Задача машинного обучения

Ксемидов Борис Сергеевич

5 марта, 2020

Задача машинного обучения

Компьютерная программа обучается при решении какой-то задачи из класса T, если ее производительность, согласно метрике P, улучшается при накоплении опыта E.

Виды задач машинного обучения

К задачам Т машинного обучения относят:

- классификация (отнесение объекта к одной из категорий)
- регрессия (прогнозирование количественного признака)
- кластеризация (разбиение множества объектов на группы на основании признаков этих объектов)

Набор данных

Под опытом E понимаются данные, и в зависимости от этого алгоритмы машинного обучения могут быть поделены на те, что обучаются с учителем и без учителя (supervised и unsupervised learning).
В задачах обучения без учителя имеется набор данных, состоящий из объектов, описываемых набором признаков.
В задачах обучения с учителем также имеется целевой признак - тот, который необходимо предсказать.

Метрика оценки P

Модель машинного обучения можно оценить следующими способами:

- accuracy - доля правильных ответов
- precision - доля верных прогнозов

Этапы решения задачи машинного обучения

1. Формализация задачи и анализ предметной области.
2. Формирование выборки данных.
3. Разведочный анализ данных (EDA - exploratory data analysis).
4. Построение модели.
5. Оценка построенной модели.
6. Использование модели.

EDA

EDA включает в себя:

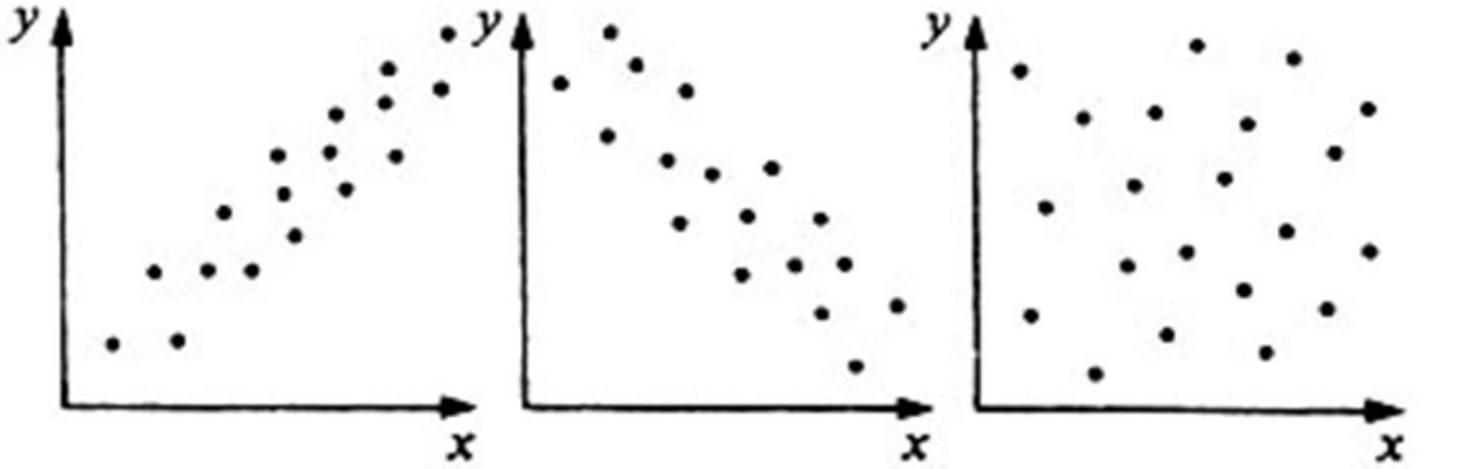
- корреляционный анализ (корреляция признаков);
- многомерное шкалирование (преобразование данных для визуализации).

Корреляция признаков

Коэффициенты корреляции:

- Пирсона;
- Кендалла (ранговый);
- Спирмена (ранговый).

