, если сайт меняет вёрстку

Global-based. Автоматизированное извлечение контента. Style-inbreeding

Берем группу документов. Несколько вариантов:

1. Пересекаем только DOM-деревья
2. Учитываем размер текста
3. Учитываем, какие именно теги попали в пересечение
 1. Тегам можно задать веса
4. Пересекаем не только теги, но и контент

Global-based. Infholder



Global-based. Infholder

1. Берём группу документов
 - По схожести до последнего сегмента, например:
<http://site.ru/a/b/c1.html> и <http://site.ru/a/b/c2.html>
 - Сад камней :)
2. Убираем все атрибуты
 - Оставляем теги и текст
3. DOM-дерево приводим к корректному xml-виду
4. Группируем соседние html-теги
 - Ориентируемся на длину текста
5. Считаем crc32 от полученных блоков
6. У каждого документа - вектор crc32, среди множества векторов находим НОП(LCS) - это обвязка

Infholder. Пример

<https://ru.wikipedia.org/wiki/%D0%90%D0%BA%D1%81%D0%BE%D0%BB%D0%BE%D1%82%D0%BB%D1%8C>

Аксолотль

Материал из Википедии — свободной энциклопедии

[\[править\]](#) [\[править вики-текст\]](#)

Текущая версия страницы пока не проверялась опытными участниками и может значительно отличаться от версии, проверенной 12 июня 2015; проверки требуют 22 правки.

Аксоло́тль (*Axolotl*) — неотеническая личинка некоторых видов амбистом, земноводных из семейства амбистомовых (*Ambystomidae*) отряда *хвостатых* (*Caudata*).

Особенность аксолотля состоит в том, что он достигает половой зрелости и становится способным к размножению, не превратившись во взрослую форму, не претерпев метаморфоз. У этих личинок хорошо развита щитовидная железа, но она обычно не вырабатывает достаточное количество индуцирующего метаморфозы гормона тироксина. Однако, если переселить аксолотля в более сухую и прохладную среду или понизить уровень воды при домашнем разведении, он превращается во взрослую амбистому. Превращение аксолотля в амбистому можно вызвать также добавлением в пищу или инъекцией гормона тироксина. Превращение может произойти в течение нескольких недель, при этом исчезнут наружные жабры аксолотля, изменится окраска, форма тела. Но вводить аксолотля в метаморфоз без поддержки специалиста опасно для жизни животного. Как правило, попытки в домашних условиях превратить аксолотля в амбистому в 99 % случаев заканчиваются смертью личинки.

Чаще всего название «аксолотль» применяют по отношению к личинке **мексиканской амбистомы** (большинство содержащихся в лабораторных или домашних условиях аксолотлей принадлежат к этому виду) или **тигровой амбистомы**, но так можно назвать личинку любой амбистомы, способной к неотении^[1].

В дословном переводе с **классического наuatля** аксолотль (*axolotl*) — «водяная собака (монстр)» (*atl* — вода, *xolotl* — собака, вместе — *axolotl*, то есть ашалотль в правильной транслитерации), что вполне соответствует его внешнему виду (аксолотль похож на крупного головастого тритона с торчащими в стороны тремя парами наружных жабр). Голова у аксолотля очень большая и широкая, несоразмерная с телом, рот тоже широкий, а глаза маленькие — создаётся впечатление, что личинка всё время улыбается. Помимо прочего, эти животные обладают способностью **регенерировать** утраченную часть тела. Общая длина — до 30 см. Как и все личинки хвостатых земноводных, аксолотли ведут хищный образ жизни.



