

Майнор «Обработка и анализ данных»

2019-21



Почему спрос на анализ данных растет?

Анализ данных стал важной частью большинства профессий в бизнесе

Сейчас аналитик данных — не математик со специализацией в методах машинного обучения, а эксперт в определенной сфере, использующий навыки анализа данных для решения конкретных задач

Анализ данных востребован не только в бизнесе, но и в науке, причем как в экономико-управленческих, так и в социальных и гуманитарных науках, владение методами обработки и анализа данных также стало важным конкурентным преимуществом при поступлении на PhD программы, исследовательские позиции.

Почему спрос на анализ данных растет?

Новые задачи:

- Выявление и объяснение трендов и паттернов (interpretable ML, explainable AI)
- Автоматизация аналитики и поддержка принятия решений
- Персонализация
- Анализ отзывов, мнений, суждений



Как студенты выбирают майноры





Денис Булыгин
преподаватель
департамента
информатики



Алина Бахитова
преподаватель
департамента
информатики



Виктор Карепин
преподаватель
департамента
информатики



Алёна Суворова
канд. физ.-мат. н.
доцент департамента
информатики



Илья Мусабиров
ст. преподаватель
департамента информатики



Александр Сироткин
канд. физ.-мат. н.
доцент департамента
математики

Этот майнор для тех, кому хочется

- научиться **применять машинное обучение** для будущей работы или научных исследований
- понять как работает **аналитика данных на практике**
- сочетать **экспертные знания в своей области** с **современными методами анализа данных** (числовых, сетевых, текстовых)
- поучаствовать в интересном **проекте** от постановки проблемы до проектирования интерактивных приложения, дэшбордов и баз данных

Чему мы будем учиться

- Исследовать данные, не имея изначальных гипотез
- Устанавливать цели анализа и задавать вопросы
- Строить статистические модели и интерпретировать результаты
- Презентовать итоги работы

Что мы будем делать

- находить скрытые связи
- объяснять
- предсказывать
- рекомендовать

...и многое другое

Какие курсы будут

1. **Программирование для анализа данных и воспроизводимые исследования** – Основы R, построение графиков и отчетов, агрегация данных
2. **Анализ данных и технологии работы с данными** – Основы обработки структурированных и неструктурированных данных, работа с текстами и сетями, базами данных
3. **Интеллектуальный анализ данных и основы машинного обучения** – изучение продвинутых моделей
4. **Приложения и практика анализа данных** – командные проекты

Нововведения следующего года

2 семестр 2 курса: **Анализ данных и технологии работы с данными** – Основы обработки структурированных и неструктурированных данных, работа с текстами и сетями

Два модуля по выбору студентов в рамках курса



Базы данных и предиктивная аналитика

Анализ для прогнозирования, принятия решения, введение в базы данных (SQL) и инструменты анализа данных (shell tools, csv/json/xml tools), визуализация для поддержки принятия решений

Сети и тексты, карты

Визуализация и анализ текстовых, сетевых, гео-данных

Ожидаем участия специалистов по компьютерной лингвистике и анализу онлайн-текстов

