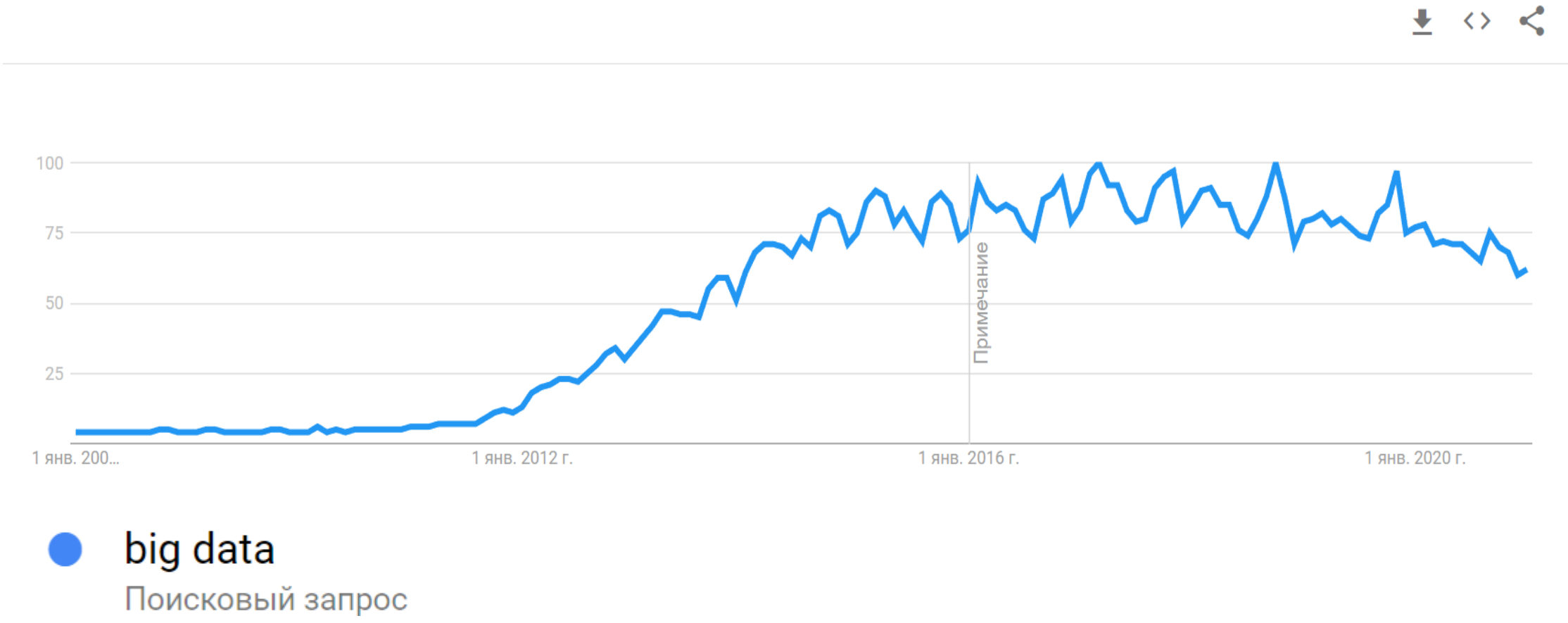
Введение в Big Data: методы, кейсы, технологии

Вячеслав Гойко

История появления Big Data

Динамика популярности запроса



Что такое Большие данные

Big Data

– разнообразные данные, которые поступают с постоянно растущей скоростью и объем которых постоянно растет (**Oracle**)

– информация, которую уже невозможно обрабатывать традиционными способами, в том числе структурированные данные, медиа и случайные объекты (**Tadviser**)

– данные огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами (**Wikipedia**)

– это такие данные, которые невозможно обрабатывать в Excel (**кто-то из коллег**)

Что такое Большие данные

Big Data

– разнообразные данные, которые поступают с постоянно растущей скоростью и объем которых постоянно растет (**Oracle**)

