



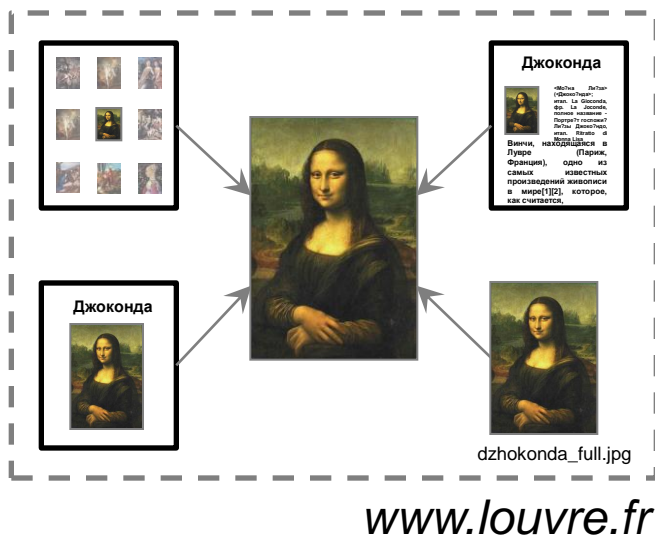
Кластеризация дубликатов в Яндекс.Картинках

Александр Крайнов
Менеджер проектов

Я.Субботник, Екатеринбург, 2 июля 2011 года

Картинки в интернете

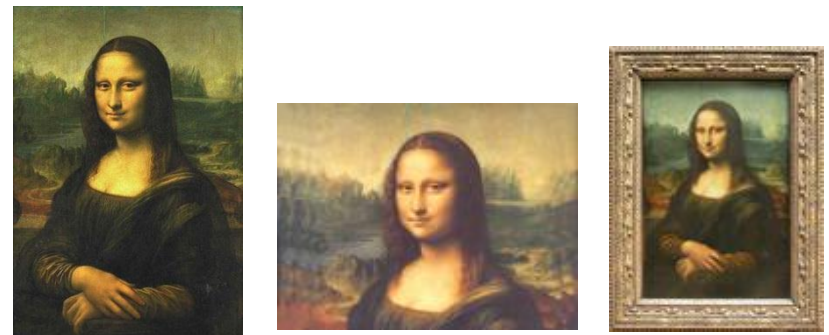
Хостовые клоны (дубликаты)



