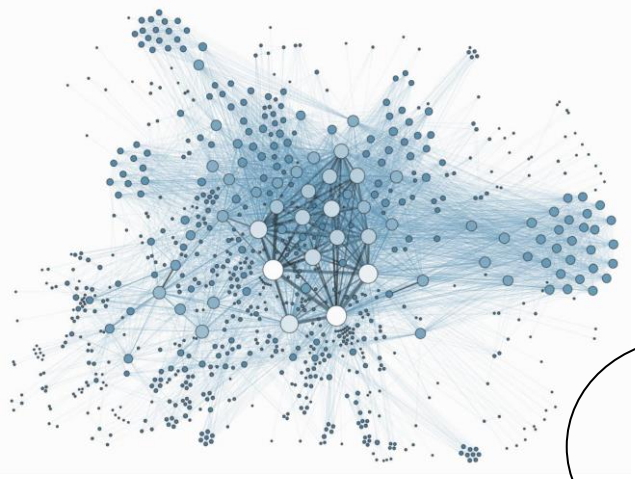
# ИССЛЕДОВАНИЕ ИНФОРМАЦИОННОГО ВЛИЯНИЯ И УПРАВЛЕНИЯ В СОЦИАЛЬНЫХ СЕТЯХ: КЛАССИФИКАЦИЯ И ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ СООБЩЕНИЙ

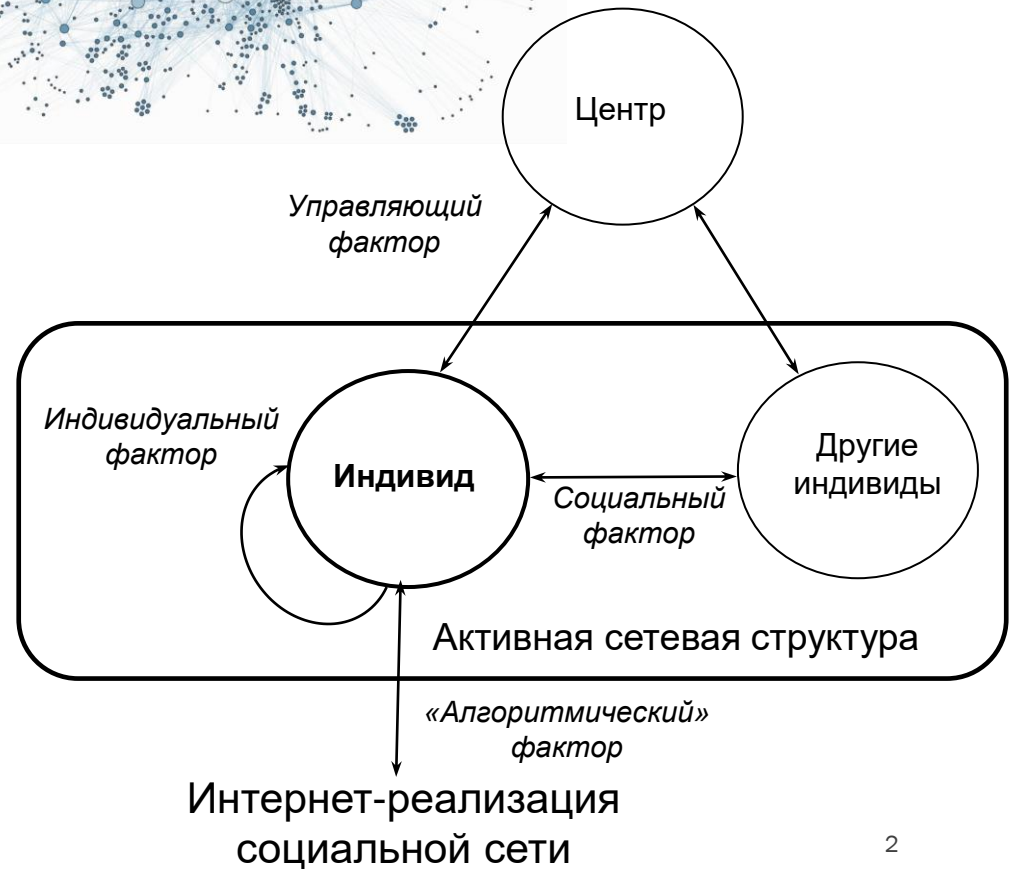
ГУБАНОВ ДМИТРИЙ АЛЕКСЕЕВИЧ
ДОКТОР ТЕХНИЧЕСКИХ НАУК
ВЕДУЩИЙ НАУЧНЫЙ СОТРУДНИК ИНСТИТУТА ПРОБЛЕМ УПРАВЛЕНИЯ РАН

МОСКВА 2023

# ПРОБЛЕМАТИКА

- Объект моделирования – *активная сетевая структура*, состоит из множества активных агентов и определенного на нем множества отношений.
  Пример – *онлайновые социальные сети*, в которых происходит размещение сообщений, высказывание мнений и реагирование на них. Второй пример – *научные сети*
- Представления агента (поведение) формируются под воздействием *информационного влияния* его окружения. Представления и поведения агентов влияют на экономику, политику и другие области деятельности общества
- Виды управления: мотивационное, институциональное, …. Наиболее «мягкое» и долгоиграющее – информационное



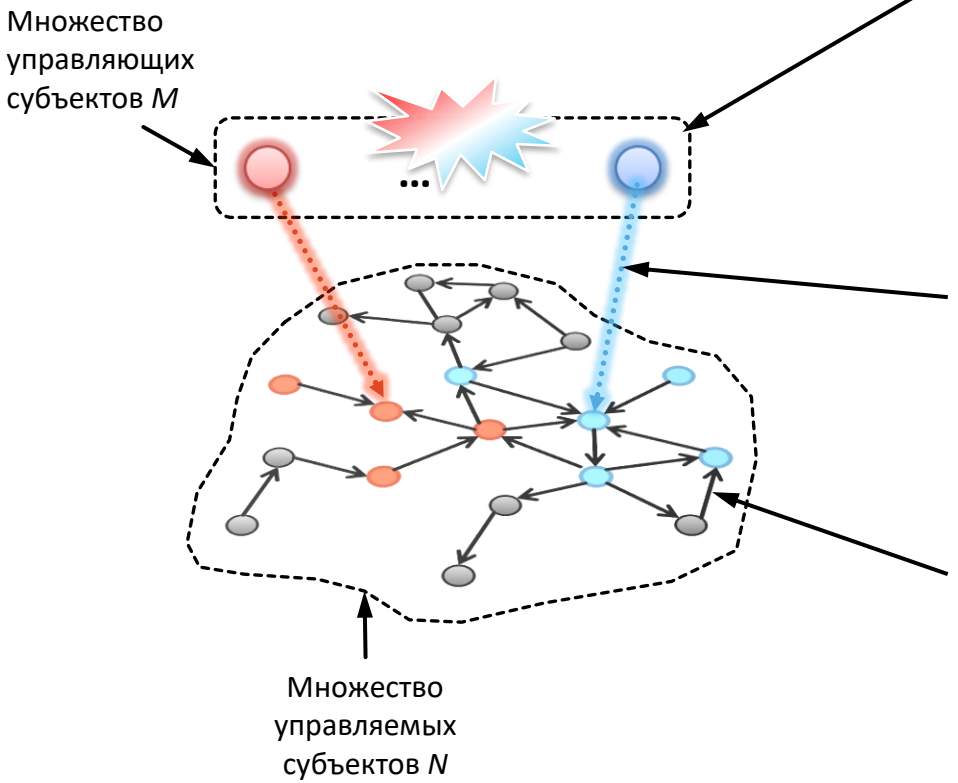


# УРОВНИ ОПИСАНИЯ И АНАЛИЗА СОЦИАЛЬНОЙ СЕТИ (ACC)*

| Уровень иерархии | Моделируемые явления/ процессы | Аппарат моделирования |
|---|---|---|
| 5 | Информационное противоборство | Теория игр, теория рефлексивных игр, теория принятия решений |
| 4 | Информационное управление | Теория оптимального управления, дискретная оптимизация |
| 3 | Информационное взаимодействие агентов | Теория динамических процессов на сложных сетях: марковские модели, конечные автоматы и др. |
| 2 | Анализ структурных свойств сети | Теория графов, теория сложных сетей |
| 1 | Анализ сети в целом | Статистические методы, методы семантического анализа и др. |

* Схема Новикова Д.А.