Геометрическая интерпретация корреляции



А. Положительная корреляция

Б. Отрицательная корреляция

В. Корреляция отсутствует

t-SNE

t-SNE - техника нелинейного снижения размерности и визуализации многомерных переменных.
Пример:

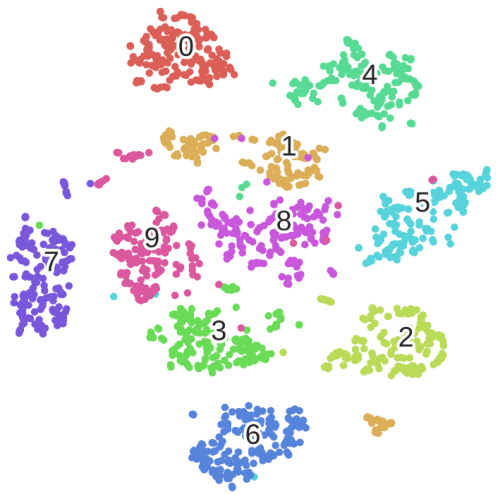


Рис.: Пример t-SNE

Построение простой модели

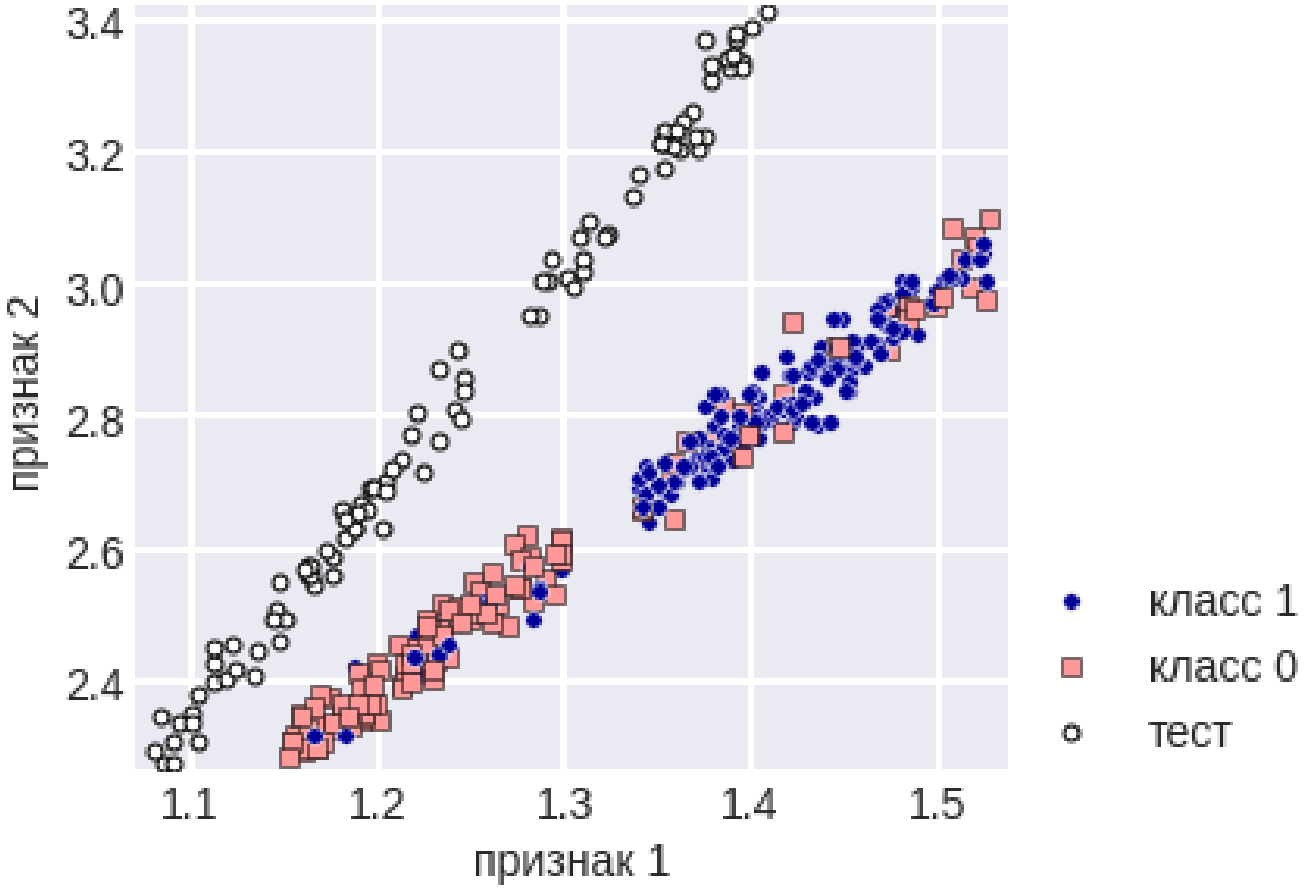


Рис.: Признаковое пространство в задаче классификации кортикограмм.

Пример предметной области

Объект изучения - коронавирус.