Тумбнейлерные дубликаты

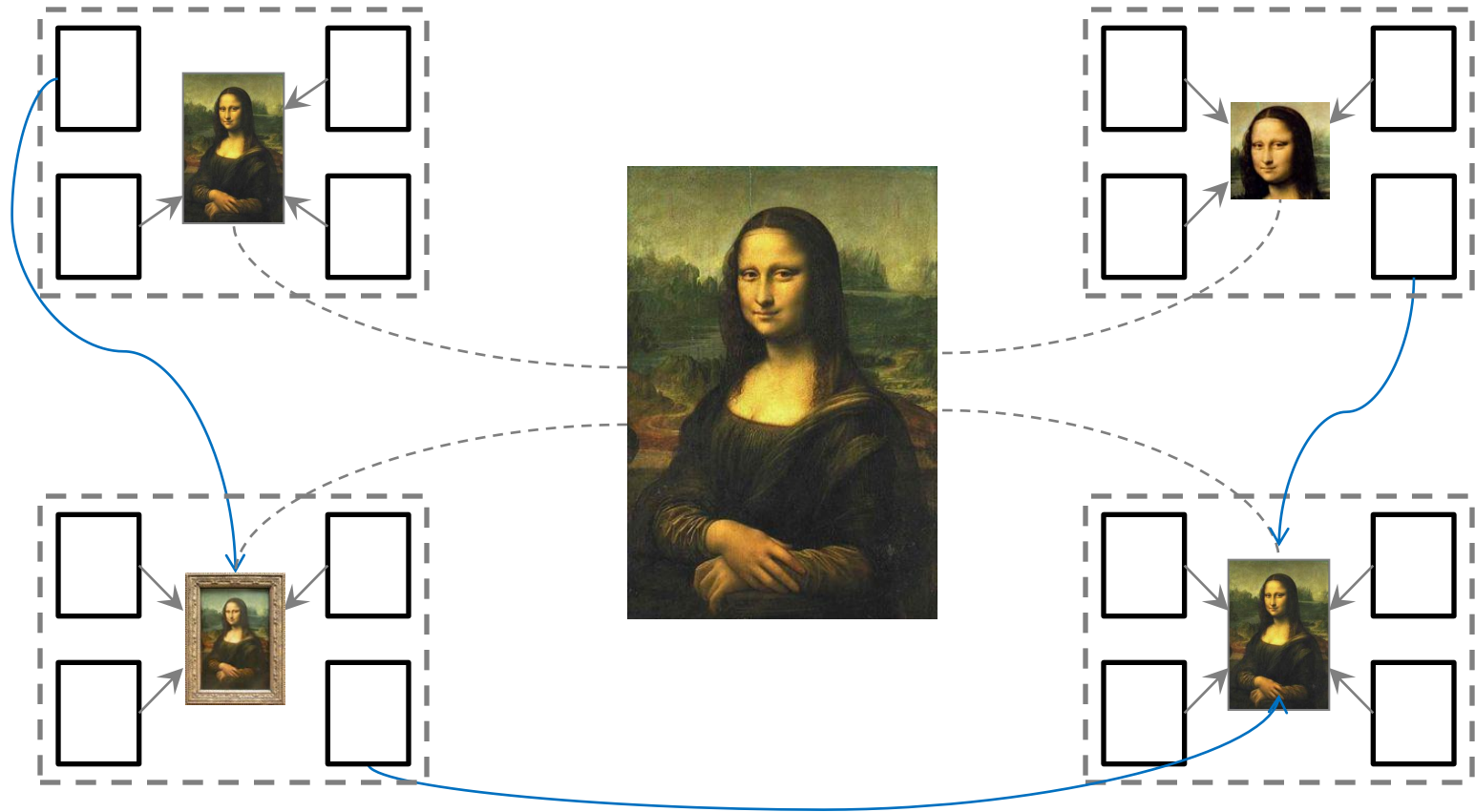


Полудубликаты



Картинки в интернете

Хостовые и межхостовые дубликаты



Тумбнейлерные полудубликаты

182 x 264



100 x 100



50 x 50



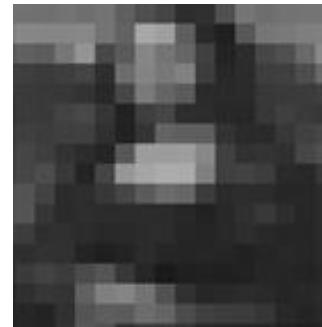
20 x 20



20 x 20, grayscale



16 x 16, grayscale



Нечеткие полудубликаты

Как их распознать?