# МОДЕЛИРОВАНИЕ ФОРМИРОВАНИЯ МНЕНИЙ В СОЦИАЛЬНЫХ СЕТЯХ




## 5. Модели информационного противоборства

Каждый игрок из множества $M$ имеет возможность влиять на начальные мнения агентов $u_{ij}$ и заинтересован в формировании итоговых мнений $X_M$.

**Задача** – найти равновесные действия игроков в игре

$$\Gamma = (M, \{U_j\}_{j \in M}, \{G_j(\cdot)\}_{j \in M}).$$

## 4. Модели информационного управления

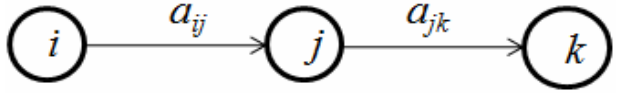
**Задача** – найти такой вектор управлений $u$, что:

$$\Phi(X, u) = H(X) - c(u) \to \max_{u \in U},$$

где $H(\cdot)$ – выигрыш, $c(\cdot)$ – затраты на управление.

## 3. Модели информационных взаимодействий

Агенты из $N$ образуют социальную сеть $G = (N, E)$.
Вектор начальных мнений $x$, конечных $X$.




$a_{ij} \geq 0$ – степень доверия $i$-го агента $j$-му, $k$-й агент косвенно влияет на $i$-го.

$$x^{k+1} = A\,[x^k + B\,u^k].$$

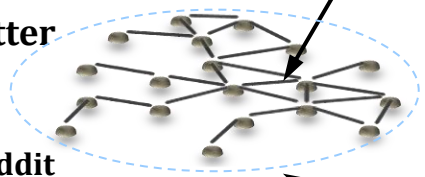
**Задача** – найти результирующее влияние одних агентов на других; найти агентов, формирующих итоговое мнение в сети.


?

## 2. Структурный анализ

VKontakte

Twitter



Reddit

Habr

## 1. Статистический анализ

# ИДЕНТИФИКАЦИЯ МНЕНИЙ ПОЛЬЗОВАТЕЛЕЙ СЕТИ

1. Сбор данных онлайновой социальной сети (ВКонтакте)
2. Разметка и предобработка исходного массива данных
3. Разработка методов классификации мнений в сообщениях
4. Обучение на размеченной выборке сообщений пользователей сети, тестирование, работа над ошибками
5. Предсказание мнений для полной выборки сообщений

-> Валидация моделей информационного влияния и управления

# ИСХОДНЫЕ ДАННЫЕ

**Источники информации –** публичные страницы ВКонтакте (информационные агентства, газеты, журналы, агрегаторы и т.п.), публикующие новости по общественно-значимым темам: РИА Новости, РБК, Москва 24, Медуза и др.

**Примеры ключевых слов для сбора постов о COVID-19**:

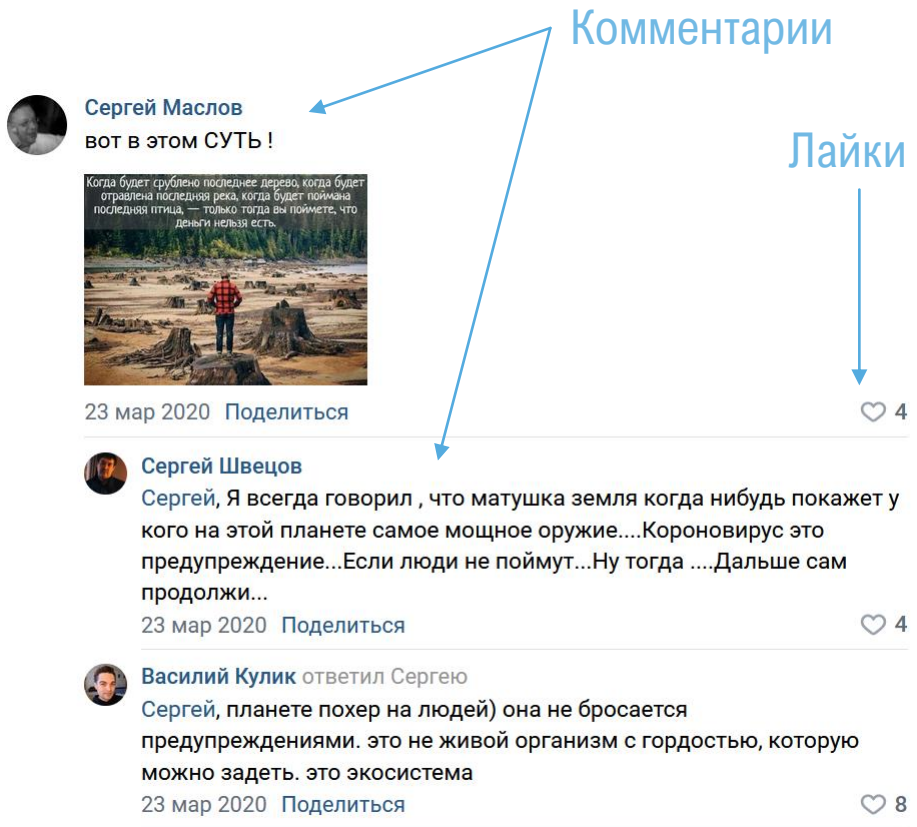
ковид, коронавирус, covid, coronavirus, карантин, удаленка, самоизоляция, пандемия, эпидемия.

**Период сбора данных**

март 2020 г. – февраль 2021 г.

**Информационные объекты**

- Посты источников – 50 тыс.
- Комментарии к постам – 2 млн.
- Лайки к постам и комментариям – 7 млн.



# ЗАДАЧА ИДЕНТИФИКАЦИИ МНЕНИЙ В СООБЩЕНИЯХ

Задача идентификации мнений, выраженных в сообщениях пользователей социальной сети:

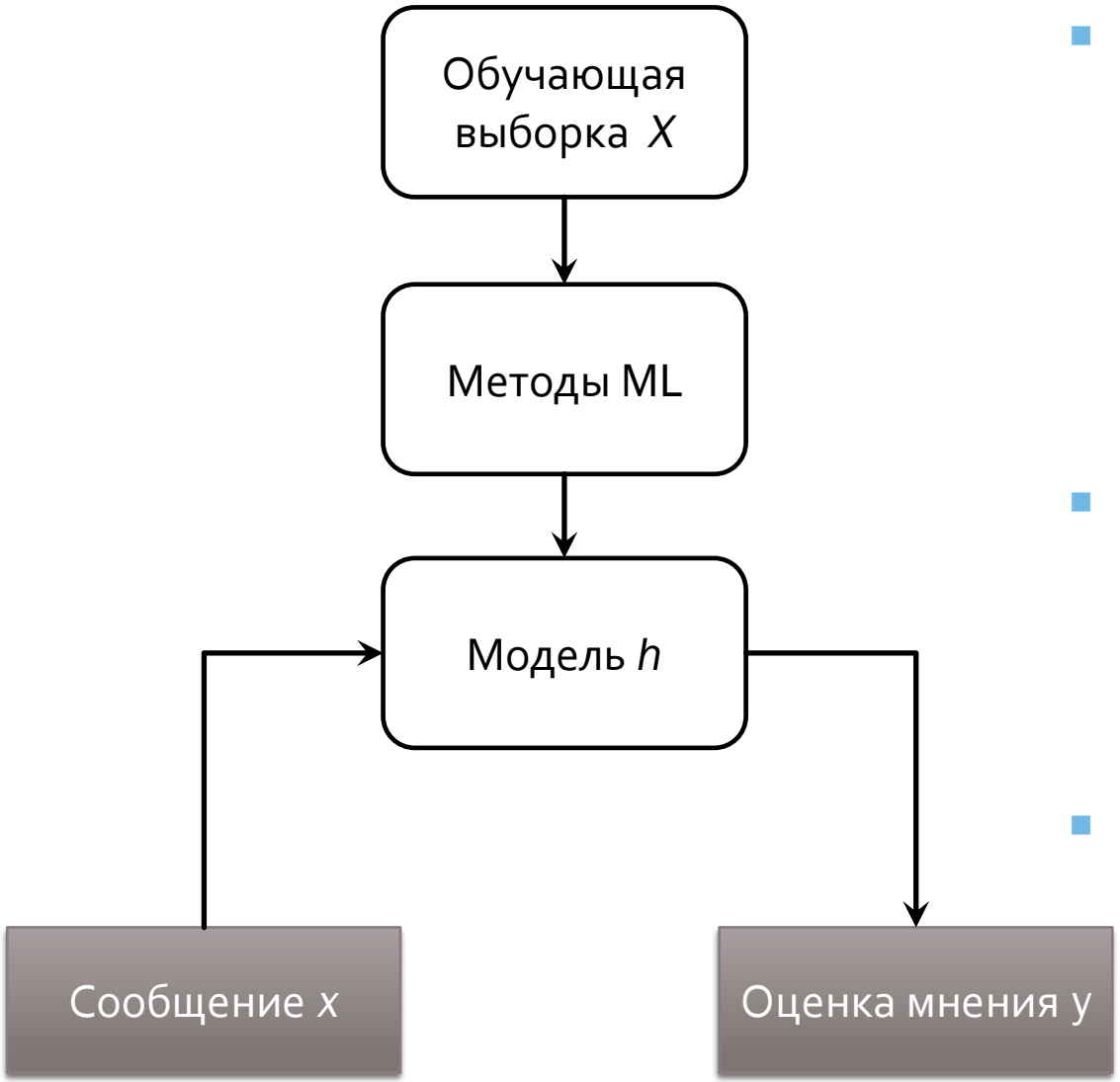
- $x$ – сообщение пользователя сети,
- $y \in \{0,1,2\}$, где 0 – мнение в сообщении против масок, 1 – за маски, 2 – нейтральное / нерелевантное.