Нововведения следующего года

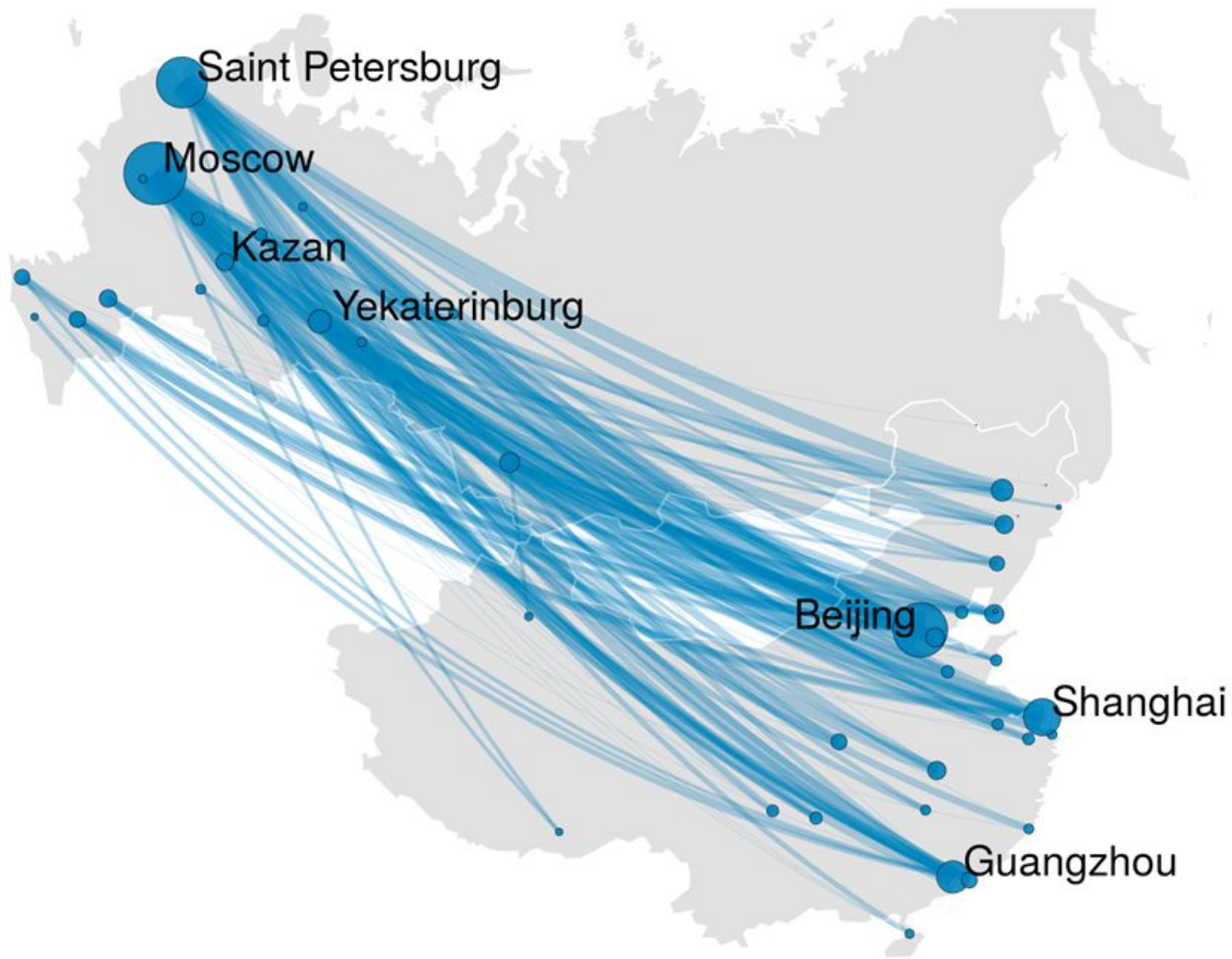


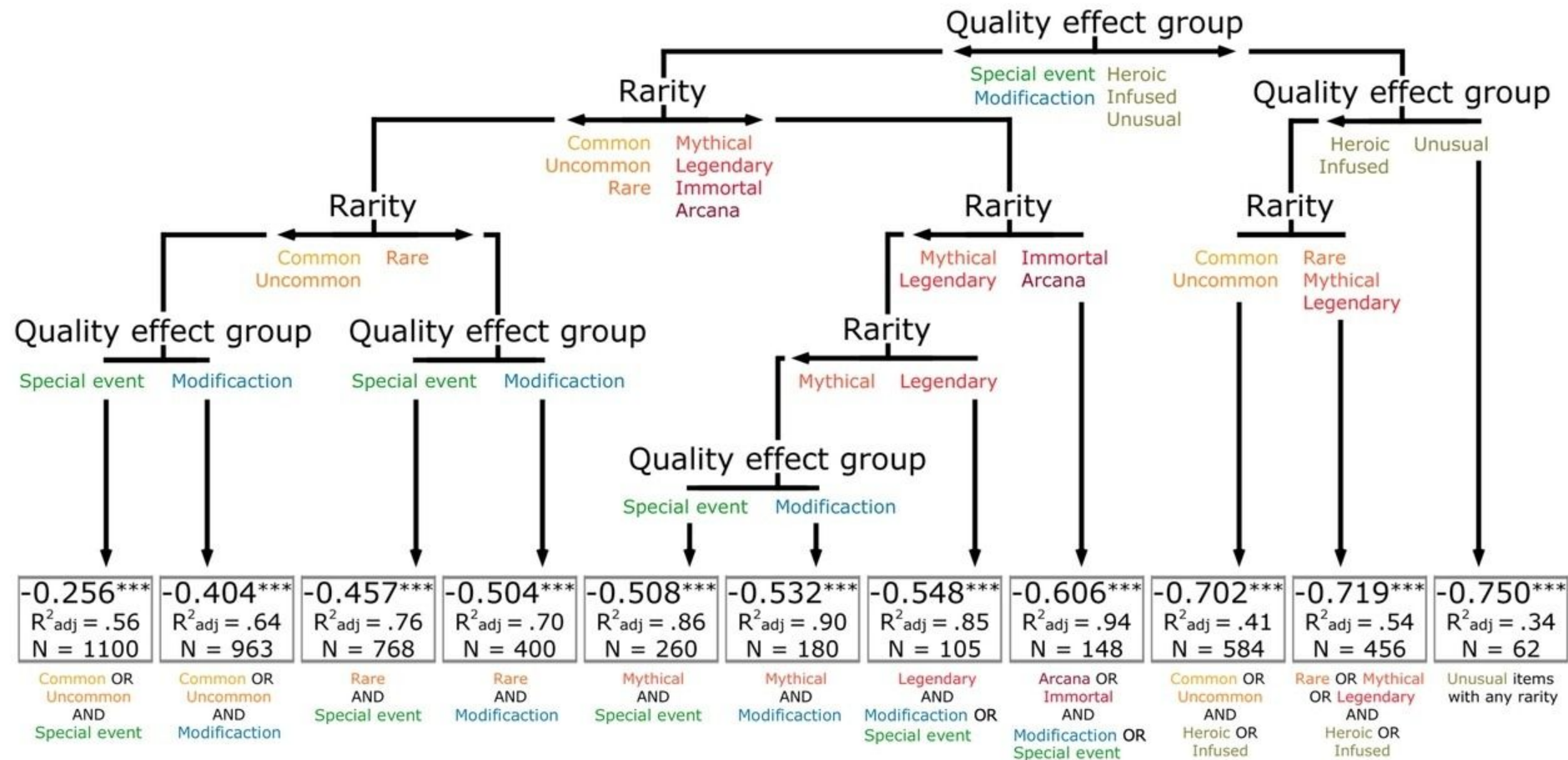
- Больше кейсов построенных вокруг реальных данных и возможностью выбора, на каких данных тренировать навыки
- Больше онлайн-тьюториалов в дополнение к онлайн-упражнениям
- Онлайн-тьюториалы выпускников майнора, работающих с анализом данных
- Возможно формирование треков на 3 курсе:

Алгоритмы машинного обучения в Python vs Анализ онлайн-данных vs Digital Humanities (согласуется, но элементы каждого блока точно будут)

Что студенты учились делать so far

- выбирать дорамы
- находить лучшую шаверму в Санкт-Петербурге
- искать лучшую страну в мире
- советовать книги и фильмы
- читать твиттер и долго думать
- понимать, как отзывы связаны с оценками отелей
- выбирать китайский ресторан в Китае
- предсказывать стоимость недвижимости
- искать самые актуальные события в Петербурге