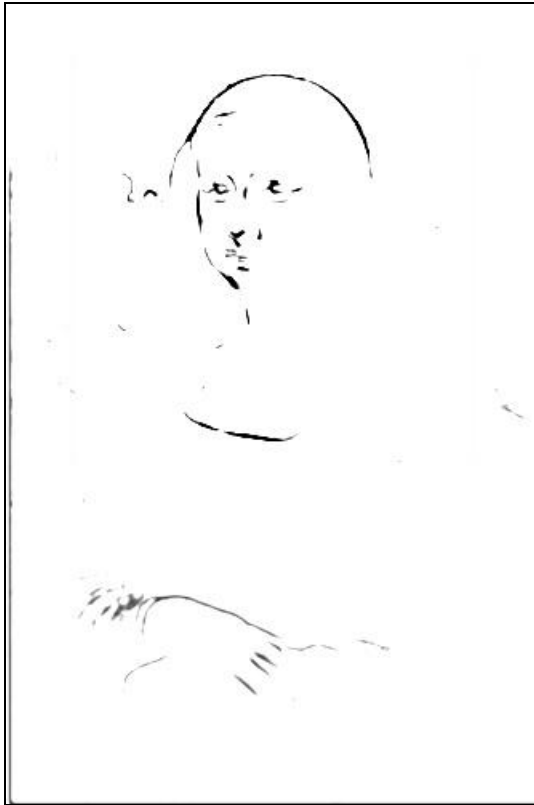
Нечеткие полудубликаты

Работаем в grayscale



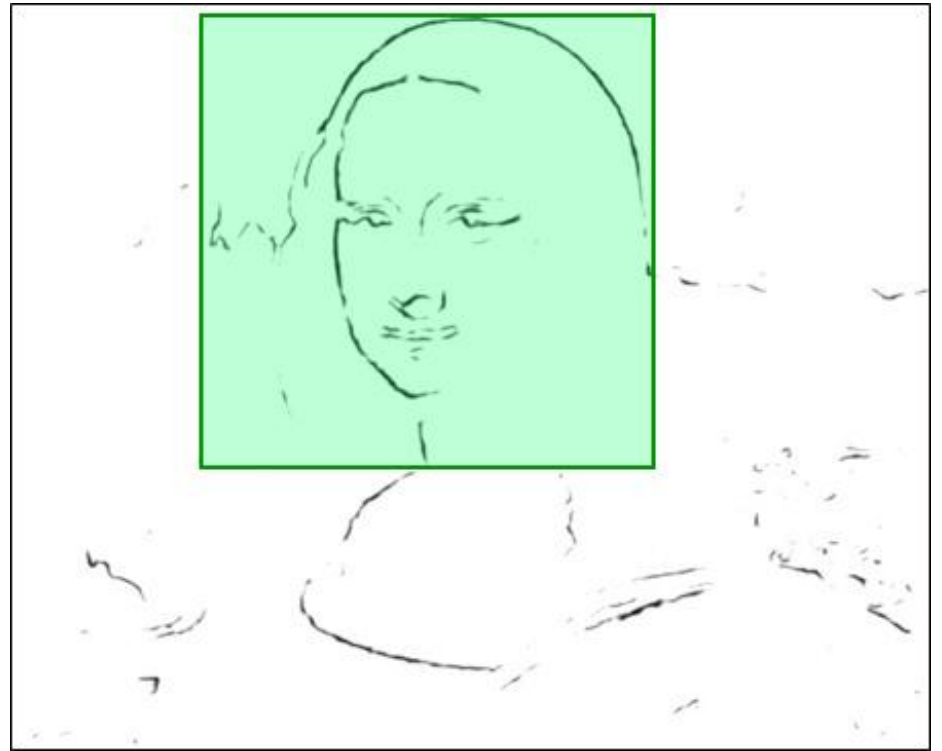
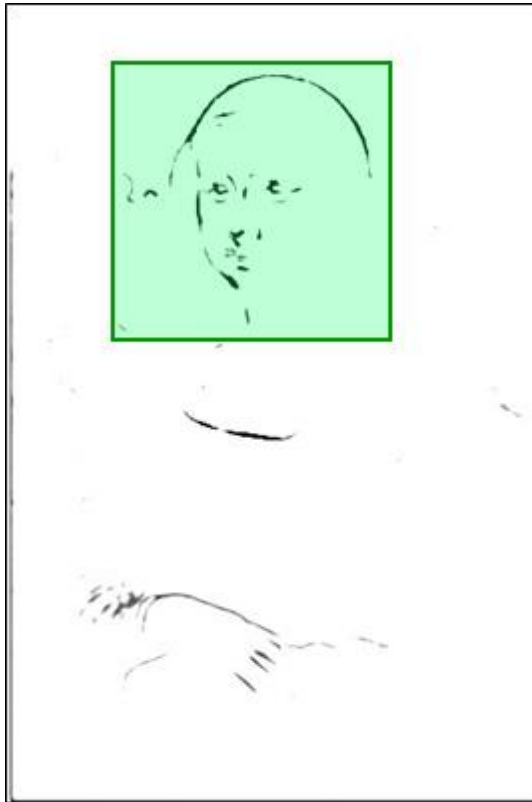
Нечеткие полудубликаты