Примеры сообщений (объектов) и мнений в них (ответов):

| # | Сообщение ($x$) | Мнение ($y$) |
|---|---|---|
| 1 | Сергей, а можно было бы этого всего избежать если бы все носили маски | 1 |
| 2 | Надо штрафовать всех, кто ходит без намордника и перчаток. | 1 |
| 3 | Про маски не забываем и руками не трогаем в транспорте поручни... | 1 |
| 4 | Я хожу без маски. Я кови-диссидент | 0 |
| 5 | Марина, Прикольно хохлушка рассуждает о масках в России)) 😂 | 2 |
| 6 | Наденьте маску, и будете вечно жить. И как раньше без масок жили. | 0 |
| 7 | Маски оказывается нужены для рабства. Они ни от чего не спасают. Эпидемии нет. Есть терроризм | 0 |

- $(x, y)$ – один обучающий пример
- $\left(x^{(i)}, y^{(i)}\right)$ – $i$-й обучающий пример
- $X = \left(x^{(i)}, y^{(i)}\right)_{i=1}^{m}$ – обучающая выборка, нужна для извлечения закономерностей из нее методами ML

# КЛАССИФИКАЦИЯ МНЕНИЙ В СООБЩЕНИЯХ




- Модель $h$ – функция, которая отображает сообщения ($x$) во множество возможных мнений ($y$)
  - Пример модели $h(x) = 2$
  - Пример модели $h(x) = \underset{k}{\operatorname{argmax}}\, \boldsymbol{\theta}_k^T \boldsymbol{x}$,
    где $\boldsymbol{\theta}$ – вектор значений параметров модели,
    $\boldsymbol{x}$ – признаковое описание объекта $x$ (вектор признаков)
- Как выбрать правильную модель для данной выборки $X$? Поиск минимума *функционала ошибки*:
  - $Q(h, X) \to \min_{h \in \mathcal{H}}$
    где $\mathcal{H}$ - множество возможных моделей
- Пример функционала – доля неправильных ответов

# ПРИЗНАКОВОЕ ОПИСАНИЕ ОБЪЕКТА (СООБЩЕНИЯ)

- Признак объекта – это число, характеризующее объект. Виды признаков: бинарные (0 или 1), вещественные (возраст), категориальные (город, пол), и т.д.
- Признаковое описание объекта (векторное представление) – совокупность всех признаков:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$$

- Как можно описать сообщение? Самый простой вариант – представление Bag-of-words (BoW), описывающее вхождение слов в сообщение. Две составляющие: а) словарь, б) мера присутствия слов в текстах.

**Словарь всего корпуса сообщений**
65 721 слов

| i | word |
|---|---|
| 6000 | актуальность |
| 6001 | актуальны |
| 6002 | актуальные |
| 6003 | актуальным |
| 6004 | актуальных |
| 6005 | акты |
| 6006 | актёр |
| 6007 | актёры |
| 6008 | акуе |
| 6009 | акунин |

**Пример векторного представления сообщения**

*x* = "*Маска предохраняет от капель слюны и пыли, на которых вирус и переносится. Потому что у него нет крыльев.*"

$\boldsymbol{x}$ = (0, 0, 0, ..., 0, 0, 0)

*d* = 65 721

Индексы ненулевых (единичных) элементов:
[10414, 23084, 25289, 25752, 27718, 29767, 31769, 32766, 36483, 38868, 43545, 44311, 48314, 53789, 63589]

Этим индексам соответствуют слова из словаря:
['*вирус*', '*капель*', '*которых*', '*крыльев*', '*маска*', '*на*', '*него*', '*нет*', '*от*', '*переносится*', '*потому*', '*предохраняет*', '*пыли*', '*слюны*', '*что*']
**Порядок слов не сохраняется!**

# ПРИЗНАКОВОЕ ОПИСАНИЕ ОБЪЕКТА (СООБЩЕНИЯ)

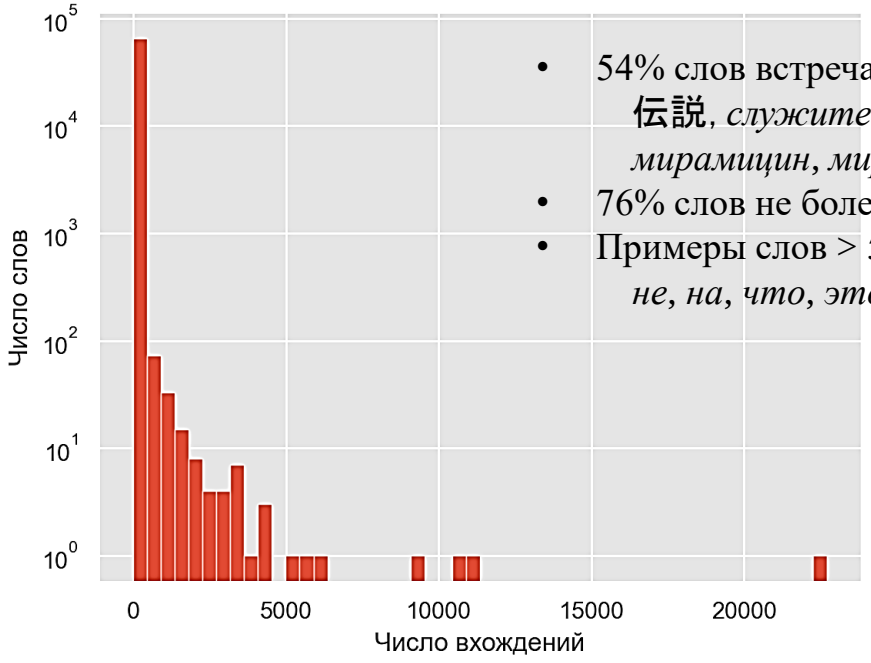
**Большое число признаков, $d$ = 65 721:**

- неэффективность расчетов, ограничения памяти
- переобучение, вероятность нахождения чрезмерно сложных и странных зависимостей между признаками и ответами

**Уменьшение размерности пространства признаков**

- Фильтрация стоп-слов (*и*, *или*, *это*, *такой*, *таким образом*, *а*, …)
- Фильтрация редко или часто используемых в сообщениях слов
- Приведение слов в нормальную форму
- Применение других методов понижения размерности (например, PCA)
- Ничего не делать с размерностью, использовать глубокое машинное обучение

Распределение числа слов по числу их вхождений в сообщения (гистограмма):




- 54% слов встречается однократно
  伝説, *служите*, *мираж*, *служило*, *мирамицин*, *миробами*, …
- 76% слов не более трех раз
- Примеры слов > 5000 раз:
  *не*, *на*, *что*, *это*, *то*, *как*, *от*, …

| i | word |
|---|---|
| 6000 | актуальность |
| 6001 | актуальны |
| 6002 | актуальные |
| 6003 | актуальным |
| 6004 | актуальных |
| 6005 | акты |

актуальный (лемма)

# ФОРМИРОВАНИЕ РАЗМЕЧЕННОЙ ВЫБОРКИ

Разметка комментариев в выборке *X*:

