



# ТЕХНОСФЕРА

## Фильтрация порнографии в поиске

Москва 2017

## План лекции

- Постановка проблемы и общие решения
- Методика фильтрации веб-страниц
- Методика фильтрации запросов
- Фильтрация картинок

## Актуальность проблемы

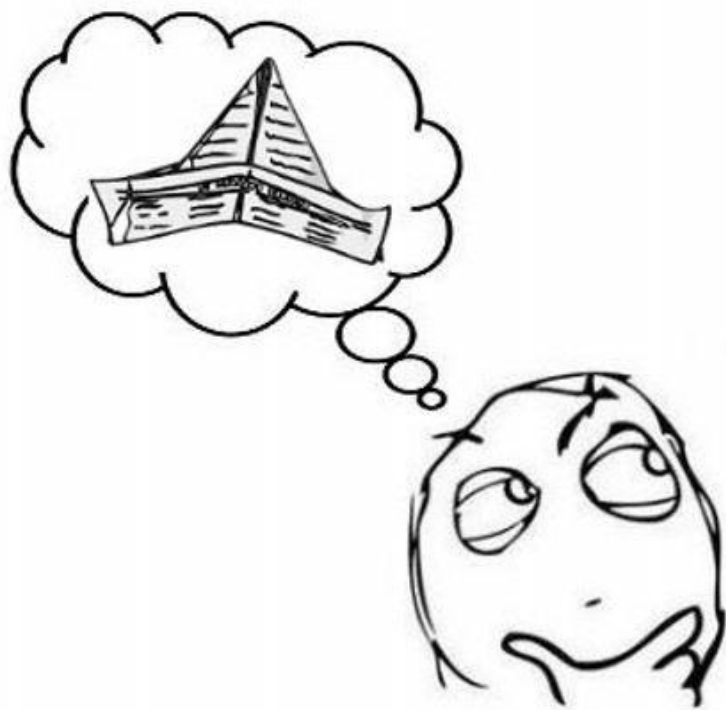
- Web поиск
- Поиск в AppStore, SmartTv (особенно топ запросы)
- Ленты соц.сетей (фото, контент)
- Рекомендательные сервисы - Пульс, Дзен и тд.
- Голосые помощники - Алиса, Маруся, Alexa, Siri
- Сайты с пользовательским контентом (pikabu, drive2.ru)
- Online игры ( Lego MMO - кто знает зачем?)
- Реклама
- И тд. (везде где есть неподобающий контент)

## Схожие проблемы - NSFW

**NSFW** (*Not safe/suitable for work* - Небезопасно/неподходяще для работы)

- То, что нельзя смотреть на работе, при родных и тд
- Порнография
- Неприличные изображения и тексты
- Нецензурная лексика
- Оскорбления (включая чувства верующих)
- Расчлененка, убийства и тд
- и другие неприятные для некоторых темы

# А в чем проблема?



# И вдруг...

поиск@mail.ru

пилотка

Найти

Открыть первые 5 результатов

Интернет

Картинки

Видео

Новости

Обсуждения

Ответы

Словари

Поиск в рунете

Поиск в мире

Поиск в Москве


Пилотка » ЭроЦех - Частная эротика, фото девушек, голые девушки...

Красотка незамедлительно оголяет аппетитные груди с возбужденными сосками, а после принимается снимать нижнюю часть купальника, обнажив шикарную попку и мокрую **пилотку**. ...

Красивые груди, упругая попка, волосатая **пилотка**...

[erosex.ru/tags/пилотка](#) копия еще с сайта

Картинки



Молоденькие » Порно, порно онлайн, смотреть порно, бесплатное...

Отодрал юную брюнетку в киску


[gorod-porno.com/molodenkie](#) копия еще с сайта

Пилотки. Женские **пилотки** крупным планом

...Смотреть эротику онлайн, домашние веб камер онлайн, Частное видео с Мобильных, а так же найти фото голых девушек, любительский секс скачать, по мимо этого женщины в возрасте, **пилотки** девушек и фото клубничка.