Используем фильтр DoG



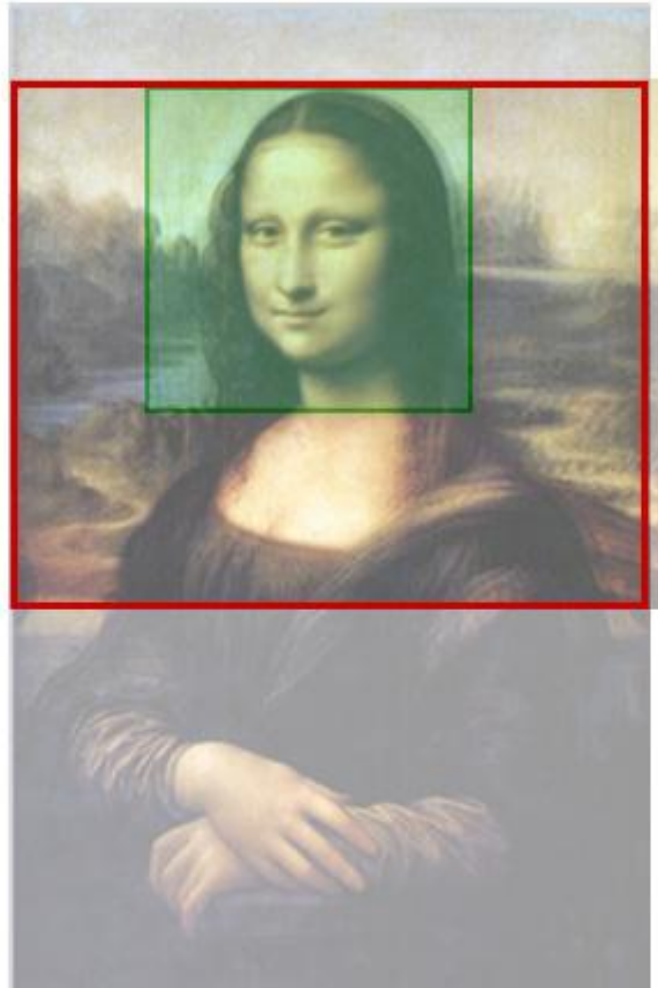
Нечеткие полудубликаты

Получаем дескрипторы



Нечеткие полудубликаты

Находим область пересечения изображений



Нечеткие полудубликаты

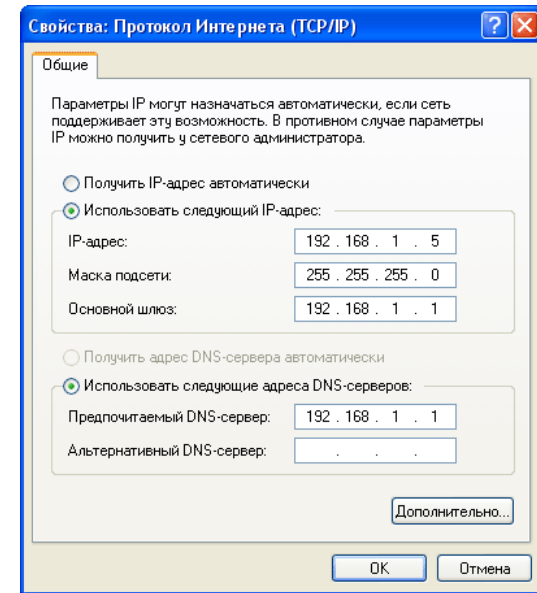
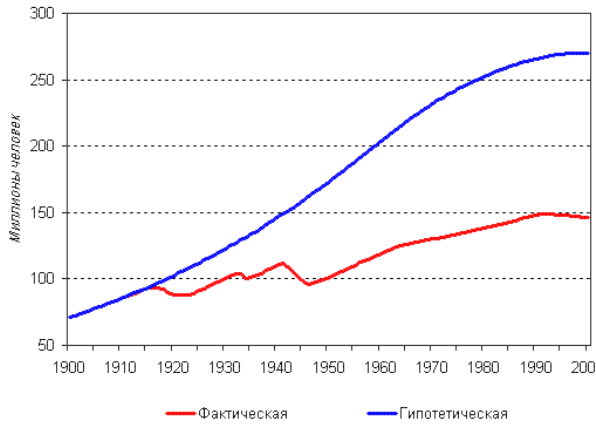
Задача свелась к предыдущей



Стадии кластеризации дубликатов

- Распределение на сотни больших пересекающихся групп по удаленности дескрипторов
- Формирование групп кандидатов в дубликаты по близости дескрипторов
- Финальная валидация

Проблемы больших групп



Кластеризация на большой базе

- Миллионы считаются на обычном компьютере за минуты
- Для сотен миллионов хватает кластера из десятка компьютеров
- Для миллиардов нужна сложная инфраструктура распределенного вычисления