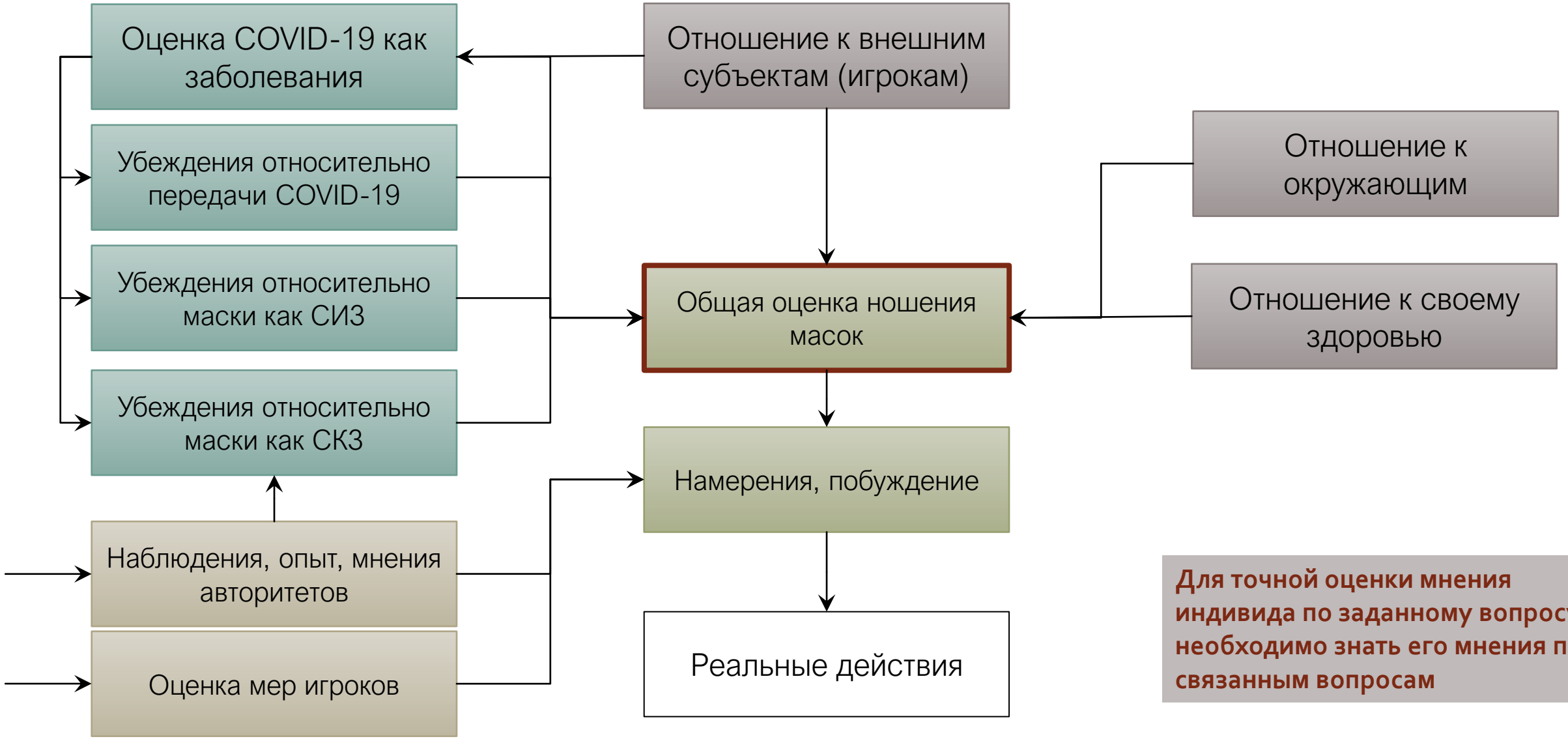
- «1» – положительное (ковид опасен, надо беречься),
- «0» – отрицательное (ковид не опасен и/или ограничения надоели / бесполезны)
- «2» – нейтральное / нерелевантное
  (действия государства и общая ситуация, нейтральные вопросы и информация)

Примеры разметки:

| Комментарий | Отношение к ношению |
|---|---|
| Зачем здоровому человеку маска? в том то и дело, что не все здоровы как им кажется | 1 |
| Маска тебе нужна чтобы твои заразные слюни не разлетались в округе когда ты чихаешь. | 1 |
| Были уже в Европе в конце февраля, а вы наверное будете маску до конца жизни носить как стадо? | 0 |
| Я 7 мес без маски хожу, и в автобусах и в магазинах, и ничего. | 0 |
| Маски не помогают, если человек здоров, но сейчас никто уже не знает, здоров он или нет. | 1 |
| Чё в масках пора спать или ещё нет? | 0 |
| Соблюдать масочный режим. Это самое главное. Берегите себя и своих родных. | 0 |

**Проблемы**: фрагментарная структура комментария, изобилие ошибок, длина комментария, отсутствие контекста

# УБЕЖДЕНИЯ И ФАКТОРЫ, ИСПОЛЬЗУЕМЫЕ В КАЧЕСТВЕ АРГУМЕНТОВ/ДОВОДОВ ЗА И ПРОТИВ НОШЕНИЯ МАСОК (ИСХОДЯ ИЗ СООБЩЕНИЙ ПОЛЬЗОВАТЕЛЕЙ)




**Для точной оценки мнения индивида по заданному вопросу необходимо знать его мнения по связанным вопросам**

# ФОРМИРОВАНИЕ РАЗМЕЧЕННОЙ ВЫБОРКИ И ПРЕДОБРАБОТКА

**Формирование размеченной выборки** $X = \left(x^{(i)}, y^{(i)}\right)_{i=1}^{m}$

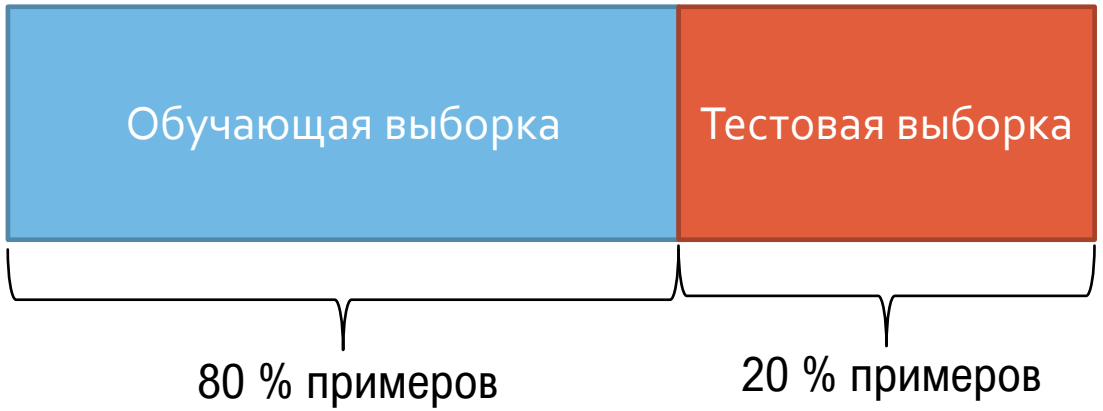
- $m$ = 8000 (комментариев)

**Предобработка текстов в размеченной выборке, в том числе:**

- Удаление обращений к собеседнику: «*[idXYZ|Серёга], попутал берега,…*»
- Удаление интернет-адресов в разных вариантах, устранение некоторых ошибок
- Приведение текстов в нижний регистр, лемматизация и т.д.
  (в зависимости от выбранного метода классификации)

# ФОРМИРОВАНИЕ ОБУЧАЮЩЕЙ И ТЕСТОВОЙ ВЫБОРКИ

- **Проблема переобучения**. На новых данных модель может показать плохое качество, т.е. она не обладает обобщающей способностью.
- **Тестовая выборка**. Как определить, что модель хорошая? Нужны дополнительные данные, отложить размеченные данные, на которые не проводится обучение.
- Как разбить исходную размеченную выборку
  - Маленькая тестовая часть – ненадежная оценка
  - Большая тестовая часть – потеря репрезентативности обучающей выборки
- Как оценить качество на новых данных:
  - Доля ошибок
  - Recall/Precision/Accuracy/F1-score