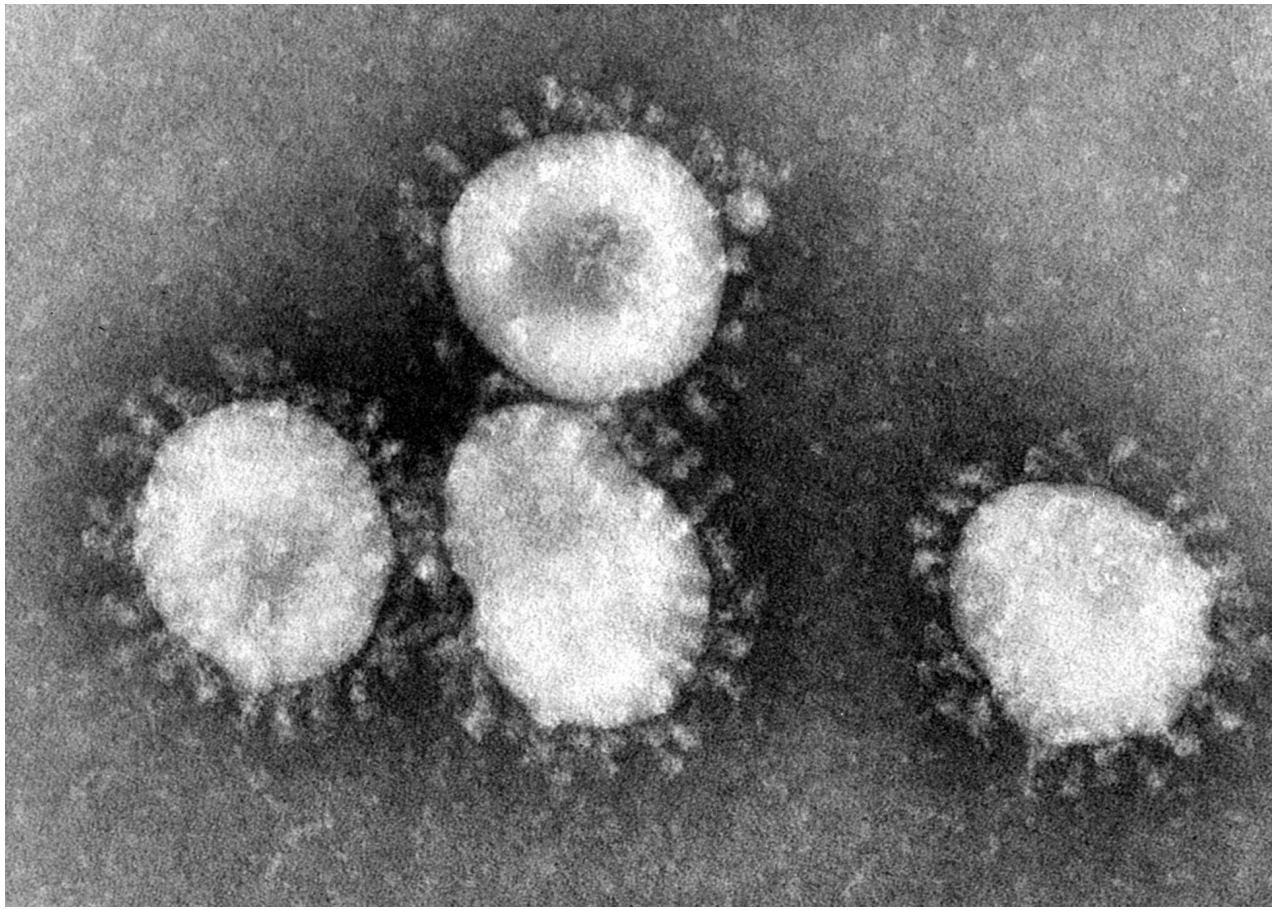


Рис.: Коронавирус

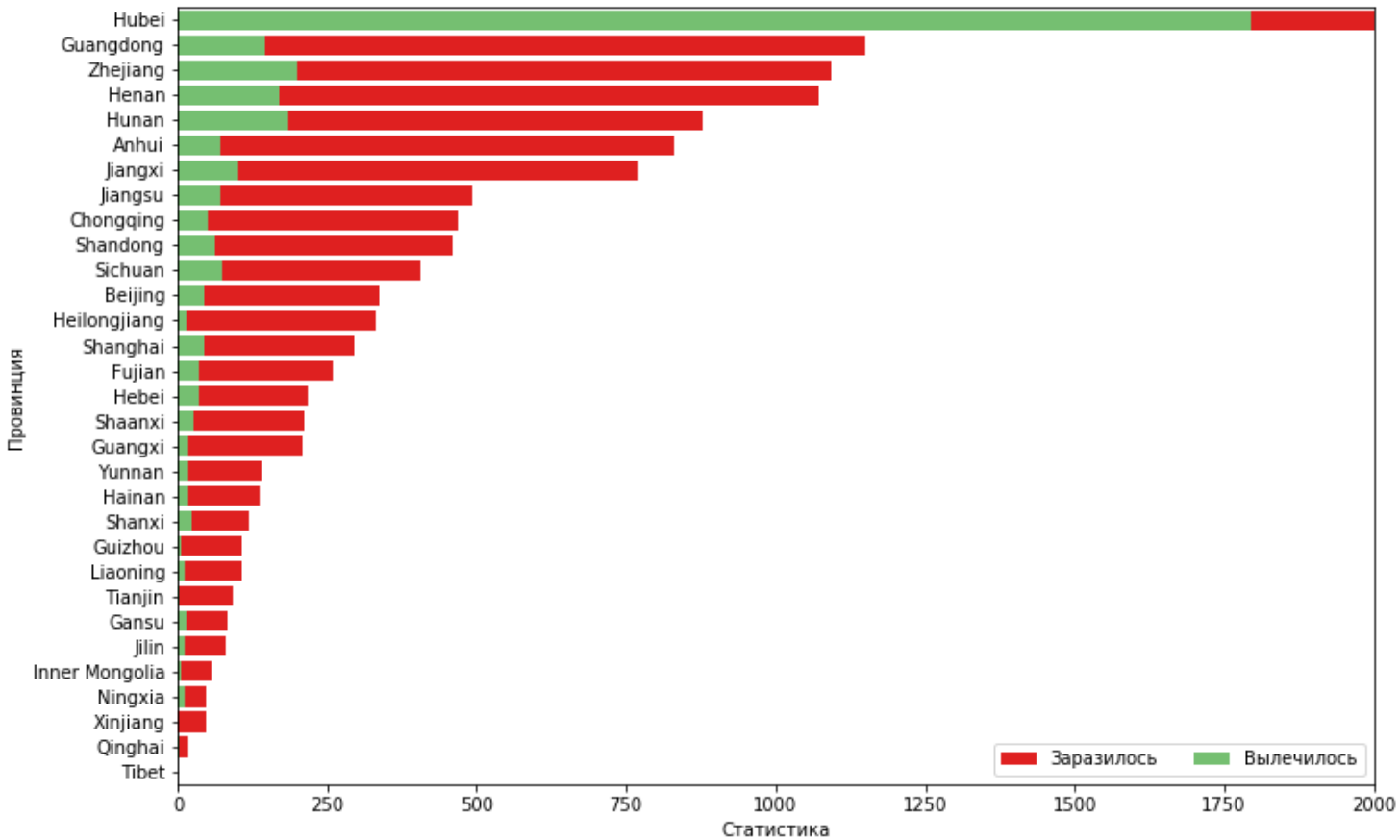
Описание признаков

- Date - дата и время наблюдения
- Province/State - наблюдаемая провинция или штат
- Country - наблюдаемая страна
- Confirmed - подтвержденное число заражений
- Deaths - число смертей зараженных
- Recovered - число выздоровевших

Пример данных

Date	Province/ State	Country	Confirmed	Deaths	Recovered
2020-01-22 12:00:00	Anhui	China	1.0	0.0	0.0
2020-01-22 12:00:00	Beijing	China	14.0	0.0	0.0
2020-01-22 12:00:00	Chongqing	China	6.0	0.0	0.0
2020-01-22 12:00:00	Fujian	China	1.0	0.0	0.0

