

Прикладные исследования в культуре

Лекция 1.

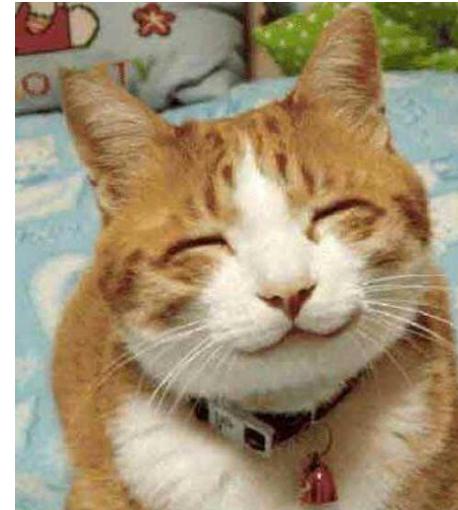
Никишина Ирина
сентябрь-октябрь 2020

О чем этот курс

- Источники данных
- Инструменты работы с данными
- Визуализация данных

О чем этот курс

- ~~Источники данных~~ Я смогу скачать интернет!
- ~~Инструменты работы с данными~~ Ура, я буду программистом!
- ~~Визуализация данных~~ Я смогу делать красивые графики!



Организация курса

- 7 лекций + 7 семинаров
- Формула оценки: **0.3*ДЗ₁ + 0.3*ДЗ₂ + 0.4*Проект**
- Посещение не учитывается

Организация курса

- Связь
 - telegram: @lilas_pourpre, @dolyasergey
 - mail: lilas.pourpre@gmail.com
 - Материалы курса:
https://github.com/lilaspourpre/hse_culture_python2020
 - Чат
 - <https://t.me/joinchat/Ckja7RlhqvMaRPn4B8mVMQ>
 - Wiki
 - http://wiki.cs.hse.ru/%D0%9F%D1%80%D0%B8%D0%BA%D0%BB%D0%B0%D0%B4%D0%BD%D1%8B%D0%B5_%D0%B8%D1%81%D1%81%D0%BB%D0%B5%D0%B4%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F_%D0%B2_%D0%BA%D1%83%D0%BB%D1%8C%D1%82%D1%83%D1%80%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D0%B8 2020 1%D0%BC%D0%BE%D0%B4%D1%83%D0%BB%D1%8C

Организация курса

- ДЗ₁ – 20 сентября (дедлайн)
- ДЗ₂ – 4 октября (дедлайн)
- Описание проекта и критерии: [тут](#)

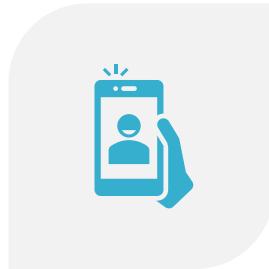
План занятий

	Лекции	Семинары
1	Данные. Основные способы представления данных.	Введение в Python. Типы данных.
2	Python как инструмент анализа данных	Функции и классы в Python.
3	Сбор и предварительная обработка данных.	Сбор данных в Python. Регулярные выражения
4	Python как инструмент визуализации данных	Основы визуализации данных в Python
5	Анализ текстовых данных	Анализ текстовых данных в Python
6	Анализ социальных сетей. Социальные графы	Анализ социальных сетей в Python. Социальные графы
7	Работа с исследовательским проектом.	Представление и презентация проекта

О чём речь пойдет сегодня



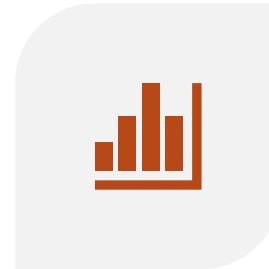
ПОНЯТИЕ DIGITAL HUMANITIES.



СФЕРЫ ПРИМЕНЕНИЯ АНАЛИЗА И
ВИЗУАЛИЗАЦИИ ГУМАНИТАРНЫХ
И СОЦИАЛЬНЫХ НАУКАХ.



ПОНЯТИЕ ДАННЫХ.

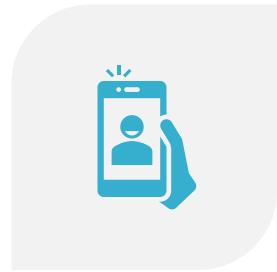


ОСНОВНЫЕ СПОСОБЫ
ПРЕДСТАВЛЕНИЯ ДАННЫХ.

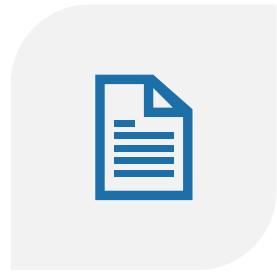
Содержание



ПОНЯТИЕ DIGITAL HUMANITIES.



СФЕРЫ ПРИМЕНЕНИЯ АНАЛИЗА И
ВИЗУАЛИЗАЦИИ ГУМАНИТАРНЫХ
И СОЦИАЛЬНЫХ НАУКАХ.



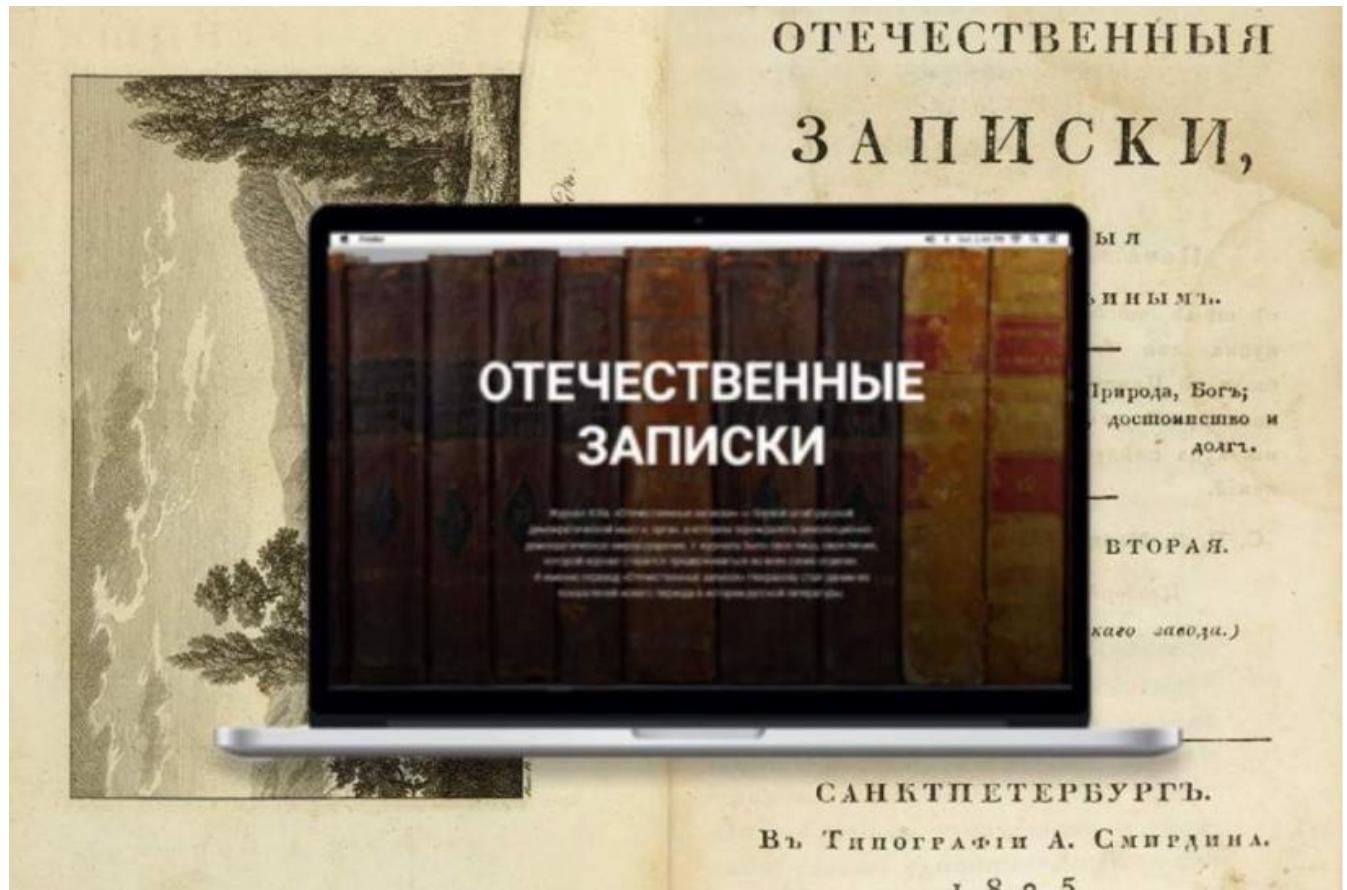
ПОНЯТИЕ ДАННЫХ.



ОСНОВНЫЕ СПОСОБЫ
ПРЕДСТАВЛЕНИЯ ДАННЫХ.

Digital Humanities

1. перевод в
электронный
машиночитаемый
формат источников,
связанных с
культурным
наследием



<https://hum.hse.ru/digital/news/374746005.html>

Digital Humanities

2. семантическая разметка значимых элементов

(не)прямая речь

Главная Золотой корпус О проекте Документация

пример разметки

"Позвольте мне вам представить жену мою", сказал Манилов. "Душенька, Павел Иванович!"

Тэги

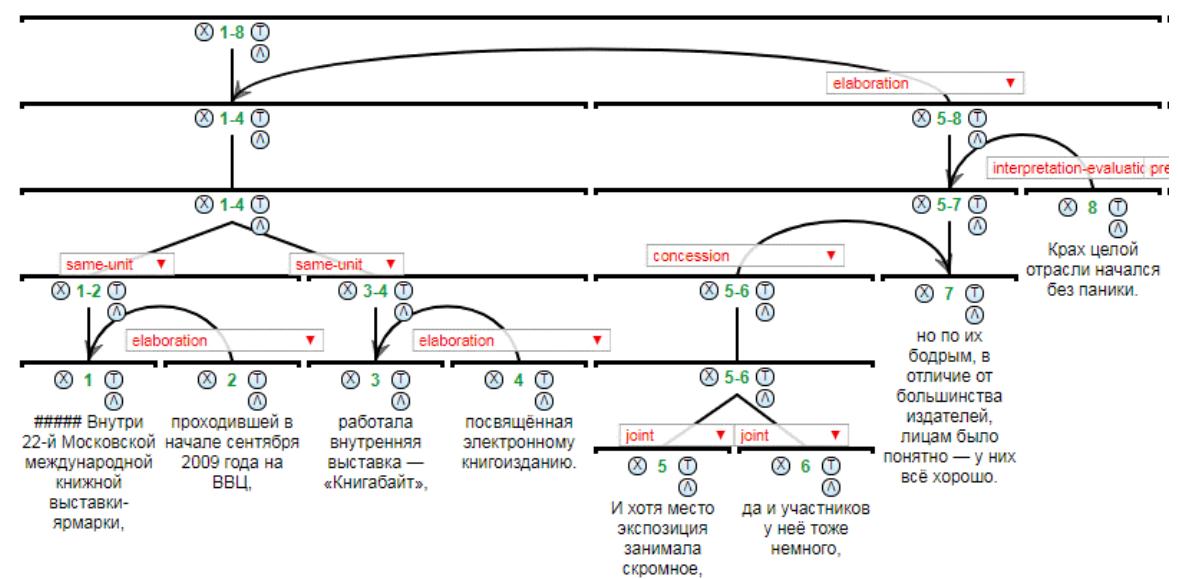
прямая речь слова автора глагол речи говорящий

Загрузить свой текст для разметки

Вы можете загрузить свой текст и работать с ним онлайн или скачать размеченный текст

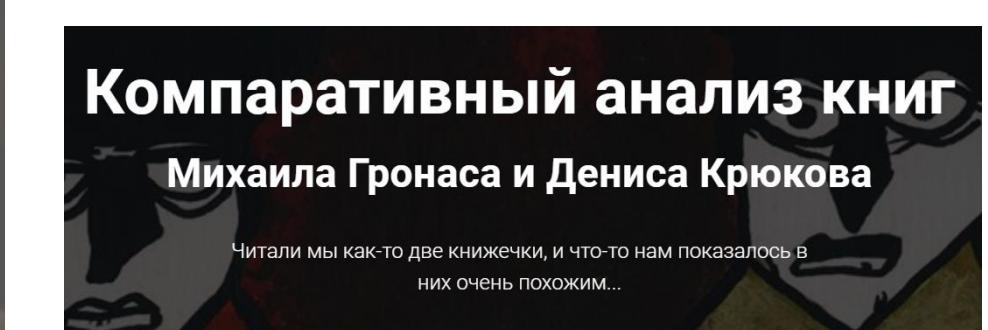
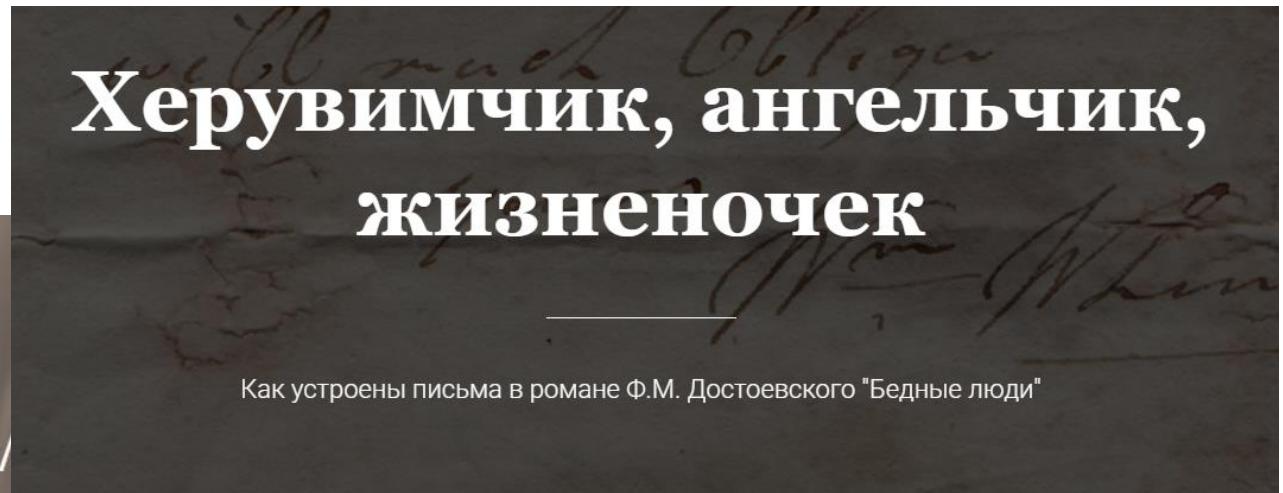
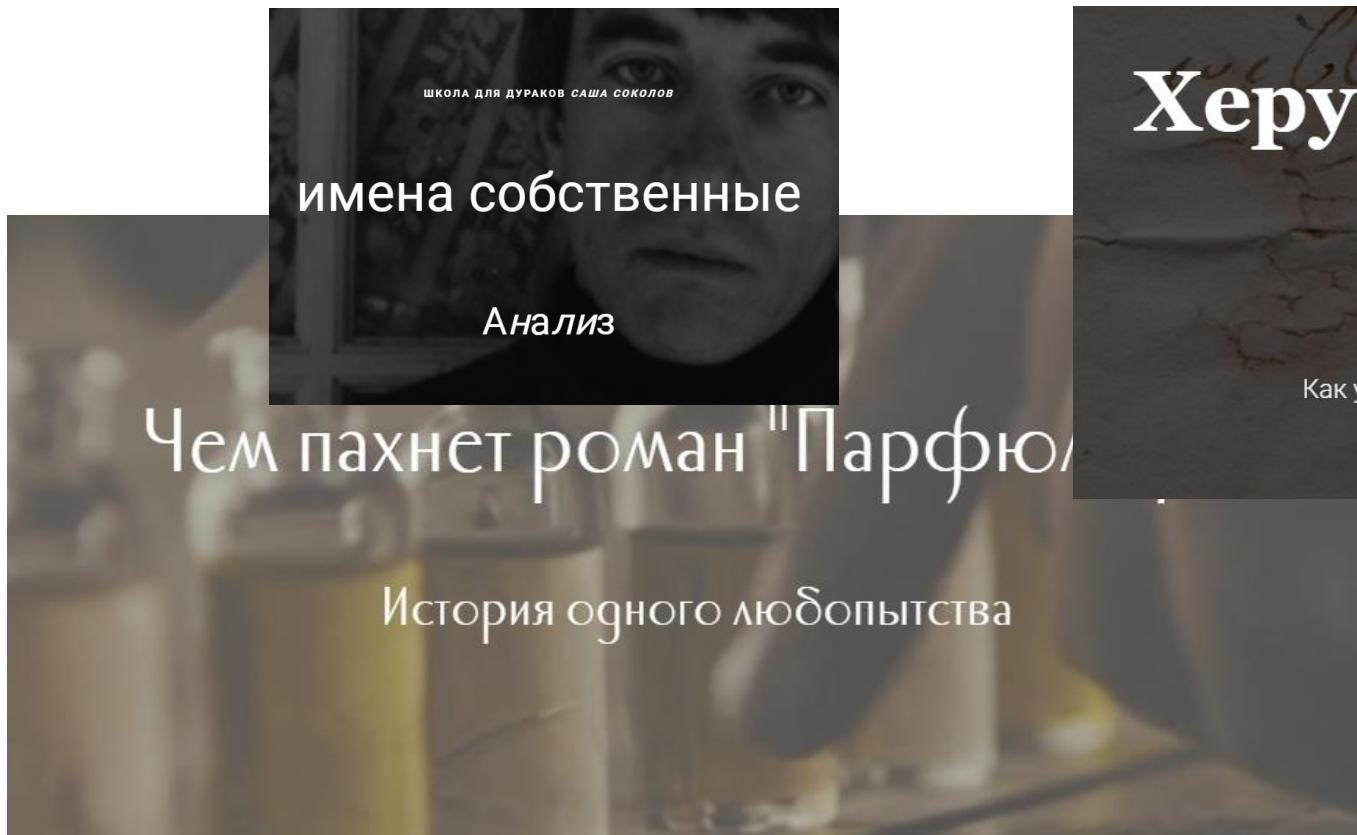
File: Choose file No file chosen