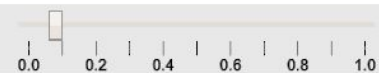




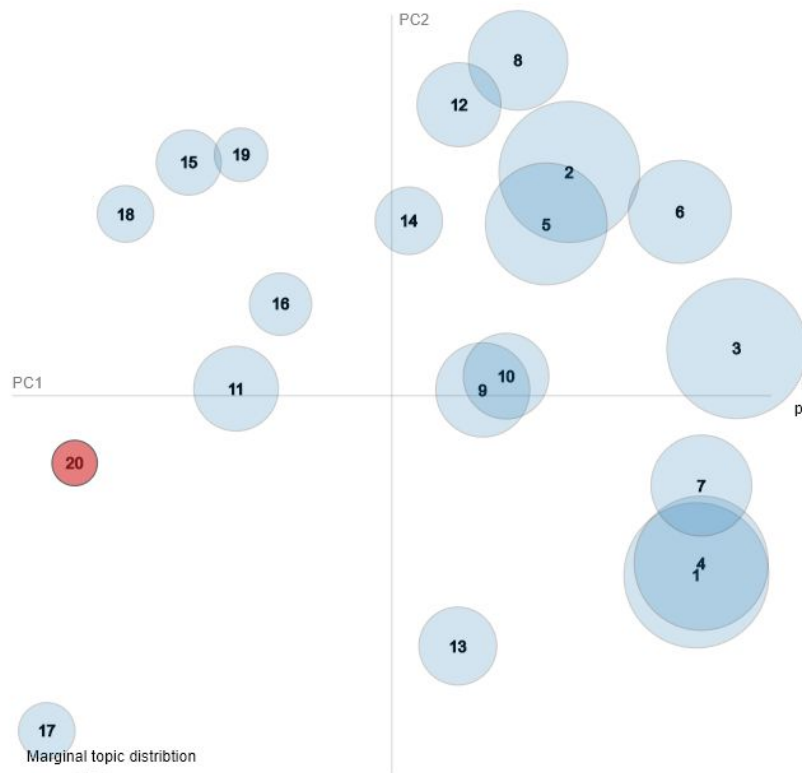
Selected Topic: 20 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:⁽²⁾

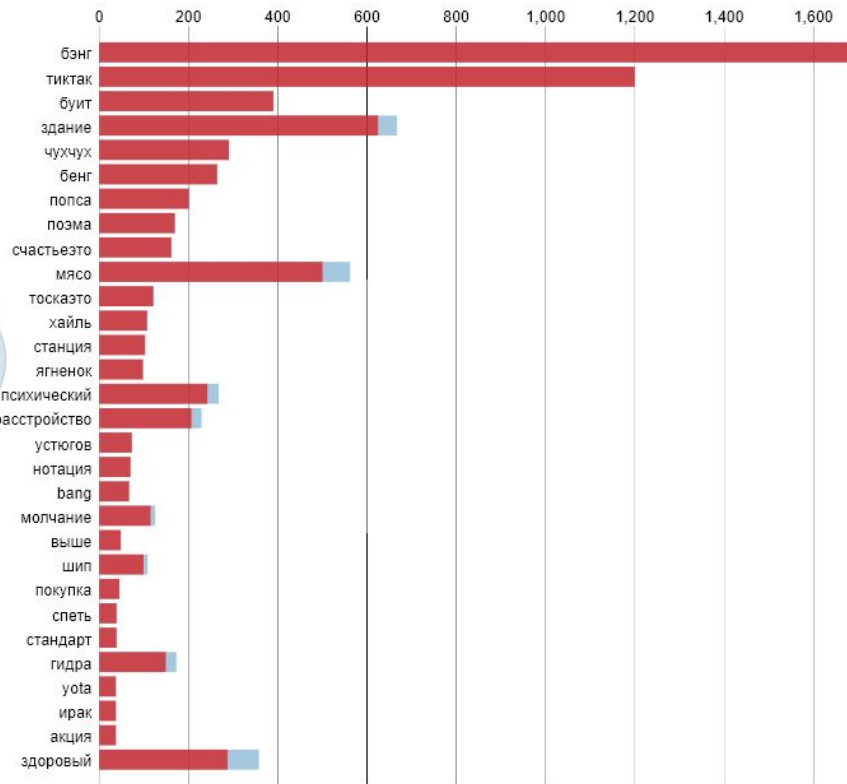
$\lambda = 0.09$



Intertopic Distance Map (via multidimensional scaling)