[pilotki-vip.ru/pilotki/page/39/](#) копия еще с сайта

Видео



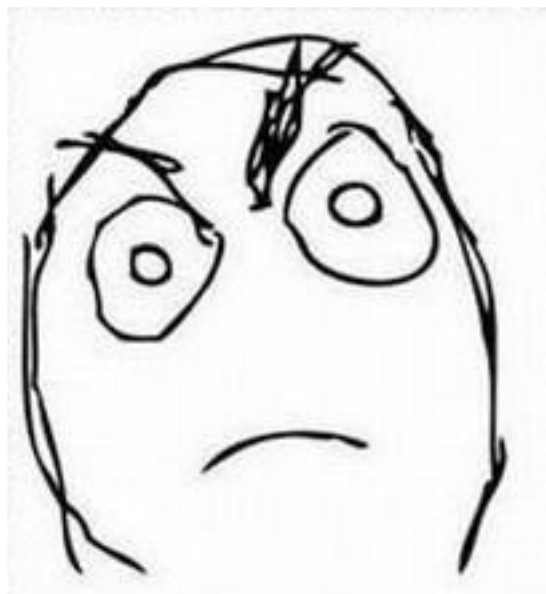
пилотка девщчек

Youtube 23.01.12

Ответы

пионерская **пилотка**. подскажите пжл ширину пионерской **пилотки**

6



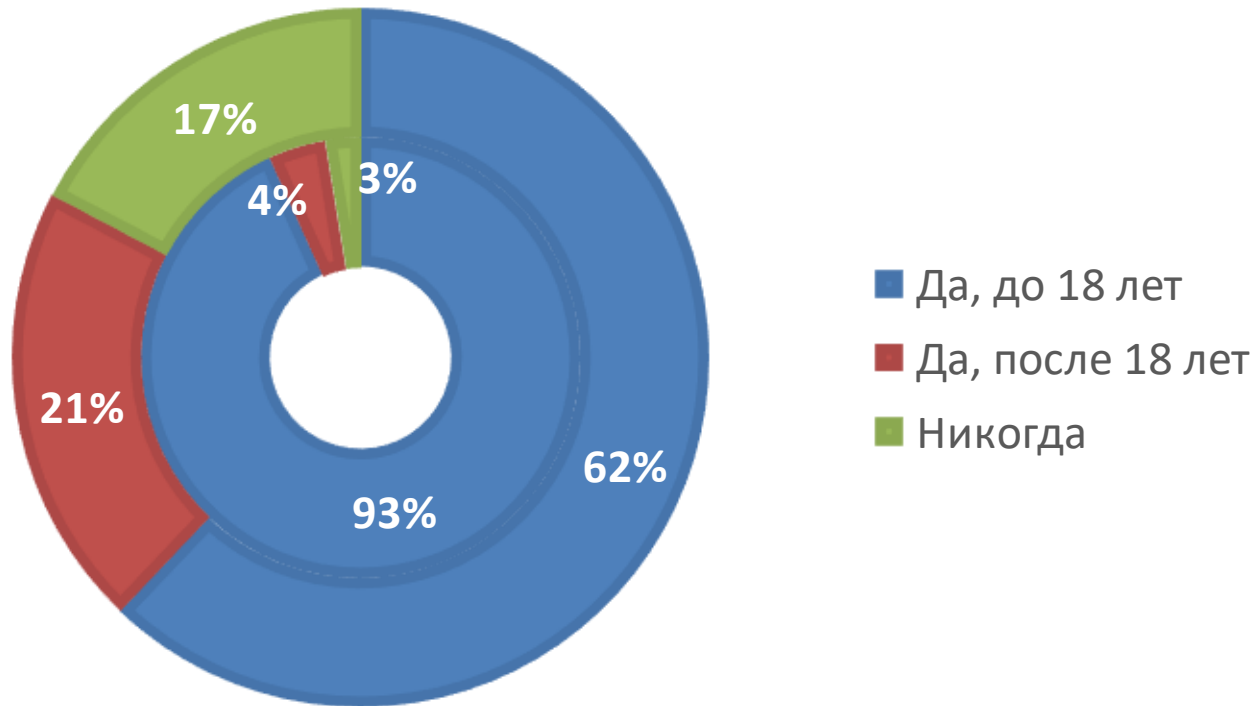
# Почему это проблема для поиска?



## Почему это проблема для поиска

1. Нормы морали и приличия и влияние на детей - нельзя показывать порно, если пользователь мог иметь ввиду, что-то другое
2. Негативно влияет на качество ранжирования - так как есть порно обо всем и на него кликают и смотрят, что сильно зашумляет поведенческие факторы

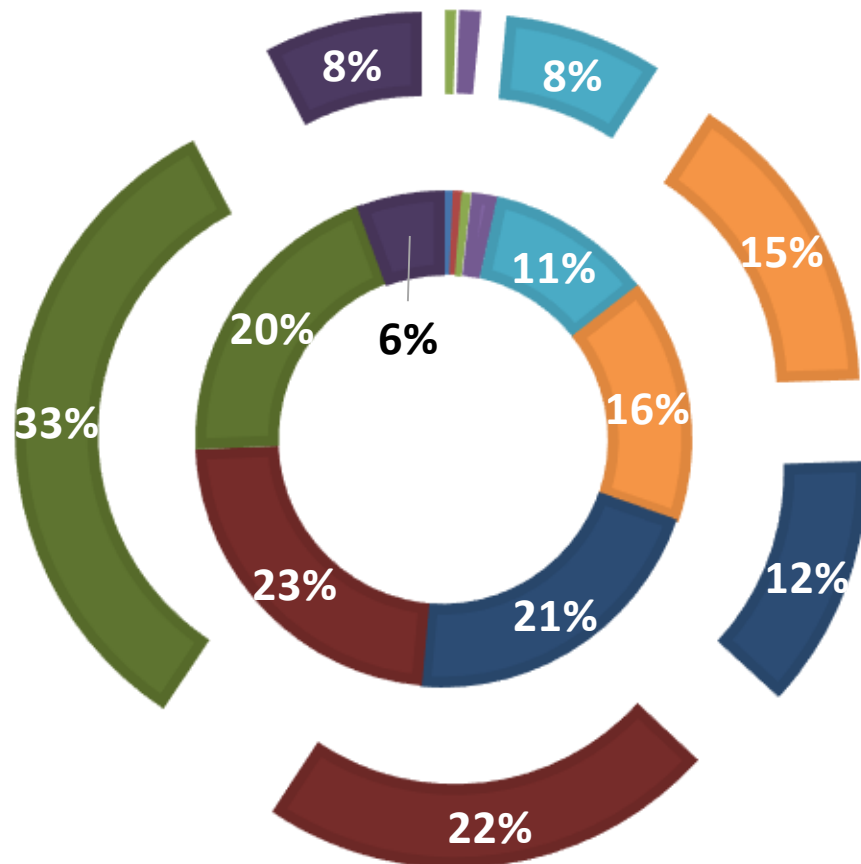
## Когда впервые увидели в интернете порнографию



| Подвергались воздействию | Юноши % | Девушки % |
|--------------------------|---------|-----------|
| Да, до 18 лет            | 93,2    | 62,1      |
| Да, после 18 лет         | 4,2     | 20,6      |
| Никогда                  | 2,6     | 17,3      |

Опрошено  
Юноши – 192  
Девушки - 371

# Возраст первого воздействия на детей



- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17

| Возраст первого воздействия | Юноши % | Девушки % |
|-----------------------------|---------|-----------|
| 8                           | 0,6     | 0         |
| 9                           | 0,6     | 0         |
| 10                          | 0,6     | 0,5       |
| 11                          | 1,7     | 1         |
| 12                          | 10,9    | 7,7       |
| 13                          | 16      | 15,3      |
| 14                          | 21,1    | 12,4      |
| 15                          | 22,9    | 22,5      |
| 16                          | 20      | 33        |
| 17                          | 5,7     | 7,7       |

## Распределение по полу

По статистике поиска женщины и мужчины почти одинаково ищут "порно". Аналогичная статистика и по соц.сетям просмотра.

Так что проблема порнографии особо не связана с полом.