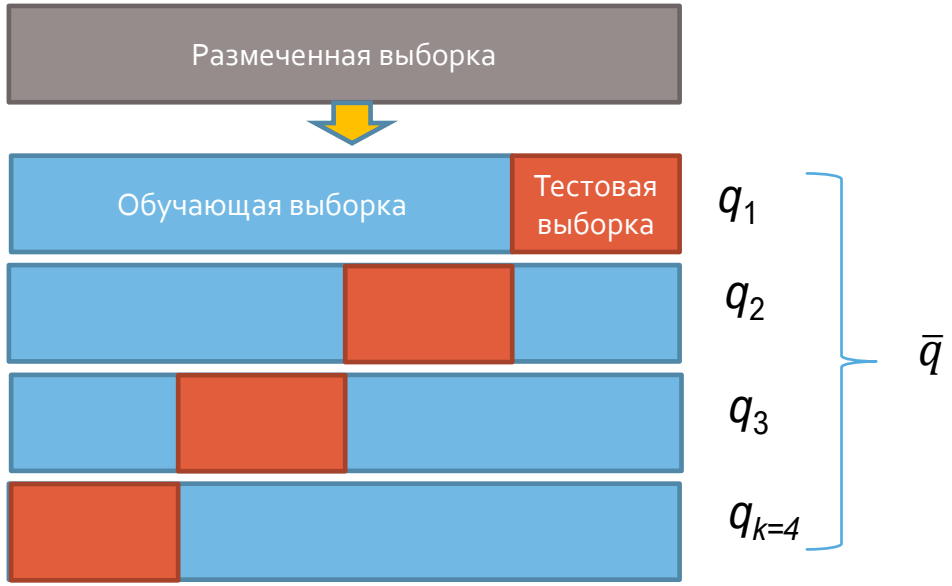


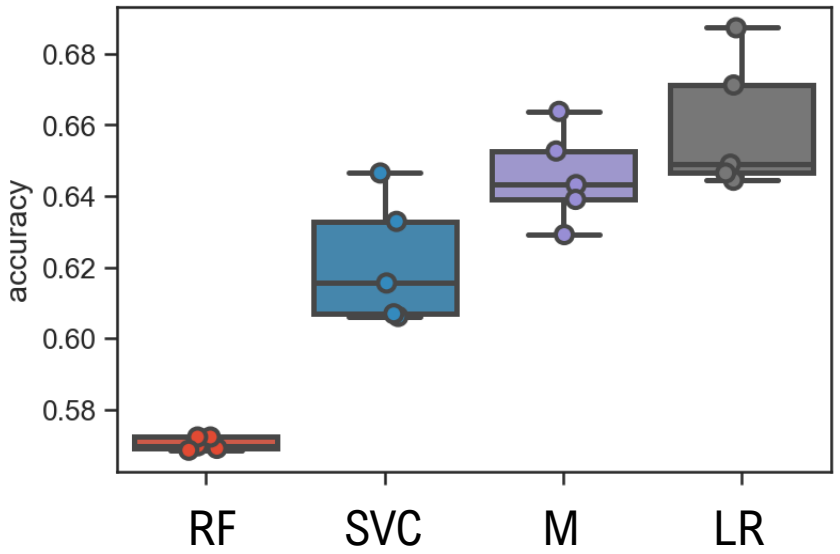

**Для отбора подходящих моделей (выбора гиперпараметров) вводят еще валидационную выборку**

# КРОСС-ВАЛИДАЦИЯ И СРАВНЕНИЕ МОДЕЛЕЙ

- **Кросс-валидация**
  - Разбиваем выборку на *k* частей
  - Каждая по очереди выступает как тестовая
- **Сравнение семейств моделей/алгоритмов** для классификации мнений в сообщениях (признаки - биграммы):
  - Случайный лес (RF)
  - Метод опорных векторов (SVC)
  - Байесовский классификатор (M)
  - Логистическая регрессия (LR)

Сравнение на основе кросс-валидации



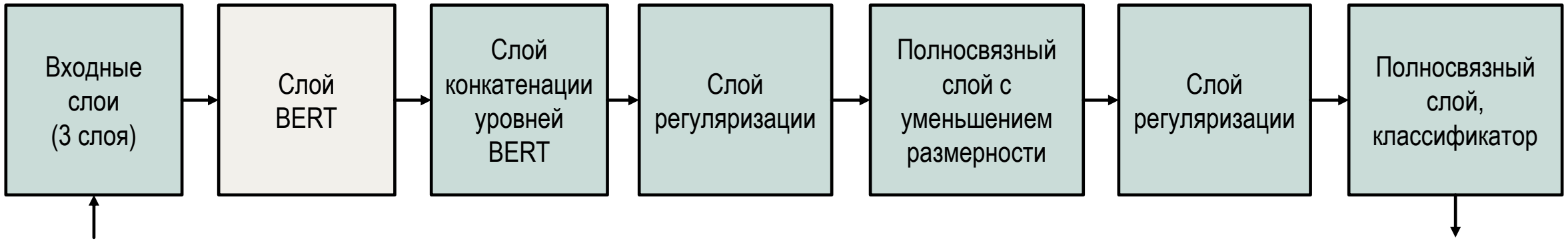



# РАЗРАБОТКА КЛАССИФИКАТОРА МНЕНИЙ

Применяется предобученная нейросетевая языковая модель BERT:

- Conversational RuBERT (размер словаря 120K, размер модели 630MB). Предобучена на данных корпусов OpenSubtitles, Dirty, Pikabu, Taiga (сегмент социальных медиа).
- Обучение состоит в восстановлении слов в предложениях и прогнозировании следующего в тексте предложения

Классификатор включает модель BERT, базовая архитектура:

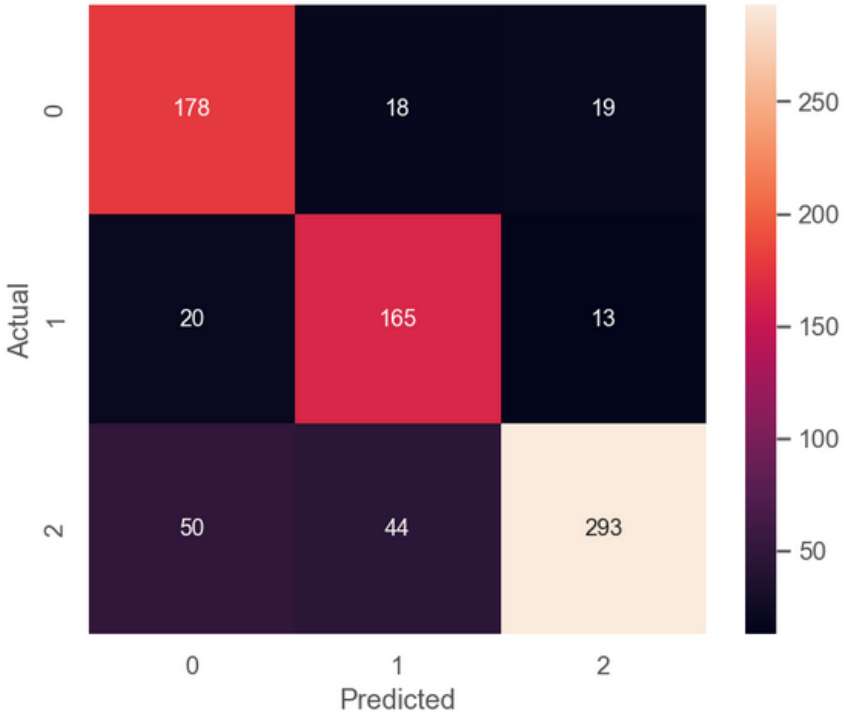



В классификаторе 180 миллионов обучаемых параметров.
В модели GPT-3 – 175 миллиардов. GPT-4 – 100 триллионов?

# РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ

- Макс. длина входной последовательности – 192 токена
- 4 эпохи дообучения: верность (accuracy) на тестовой выборке 0,8.