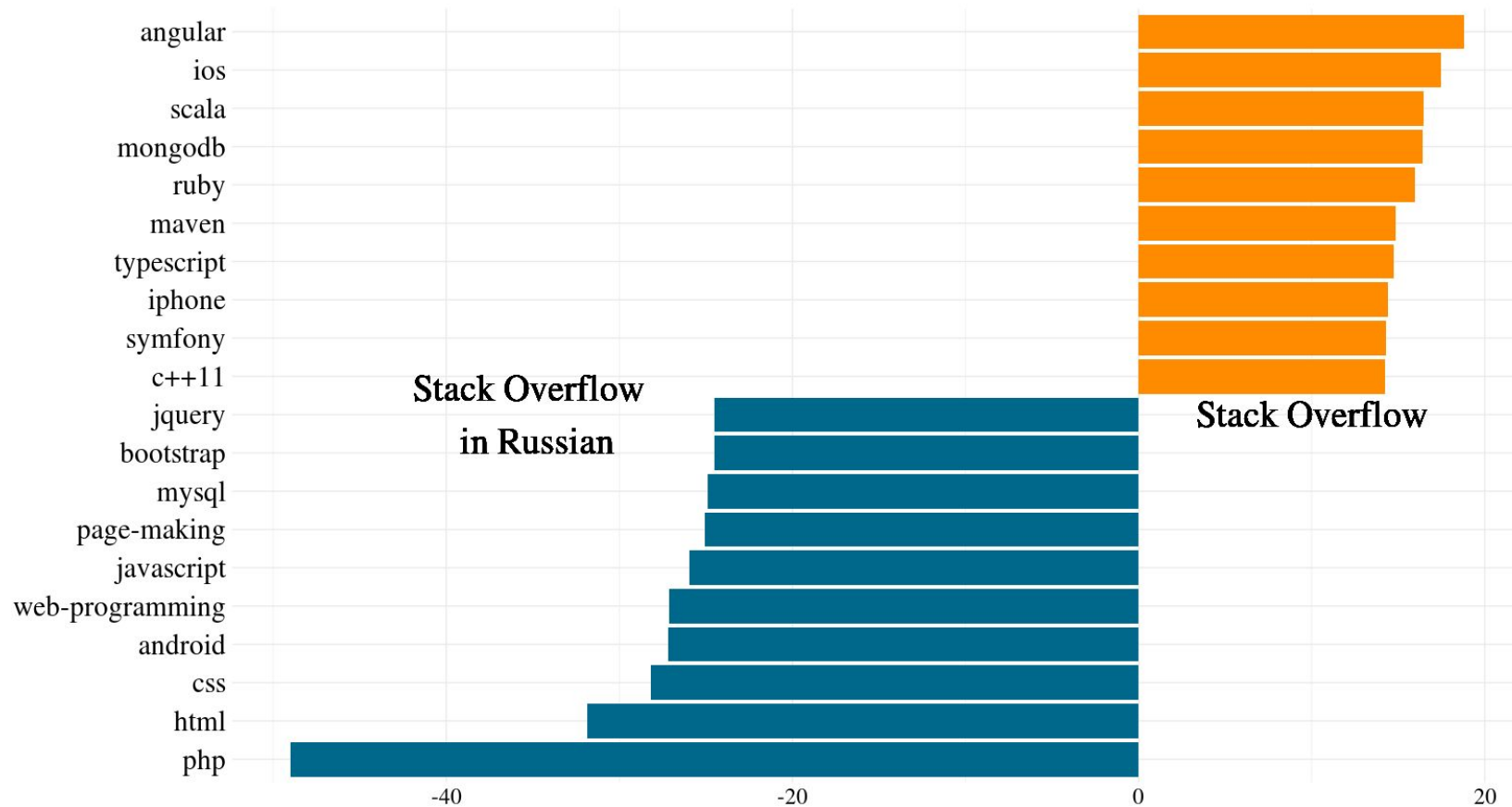


Top-30 Most Relevant Terms for Topic 20 (1.1% of tokens)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))]] for topics t; see Chuang et. al (2012)