Аксолотль **мексиканской амбистомы** (*Ambystoma mexicanum*)

Содержание [\[править\]](#) [\[править вики-текст\]](#)

Содержание аксолотлей мексиканской и тигровой амбистомы в домашних условиях может быть сопряжено с рядом проблем. В частности, это связано с трудностями поддержания нужного температурного режима в условиях квартиры, особенно летом. Для нормального самочувствия и стабильной работы иммунной системы

Infholder. Пример

Аксолотль

, проверено 12 июня 2015; проверка требует [13 правок](#)

Аксолотль [мексиканской амбистомы](#) (*Ambystoma mexicanum*)

Аксолотль (*Axolotl*) — [неестественная](#) личинка некоторых видов амбистом, [земноводных](#) из семейства [амбистомовых](#) (*Ambystomidae*) отряда [хвостатых](#) (*Caudata*).

Особенность аксолотля состоит в том, что он достигает половой зрелости и становится способным к размножению, не превратившись во взрослую форму, не претерпев [метаморфоз](#). У этих личинок хорошо развита [циклическая железа](#), но она обычно не вырабатывает достаточное количество индуцирующего метаморфозы [гормона тироксина](#). Однако, если переселить аксолотля в более сухую и прохладную среду или повысить уровень воды при домашнем разведении, он превращается во взрослую амбистому. Превращение аксолотля в амбистому можно вызвать также добавлением в воду или инъекцией [гормона тироксина](#). Превращение может произойти в течение нескольких недель, при этом исчезнут наружные [жабры](#) аксолотля, изменится окраска, форма тела. Но вводить аксолотля в метаморфоз без поддержки специалиста-герпетолога опасно для жизни животного. Как правило, попытки в домашних условиях превратить аксолотля в амбистому в 99 % случаев заканчиваются смертью личинки.

Чаще всего название «аксолотль» применяют по отношению к личинке [мексиканской амбистомы](#) (большинство содержащихся в лабораторных или домашних условиях аксолотлей принадлежит к этому виду) или [тигровой амбистомы](#), но так можно назвать личинку любой амбистомы, способной к нестению.

В дословном переводе с [класического нгуатла](#) аксолотль (*axolotl*) — «водяная собака (монстр)» (на яз. науатль, *atl* — вода, *xolotl* — собака, что вместо даёт *axolotl*, то есть асхалотль в правильной транслитерации), что вполне соответствует его внешнему виду (аксолотль похож на крупного головастого тритона с торчащими в стороны тремя парами наружных жабр). Голова у аксолотля очень большая и широкая, несопоставимая с телом, рот тоже широкий, а глазки маленькими — создаётся впечатление, что личинка всё время унывает. Помимо прочего, эти животные обладают способностью [регенерировать](#) утраченную часть тела. Общая длина — до 30 см. Как и все личинки хвостатых земноводных, аксолотли ведут хищный образ жизни.

Содержание аксолотлей мексиканской и тигровой амбистомы в домашних условиях может быть сопряжено с рядом проблем. В частности, это связано с трудностями поддержания нужного температурного режима в условиях квартиры, особенно летом. Для нормального самочувствия и стабильной работы иммунной системы аксолотлям требуется вода с температурой от 18 до 20 градусов по Цельсию. Содержание этих животных при более высоких температурах ведет к частым заболеваниям и смерти, быстрой, ил медленной, в зависимости от вида.

Кормление аксолотлей также имеет ряд своих особенностей: например, их нельзя кормить обычным кормом для рыб или мясом (субпродуктами) теплокровных животных. Для кормления подойдут свежемороженые морские коктейли, сырая морская рыба или треска, миксина черня, некоторые аквариумные рыбы (гуппи, неоны, расборы, даяно), креветки-черны.

Для содержания аксолотлей требуется чистая вода, свободная от хлора, нейтральная или чуть жесткая. Около 40 литров на одну особь.