P.S. Но вот при исследованиях, где опрашивают людей статистика отличается сильно ;)

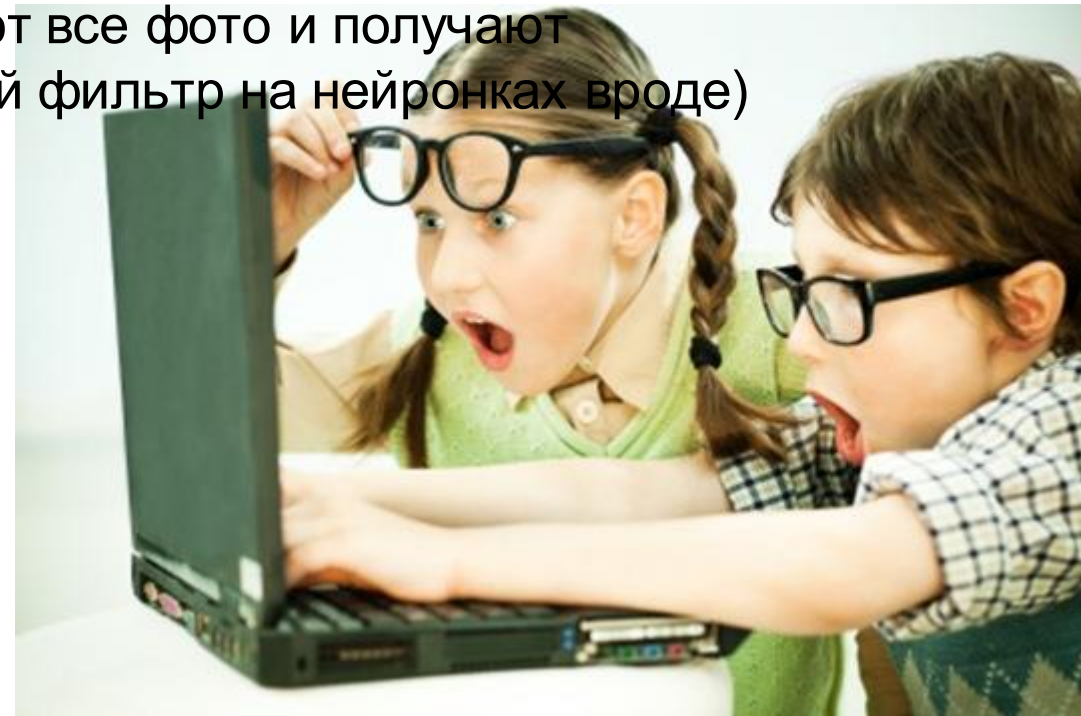
# Варианты решений

## 1. Заставить пользователей самих фильтровать контент

- Ассесоры в Поиске
- Редакторы на сайтах, Модераторы в играх, рекламе и тд
- В одноклассниках пользователи модерируют все фото и получают внутреннюю валюту. (есть предварительный фильтр на нейронках вроде)

## 2. Автоматические методы

- Черные и белые списки
- Блокировка по ключевым словам
- Системы рейтингов
- Классификация страниц
- Классификация картинок



## Где встречается порно в Поиске

- Запросы
- Документы (сайты)

## Зачем детектировать порно запросы

- Другая логика ранжирования (нет штрафа порно)
- Такие запросы не добавляются в "подсказки" и топы
- Не исправляют опечатки, если результат "порно" запрос
- Можно "прикрывать" картинки и видео по таким запросам

## Зачем детектировать порно документы

- Убирать из выдачи по не "порно" запросам
- Иначе отображать в выдаче - прикрывать картинки и тд
- Фактор для ранжирования



## Типы порно контента в сети

- Порно сайты с фото, видео (самые легкие для детекции)
- Что еще?)

## Типы порно контента в сети

- Порно сайты с фото, видео (самые легкие для детекции)
- Сайты "эскорта", "массажа", проституток, "типа знакомств"
- Эротические рассказы
- Sex-shop
- Взрослые разделы на "приличных" сайтах - например раздел "хентай" на аниме ресурсе
- Обычный сайт с кучей "порно" тизеров(рекламы)
- И тд.

## Как обычно решают

1. Вначале банят самое серьезное по ключевикам быстро
2. Создаются списки плохих сайтов и через граф ссылок расширяется
3. Анализ контента и machine learning

## Классификация документов

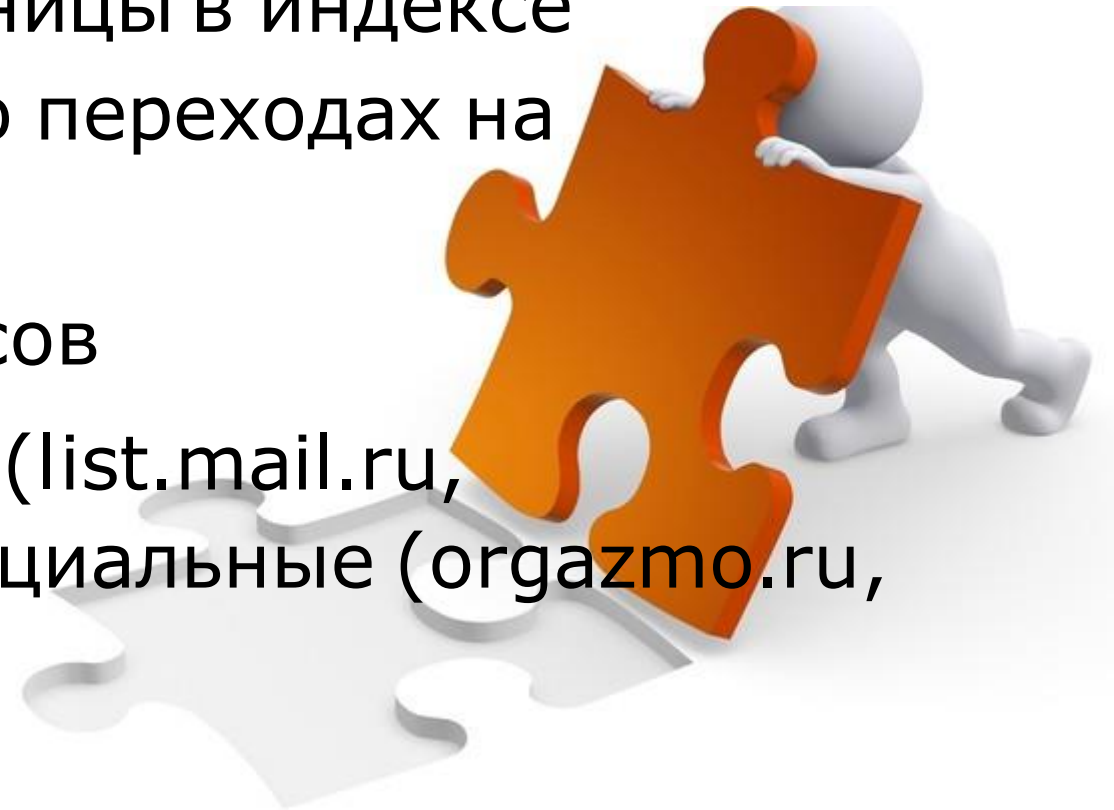
- Выделяем набор характеристик, которые отличают порносайт от обычного
- Порносайты не маскируются, и это хорошо
- Проблема - много документов в “серой” зоне