Загрузить файл



Digital Humanities

3. квантиативный анализ художественных текстов



"Поэт, мечтатель, хиромант..." :
исследование корпуса текстов П. Д.
Когана

Digital Humanities

4. популяризация гуманитарного знания через визуализации и разработку электронных продуктов



From: Толстой

944 members

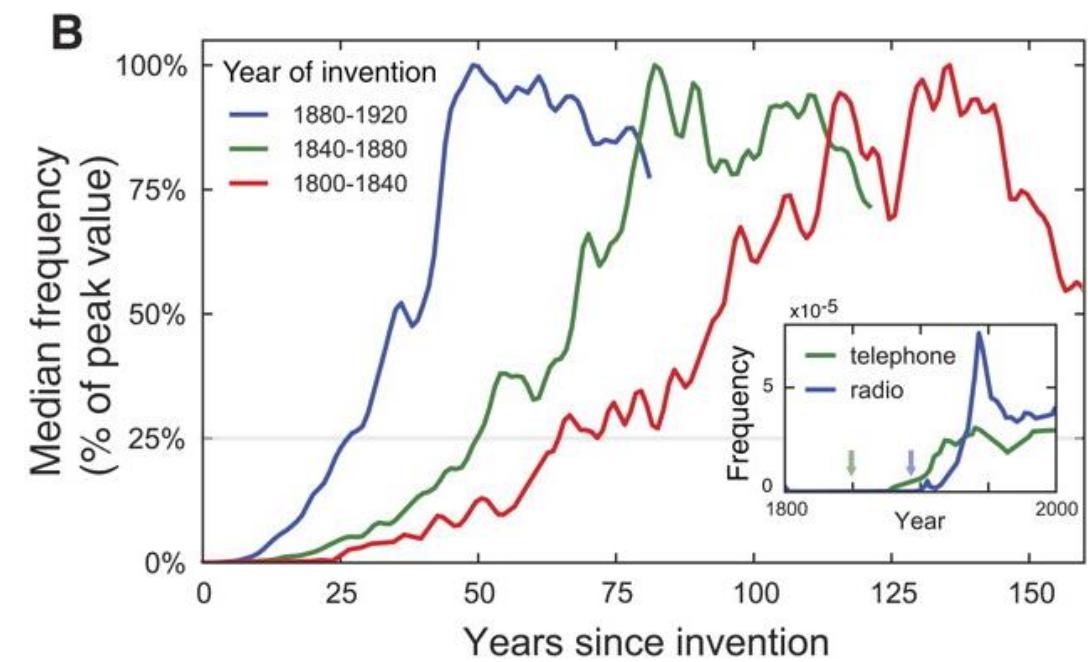
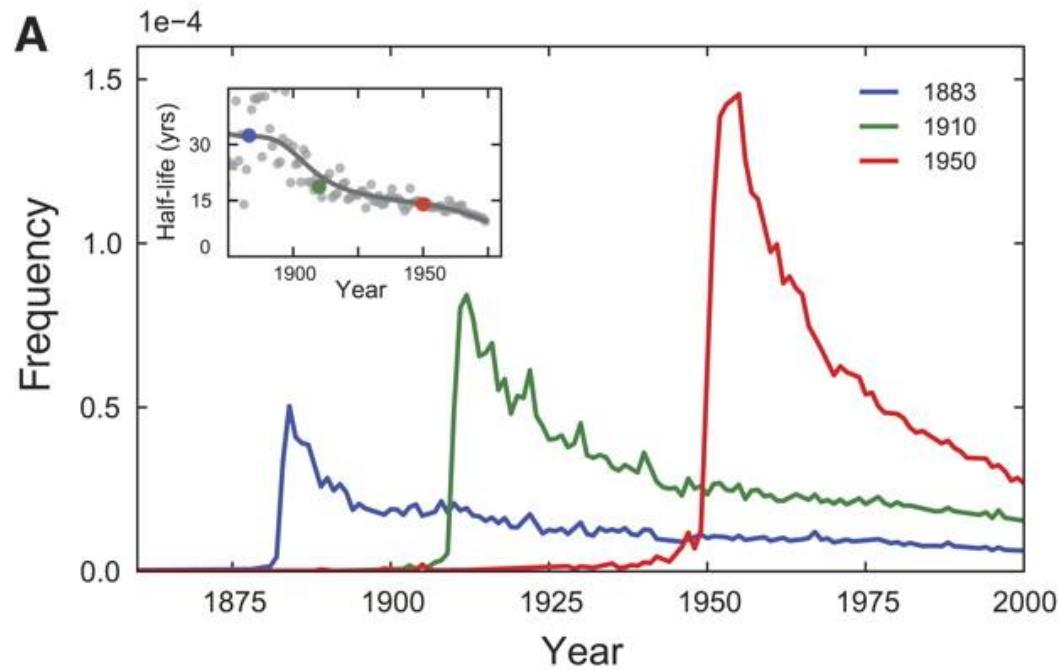
Лев Толстой как хипстер, любитель ЗОЖ,
тусовщик, прокрастинатор и
оппозиционер.

В письмах великого писателя – его...

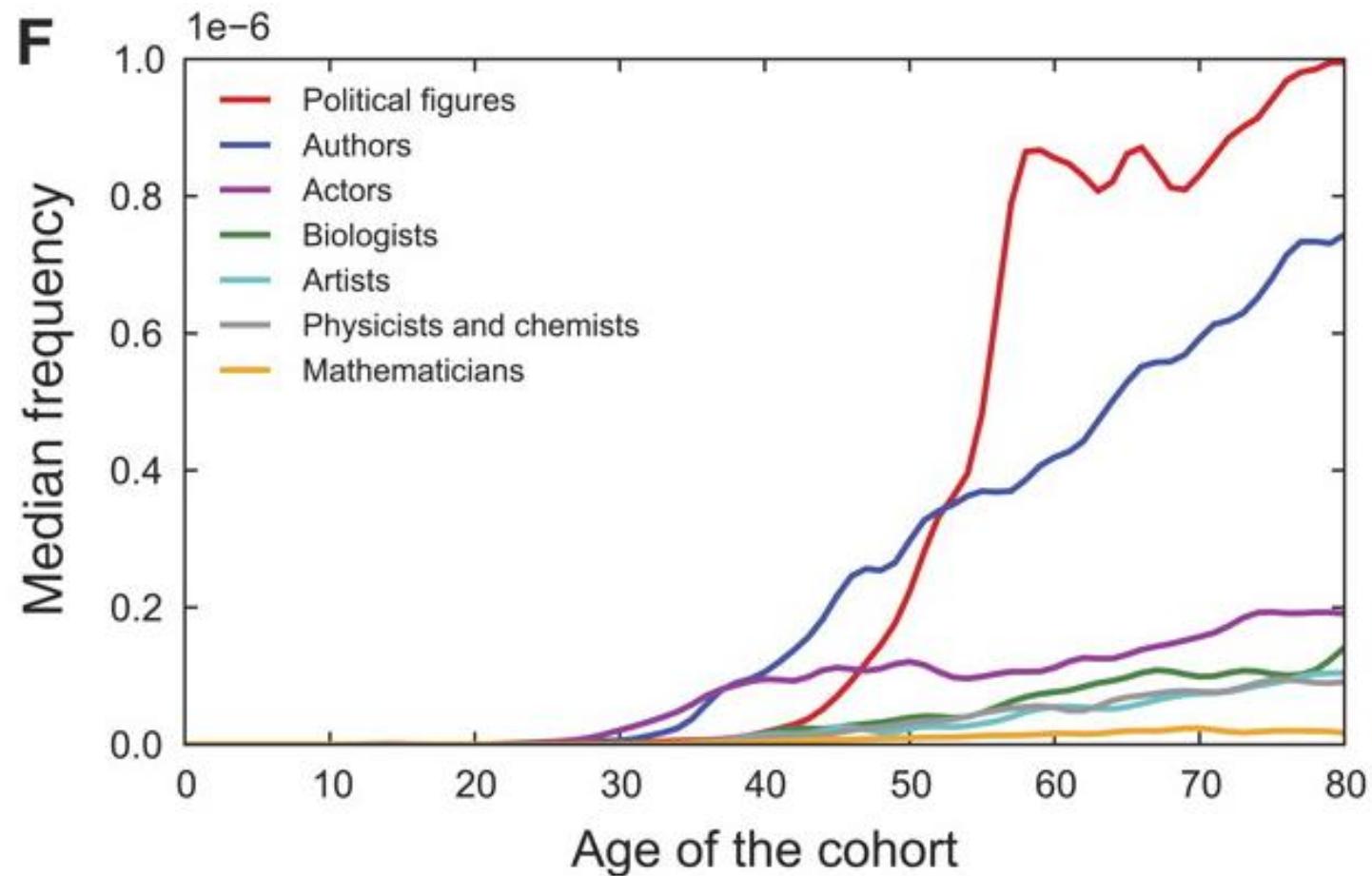
[VIEW IN TELEGRAM](#)

[OPEN IN WEB](#)

Ускорение истории

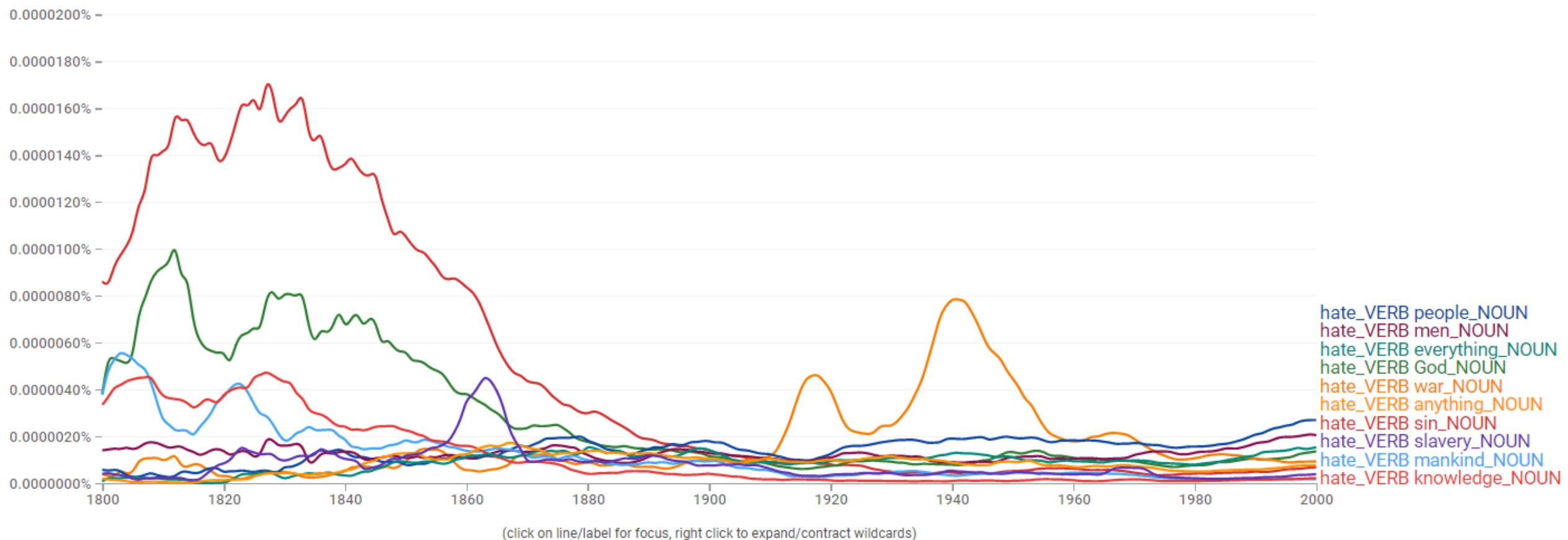


Возраст славы в разных профессиях



Google Books Ngram Viewer

⋮

 hate_VERB *_NOUN[X](#) [?](#)[1800 - 2000](#) ▾[English \(2019\)](#) ▾[Case-Insensitive](#)[Smoothing](#) ▾

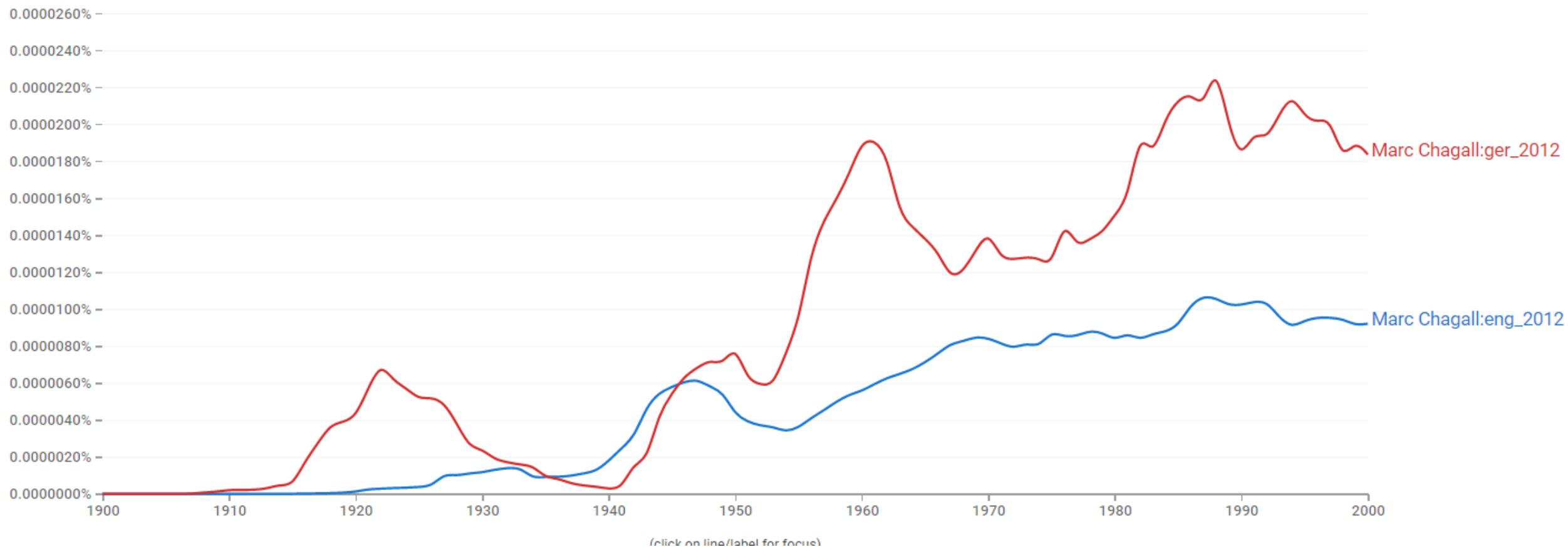
Marc Chagall:eng_2012,Marc Chagall:ger_2012