Матрица ошибок:

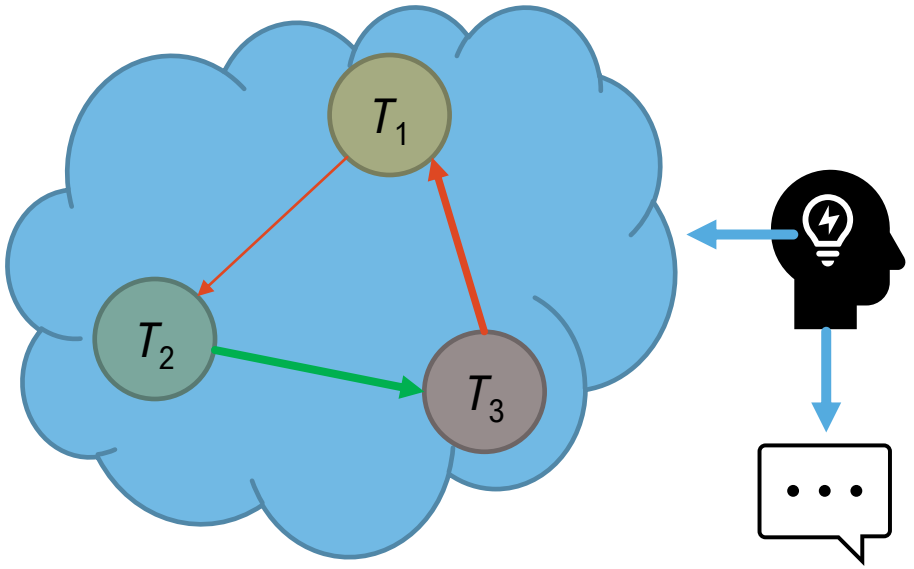



Примеры неверно классифицированных комментариев

| $y$ | $\hat{y}$ | Текст комментария |
|---|---|---|
| 0 ? | 1 | к тому же, вон весь тернет тут вопит, что лишь бараны в масках ) выходит, кто вопит, сам и баран или провокатор, желающий именно распространения у нас вирусняка, призывая не соблюдать режим самоизоляции. искать не надо, вот тут на стеночке каждый день стадо вопит, чтобы не подчинялись властям. |
| 1 | 0 | вообще почти никого не видел кто б простые маски элементарно подгонял плотно вокруг носа начни с теории чтоб пользоваться чем то разумно и маски фуфловые не бери чтоб защита была адекватная. |
| 0 | 1 | да по фиг!! как сказали носить маску, так я ее и ношу! правда в нагрудном кармане рубашки! и не потому - что я против чего - то там!! а просто достали уже!! хотите что - бы я носил маску!! да флаг вам в руки!! ношу уже месяц! в кормане!! |

# ТЕМЫ COVID-19

- Какие темы и аспекты обсуждений есть еще (помимо масок)? Насколько они массово обсуждаются? Какова взаимосвязь между темами? Как на мнение индивида по одному вопросу влияют мнения по другим вопросам?
- Возможные темы:
  - Происхождение коронавируса (в т.ч. конспирология)
  - Заражение ковидом, течение болезни
  - Профилактика и лечение (вакцина, вакцинация, масочный режим)
  - Тестирование и диагностика
  - Ограничения при коронавирусе (карантин, самоизоляция, …)
  - Осложнения при коронавирусе
  - Социальные проблемы и настроения
  - Воздействие на экономику, государственные меры поддержки
  - …
- Тематическое моделирование!

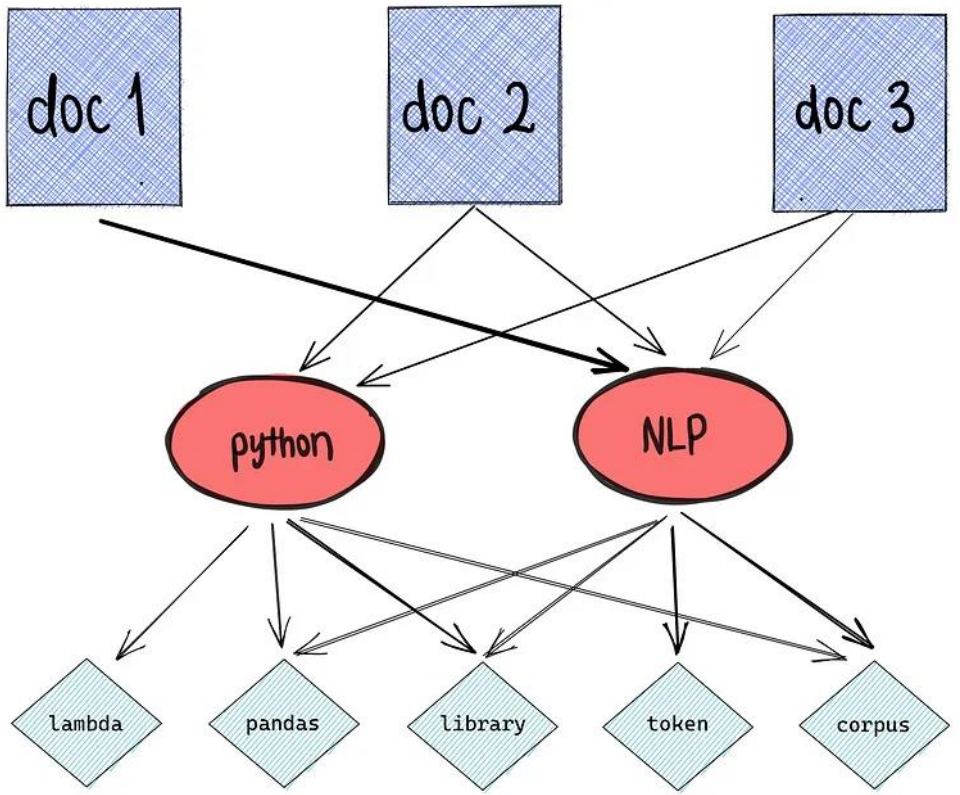


# ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

- *Тематическое моделирование* (topic modeling) — способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов […]
- *Тематическая модель* (topic model) коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему […]
- Поиск скрытых тем:
  - Вероятностные тематические модели
  - Методы кластерного анализа документов
  - Методы сингулярного разложения
  - …

# ЛАТЕНТНОЕ РАЗМЕЩЕНИЕ ДИРИХЛЕ (ВЕРОЯТНОСТНЫЕ МОДЕЛИ)

- LDA (Latent Dirichlet Allocation)
- *Документ – смесь тем*, каждое *слово* порождается одной из тем в этой смеси.
  - Существуют скрытые *темы*, отражающие содержание документа
  - Каждая тема – это распределение вероятностей на словах (мешок слов)
  - Каждый документ – это смесь тем, т.е. распределение вероятностей на темах
- Вероятностные модели можно представлять в виде порождающих процессов. Как порождаются слова в документах?




*Документы, темы, слова […]*

# ЛАТЕНТНОЕ РАЗМЕЩЕНИЕ ДИРИХЛЕ

Порождающий процесс:

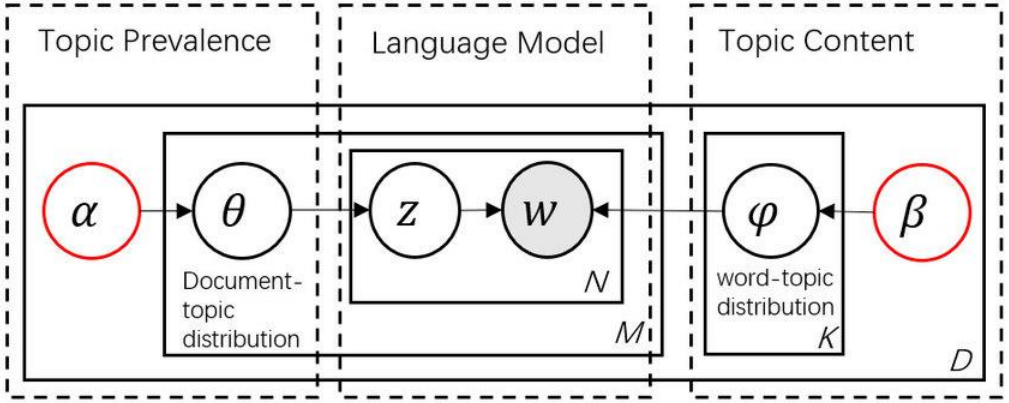
1. Для каждой темы $k \in \{1, \dots, K\}$ нужно выбрать вектор $\varphi_k \sim \mathrm{Dir}(\beta)$.
2. Для каждого документа $i \in \{1, \dots, M\}$ длины $N_i$:
   - выбрать вектор $\theta_i \sim \mathrm{Dir}(\alpha)$ — вектор «степени выраженности» каждой темы в этом документе $i \in \{1, \dots, M\}$;
   - для каждой позиции $j \in \{1, \dots N_i\}$ выберем слово *w*:
     - выбрать тему $z_{i,j}$ по распределению $z_{i,j} \sim \mathrm{Cat}(\theta_i)$;
     - выбрать слово $w_{i,j} \sim \mathrm{Cat}(\varphi_{z_{i,j}})$.