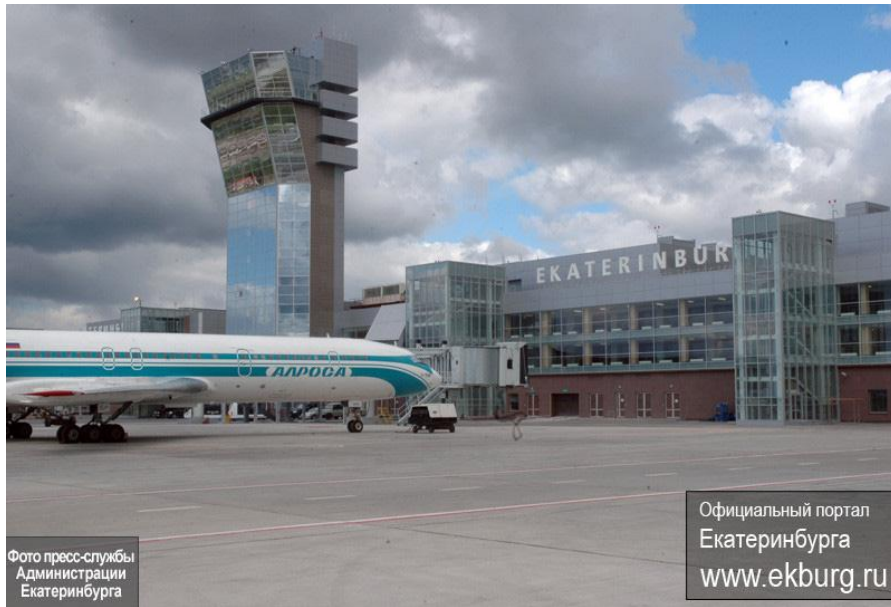
Кластеризация на маленькой базе

1 雲 2 女 3 雲
4 星 5 雲 6 和
7 和 8 命 9 神

Кластеризация на большой базе

	1	2	3	4	5	6	7	8	9
1	生	雨	父	月	日	水	土	宙	神
2	赤	生	愚	女	山	犬	花	月	殺
3	友	母	星	雲	和	天	男	赤	卒
4	芸	愛	和	命	神	死	愚	生	己
5	勇	法	毒	卒	戰	生	殺	珠	日
6	珠	愚	幻	宙	危	仰	鬼	刀	赤
7	獸	寿	蛇	和	己	仙	識	命	宙
8	和	土	愚	星	毒	卒	日	神	犬
9	命	父	月	愚	幻	宙	危	命	毒

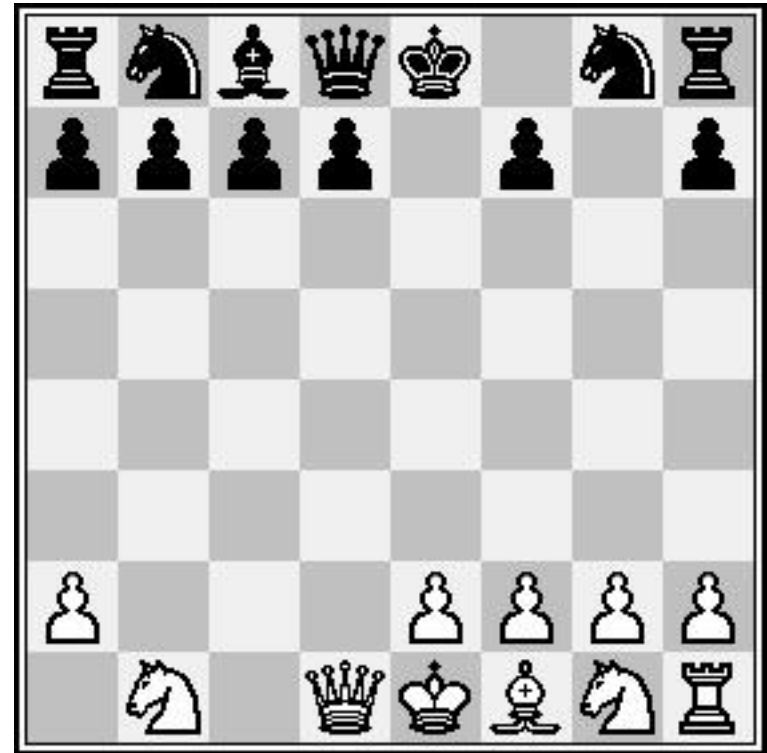
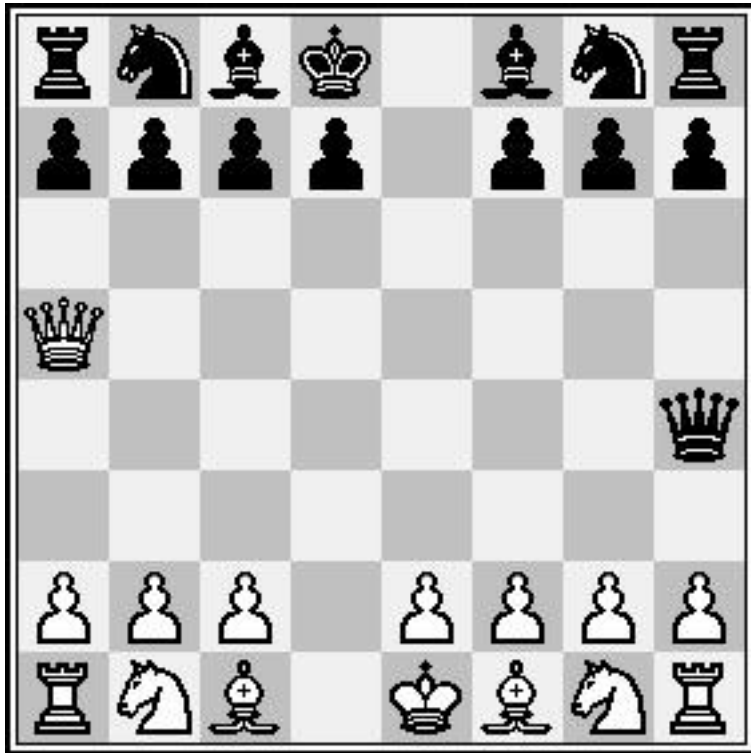
Что считать дубликатами?