1900 - 2000 ▾

English (2012) ▾

Case-Insensitive

Smoothing ▾



Google Books Ngram Viewer

Ленин, Сталин

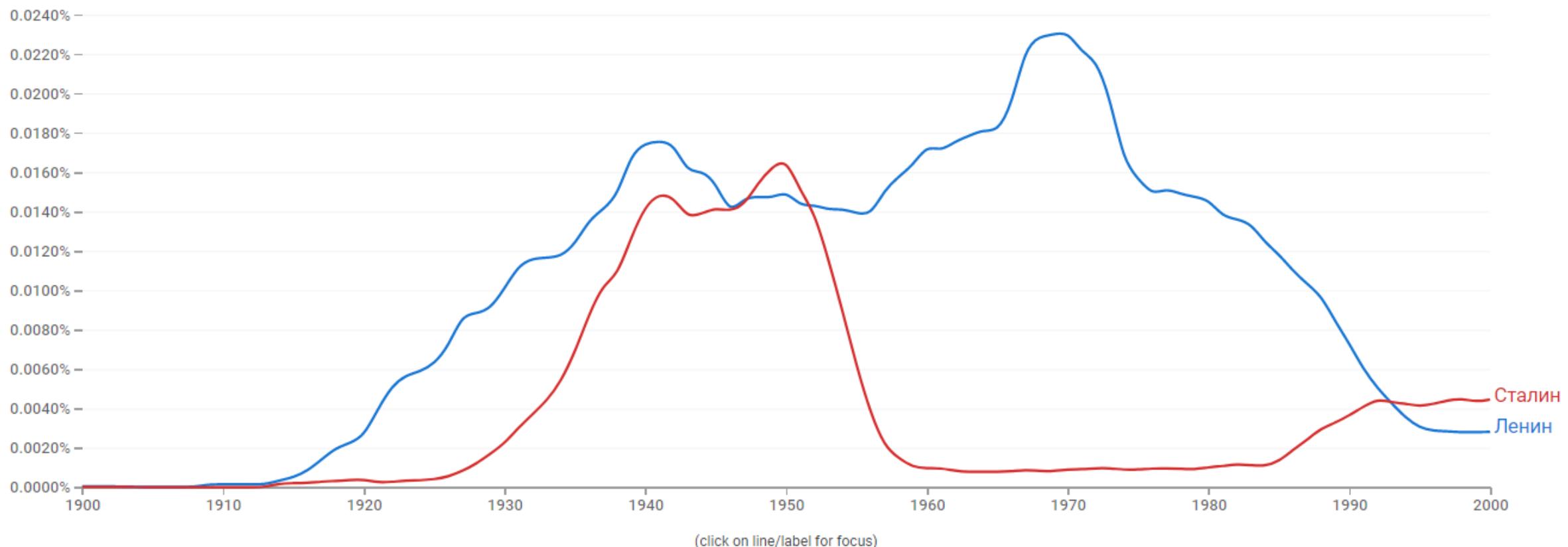
X ?

1900 - 2000 ▾

Russian (2012) ▾

Case-Insensitive

Smoothing ▾



Стилеметрия / стилометрия

У авторского стиля есть
осознаваемая и
неосознаваемая
составляющая

Неосознаваемая
составляющая может быть
измерена и служить
«отпечатком пальца»
автора

«Уже двести лет не прекращается дискуссия о том, что представляет собой «Слово о полку Игореве», — подлинное древнерусское произведение или искусственную подделку под древность, созданную в XVIII веке. <...>

Гибель единственного списка этого произведения лишает исследователей возможности произвести анализ почерка, бумаги, чернил и прочих материальных характеристик первоисточника.

Наиболее прочным основанием для решения проблемы подлинности или поддельности «Слова о полку Игореве» оказывается в таких условиях язык этого памятника».

— А.А. Зализняк. "Слово о полку Игореве": взгляд лингвиста.

А какие случаи
спорного или
поддельного авторства
знаете вы?

Первопроходец стилометрии

Lorenzo Valla (1407 – 1457)

Итальянский гуманист, риторик,
католический священник

В 1439 году доказал, что
Константинов дар является
подделкой

Показал, что она не могла быть
написана в эпоху Константина I (IV
век), так как её стиль датируется XIII
веком



Где может быть отпечаток пальца?



Что можно узнать, кроме авторства текста?



Датировка



Сравнение
жанров



Сравнение мужских
и женских текстов



Сравнение
оригиналов
и переводов



Исследования
«стилома» человека
(т.н. идиостиль)



Лингвистическая
экспертиза

Содержание



ПОНЯТИЕ DIGITAL HUMANITIES.



СФЕРЫ ПРИМЕНЕНИЯ АНАЛИЗА И
ВИЗУАЛИЗАЦИИ ГУМАНИТАРНЫХ
И СОЦИАЛЬНЫХ НАУКАХ.

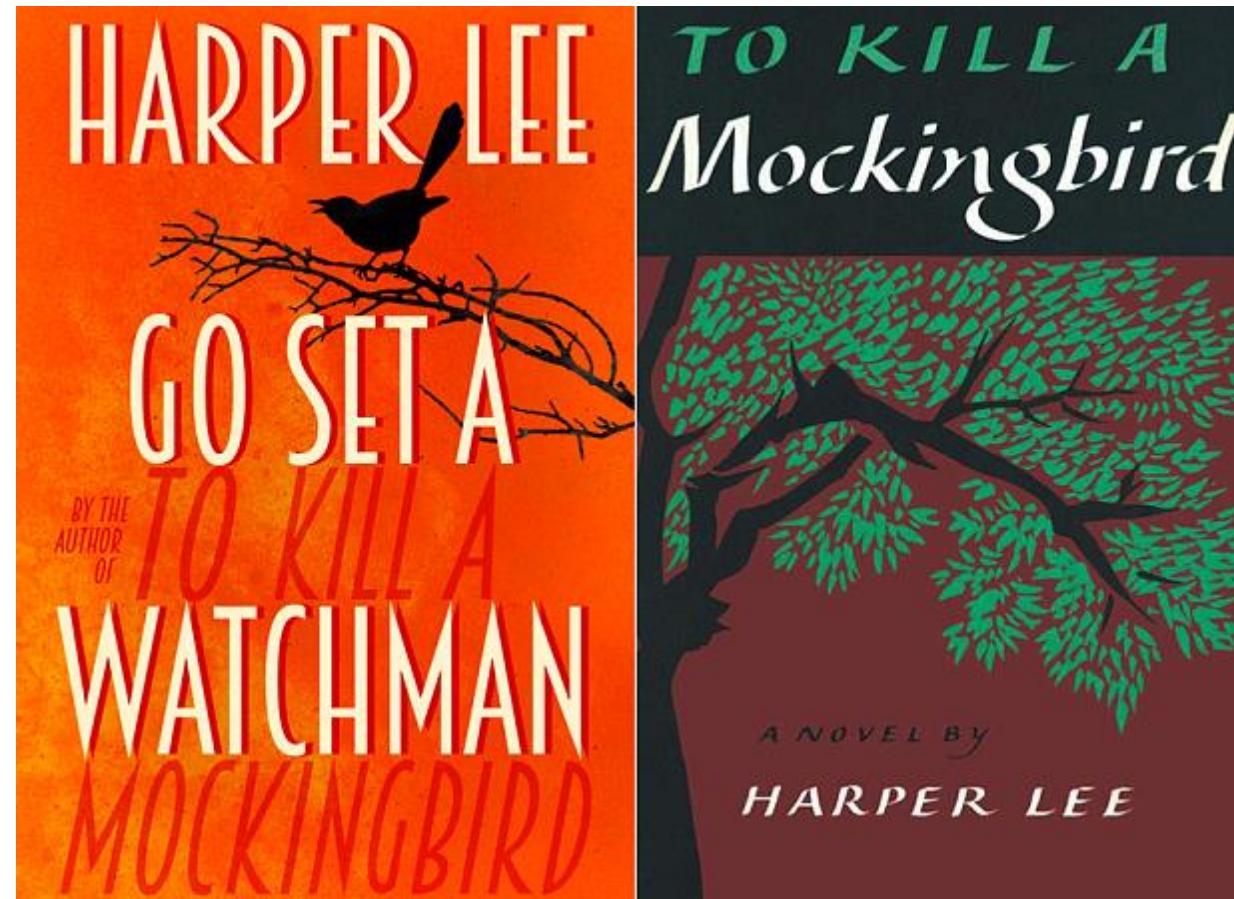


ПОНЯТИЕ ДАННЫХ.



ОСНОВНЫЕ СПОСОБЫ
ПРЕДСТАВЛЕНИЯ ДАННЫХ.

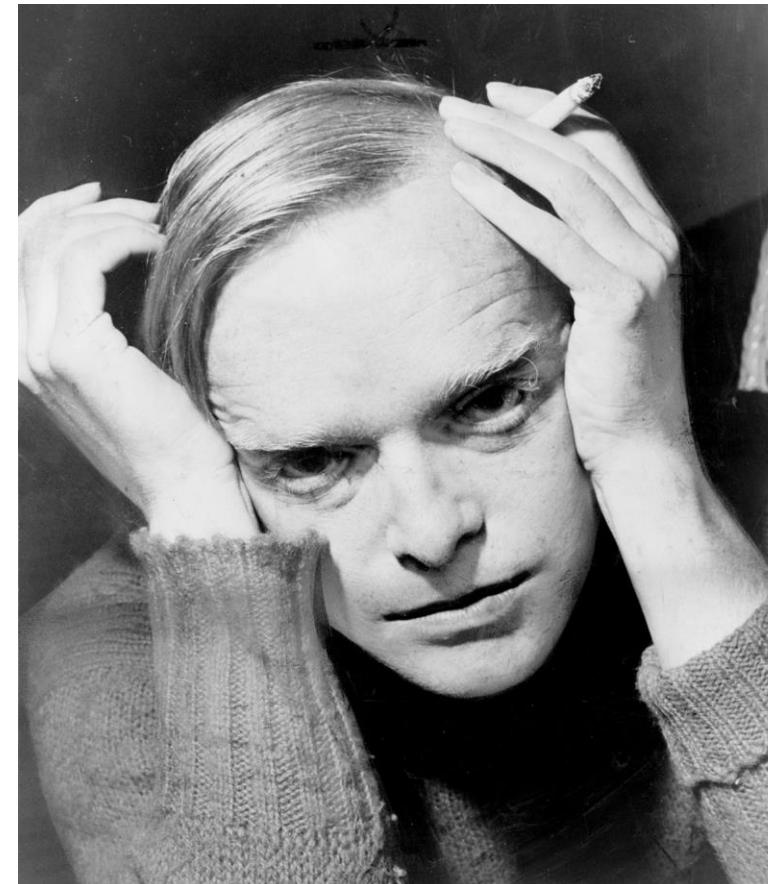
Кто написал «Убить пересмешника»?