Совместное распределение:

$$P(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} | \alpha, \beta) = \prod_{k=1}^{K} P(\phi_k | \beta) \prod_{i=1}^{M} P(\theta_i | \alpha) \prod_{j=1}^{N} P(z_{i,j} | \theta_i) P(w_{i,j} | \varphi_{z_{i,j}})$$

- Необходимо найти вектора $\boldsymbol{\theta}$ и вектора $\boldsymbol{\phi}$:
  — получить для каждого документа список встречающихся тем,
  — получить для каждой темы список характерных слов

*Граф вероятностной модели […]:*



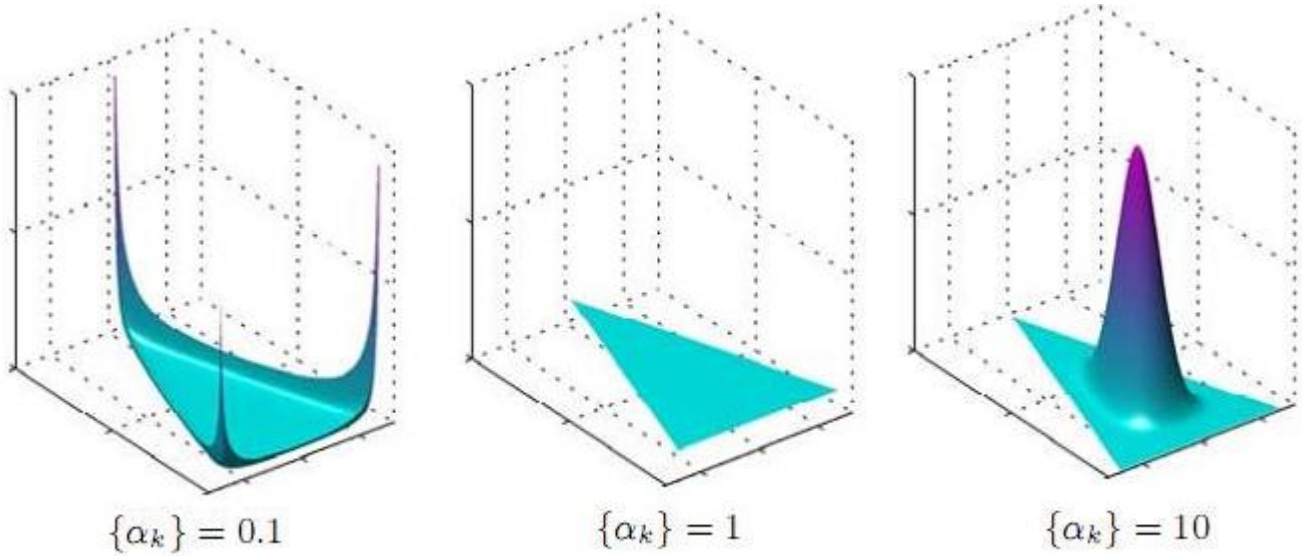
# РАСПРЕДЕЛЕНИЕ ДИРИХЛЕ

- Плотность распределения:

$$\mathrm{Dir}(\boldsymbol{\alpha}) \rightarrow \mathrm{p}(\,\boldsymbol{\theta} \mid \boldsymbol{\alpha}\,) = \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_{\mathrm{i}}\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \boldsymbol{\theta}_{\mathrm{i}}^{\alpha_{\mathrm{i}}-1}$$

- где *k* – количество элементов (например, тем),
- $\boldsymbol{\alpha}$ – параметры распределения (вектор размерности *k*)

*Графики плотности распределения Дирихле при различных параметрах, k = 3*



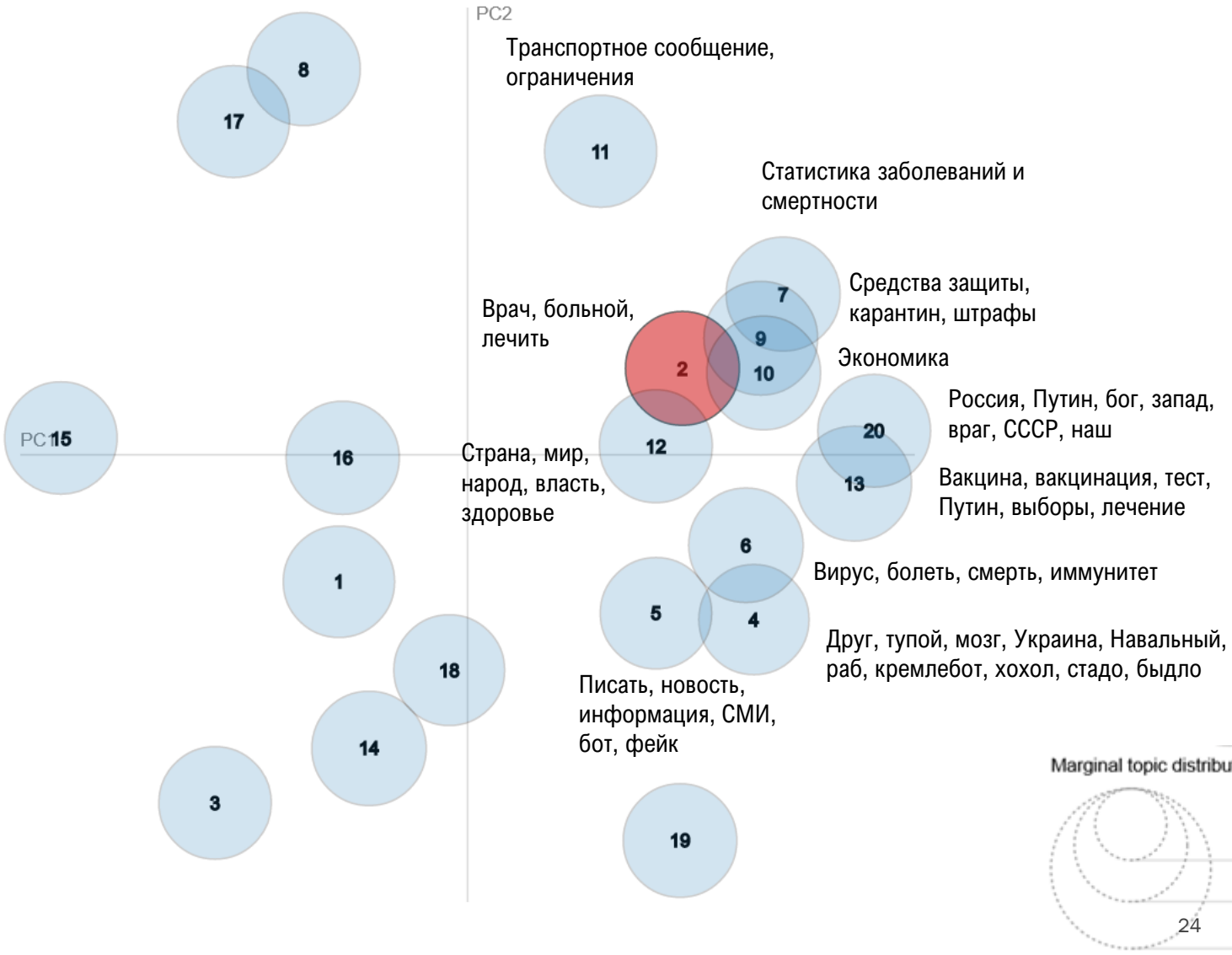

Распределение Дирихле является сопряженным к категориальному

# LDA МОДЕЛИРОВАНИЕ КОММЕНТАРИЕВ ПО ТЕМЕ COVID-19

- Всего 2,1 млн неразмеченных комментариев к «ковидным» постам
- **Процесс обработки**:
  1. Предобработка комментариев (напр., удаление обращений к собеседнику)
  2. Токенизация комментариев (текст как последовательность токенов)
  3. Фильтрация комментариев
     - Удаление мусорных токенов
     - $\leq$ 100 токенов в тексте (99% текстов) и $\geq$ 3 токенов в тексте
  4. Лемматизация комментариев (текст как последовательность лемм)
  5. Фильтрация комментариев по PoS (результат 1,6 млн. комментариев)
     - В тексте есть существительное
     - В тексте есть две «содержательные» леммы (существительное, глагол, прилагательное и т.д.)
  6. Формирование словаря (не меньше, чем в 2 документах, не больше чем в 50% документов)
  7. Формирование векторного представления текстов (мешок слов)
  8. Тематическое моделирование текстов (векторное представление), поиск матриц $\Theta$ и $B$