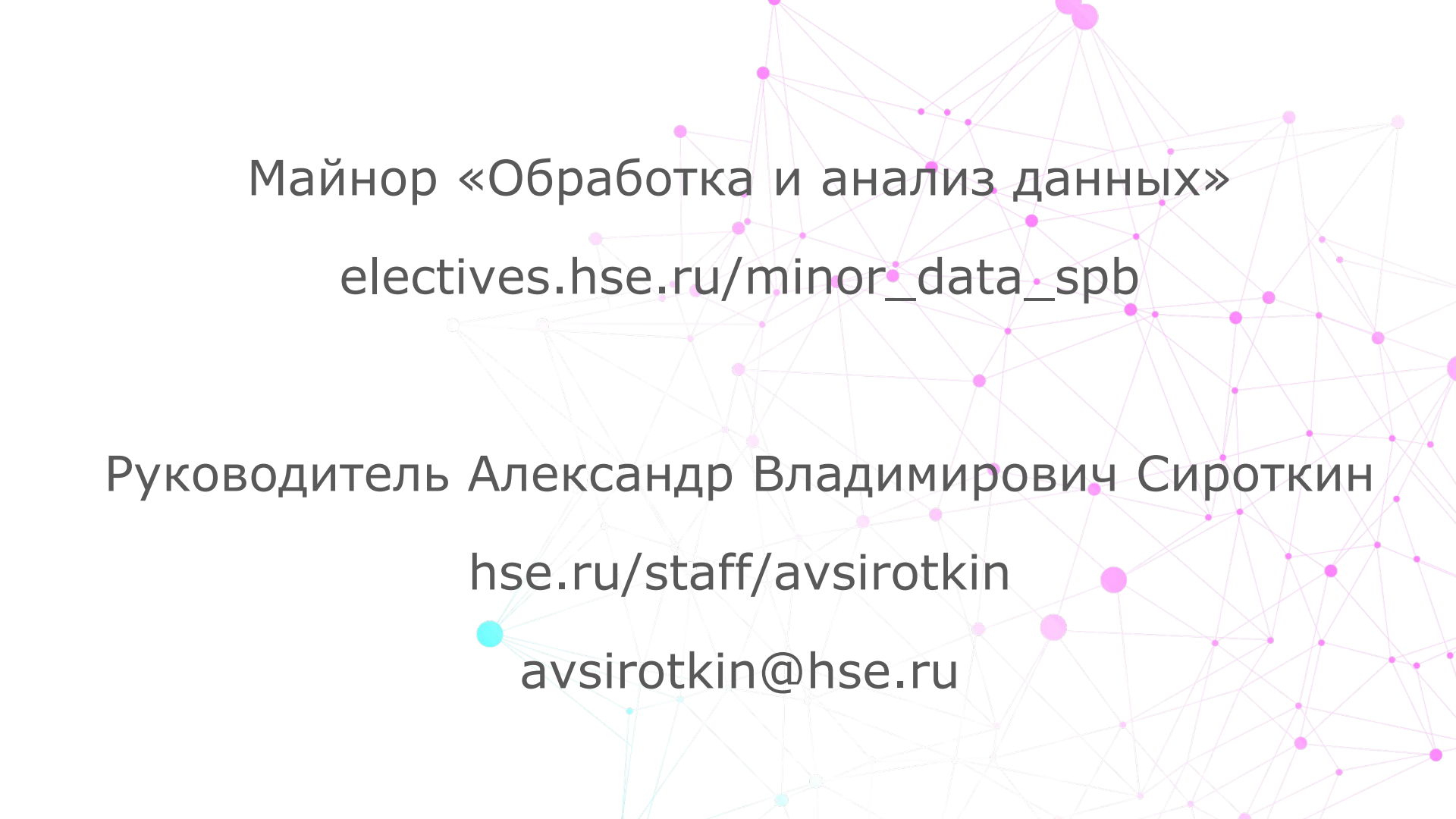
2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)



“It is said that despite its many glaring (and occasionally fatal) inaccuracies, the Hitchhiker's Guide to the Galaxy itself has outsold the Encyclopedia Galactica because it is slightly cheaper, and because it has the words

'DON'T PANIC'

in large, friendly letters on the cover.”



Майнор «Обработка и анализ данных»

electives.hse.ru/minor_data_spb

Руководитель Александр Владимирович Сироткин

hse.ru/staff/avsirotkin

avsirotkin@hse.ru