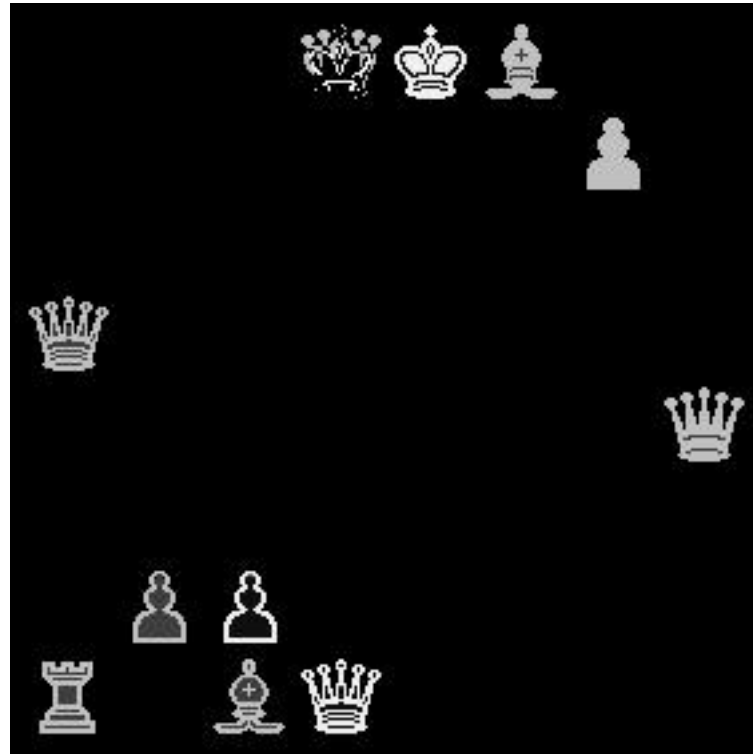
Что считать дубликатами?



Что считать дубликатами?



Что считать дубликатами?



Клоны – кто они для поиска?

Враги?



...или друзья?