– информация, которую уже невозможно обрабатывать традиционными способами, в том числе структурированные данные, медиа и случайные объекты (**Tadviser**)

– данные огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами (**Wikipedia**)

– это такие данные, которые невозможно обрабатывать в Excel (**кто-то из коллег**)

– серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов (**Хабр**)

Основные свойства Big Data

Volume (Объем)

Velocity (Скорость)

Variety (Разнообразие)

Основные свойства Big Data

Volume (Объем)

Velocity (Скорость)

Variety (Разнообразие)

Veracity (Достоверность)

Value (Ценность)

Примеры источников Big Data

- Социальные сети
- Интернет вещей (IoT), промышленные и бытовые приборы
- Банковские транзакции
- Медицинские данные
- Собственные данные компаний
- Спутниковые снимки

Основные методы анализа

Data Mining (интеллектуальный анализ данных) – обнаружение в данных новых знаний и закономерностей

Краудсорсинг – ручная разметка и обогащение данных с привлечением широкого круга лиц

Машинное обучение и нейронные сети

Обработка естественного языка

Анализ социальных сетей (Social Network Analysis)

Распознавание образов

Статистический анализ

Моделирование и предиктивная аналитика

Визуализация аналитических данных

Применение Больших данных

Индивидуализация рекомендаций товаров и услуг

Прогнозирование академической успеваемости

Моделирование когнитивных и поведенческих особенностей пользователей

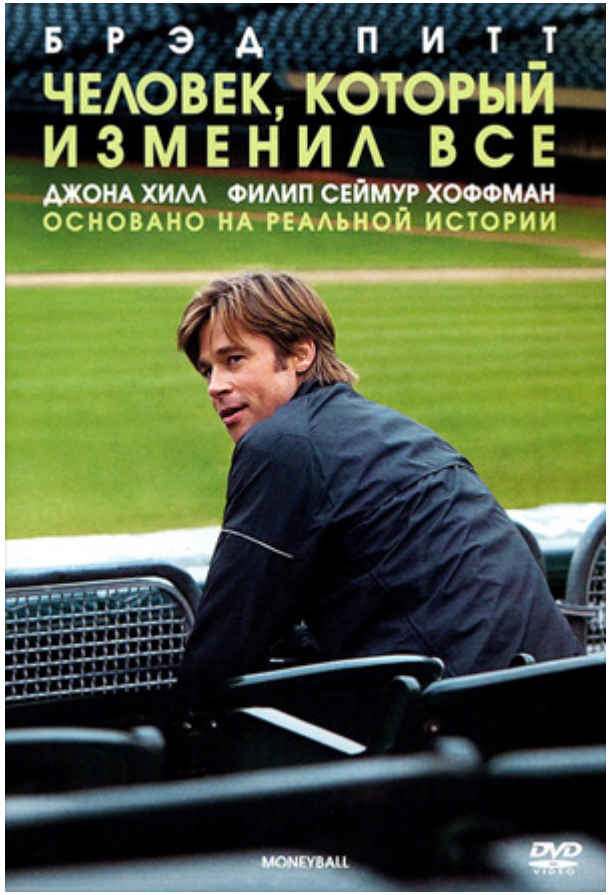
Анализ потребностей рынка труда

Поиск мошенников на основе анализа банковских транзакций

Оптимизация логистики

Прогнозирование поломок техники и аварийных ситуаций на производстве

Применение в спорте



Зачем Вам нужны Большие данные?

1. Аналитик данных - актуальная и востребованная профессия на рынке труда

Зачем Вам нужны Большие данные?

1. Аналитик данных - актуальная и востребованная профессия на рынке труда

2. Современные вызовы общества - междисциплинарные задачи на стыке различных научных областей и компьютерных наук.

Аналитика данных повышает вероятность принятия Ваших публикаций в журналы Q1-Q2

Зачем Вам нужны Большие данные?

1. Аналитик данных - актуальная и востребованная профессия на рынке труда

2. Современные вызовы общества - междисциплинарные задачи на стыке различных научных областей и компьютерных наук.