## Классификация запросов

- Необходимо понять, что показывать
- Запросы короткие и зачастую имеют много смыслов
- Некоторые из них содержат опечатки

## Источники данных

- Неразмеченные html-страницы в индексе
- Логи запросов с данными о переходах на страницы
- Переформулировки запросов
- Внешние каталоги: общие ([list.mail.ru](http://list.mail.ru), [liveinternet.ru/rating](http://liveinternet.ru/rating)) и специальные ([orgazmo.ru](http://orgazmo.ru), [nulist.biz](http://nulist.biz)).



# Классификация документов

## Задача текстовой классификации

$T$  - множество слов документов словаря  $|V|$

$Y$  – множество меток классов

$T^m = \{(t_1, y_1), (t_2, y_2), \dots, (t_m, y_m)\}$  -обучающая выборка

$A : T \rightarrow Y$  - алгоритм классификации,

классифицирующий документ произвольный документ  $D$



# Наивный байесовский классификатор

Байесовский классификатор

$$p(C_i | D) = \frac{\prod_{j=1}^n p(t_j | C_i) \cdot p(C_i)}{p(D)}$$

- Документ рассматривается как набор независимых слов
- По обучающему множеству составляются словари с весами
- Находим слова классифицируемого документа в словарях и смотрим, вес слов какого класса больше

## Недостатки байесовского подхода

- Не учитываем разметку документа, хотя она может быть важна
- Не учитывает значимость слов в разных частях документа (title, url, keywords, a href)
- Тяжело составить правильное обучающее множество.
- Проблема редких слов, как их представлять в признаках
- Представление различных тематик (рассказы, фото видео хостинги, сайты знакомств и т. п.)
- Короткие документы, документы без слов

## Другие подходы

- Метрические - ищут близость в образцу
- Решающие правила
- Регрессия
- SVM
- Деревья решений
- Нейронные сети

# Примеры ошибок классификатора

**Фазиль Искандер «Рассказы о Чике».**

- Много слов, неизвестных классификатору
- Некоторое количество слов, употребляемых на порносайтах (top в списке справа)

*Результат — документ попал в «серую» зону.*

толстенькими 5.33898  
поскуливала 4.89884  
дырочках 4.69508  
зрелые 4.62396  
всовывал 4.56707  
жесткое 4.45139  
отсосав 4.40743  
раздвигал 4.37484  
юбкой 4.3215  
лизала 4.2573  
кончала 4.17881  
чулках 4.16239  
щекотали 4.13878  
всунув 4.1025  
пахучую 4.06708  
задвигалась 4.0397  
блаженном 3.99227  
упругую 3.9481  
аппетитную 3.92286  
глотала 3.90956  
извивалась 3.88  
покачивались 3.85739  
оттопырила 3.8235  
бритую 3.81365

## Трудные случаи

- Мало текста
- Текст является навигационной обвязкой
- Текст не имеет отношения к картинкам
- На странице только картинка



# Трудные случаи

Порнотизеры

Часто занимают большую часть  
страницы

Сильно привлекают внимание

В html-коде выглядят как часть  
скрипта, запрос к соответствующей  
тизерной сети:

```
<script  
  src="http://camo4ek.net/effect.php?informer=101"  
  type="text/javascript">
```



Хохотать будете до  
изнеможения! Не  
пропустите (онлайн)



Убит глава КНДР Ким Чен  
Ын



Жуткая первая брачная  
ночь для 13-летней  
невесты стала последней!



Это видео поразило всех! Смотри бесплатно



Камерон Диаз показала всем  
свою сногшибательную  
фигуру. Фото



Пэрис Хилтон в купальнике:  
красиво ли? Фото!

## Добавляем данных

- Порносайты часто имеют URL, содержащий определенные подстроки (xxx, porno, adult, sex, erotic)
- Заголовки и ключевые слова будем обрабатывать по другому (посчитаем встреченное количество слов из порнословаря)
- Крупных тизерных сетей, отдающих порно, не так много, будем искать обращения к ним в коде страницы
- Будем считать количества переходов на страницу по порнозапросам: раз пользователи ходят на сайт за порнографией, наверное, она там есть.

## Машинное обучение

$X$  - множество признаков документа

$Y$  – множество меток классов

$X^m = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$  -обучающая выборка

$A : X \rightarrow Y$  - алгоритм классификации, классифицирующий документ произвольный документ  $D$



# Машинное обучение

- Gradient boosting oblivious trees
- Обучали на небольшом, но точном наборе документов (~ 8 тыс.)

## Процесс построение модели

- Находим ошибки классификации
- Выделяем свойство, которое плохо распознается
- Находим примеры, добавляем в обучающее множество
- Выделяем доп. признаки при необходимости
- Переобучаемся
- Проверяем тестом, что ничего не «отъехало»

## Что получилось?

### Признаки документа, по степени значимости

bayes 2.05685

porn\_clicks 0.8613

keywords 0.7252

title 0.1173

url 0.1018

img\_num 0.028

script\_num 0.017

teasers 0.00013

Точность — 98.3, полнота — 95.5 и F1 — 96.5

# Классификация запросов

## Цель: разделить запросы на

- Запросы не порнографического характера
- Запросы, которые в большей мере имеют порнографический смысл

## Фильтрация порнографии в выдаче

- По обычным запросам
  - *(мультфильмы, рассказы, видео, фото)*
- По неявным «взрослым» запросам
  - *(азиатки, мама и сын, девушка с конем)*

## Показываем все как есть

- По явным порно запросам
  - (*эротика, порно смотреть онлайн, проститутки в москве*)
- По навигационным и точно попадающим в тему запросам
  - (*саша грей видео, redtube, gexx.com*)

А что делать с "каштанкой"?