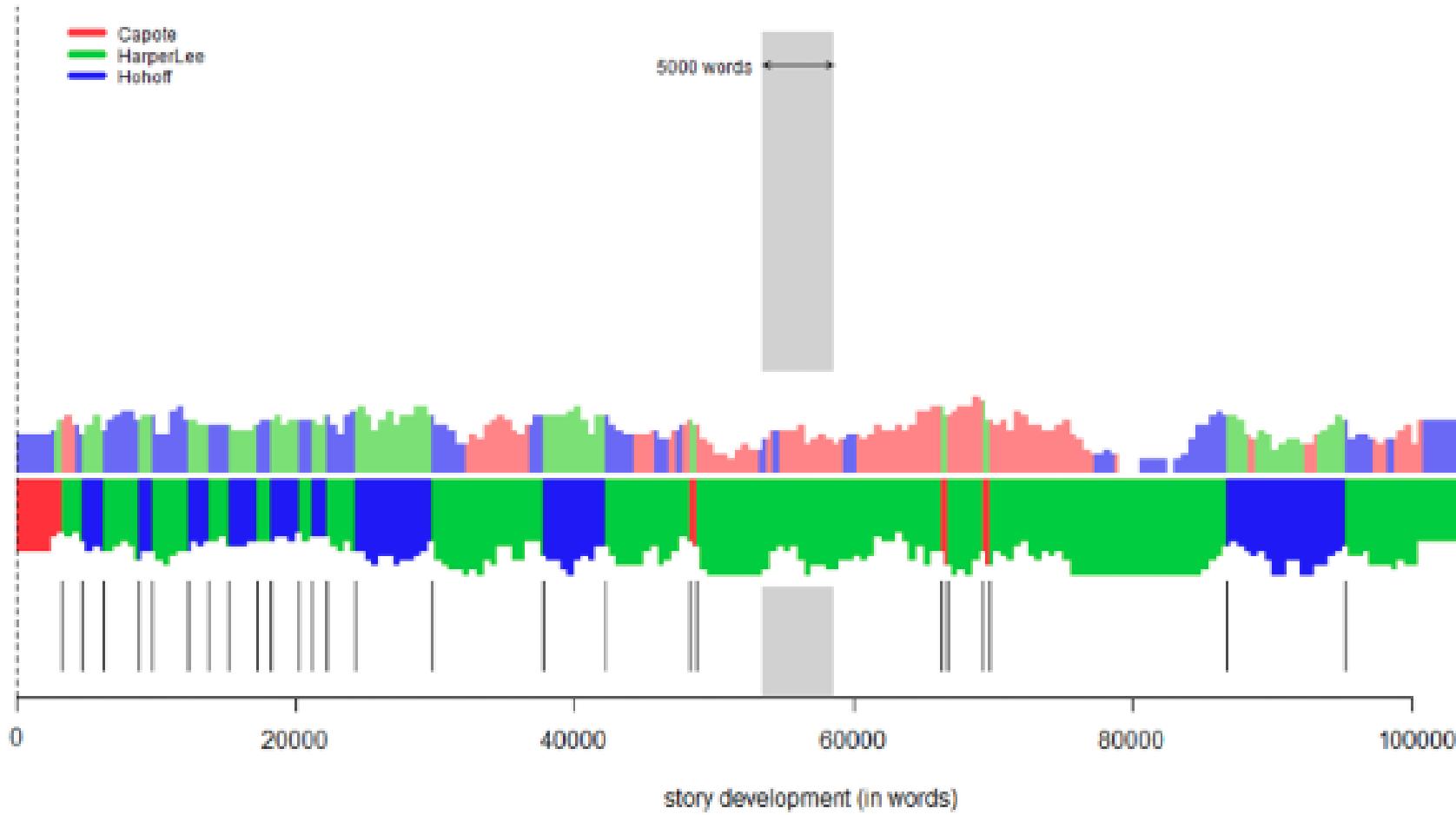
Харпер Ли и другие



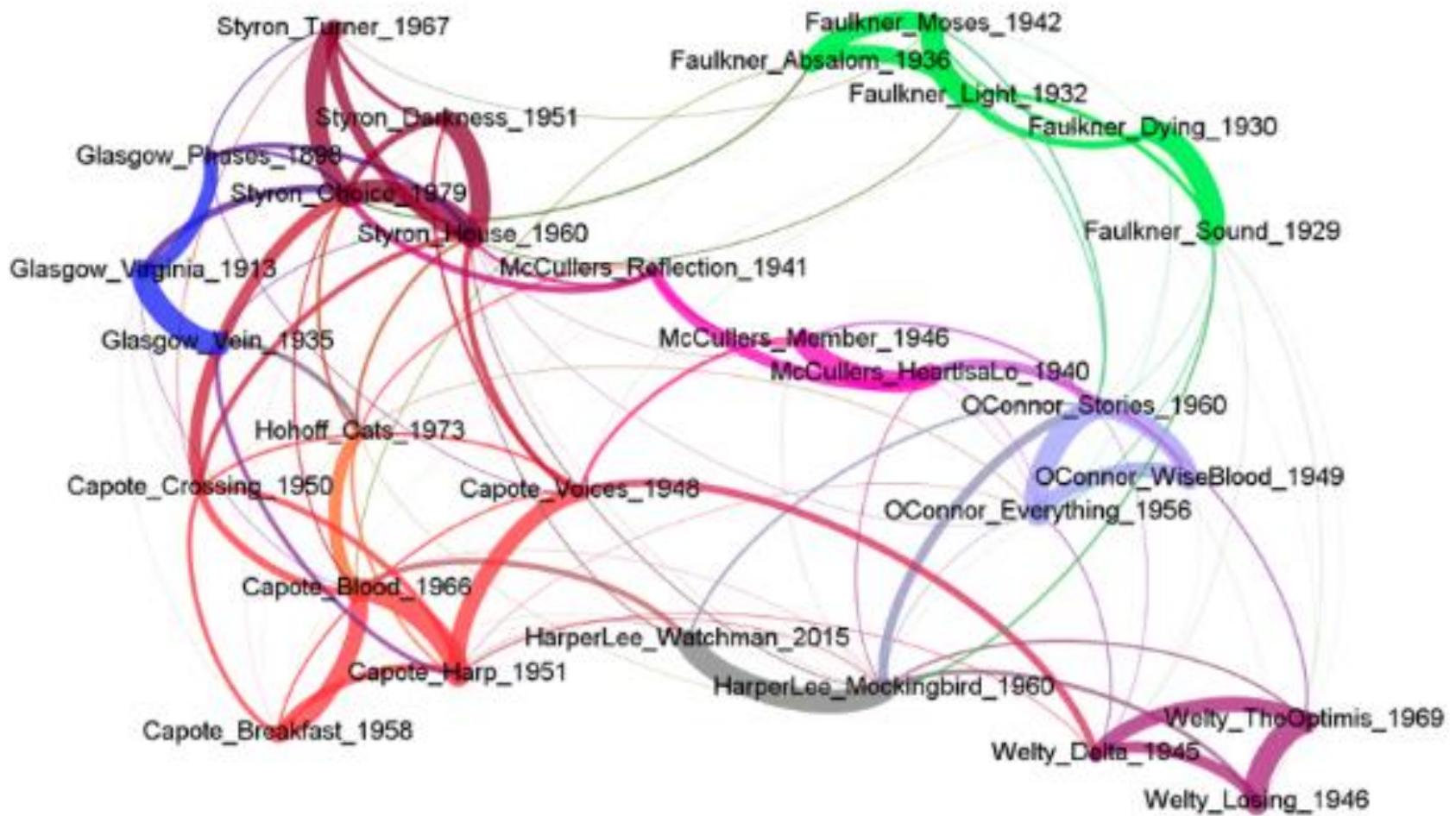
Tay Hohoff, редактор



Truman Capote, писатель и друг Харпер Ли



Maciej Eder, Jan Rybicki (2016)

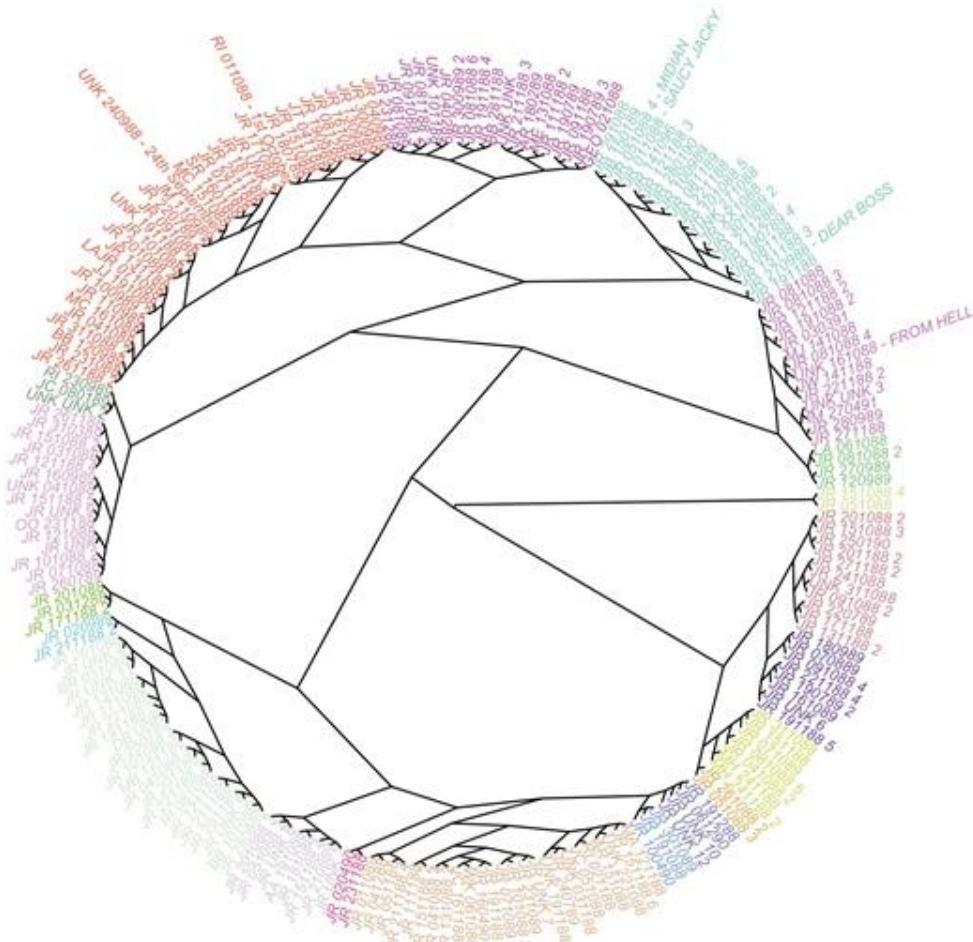


Maciej Eder, Jan Rybicki (2016)

Статистическая стилометрия и литературный рынок

- http://www.cs.stonybrook.edu/~songfeng/papers/emnlp2013_success.pdf

Потрошите письма “Джека Потрошителя”



<https://vk.com/@sysblok-potroshim-pisma-dzheka-potroshitelya>

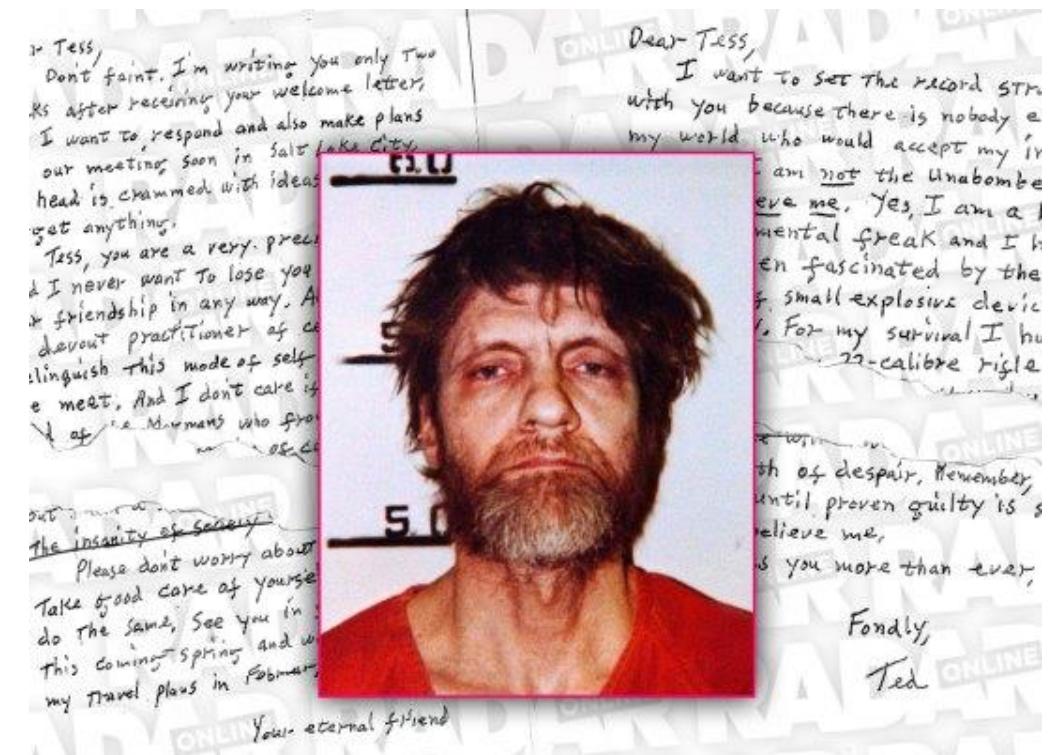
За границей литературы

Теодор Казински

В период с 1978 по 1995 год совершил ряд взрывов бомб в университетах и авиалиниях

Обещал прекратить, если его антииндустриальный "манифест" будет опубликован в крупных газетах

Отличительный стиль письма, обороты и фразы позволили его идентифицировать



Ничего святого: отделить Маккартни от Леннона

<https://vk.com/@sysblok-nichego-svyatogo-matematiki-otdelili-makkartni-ot-lennona>

Даже стиль кода!

De-anonymizing Programmers via Code Stylometry

Aylin Caliskan-Islam

Drexel University

Arvind Narayanan

Princeton University

Richard Harang

U.S. Army Research Laboratory

Clare Voss

U.S. Army Research Laboratory

Rachel Greenstadt

Drexel University

Andrew Liu

University of Maryland

Fabian Yamaguchi

University of Goettingen

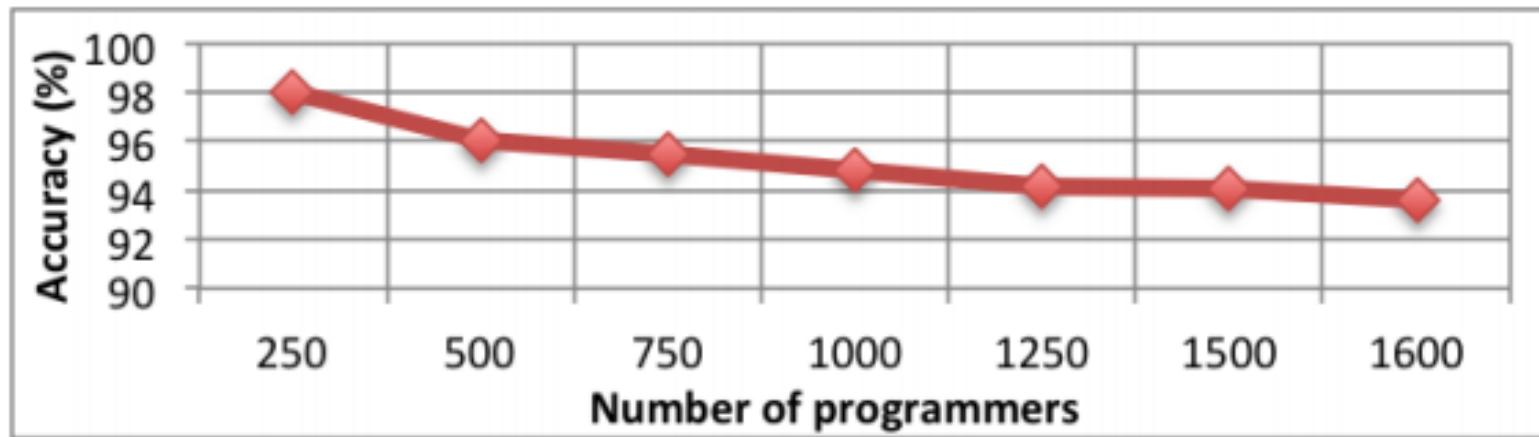
Abstract

Source code authorship attribution is a significant privacy threat to anonymous code contributors. However, it may also enable attribution of successful attacks from code left behind on an infected system, or aid in resolving copyright, copyleft, and plagiarism issues in the programming fields. In this work, we investigate machine learning methods to de-anonymize source code authors of C/C++ using coding style. Our Code Stylometry Fea-

grammer was sentenced to death in 2012 for developing photo sharing software that was used on pornographic websites [31].

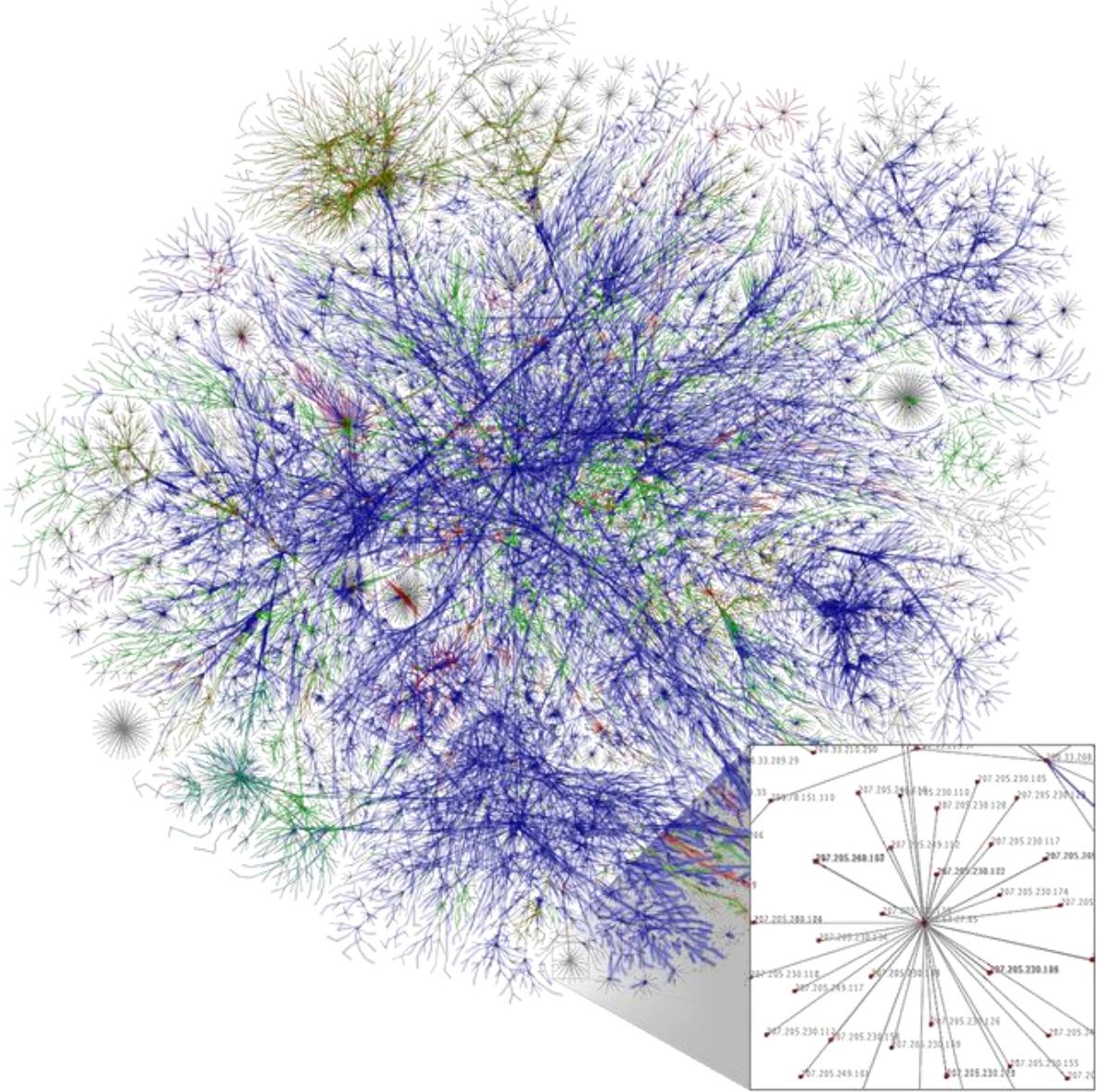
The flip side of this scenario is that code attribution may be helpful in a forensic context, such as detection of ghostwriting, a form of plagiarism, and investigation of copyright disputes. It might also give us clues about the identity of malware authors. A careful adversary may only leave binaries, but others may leave behind code written in a scripting language or source code down-