   Аналитика данных повышает вероятность принятия Ваших публикаций в журналы Q1-Q2

3. Данные лежат в основе цифровизации организаций и управления на основе данных.
   Знания в области аналитики данных помогут Вам в реализации управленческих задач и в формировании необходимой команды специалистов

Кейс №1

Поиск и привлечение талантливых абитуриентов

Университетский консорциум исследователей больших данных

data-diving

Национальный исследовательский Томский государственный университет

ЧГУ ГРОЗНЫЙ

Поиск и привлечение абитуриентов в социальных сетях

Модель целевого абитуриента

Цифровой след в социальной сети

- Цифровые следы абитуриентов в соцсетях
- Открытые пользовательские данные
- Анализ данных с помощью методов машинного обучения и психологии*
- SMM+контент маркетинг

* Большая пятёрка OCEAN: экстраверсия, доброжелательность, добросовестность, нейротизм и открытости опыту.

Методология исследования

Образовательные интересы

- выявление сообществ-маркеров (150+)
- лингвистические маркеры для текстов на стене пользователя

Признаки одарённости

- профдиагностика потенциальных абитуриентов
- выявление корреляции между результатами диагностики и «следом» ВКонтакте

Данные

- Профили ВК 126 000 потенциальных абитуриентов СФО
- Классификация тематических сообществ (100 000+)
- Профдиагностика школьников 3000+

A. Feshchenko, V. Goiko, A. Stepanenko Recruiting university entrants via social networks//EDULEARN17 Proceedings 9th International Conference on Education and New Learning Technologies July 3th-5th, 2017, Barcelona, Spain. – P. 6077 – 6082.

Feshchenko, A., Goiko, V., Matsuta, V., Stepanenko, A., Kiselev, P. Modelling of an educational profile of a student by analyzing public user data from social networks// 12th international technology, education and development conference (INTED), 2018, . – P. 640-646

Результаты проекта

За 2017-2021 гг. проанализированы цифровые следы **1,6 млн.** абитуриентов

14 125 приглашены в ТГУ → **1 918** подали заявления → **1 049** поступило на 77 направления