Аксолотли некоторых амбистом

- Аксолотль [мексиканской амбистомы](#) (*Ambystoma mexicanum*) ([альбинос](#))
- Аксолотль [тигровой амбистомы](#) (*Ambystoma tigrinum*)
- Аксолотль [красноухой амбистомы](#) (*Ambystoma andersoni*)

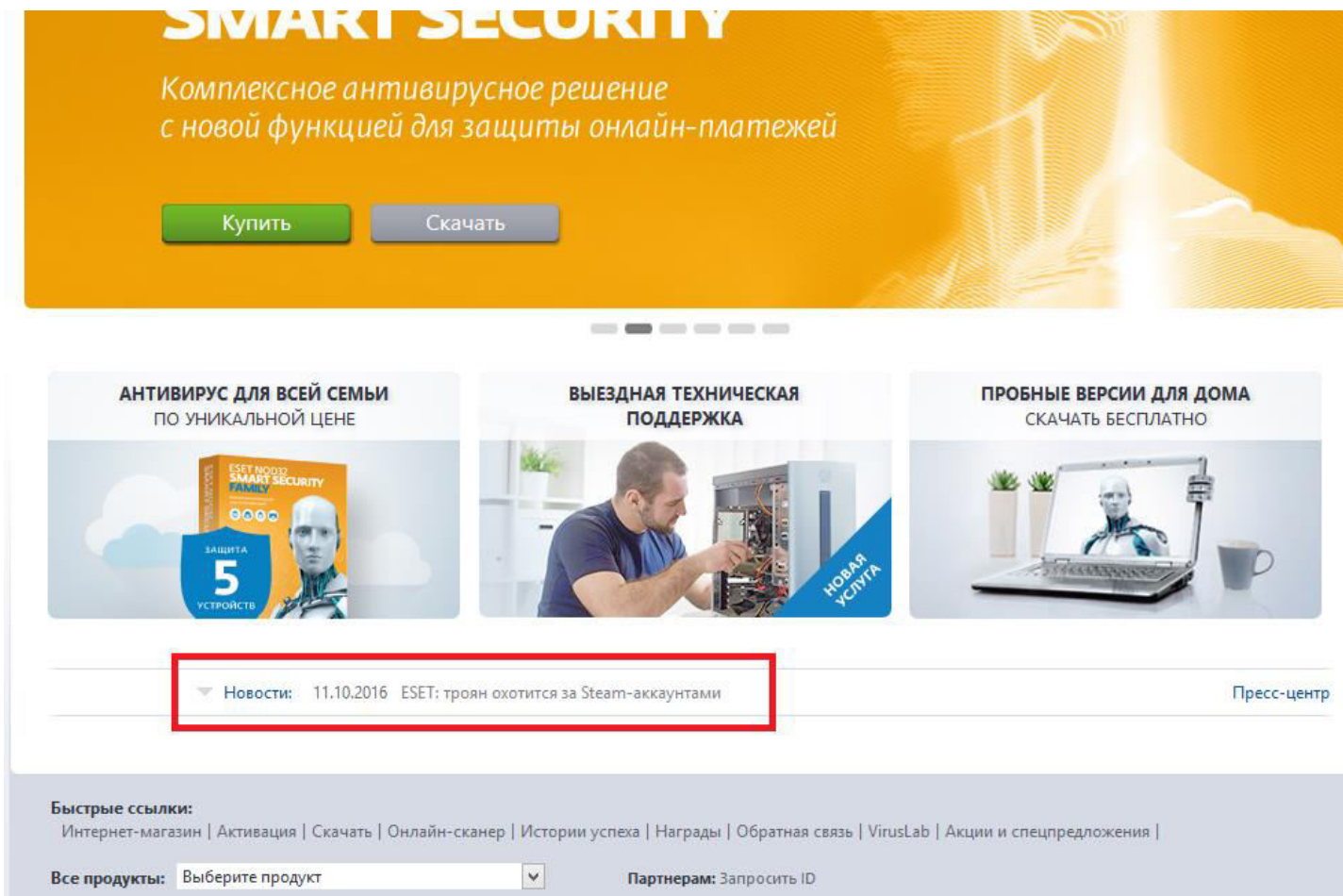
Ссылки[править | править вики-текст]

- [Принципы аксолотли](#) Денис Туликов «Популярная микология» № 1, 2015 [Архив](#)
- [Аксолотль](#)
- [Axolotl.org](#)
- [Сайт любителей аксолотлей](#)

1. [Т. Даревский Н. С., Орлов Н. Л.](#) Редкие и исчезающие животные. Земноводные и пресмыкающиеся / под ред. В. Е. Соколова. — М.: Высш. шк., 1988. — С. 76. — 190 000 экз. — ISBN 5-06-001429-0.