Деанонимизация программистов

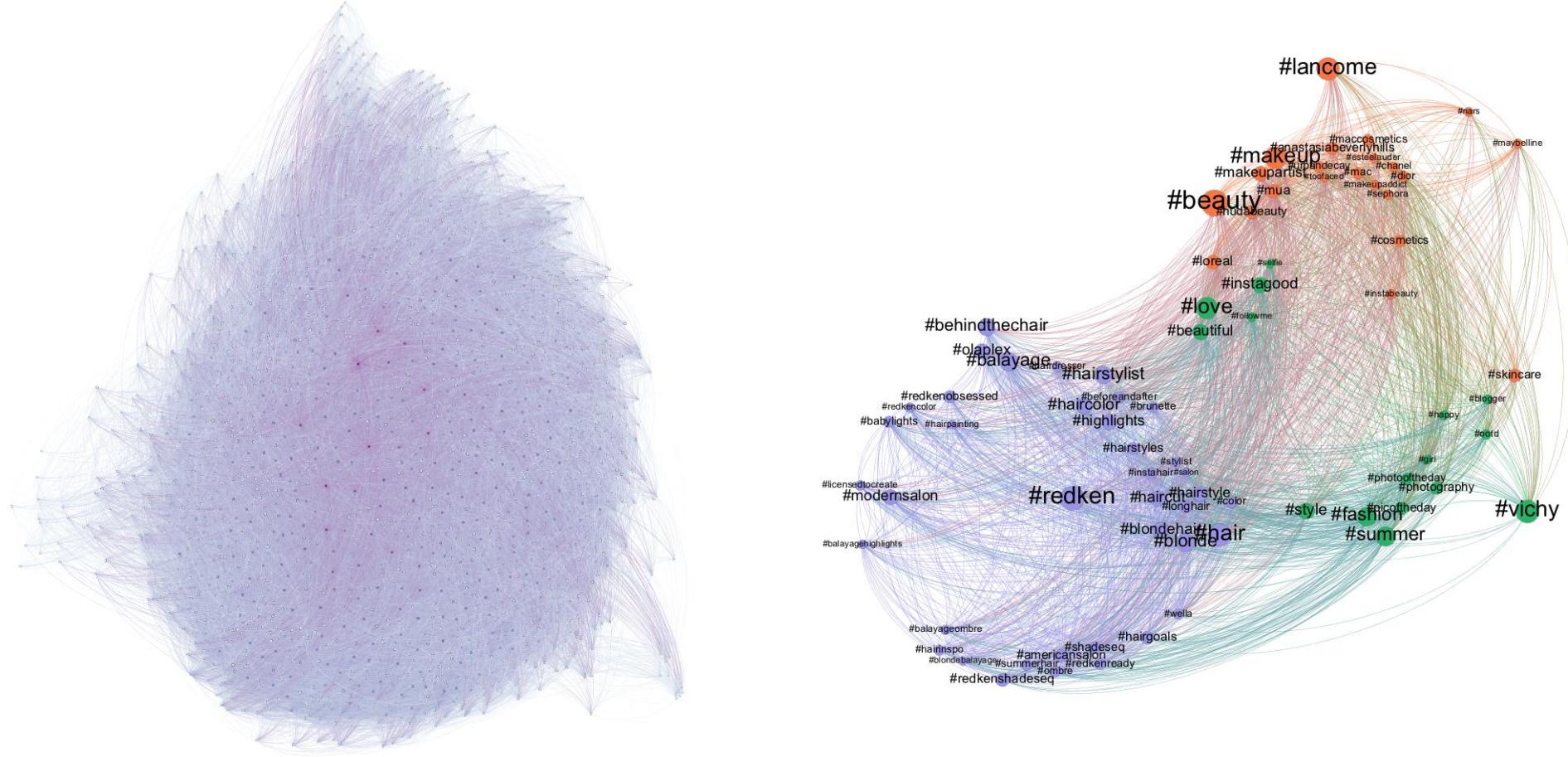


<https://habr.com/ru/post/406297/>

<https://habr.com/ru/post/274533/>

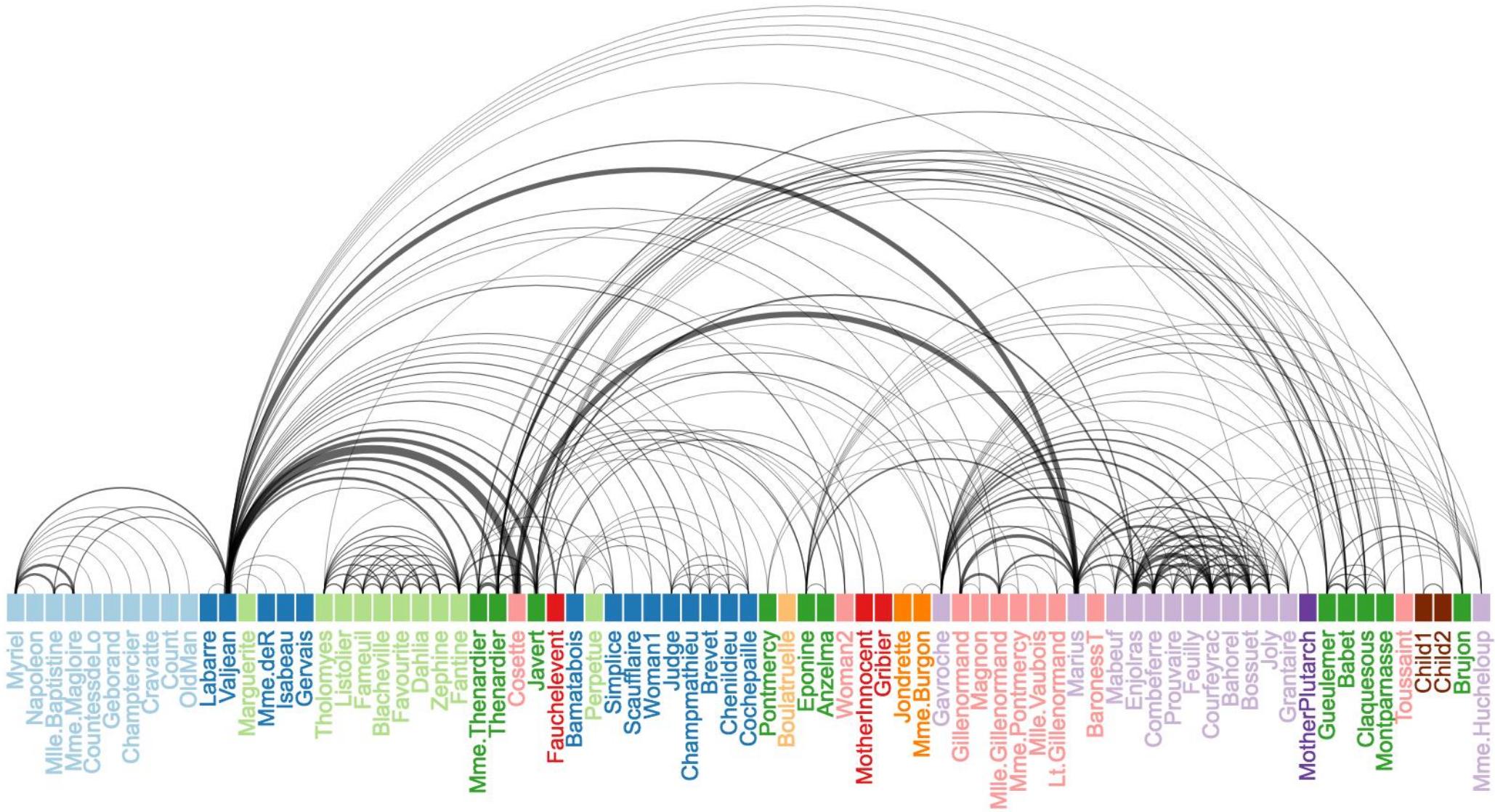


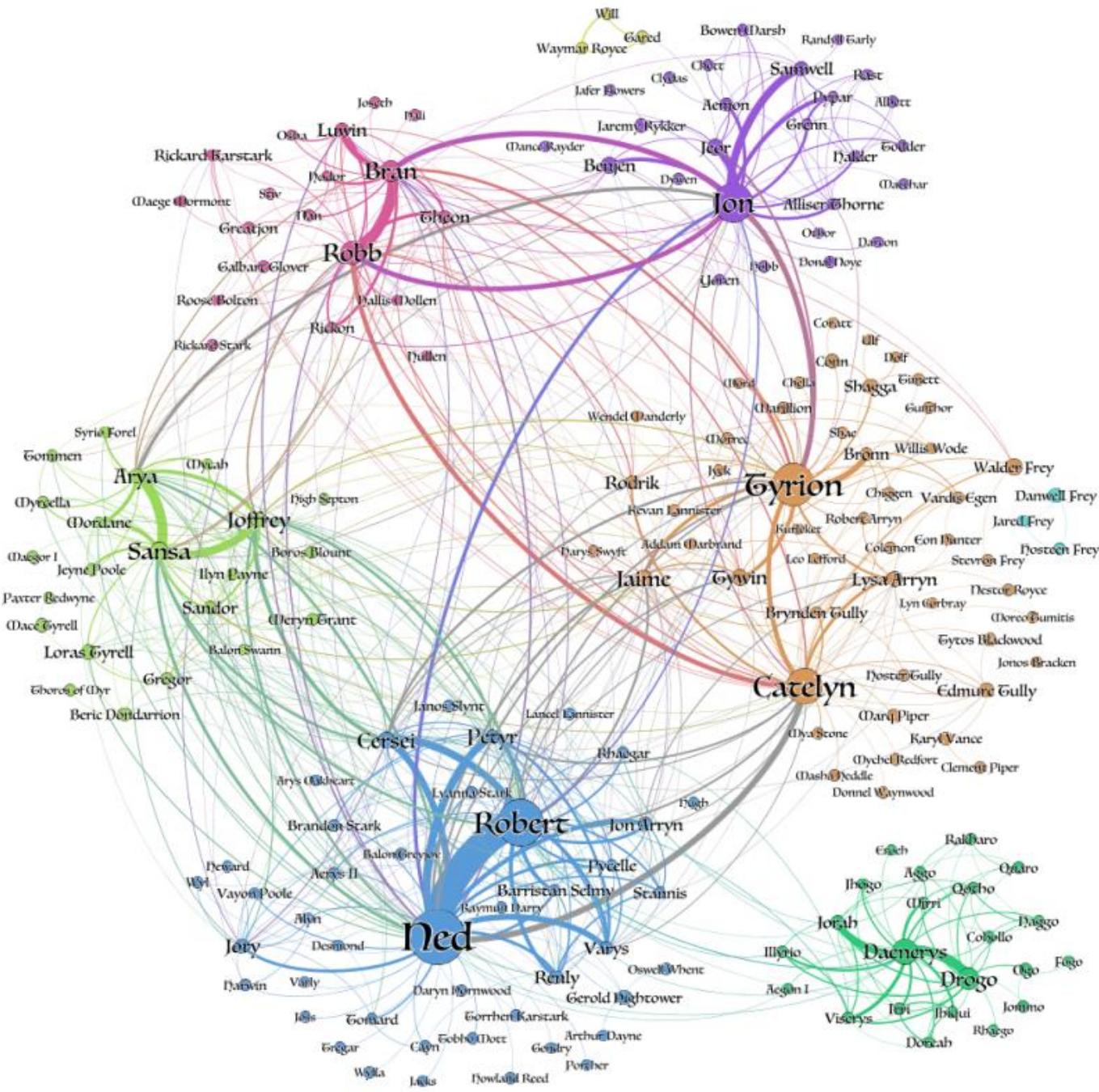
Анализ социальных сетей



<https://towardsdatascience.com/social-network-analysis-of-related-hashtags-on-instagram-using-instacrawler-46c397cb3dbe>

Character Co-occurrence in Les Miserables





Содержание



ПОНЯТИЕ DIGITAL HUMANITIES.



СФЕРЫ ПРИМЕНЕНИЯ АНАЛИЗА И
ВИЗУАЛИЗАЦИИ ГУМАНИТАРНЫХ
И СОЦИАЛЬНЫХ НАУКАХ.



ПОНЯТИЕ ДАННЫХ.



ОСНОВНЫЕ СПОСОБЫ
ПРЕДСТАВЛЕНИЯ ДАННЫХ.

Данные



Данные (по организации)

Структурированные

- Таблицы
- Базы данных
- Данные в формате json/yaml
- Файлы с разметкой (xml)

Неструктурированные

- Тексты (статьи, книги)
- Веб-страницы
- Диалоги
- Аудио/видео

Извлечение информации

Извлечение информации

Information Extraction Example

Salvadoran President-elect Alfredo Cristiani condemned the terrorist killing of Attorney General Roberto García Alvarado and accused the Farabundo Martí National Liberation Front (FMLN) of the crime. ... García Alvarado, 56, was killed when a bomb placed by urban guerillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador. ... According to the police and García Alvarado's driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.

Incident: Date 19 Apr 89

Incident: Location El Salvador: San Salvador

Incident: Type Bombing

Perpetrator: Individual ID "urban guerillas"

Perpetrator: Organization ID "FMLN"

Human Target: Name "Roberto García Alvarado"

...

Mari A. Hearst
UW/MS CM Workshop, 1997

Example: Information Extraction from Twitter



Агрегация новостей

Ротенберг уверен, что Крымский мост простоят сто лет

РИА Новости вчера в 17:28



Председатель совета директоров компании "Стройгазмонтаж" Аркадий Ротенберг уверен, что Крымский мост простоят 100 лет. Читать далее



Заголовки

Ставленник Путина побеждает в Приморье. Его соперник объявил голодовку

BBC Русская служба • час назад

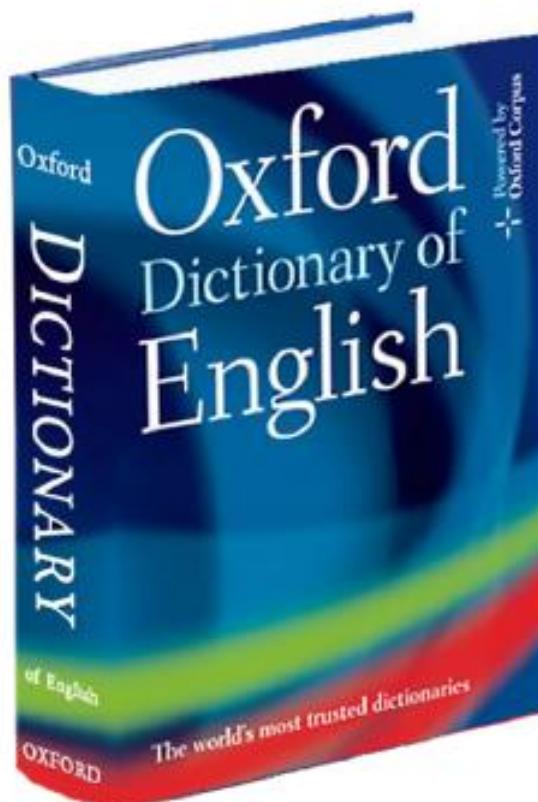
- Кандидат от КПРФ Ищенко выигрывает выборы у врио главы Приморья Интерфакс • сегодня
- Разрыв между кандидатами на выборах главы Приморья сократился РИА НОВОСТИ • 5 ч. назад
- Кандидат от КПРФ стал лидером во втором туре выборов в Приморье РБК • сегодня
- Побеждающий в Приморье кандидат от КПРФ Ищенко обещает продолжить политику Путина Газета.Ru • сегодня

Взгляд с разных сторон

- Участвующие в строительстве Крымского моста компании выплатили 120 млрд рублей налогов
- Ротенберг заявил, что Крымский мост простоят еще 100 лет
- Стало известно, сколько налогов заплатили строители Крымского моста
- Ротенберг сказал, сколько простоят Крымский мост
- Ротенберг уверен, что Крымский мост простоят сто лет

Зачем структурировать данные?