Распределение заражённых по Китаю



Количество заражённых коронавирусом

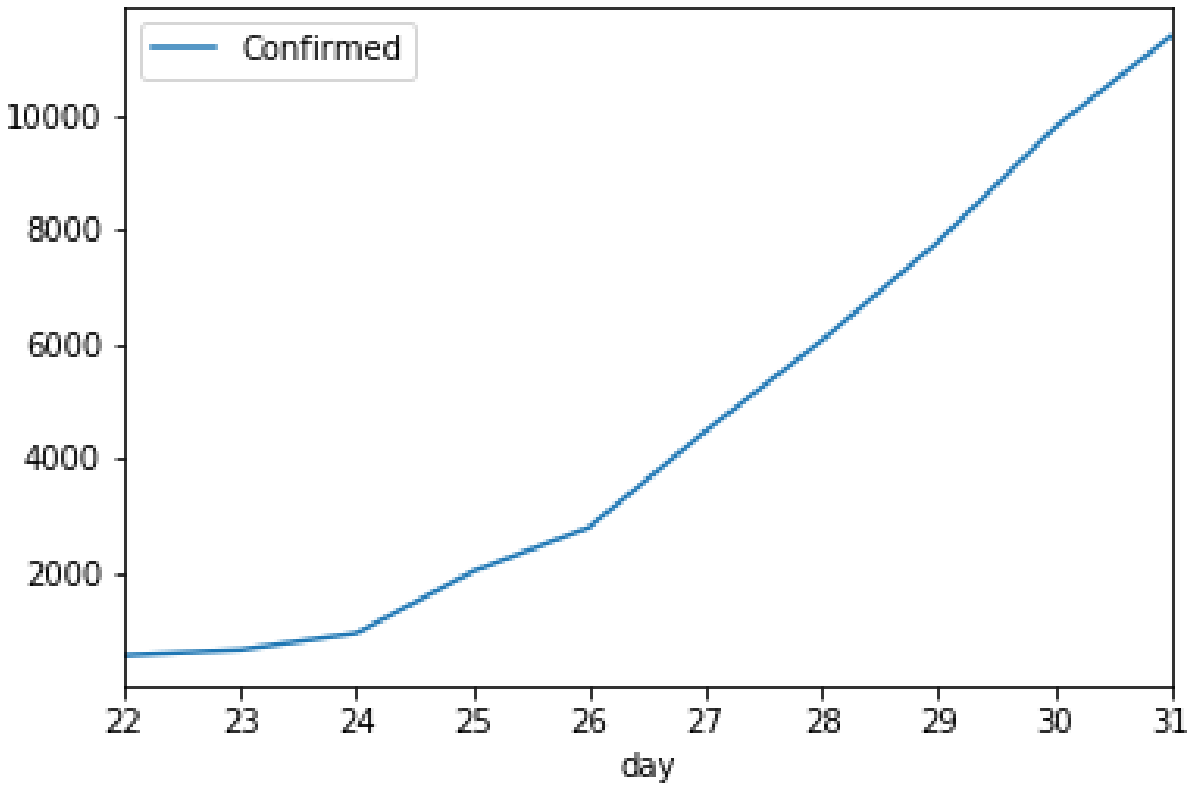


Рис.: Январь, 2020

Количество заражённых коронавирусом