## Фильтрация по спискам

Составляем большие списки «плохих» и «хороших» запросов.

### *Недостатки*

- Слишком много форм
  - (*порно, порнушечка, порево, порноонлайн*)
- Морфология часто мешает
  - (*вафли → вафлить*)
- Можно забанить хорошие запросы:
  - xxx – Олимпиада XXX
  - индивидуал(ки) – средства индивидуальной защиты



## Регулярные выражения

Составлять списки регулярных выражений для паттернов в запросах

*Плюс*

✚ Не нужна морфология

*Минус*

— Потеряли из саджестов *чЕБурашку, аЭРОфлот, оПОРНый прыжок*

## Переформулировки запроса

- Смотрим был ли переход на порно сайт по запросу
- Смотрим на предыдущий запрос
- Оцениваем вероятность порнографичности начального запроса, через количество переформулировок в порно запрос ( $p(q_{porn}) = \frac{n_{ref}}{n_q}$ )
  - *видео → порноонлайн, youtube, эровидео, sexvideo*

*Недостатки: Большое  $n_q$  маленькое  $n_{ref}$*

## Оценка запроса по выдаче

Запрос оцениваем через относительную порнографичность выдачи взвешивая количество порно документов по запросу

$$f_q = \left( \frac{n_{dporn}}{n_{dporn} + n_{dqood}} \right)$$

$n_{dporn}$  - количество документов по запросу  $q$ , размеченных как порно

$n_{dqood}$  - количество документов по запросу  $q$ , размеченных как не порно

### *Пример*

По запросам *мулатки,малолетки,бесплатное видео* доля страниц из выдачи определяющихся как порно равна 0.8

## Оценка выдачи по запросу

Выдача дает много информации о запросе, но есть минус:

*Словарь «запрос — доля порнодокументов» строится оффлайн и содержит не все возможные запросы.*

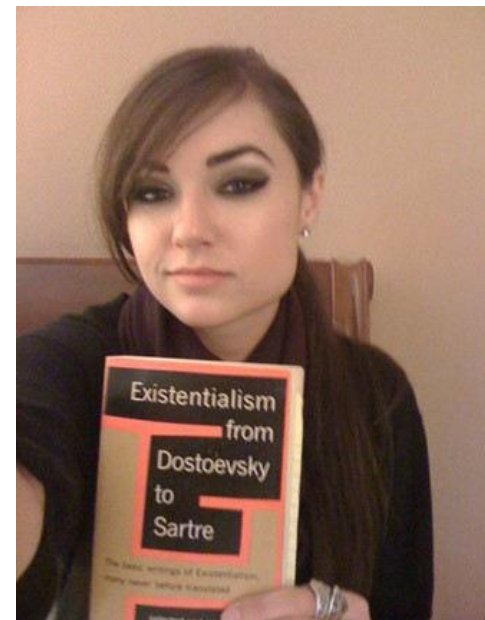
Все равно требуется онлайн поддержка

## Саша Грей.



Оценка запроса по  
выдаче 0.85 porn

VS



Sasha Grey and  
Existentialism

**Экзистенциализм** — направление философии, главным предметом изучения которого стал человек, его проблемы, трудности, существования в окружающем мире.

## Использование словарей онлайн фильтрации

Составляем небольшой словарь (около 200 вхождений) регулярных выражений (плохая полнота, но хорошая точность).

*Этот же словарь используется при классификации документов.*

Составляем словарь «эвфемизмов» — обычных слов, в некотором контексте придающих запросу порносмысл ( *девочки, секретарши, бесплатно, зрелые* )

## Как происходит классификация

- Ищем, соответствует ли запрос выражениям из ручного «черного» списка (если да, запрос порнографичен).
- Если нет, ищем его в автоматически составленном словаре, проверяя, сколько документов из выдачи порнографические.
  - Если меньше некоторого порога, запрос чист.
  - Если больше — удаляем из запроса все «эвфемизмы» и проверяем (также по выдаче) оставшуюся часть запроса.