Где быстрее найти слово? Почему?



Базы знаний

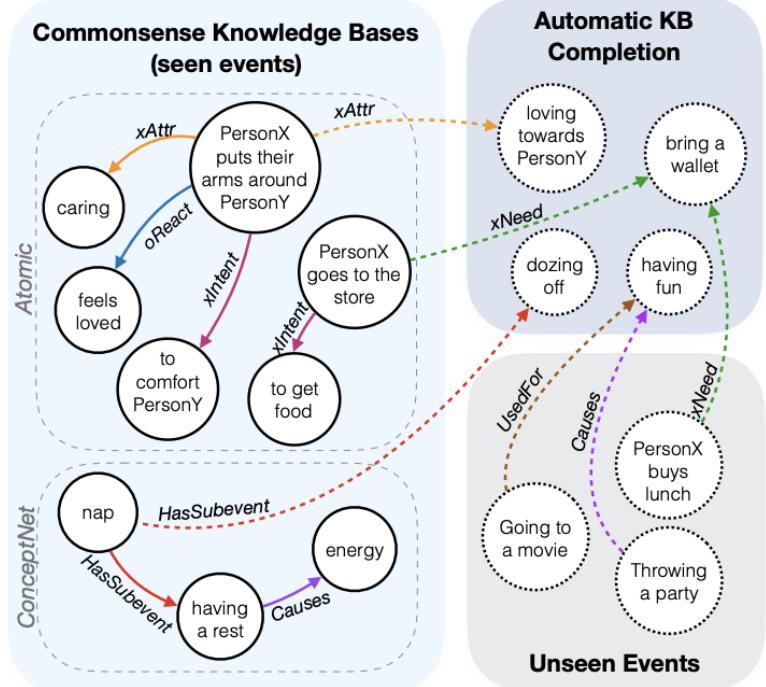
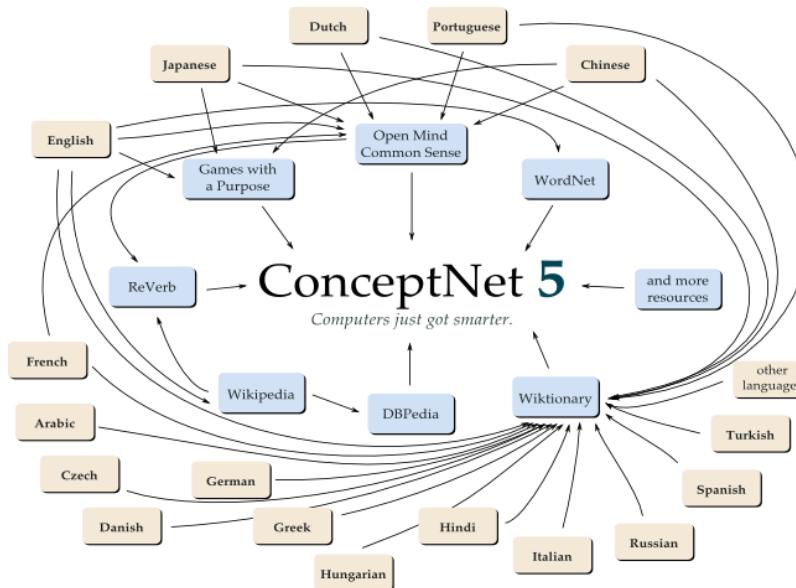


Figure 1: COMET learns from an existing knowledge base (solid lines) to be able to generate novel nodes and edges (dashed lines).



Данные (по доступности)

- Публичные (public)
- Открытые (open)
- Приватные/личные (private)

Почитать о разнице можно тут:

<https://blog.thinkdataworks.com/open-data-vs-public-data>

Открытые данные

Data.gov.ru
открытые данные России

Данные Библиотека Сообщества Сервисы Вход

Портал открытых данных Российской Федерации

Количество наборов открытых данных на портале: 23858 Добавить набор данных >

Добро пожаловать на Портал открытых данных Российской Федерации! Наш гид поможет вам ознакомиться с особенностями Портала и быстрые

DATA.GOV

DATA TOPICS RESOURCES STRATEGY DEVELOPERS CONTACT

The home of the U.S. Government's open data

research, develop web and mobile

please visit [Coronavirus.gov](#).

ГОСУДАРСТВЕННЫЙ ПОРТАЛ ОТКРЫТЫХ ДАННЫХ
Правительства Москвы

ПОИСК

ДАННЫЕ СПРАВОЧНИКИ ПРИЛОЖЕНИЯ НОВОСТИ ИНФОРМАЦИЯ ФОРУМ

СТАНДАРТЫ Разработчикам ENG

город начинается здесь mos.ru

ПОИСК по 1115 наборам данных и материалам портала

Sitemap Legal notice Contact English (en)

EU Open Data Portal Access to European Union open data

EUROPA > EU Open Data Portal > Home

Home Data Applications Linked data Visualisations Developers' corner About Share

The European Union Open Data Portal (EU ODP) gives you access to open data published by EU institutions and bodies. All the data you can find via this catalogue are free to use and reuse for commercial or non-commercial purposes.

Show results with:
○ all of these words | ● any of these words | ⚪ the exact phrase

Search for metadata using our SPARQL endpoint query editor or access the API.

DATA TOPICS RESOURCES STRATEGY DEVELOPERS CONTACT

The home of the U.S. Government's open data

research, develop web and mobile

please visit [Coronavirus.gov](#).

SETS

SEARCH

Что еще?

- Персональные данные
- Биометрические данные
- Пространственные данные
- **Большие данные**

Big data

Что такое «Big Data»?

Мифы и легенды про Big Data

Если кратко:

— А почему тогда все на конференциях говорят про Big Data?

Потому что тренд модный, слово любят журналисты. Даже если вы обработаете 50 тысяч записей интернет-магазина и назовёте это Big Data — это будет достаточно пафосно, чтобы написать в пресс-релизе.

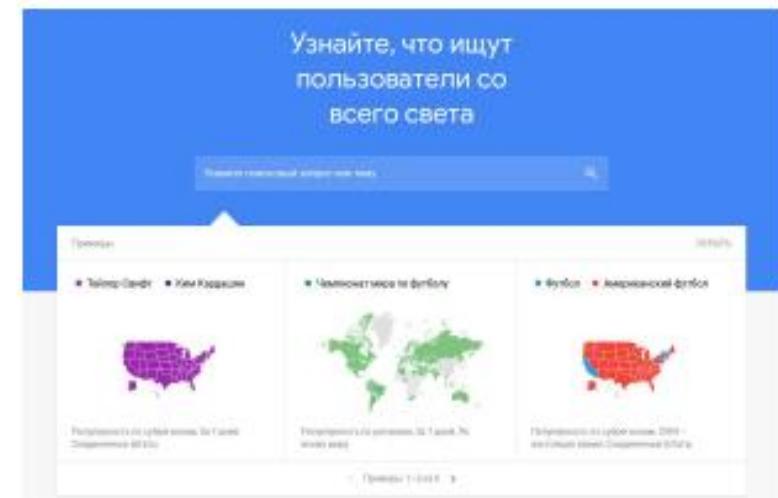


Где искать данные?

- Google
- Yandex
- Facebook
- Vk.ru
- Twitter

Государственные наборы данных:

- <http://data.mos.ru/> — портал открытых данных Правительства Москвы. База данных об учреждениях досуга и отдыха, ЖКХ, здравоохранения, образования и т.д., а также приложения для мобильных устройств на основе этих данных. Получить доступ к использования данных с этого портала можно с помощью API портала открытых данных;
- <http://data.gov.ru/> — портал открытых данных Российской Федерации. На портале можно найти данные по разным темам, а также сделать запрос на открытие данных;



Содержание



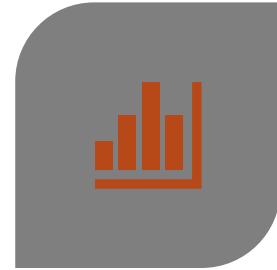
ПОНЯТИЕ DIGITAL HUMANITIES.



СФЕРЫ ПРИМЕНЕНИЯ АНАЛИЗА И
ВИЗУАЛИЗАЦИИ ГУМАНИТАРНЫХ
И СОЦИАЛЬНЫХ НАУКАХ.

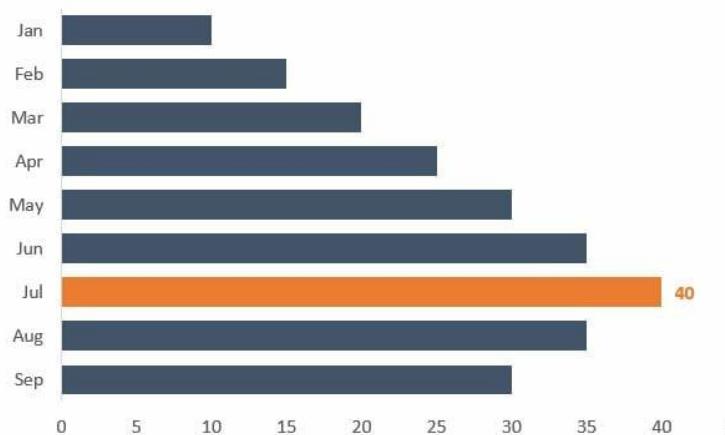


ПОНЯТИЕ ДАННЫХ.

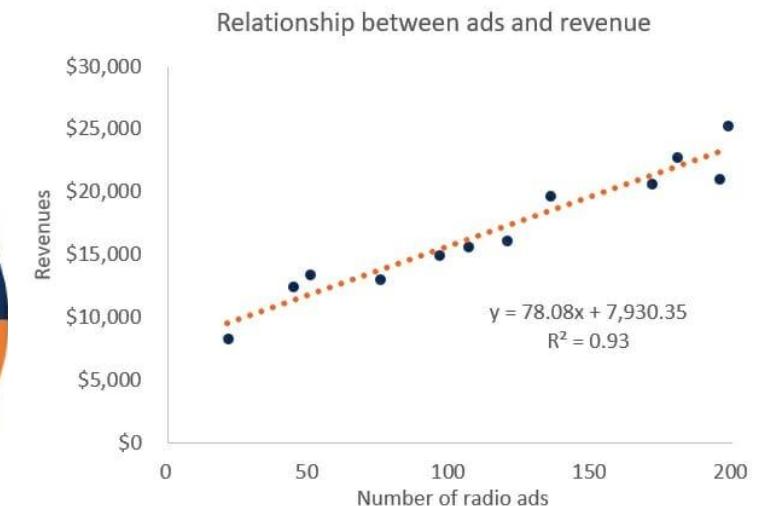
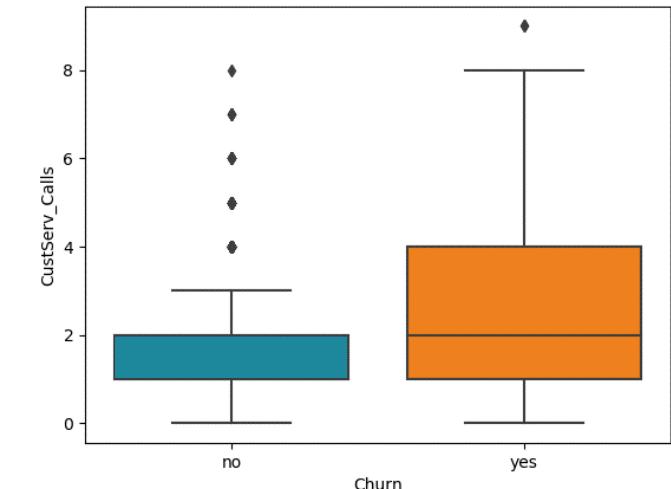
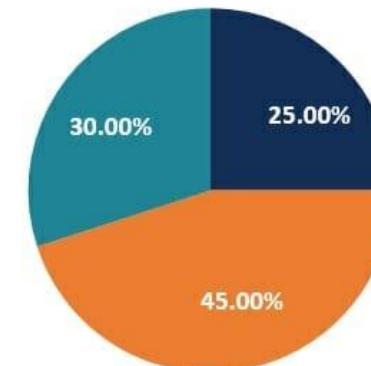


ОСНОВНЫЕ СПОСОБЫ
ПРЕДСТАВЛЕНИЯ ДАННЫХ.