# LDA МОДЕЛИРОВАНИЕ КОММЕНТАРИЕВ ПО ТЕМЕ COVID-19

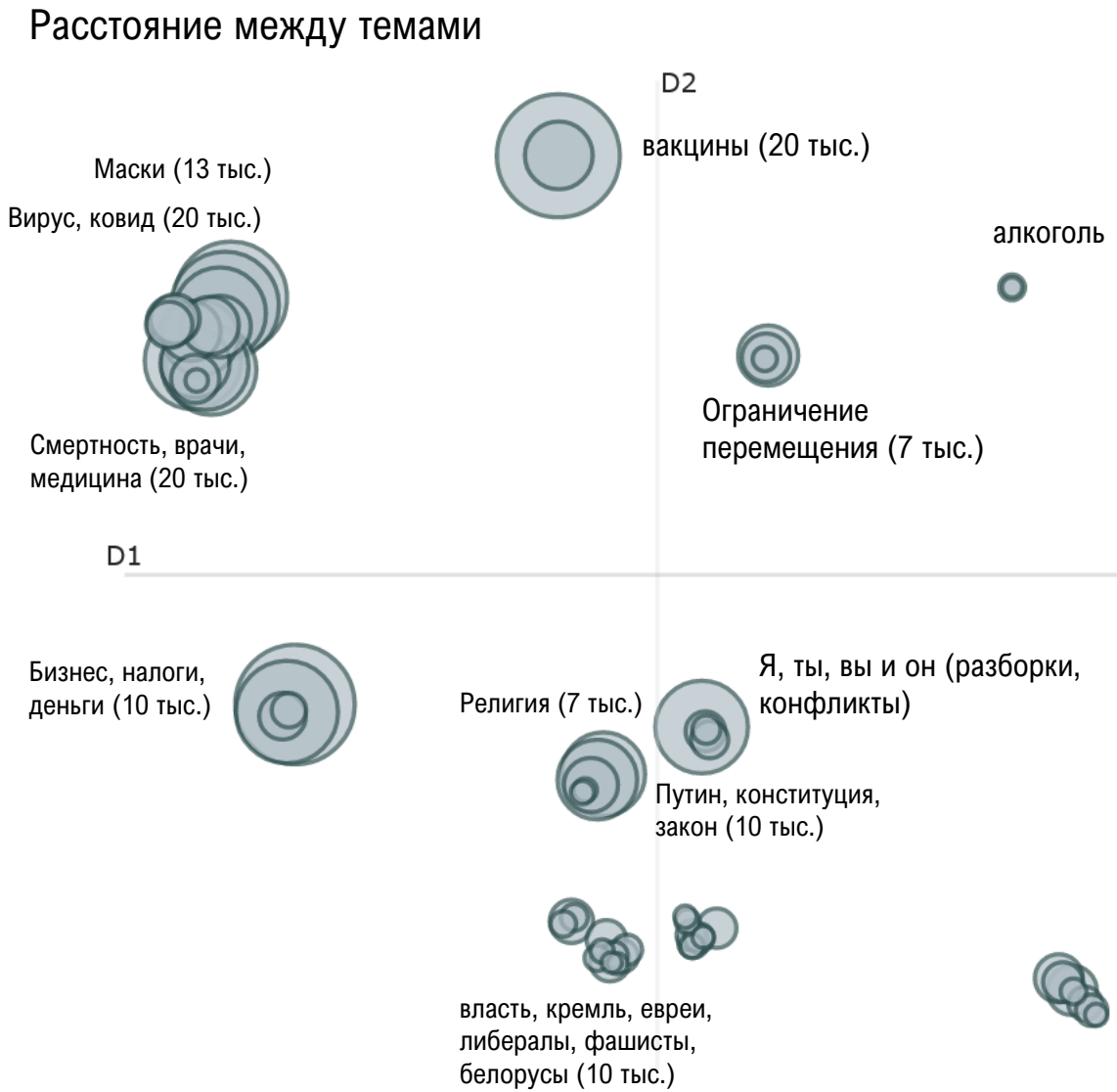
- Результаты тематического моделирования
- 20 тем (?),
  когерентность 0,57 (от 0 до 1)

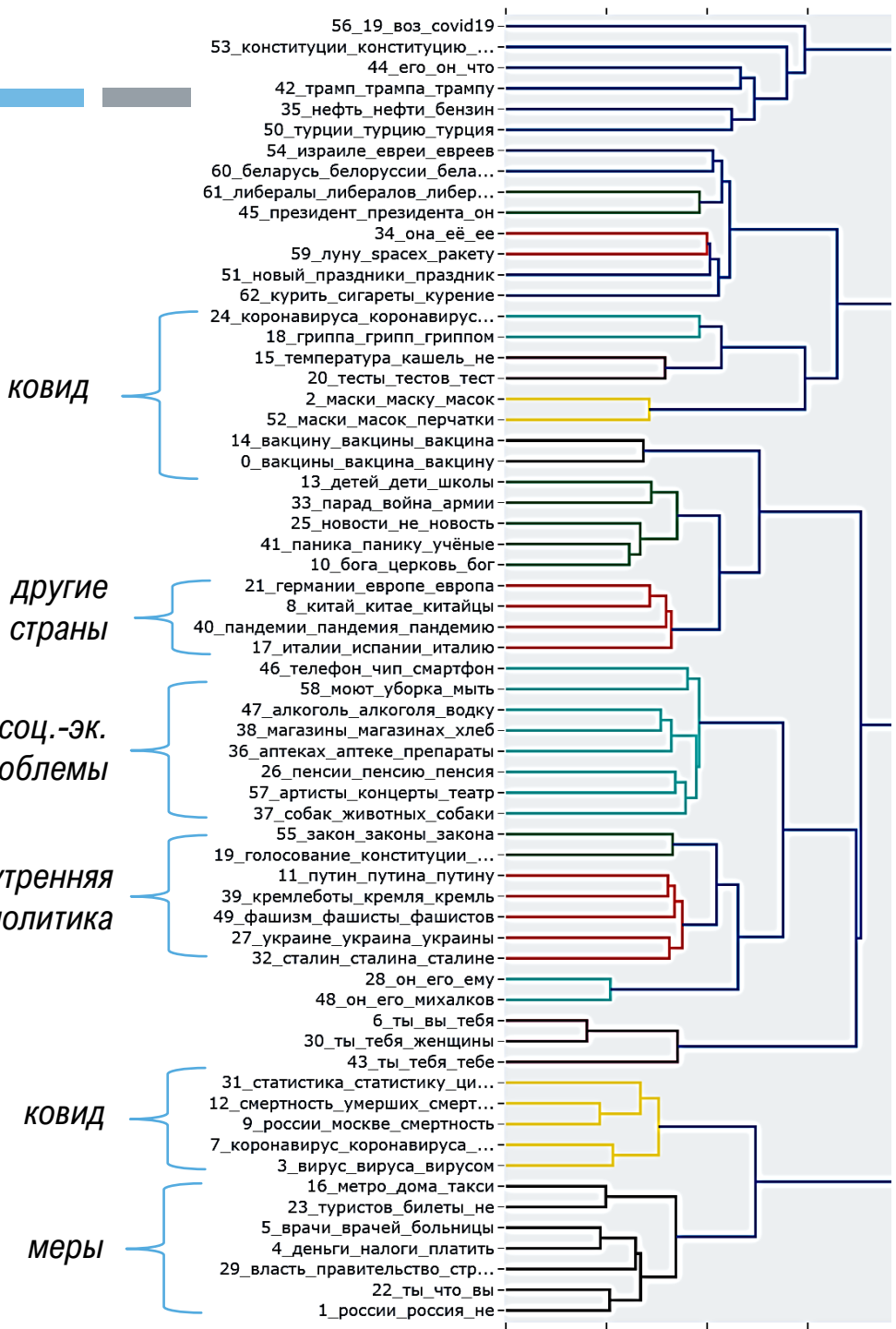


# ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ: ТРАНСФОРМЕРЫ И КЛАСТЕРИЗАЦИЯ

- *BERTopic*:
  1. Векторное представление текстов (трансформеры)
     - Текст в вектор (размерность 400) на основе предобученных нейросетевых языковых модели
  2. Понижение размерности векторных представлений (PCA, UMAP)
     - 400 → 5 компонент
  3. Кластеризация точек в пространстве (k-means, кластеризация по плотности HDBSCAN)
     - Число кластеров (тем) может быть ограничено
  4. Представление тем (c-tf-idf)
- Дополнительная фильтрация корпуса комментариев:
  - Число токенов не меньше 20
  - 400 тыс. комментариев из 1,6 млн

# ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ: BERTOPIC



# ДАЛЬНЕЙШИЕ ШАГИ

- Извлечение устойчивых и осмысленных кластеров
- Формирование согласованного перечня значимых дискуссионных тем по COVID-19
- Формирование набора релевантных комментариев по каждой теме
- Разметка набора комментариев на предмет отношения к заданной теме
- Классификация мнений пользователей социальной сети
- Исследование взаимосвязей: *темы – индивиды – мнения в обществе*
- Разработка математических моделей динамики тем и многомерных мнений в социальных сетях

# СПАСИБО ЗА ВНИМАНИЕ!