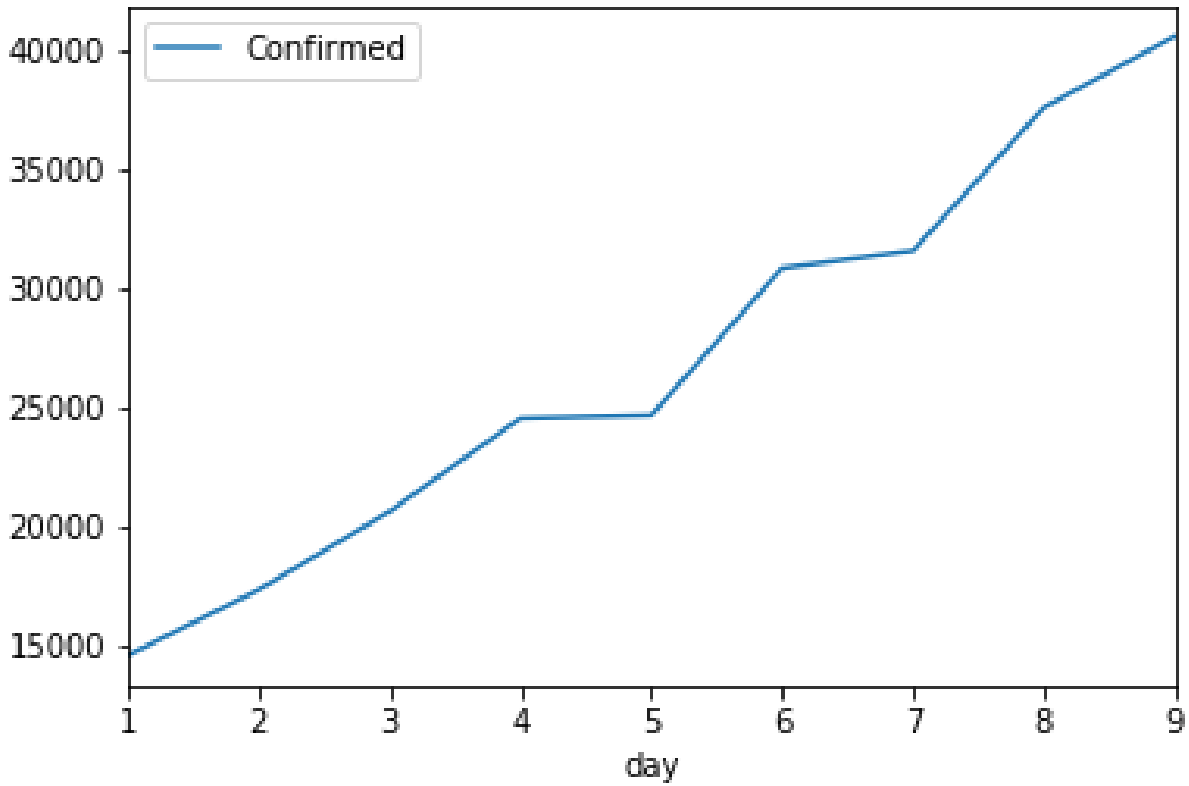


Рис.: Февраль, 2020

Летальность

- общее количество заражённых - 40536;
- общее количество смертей - 910;
- общее количество выздоровевших - 3312.