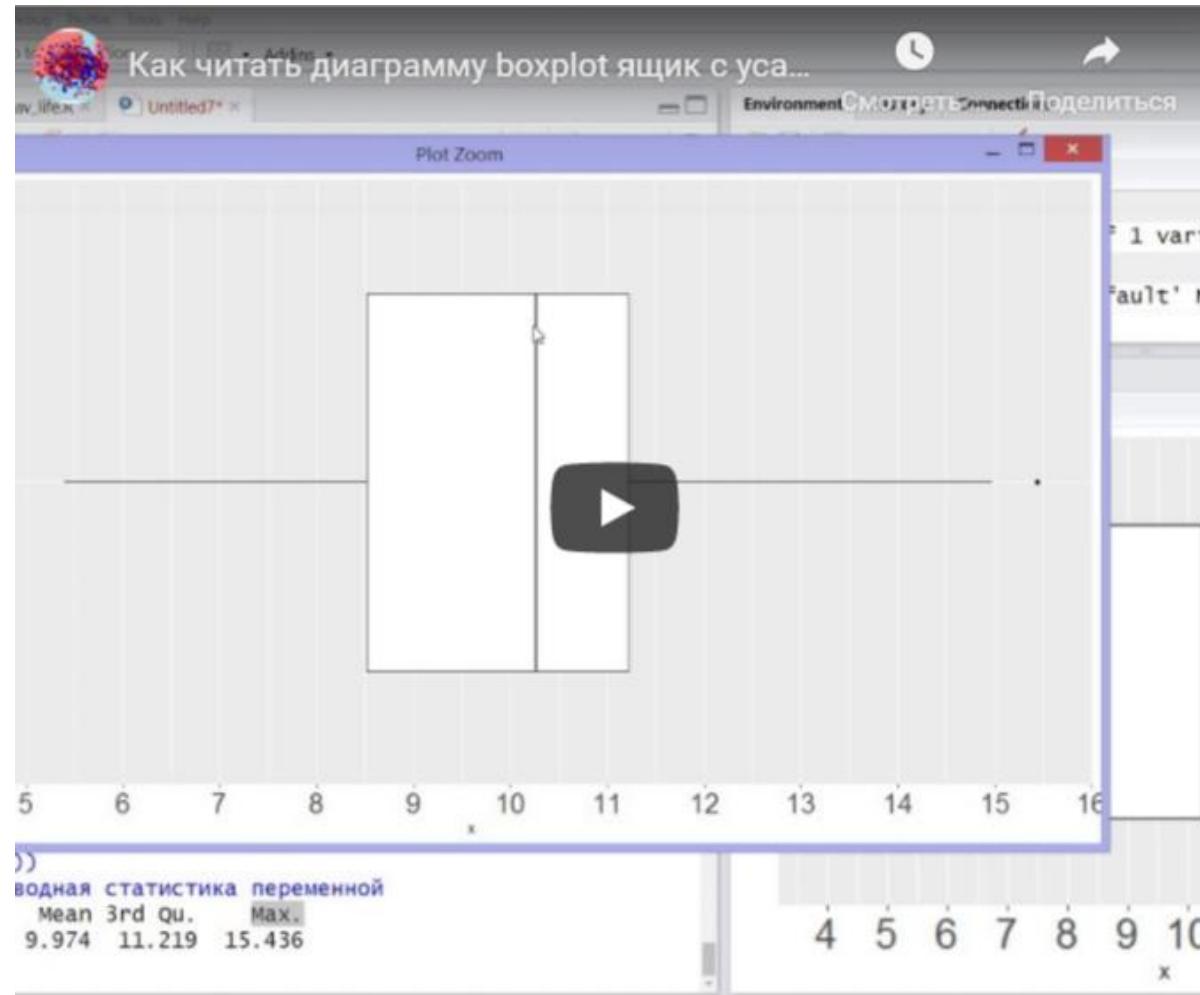
Визуализация данных



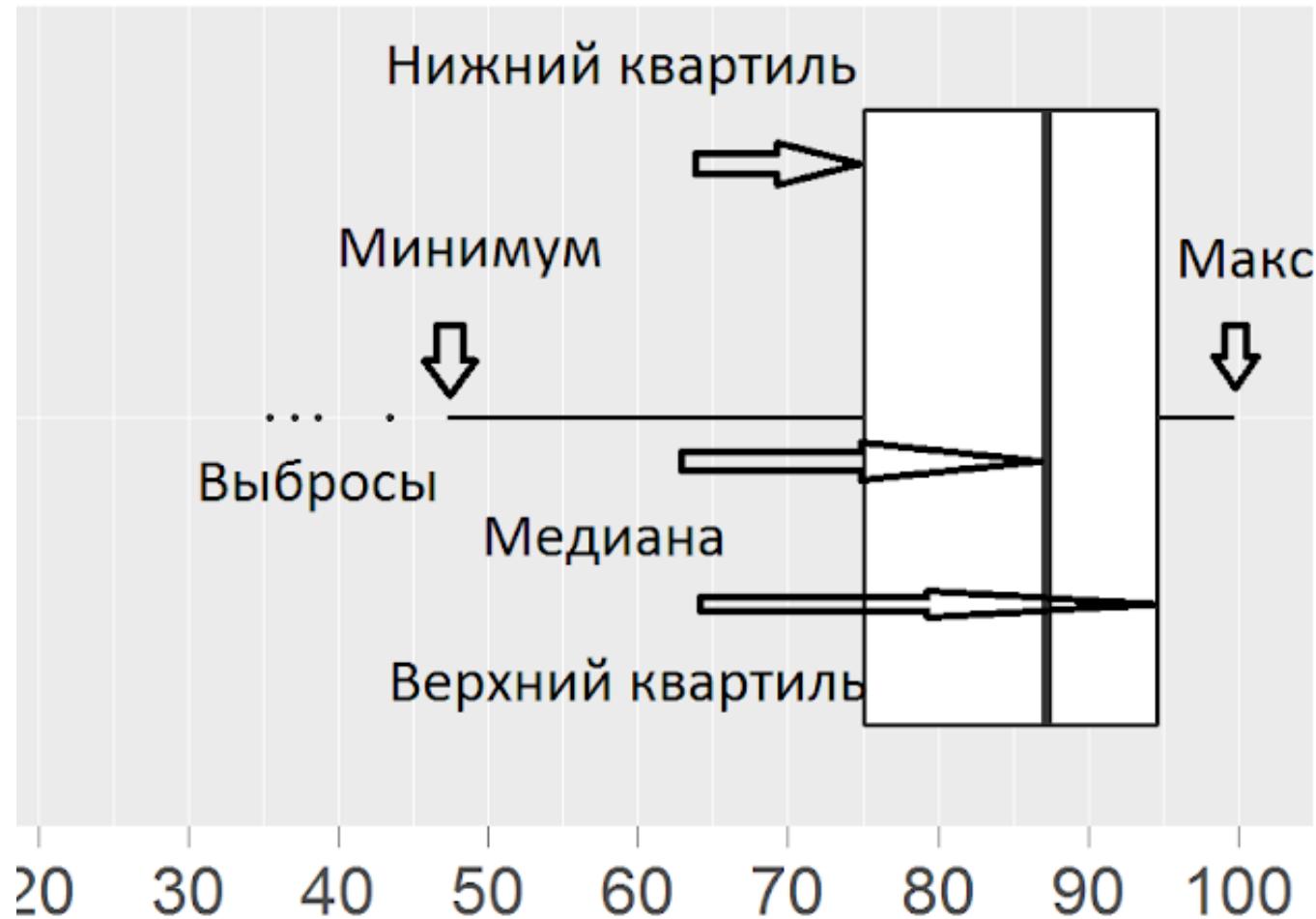
- График (index chart)
- Столбчатая диаграмма (bar chart)
- Ящик с усами (box plot)
- Диаграмма рассеяния (scatter plot)
- Круговая диаграмма (pie chart)



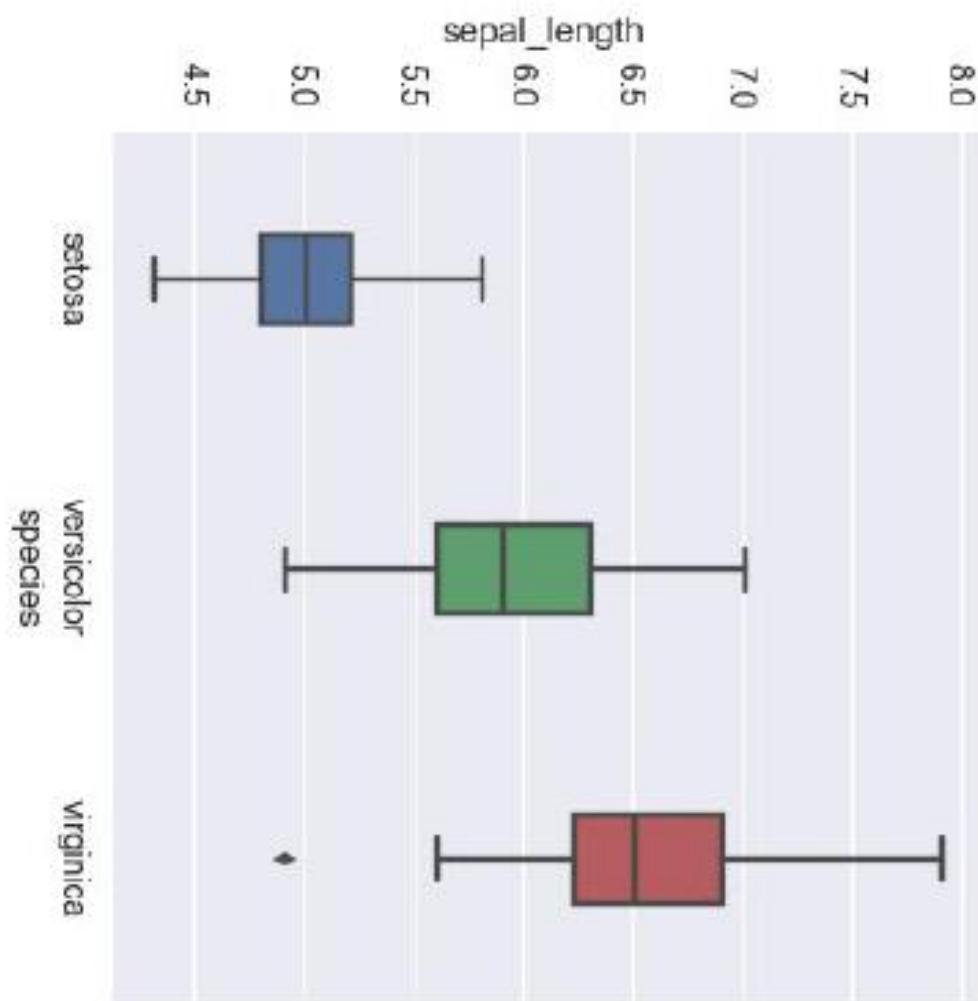
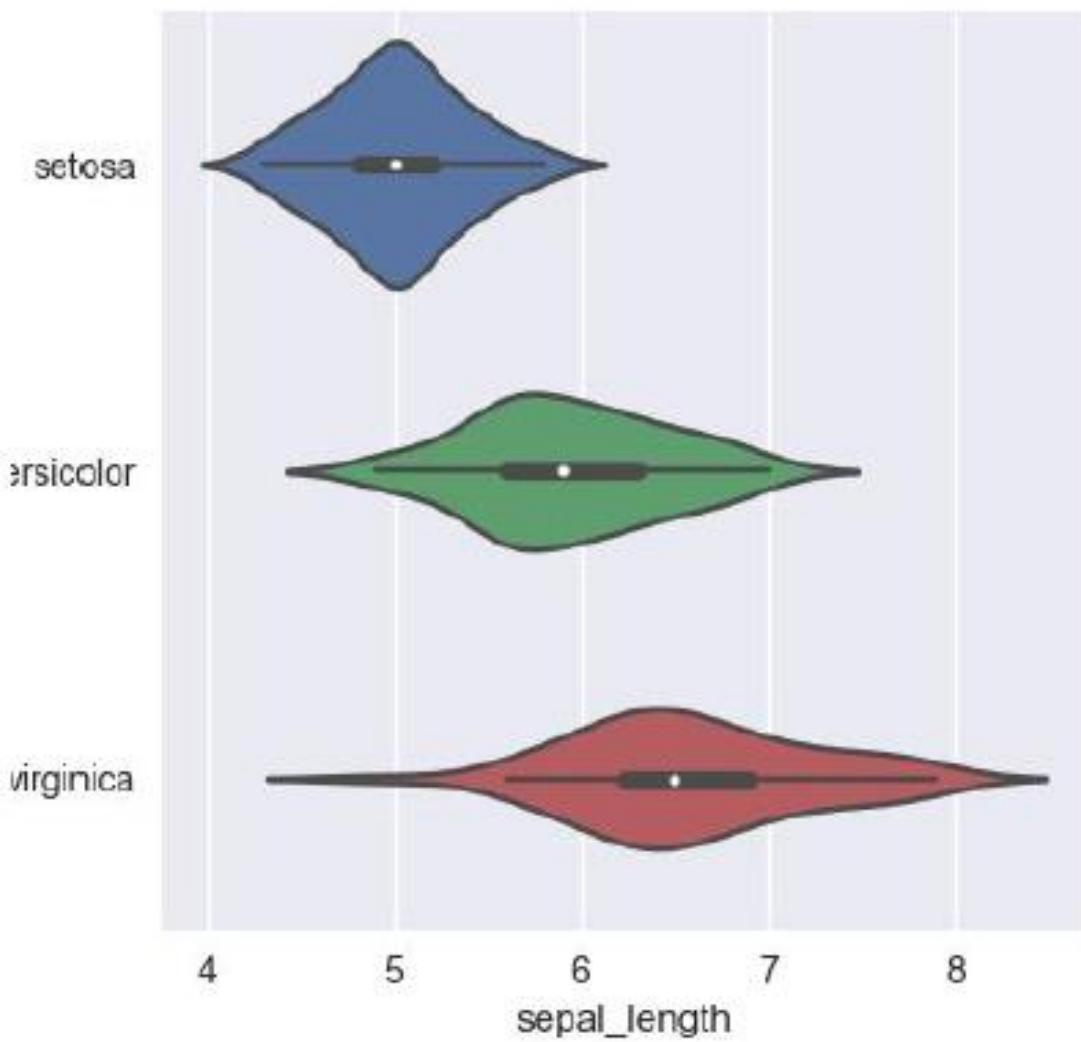
Ящик с усами (box plot)



Ящик с усами (box plot)