4+ повышение среднего балла **ЕГЭ**

34% повышение доли **поступивших** от подавших заявления

50% снижение **отчислений** в первый год обучения

36% снижение количества **«троечников»** в 1 год обучения

* В сравнение с абитуриентами, привлекаемыми традиционным рекрутингом

Аналитика для каждого аккаунта

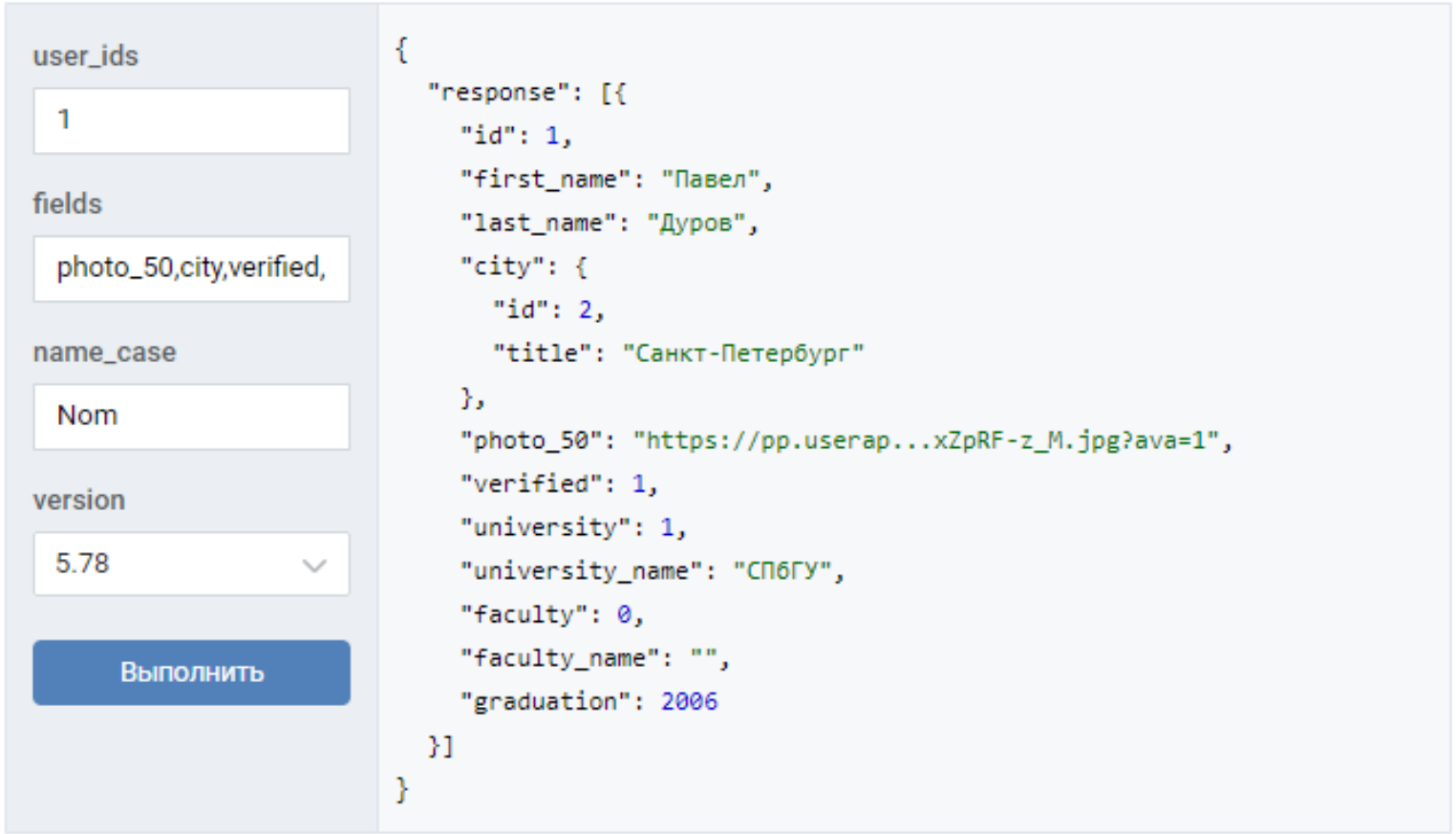
Имя Фамилия:
Всего подписок: 377
Классифицированных: 7.16%

Немного о сборе данных

API (application programming interface)

программный интерфейс приложения, интерфейс прикладного программирования

Пример запроса



Пример вызова метода получения данных профиля пользователя с айди 1.

Формат данных – JSON (JavaScript Object Notation).

Подробная информация по ссылке *https://vk.com/dev/first_guide*

Ограничения

- Access token привязывает сбор данных к ip и к идентификатору пользователя. Масштабирование сбора затрудняется
- Ограничение на 3 запроса в секунду к API Вконтакте
- “Грязные данные”. Существенная доля контента бесполезно для дальнейшего анализа
- Неявные ограничения на многие методы (методы поиска, сбора друзей пользователя и т.д.)

Реально большие данные…

На данный момент в ВК 600 млн аккаунтов

Для выгрузки профилей этих пользователей с 1 айпи нужно:

200 000 000 секунд

Реально большие данные…

На данный момент в ВК 600 млн аккаунтов

Для выгрузки профилей этих пользователей с 1 айпи нужно:

200 000 000 секунд

3 333 333 минут

55 555 часов

2 314 дней

~ 6 лет ☺

Выявление личностных характеристик

Исследование #1: личность через текст

Schwartz H. A. et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. 2013.

Была прослежена зависимость между психологическими чертами личности и используемым лексиконом. Метод Differential Language Analysis (DLA), основанный на подходе открытого словаря.

Например, люди, говорящие о совместном досуге (спорте, отдыхе, пляже), отличаются высокой эмоциональной стабильностью. Интроверты чаще увлекаются современной японской культурой (аниме, манга, используют каомодзи: ^_^).