Летальность - 20%.

Визуализация очагов заражения

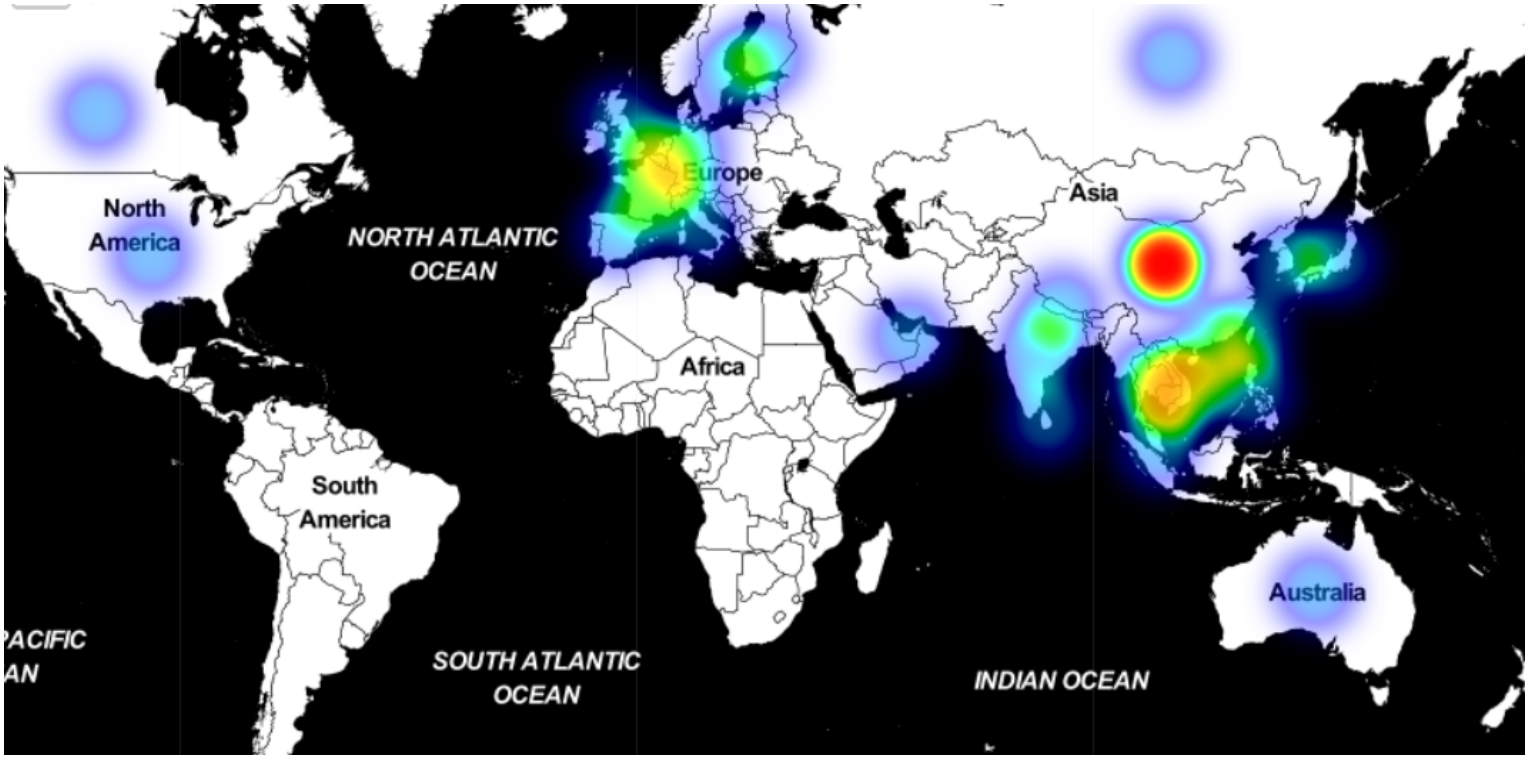


Рис.: Тепловая карта