Пример: *redtube видео , блондинка в лесу*

## Результаты

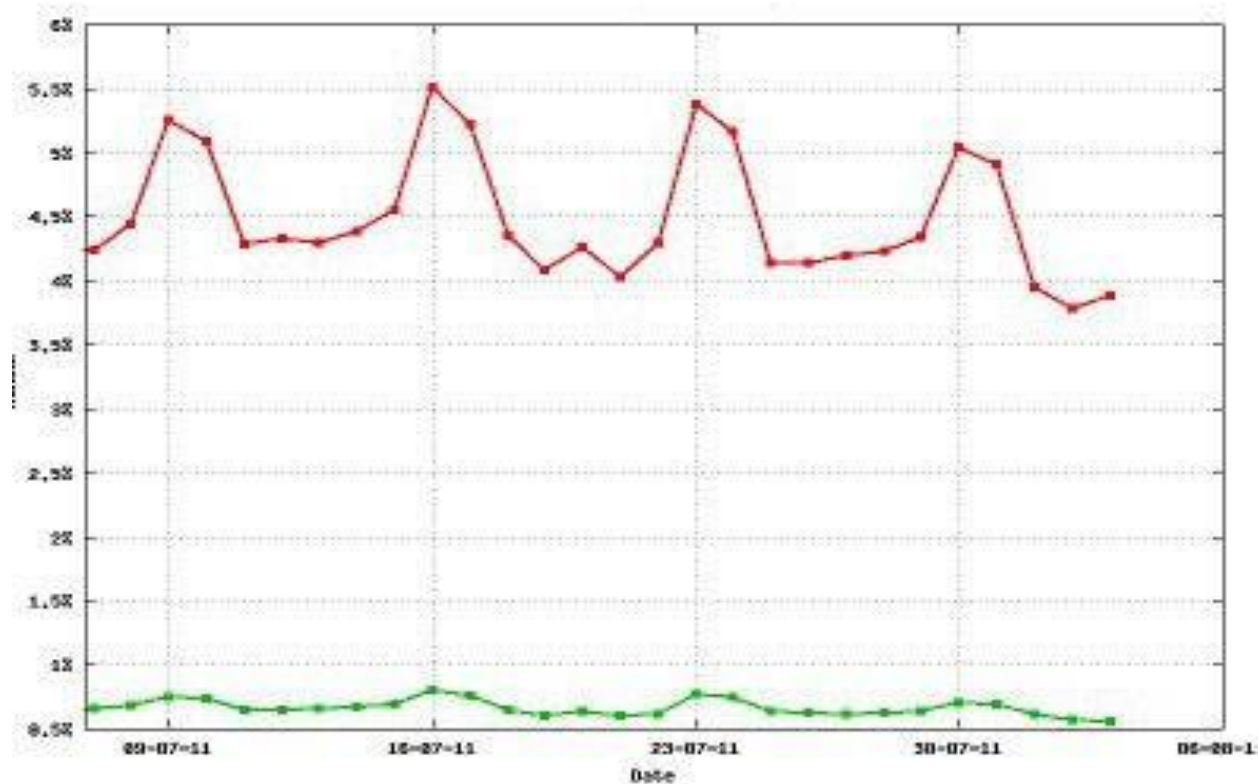
*Точность 96%, полнота 85%.*

### Ошибки

- Наличие слова *порно* не всегда говорит о «взрослом» запросе  
(*Зак и Мири снимают порно, незаконное распространение порнографии*)
- Не всегда по навигационным запросам в выдаче много порно



# Процент порнозапросов в потоке



**Красная линия** — то, что нашлось по списку.  
(будни 4.5%, выходные 5.5%)

**Зеленая** — по выдаче (0.7%)

# Классификация картинок

## В чем проблема?

Не всегда возможно определить контент страницы по тексту

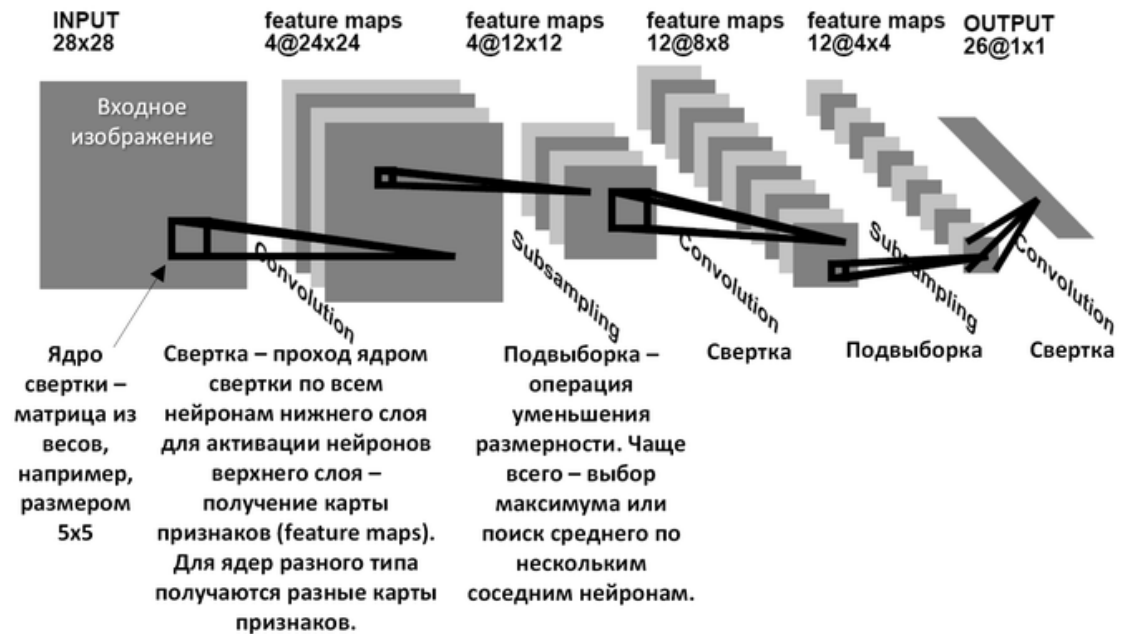
- Мало текста
- Текст не имеет отношения к картинке
- Страница не имеет текста, только картинки

## Подход: Skin detection

- Большинство картинок с людьми имеют открытые участки тела
- Эти открытые участки могут быть определены по цветовой палитре
- Задача:
  - Выделить большие участки с заданной цветовой палитрой на картинке
  - По конфигурации областей определить класс картинки
- Недостатки: Низкая точность  $< 0.8\%$