Исследование #2 : психотип через лайки

Kosinski M. et al. Manifestations of user personality in website choice and behaviour on online social networks. 2014.

Выявлен феномен «цифровой тени».
Для исследования был использован массив данных 350 тысяч американских пользователей Facebook, участвовавших в проекте myPersonality. Модель «Большой Пятёрки» (OCEAN).

С наибольшей точностью можно предсказать возраст, далее идут экстраверсия, нейротизм, добросовестность и открытость новому опыту. Предсказать уровень доброжелательности оказалось сложнее всего.

Исследование #3: одарённость и подписки

Гойко В., Киселев П., Мацута В., Фещенко А. Исследование потенциала социальных сетей для выявления одаренных старшеклассников *.

Многомерные модели одаренности Дж. Рензулли, К. Хеллера.

Методы сбора данных: психологическое тестирование.

Методы обработки и анализа данных: процентильная (нелинейная) нормализация, машинное обучение (бинарная классификация на основе моделей: метод опорных векторов, случайные леса и градиентный бустинг).

* Исследование потенциала социальных сетей для выявления одаренных старшеклассников/ Психология и Психотехника. — 2017. - № 4.

Алгоритмы классификации школьников по результатам тестов на основе анализа их подписок

интеллект

креативность

мотивация

Шкала субтеста	Метод опорных векторов	Случайные леса	Градиентный бустинг
Аналогии	0,78	0,72	0,76
Дивергентный стиль	0,64	0,60	0,62
Конвергентный стиль	0,70	0,63	0,67
Беглость	0,67	0,61	0,65
Семантическая гибкость	0,69	0,64	0,69
Оригинальность	0,71	0,64	0,69
Креативное поведение	0,71	0,62	0,68
Инициативность	0,69	0,65	0,70
Решительность	0,71	0,67	0,70
Настойчивость	0,68	0,60	0,65
Мотив самореализации	0,76	0,68	0,72
Социальные навыки	0,74	0,69	0,73

В КАКОМ СООБЩЕСТВЕ СОСТОИТ БОЛЬШЕЕ КОЛИЧЕСТВО ШКОЛЬНИКОВ С ВЫСОКИМ УРОВНЕМ ИНТЕЛЛЕКТА?

1. ИНСТИТУТ БЛАГОРОДНЫХ РОКЕРШ
2. НАУКА ДНЯ

В КАКОМ СООБЩЕСТВЕ СОСТОИТ БОЛЬШЕЕ КОЛИЧЕСТВО ШКОЛЬНИКОВ С ВЫСОКИМ УРОВНЕМ ИНТЕЛЛЕКТА?

ИНСТИТУТ БЛАГОРОДНЫХ РОКЕРШ

НАУКА ДНЯ

СТАРШЕКЛАССНИКИ ИЗ КАКОЙ ГРУППЫ БОЛЕЕ КРЕАТИВНЫ?

1 БЕЗ КОТА И ЖИЗНЬ НЕ ТА

2 ПСИХОЛОГИЯ ТИШИНЫ

СТАРШЕКЛАССНИКИ ИЗ КАКОЙ ГРУППЫ БОЛЕЕ КРЕАТИВНЫ?

БЕЗ КОТА И ЖИЗНЬ НЕ ТА

ПСИХОЛОГИЯ ТИШИНЫ

В КАКУЮ ГРУППУ ВСТУПАЮТ БОЛЕЕ МОТИВИРОВАННЫЕ ДЕТИ?

1. АКАДЕМИЯ ПОРЯДОЧНЫХ ДЕВИЦ
2. ИСКУССТВО РЕАЛЬНОСТИ

В КАКУЮ ГРУППУ ВСТУПАЮТ БОЛЕЕ МОТИВИРОВАННЫЕ ДЕТИ?

АКАДЕМИЯ ПОРЯДОЧНЫХ ДЕВИЦ

ИСКУССТВО РЕАЛЬНОСТИ

Спасибо за внимание!

Вячеслав Гойко
goiko@data.tsu.ru

data-diving

Университетский консорциум исследователей больших данных