Описания изображений на сайтах

Сопоставление описаний



«запорожец»

15 картинок

«синий запорожец»

10 картинок

«зеленый запорожец»

5 картинок

«лимузин»

10 картинок

Степень правдоподобия описаний:

- запорожец – 0,75 (30 картинок из 40)
- синий – 0,25 (10 картинок из 40)
- лимузин – 0,25
- зеленый – 0,13 (5 картинок из 40)



синий запорожец
запорожец лимузин

Разнообразие выдачи без кластеризации дубликатов

Яндекс
картинки

ipad

Найти

в найденном

расширенный поиск

[Настройка](#)
[Мои находки](#)

[Помощь](#)

Любые [Обои](#) [Большие](#) [Средние](#) [Маленькие](#) [Портреты](#)



Найдено картинок: **708 132**

Включен [умеренный фильтр](#)

Поиските также: [электронные книги](#) [ipad 3g](#) [apple ipad](#) [ещё](#)

Яндекс.Видео



Zoom CNews тестирует iPad

[Еще по запросу: ipad»](#) 3424



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[apple.com](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[i-ekb.ru](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[www.3angels.ru](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[komjet.ru](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[www.fantasiya.net](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[www.wallpaperweb.org](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[cm.dk](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[www.danlynchonline.com](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[paujak.blogas.lt](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[www.logster.ru](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[appletabletreview.net](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[free-ipads.org](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[bz9.com](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[withapple.ru](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[prowriterslab.com](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[gozamos.com](#)



apple ipad wifi 16 gb,32 gb,64 gb
gb ipad 3g

[www.studio980.gr](#)

Разнообразие выдачи с кластеризацией дубликатов

Яндекс
картинки

ipad

Найти

в найденном

[расширенный поиск](#)

[Настройка](#)

[Мои находки](#)

[Помощь](#)

Любые [Обои](#) [Большие](#) [Средние](#) [Маленькие](#) [Портреты](#)



Найдено картинок: **708 132**

Включен умеренный фильтр

Поиските также: [электронные книги](#) [ipad 3g](#) [apple ipad](#) [ещё](#)

[Яндекс.Видео](#)



Zoom CNews тестирует iPad
[Еще по запросу: ipad](#) 3424



23 июля про iPad узнают еще девять стран.
[450x285 ... 1000x619](#)
[www.freshdev.info](#) [похожие](#)



apple ipad wifi 16 gb,32 gb,64 gb ipad 3g
[460x276 ... 1680x1050](#)
[www.telewood.com](#)



apple ipad. steve jobs with his ipad.
[300x228 ... 630x563](#)
[www.g-g4u.com](#)



Apple (Republic of Ireland) - iPad hero 3.
[978x624 ... 1440x900](#)
[www.apple.com](#)



В отличие от Apple iPad, китайская копия.
[478x520 ... 500x438](#)
[prog-i-ot-gri.ucoz.ru](#)



Продам Apple iPad Wi-Fi 32GB, Челябинск.
[75x56 ... 1024x768](#)
[prodam.slando.chel.ru](#)



При этом стоимость iPad в Украине минимум.
[300x210 ... 1000x750](#)
[www.kazok.net](#)



Nu kan du se Apples iPad-event på video.
[300x293 ... 800x600](#)
[maczonen.dk](#)



Японский блог рассказал о следующем iPad.
[98x73 ... 904x600](#)
[ubr.ua](#) [похожие](#)



iPad Glare-Free Screen Protector.
[200x200 ... 600x490](#)
[new.ukdvd.co.uk](#)



iPad Wi-Fi (16 GB) Европа MB292
[40x40 ... 501x566](#)
[www.cplaza.ru](#) [похожие](#)



Send VPS.Net a Postcard and Win an iPad!
[320x483 ... 512x512](#)
[cloud.com](#)



5 Amazing iPad Apps For Business.
[322x241 ... 750x750](#)
[www.ivankristianto.com](#)



Apple Launches iPad - POSH Media Inc.
[350x300 ... 415x410](#)
[blog.poshmedia.ca](#)
[похожие](#)



iPad, 16GB, 32GB, 64GB Wi-Fi + 3G
[70x70 ... 1024x768](#)
[www.torg.uz](#)



iPad Sells 90,000 Units on Day 1.
[450x300](#)
[www.milehighautomation.com](#)
[похожие](#)



В iPad обнаружено место под камеру iSight?
[450x286 ... 620x395](#)
[applemasters.ru](#)

Применение дубликатов

Для чего используется

- Разнообразиие выдачи
- Точность поиска:
 - популярные изображения
 - сопоставление описаний
- Уточнение порно-классификатора
- Улучшение поиска «зеркал» и сайтов-клонов



Александр Крайнов

Менеджер проектов

krainov@yandex-team.ru