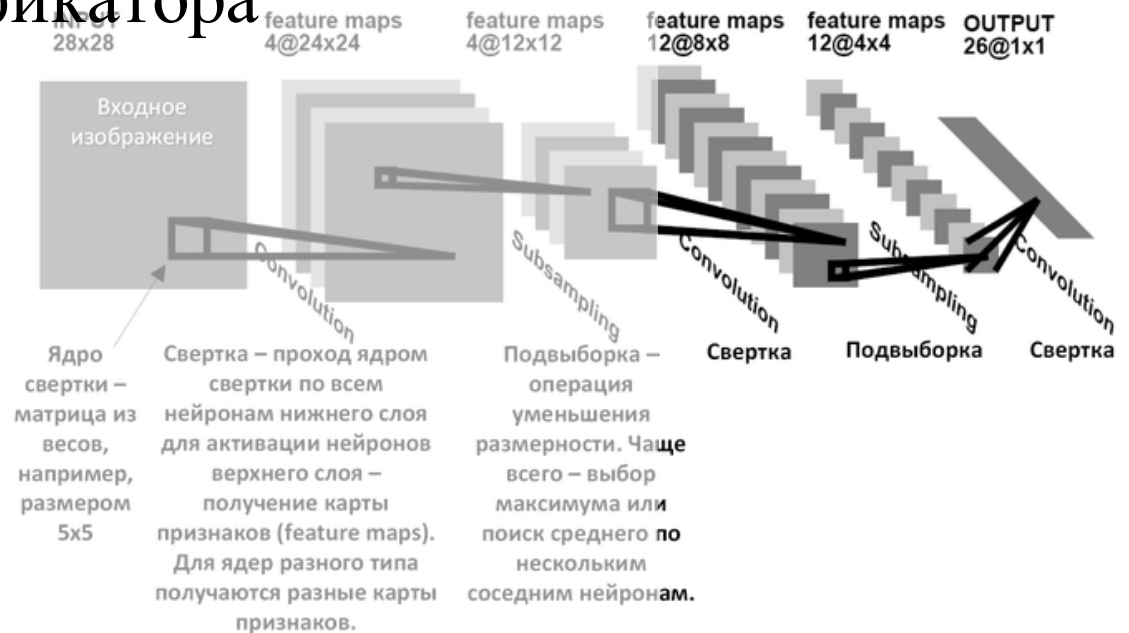
# Подход: Глубокое обучение вариант I

- Сверточная нейронная сеть
- Размеченное обучающее множество
  - порно
  - не порно
- Строим классификатор
- *Проблема:*
  - Набрать обучающее множество
  - Обучить сеть достаточной глубины



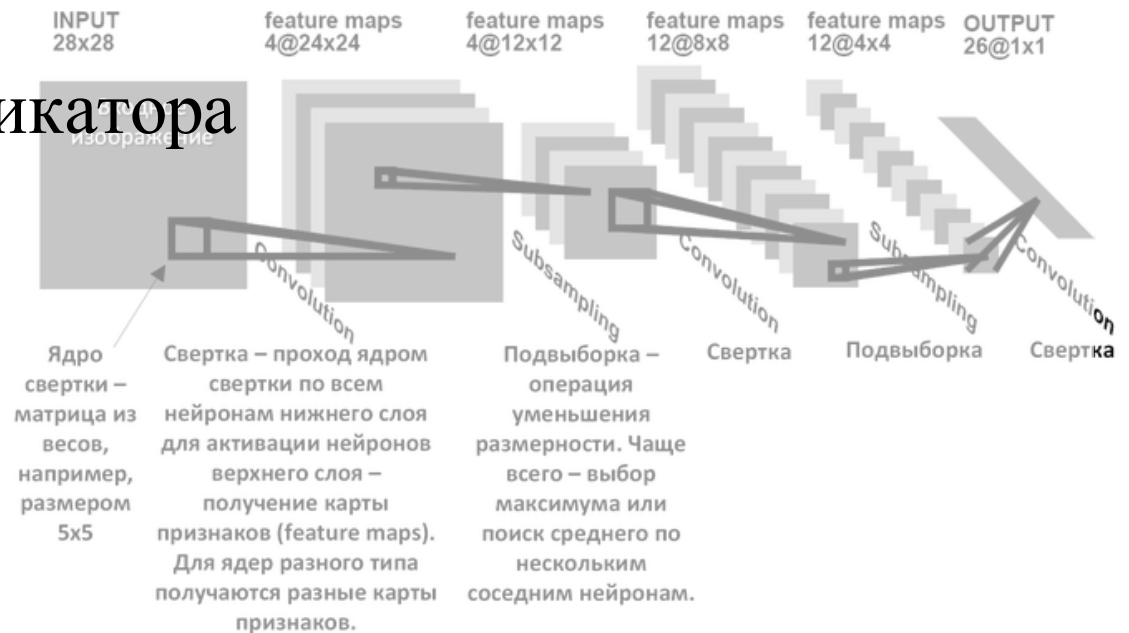
## Подход: Глубокое обучение вариант II

- Обученная сверточная нейронная сеть, например, на ImageNet
  - <http://www.image-net.org/>
- Размеченное обучающее множество
- Выходной слой сверки - вход классификатора
- *Плюсы:*
  - Не нужно обучать сеть
- *Проблема:*
  - Набор обучающего множества



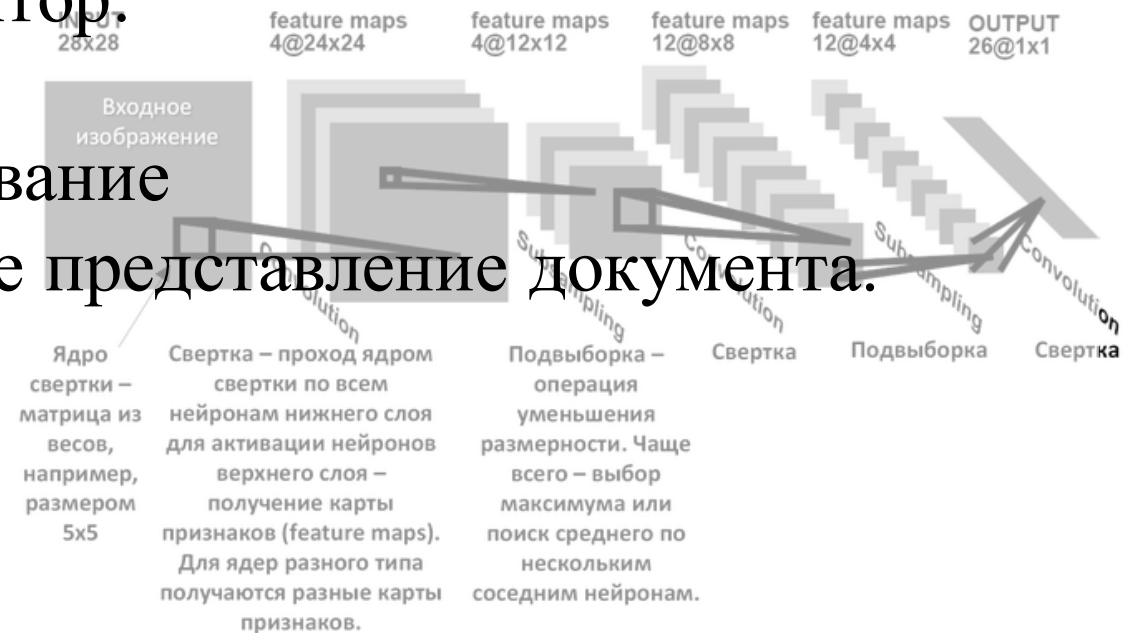
## Подход: Глубокое обучение вариант III