Infholder. Плохой пример

<https://www.esetnod32.ru/>



SMART SECURITY

Комплексное антивирусное решение
с новой функцией для защиты онлайн-платежей

Купить Скачать

АНТИВИРУС ДЛЯ ВСЕЙ СЕМЬИ
ПО УНИКАЛЬНОЙ ЦЕНЕ

ВЫЕЗДНАЯ ТЕХНИЧЕСКАЯ
ПОДДЕРЖКА

ПРОБНЫЕ ВЕРСИИ ДЛЯ ДОМА
СКАЧАТЬ БЕСПЛАТНО

Новости: 11.10.2016 ESET: троян охотится за Steam-аккаунтами

Пресс-центр

Быстрые ссылки:
Интернет-магазин | Активация | Скачать | Онлайн-сканер | Истории успеха | Награды | Обратная связь | VirusLab | Акции и спецпредложения |

Все продукты: Выберите продукт

Партнерам: Запросить ID

Очень сложный пример

<http://forum.ixbt.com/topic.cgi?id=66:11432>

Как удалить здесь обвязку?

План занятия

1. Поиск дубликатов в больших коллекциях
2. Подготовка текста
3. Использование знаний о дубликатах

Зачем искать дубликаты?

Вспомним прошлую лекцию

Капибара, или водосвинка

cyclowiki.org/wiki/Kapibara

Статья Обсуждение

Капибара

Капи́бара, 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, калуга, 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.

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

исключенные | исключенные | с незначительными рисками

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 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 броушных 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 Звёздочер»: Этот гигантский грызун представляет собой жирного зверька с продолговатым телом, покрытым жесткой лохматой шерстью густой коричневой расцветки. Передние лапы у капибары длиннее задних, массивный оступ не имеет хвоста, и поэтому у нее всегда такой вид, будто она вот-вот собирается сесть. У нее крупные лапы с широкими перепончатыми пальцами, а когти на передних лапах, короткие и тупые, удивительно напоминают миниатюрные копыта. Вид у нее весьма аристократический: ее 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 легкостью и проворством. Капибара — флегматичный добродушный вегетарианец, лишенный ярких индивидуальных черт, присущих некоторым его сородичам, но этот недостаток восполняется у нее 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).

КПД=1/3

Капи́бара^[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)^[5], 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. Молодичок окрашен светлее. У половозрелых самцов на верхней части морды расположен участок кожи с многочисленными крупными сильными железами. У самок имеется 6 пар броушных 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^[6]. 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 Звёздочер»:

«Этот гигантский грызун представляет собой жирного зверька с продолговатым телом, покрытым жесткой 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

Что делать с дубликатами в индексе?

1. Удалять из индекса
2. Оставлять в индексе, но в выдаче ранжировать сильно ниже основного результата

Зачем оставлять >1 документа?

1. Люди привыкли к определенным ресурсам
2. Контент перепечатывается - добавляется дополнительная информация (например, комментарии)
3. Контент не вечен - удаление страницы, недоступность сайта

Как определить "главный" документ?

Как определить "главный" документ?

1. Появился в сети раньше
2. Популярность ресурса
3. Дополнительные метрики: количество рекламы и объем обвязки на странице, всплывающие баннеры (чем меньше страница раздражает людей, тем она лучше)

Вместо резюме

Серебряной пули не существует

Выбор алгоритма зависит от наших возможностей и целей (в контексте этой лекции это касается и поиска дубликатов, и удаления обвязки)

Спасибо за внимание

Вопросы?