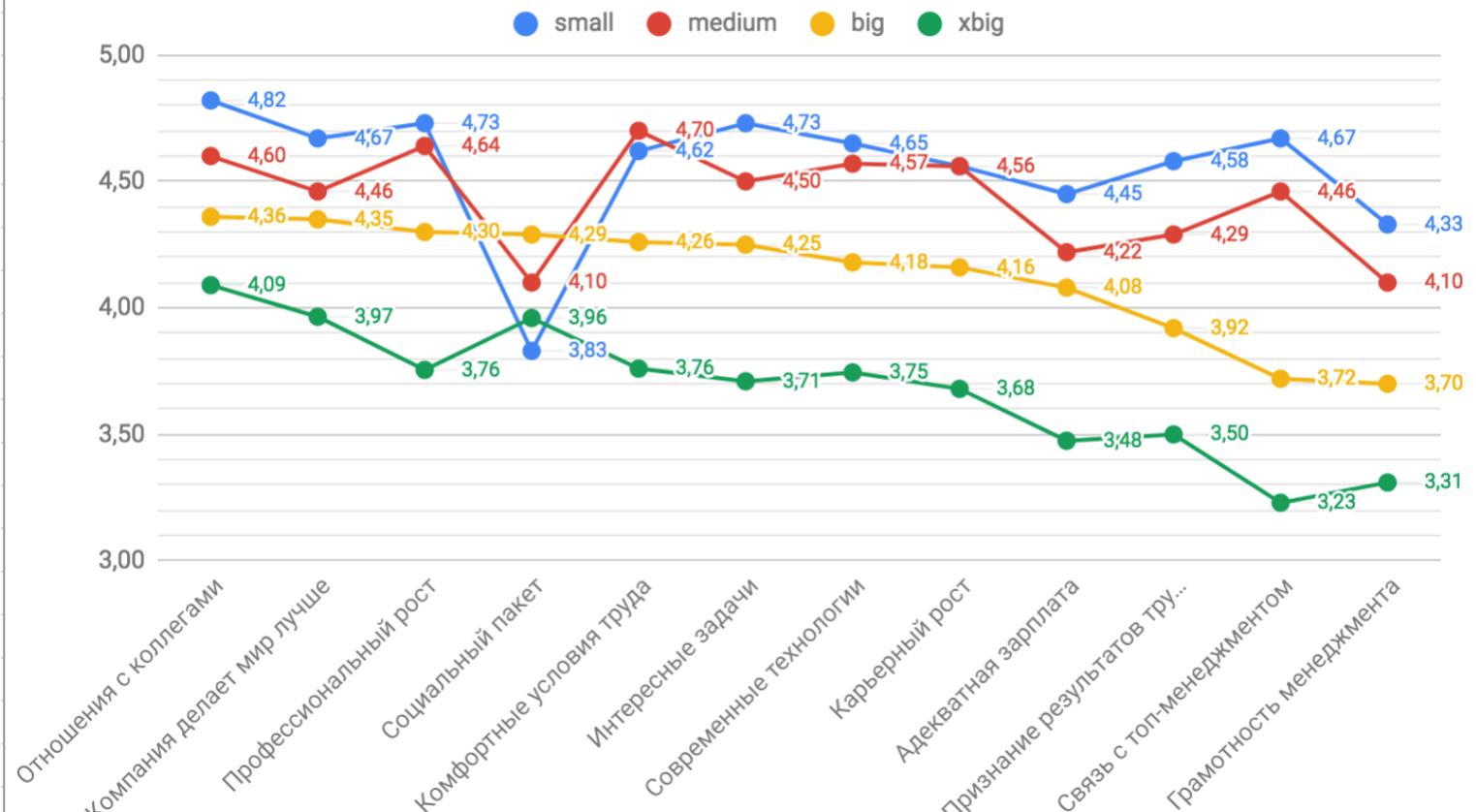
Violin plot



Плохие примеры

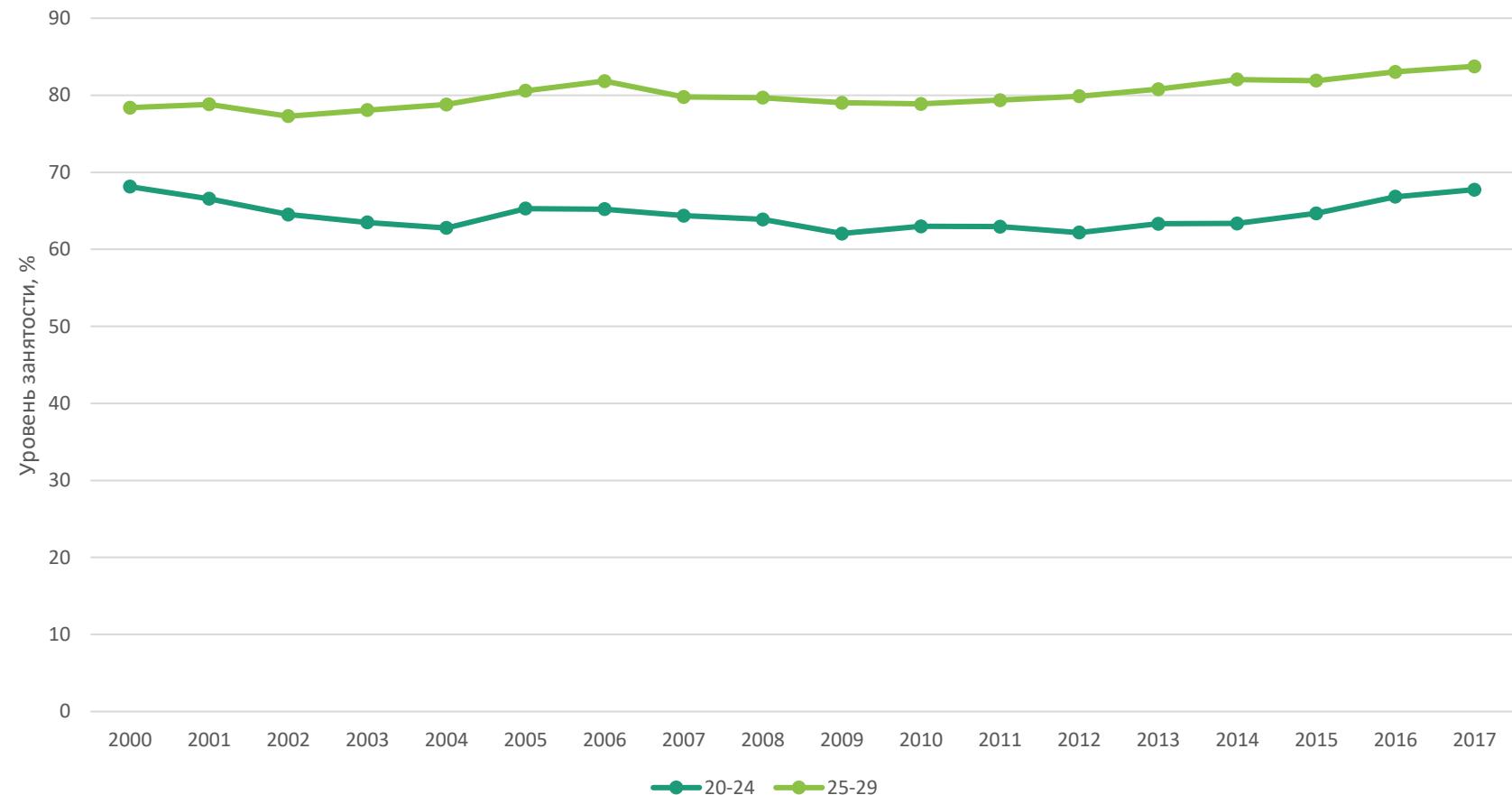
Графики

Оценка компаний по 12 критериям в зависимости от их размеров



https://habr.com/ru/company/habr_career/blog/418047/

График 2.1. Доля занятости населения 20-29 лет с 2000 по 2017 год



Графики

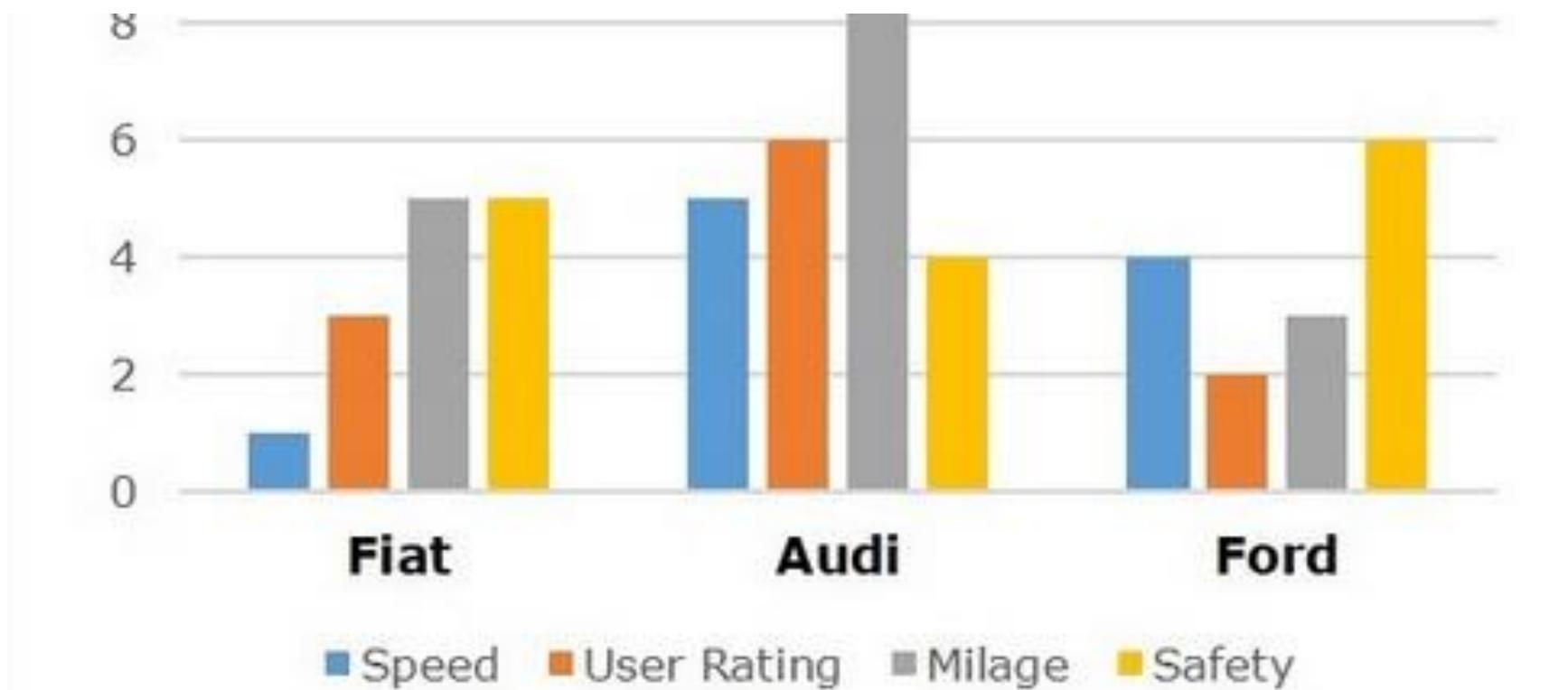
<https://www.hse.ru/edu/vkr/219185590>

Графики



<https://www.hse.ru/edu/vkr/219185590>

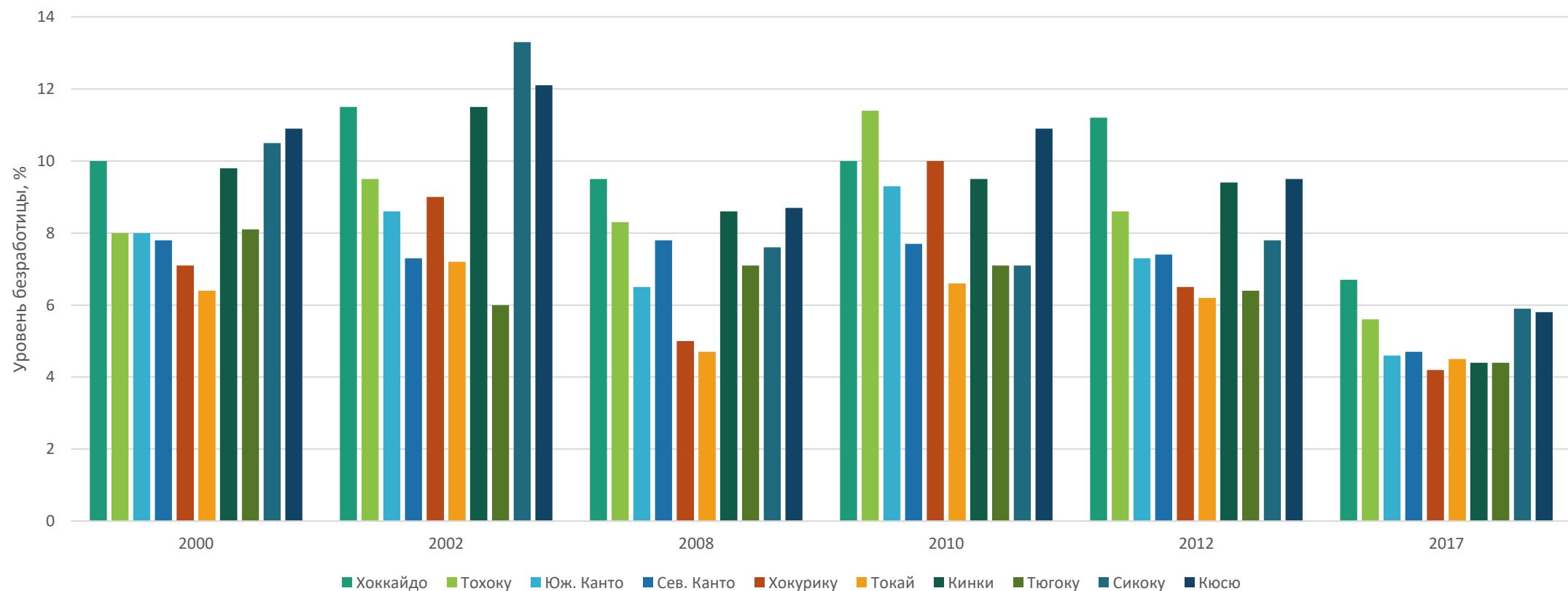
Столбчатая диаграмма (bar plot)



<http://www.knowledgewala.com/ibm-curam-birt-report-bar-chart-development-example/>

Столбчатая диаграмма (bar plot)

График 3.1. Уровни безработицы в регионах (20-24 года)





Пицца – это круговая диаграмма,
показывающая, сколько
у тебя осталось пиццы.



101



6



Vadim Nikolaenko
Пирог тоже
today at 8:42

2



Stanislav Olegovich
Умно
today at 8:46

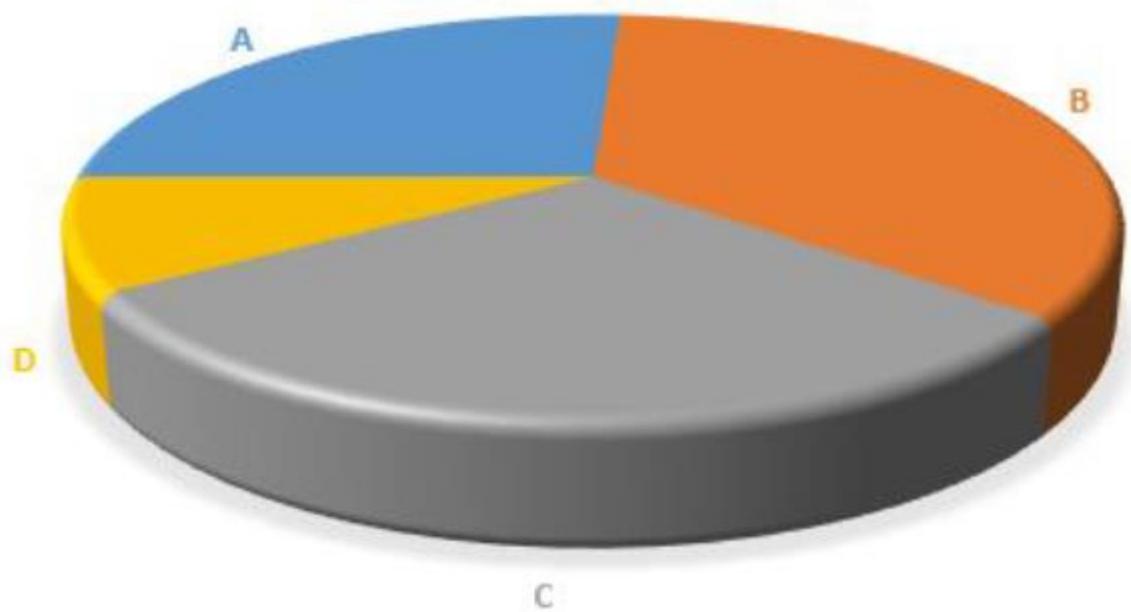


Maxim Kapralov
Вадим, пирог не показывает, сколько у тебя осталось пиццы. Так что - нет,
не то-же.
today at 8:46 to Vadim

24

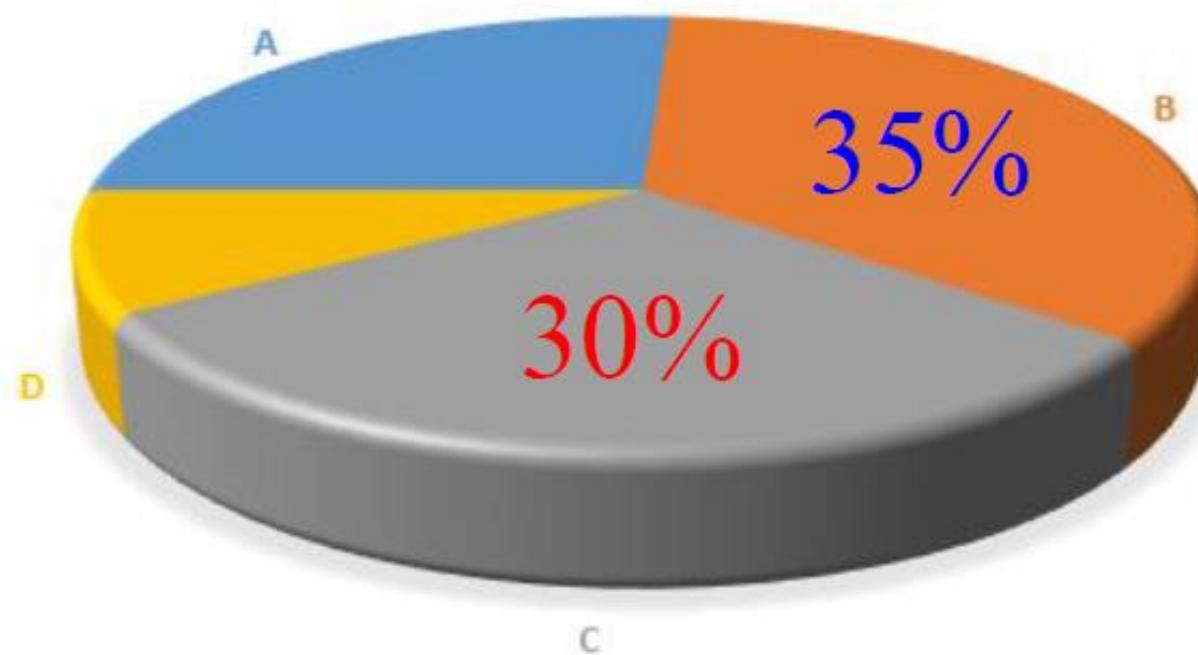
Pie chart

- На сколько процентов С больше В?



Pie chart

- На сколько процентов С больше В?



Полезные ссылки

- [Storytelling with Data](#)
- [Visualization Zoo](#)
- [Flowing data](#)
- [Каталог визуализации данных](#)
- [Points of view: bar charts and box plots](#)
- [Points of significance: visualizing samples with box plots](#)
- [Coblis](#) - a tool to stimulate colour-blind vision and correct your graphics for colour-blind viewers.
- [Points of view: color coding](#)
- [Points of view: avoiding color](#)
- [Elements of visual style](#)

ДЗ₀

- Установить [Anaconda](#)
- Принести ноутбук на следующее занятие

Семинар 1.

Знакомство с Python