- Обученная сверточная нейронная сеть, например, на ImageNet
  - <http://www.image-net.org/>
- Обучающее множество:
  - Кластеризуем порно и не порно запросы
  - Отбираем из кластеров картинки для ОМ
- Выходной слой сверки - вход классификатора
- *Плюсы:*
  - Не нужно обучать сеть
  - Автоматически набираем ОМ



## Подход: Глубокое обучение вариант IV

- Сверточная нейронная сеть
- Обучающее множество:
  - Набор документов, размеченный порно классификатором
- Тексты документов превращаем в вектор:
  - используя word2vec
  - используя тематическое моделирование
- Обучаем НС предсказывать векторное представление документа.



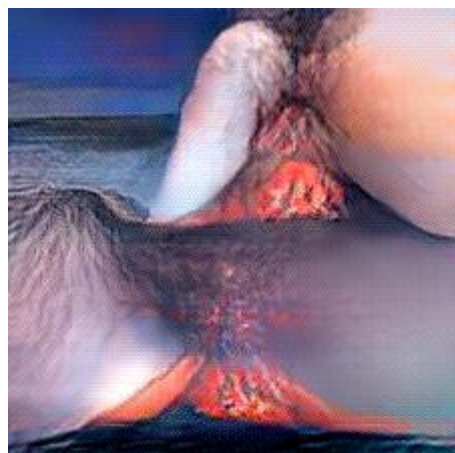
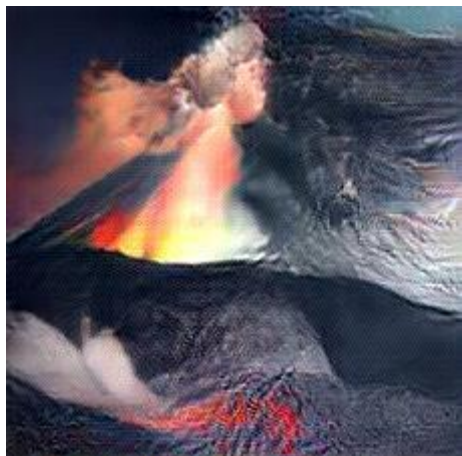


## Подход: Глубокое обучение - bonus

Yahoo выложил нейросеть, которую они используют для классификации порнографии в поиске

Архитектура обычная - resnet с оптимизациями по скорости

При помощи нее научились генерировать порно картинки, а точнее превращать приличные картинки в порнографически - аля artistic style (prisma, artistico, vinci)





# ТЕХНОСФЕРА

**Спасибо!**

**Вопросы?**



# ТЕХНОСФЕРА

## Семинар

## Наивный Байесовский классификатор

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(C|D) = P(C|x_1 \dots x_n) = \frac{P(x_1 \dots x_n|C)P(C)}{P(D)}$$

Как используют на практике:

$P(C|x_1 \dots x_n)$  - где  $x_1 \dots x_n$  какие-то признаки объекта в большом количестве

В случае текстовой классификации - это слова, но могут быть любые признаки

# Наивный Байесовский классификатор

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(C|D) = P(C|x_1..x_n) = \frac{P(D|C)P(C)}{P(D)} = \frac{P(x_1..x_n|C)P(C)}{P(D)}$$

$$P(x_1..x_n|C) = \prod_i^n P(x_i|C)$$

Последнее утверждение и говорит о "наивности" - мы считаем, что все признаки независимы между собой.

В случае классификации текстов это слова, но могут быть любые признаки.

## Наивный Байесовский классификатор

$$P(C|D) = \frac{\prod_i^n P(x_i|C)P(C)}{P(D)}$$

$$P(\neg C|D) = \frac{\prod_i^n P(x_i|\neg C)P(\neg C)}{P(D)}$$

$$\frac{P(C|D)}{P(\neg C|D)} = \frac{\prod_i^n P(x_i|C)P(C)}{\prod_i^n P(x_i|\neg C)P(\neg C)}$$

$$\ln \frac{P(C|D)}{P(\neg C|D)} = \ln \frac{P(C)}{P(\neg C)} + \sum_i^n \ln \frac{P(x_i|C)}{P(x_i|\neg C)} = \ln \frac{P(C)}{P(\neg C)} + \sum_i^n \ln \frac{classC \ freq X_i}{classNotC \ freq X_i}$$

$$P(x_i|C) = classC \ freq X_i / total$$

$$P(x_i|\neg C) = classNotC \ freq X_i / total$$

$$\ln \frac{P(C)}{P(\neg C)} = const$$

$$P(C|D) + P(\neg C|D) = 1$$

## Наивный Байесовский классификатор

$$\ln \frac{P(C|D)}{P(\neg C|D)} = bias + \sum_i^n \ln \frac{classC\ freq X_i}{classNotC\ freq X_i}$$
$$P(C|D) + P(\neg C|D) = 1$$

Соответственно из этих формул и вытекает, что в случае, если документ Принадлежит классу С, то логарим отношений будет положительный и следовательно больше нуля - иначе меньше.

## Наивный Байесовский классификатор практические трюки

- Обычно сумму делят на кол-во слов в документе, чтобы отношение логарифмов не достигало бы больших значений.

$$\ln \frac{P(C|D)}{P(\neg C|D)} = bias + \sum_i^n \ln \frac{classC\ freqX_i}{classNotC\ freqX_i} / wordsInDocument$$

- Выкидывают stop-слова - предлоги и тд
- Выкидывают слишком редкие слова
- Выкидывают слова встретившиеся только в одном классе
- Выкидывают слова встретившиеся близкое кол-во раз - логарифм близок к нулю



## Данные для семинара

Регистрация в соревновании:

[https://kaggle.com/join/antoporn\\_bayes](https://kaggle.com/join/antoporn_bayes)

Данные

<https://inclass.kaggle.com/c/antiporn-infopoisk/data>