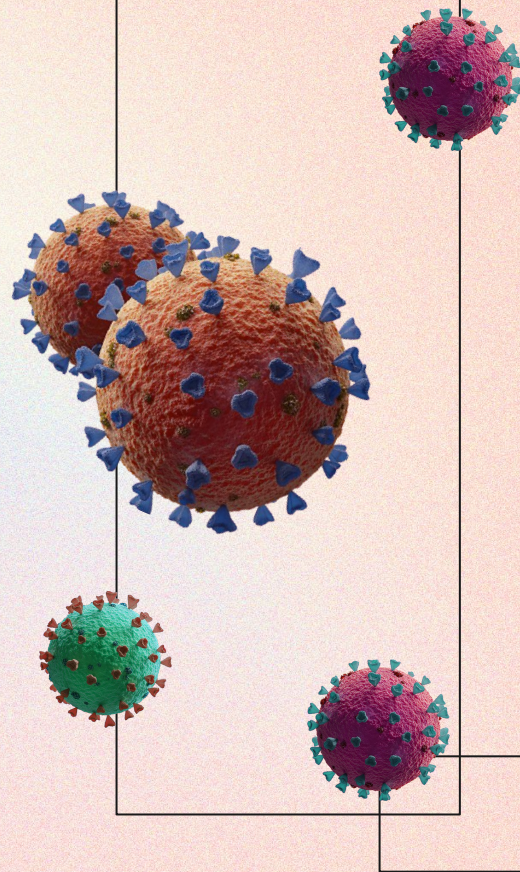




Карта динамики распространения коронавируса по данным телеграм- канала BBC RUSSIA

Криволап Мария
Кузьмина Александра
Харская Стефания
Чан Тхюи Хуен
ФИКЛ 2022, БКЛ 213

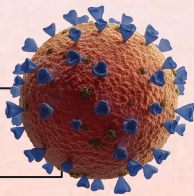
Куратор: Бузанов Антон



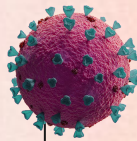
Цель: создать карту, с помощью которой можно будет увидеть:



Состояние эпидемии в мире (где распространился вирус) на конкретный момент времени, что будет показано через закрашивание маркеров на карте



Насколько неблагоприятна в стране/регионе/области эпидем-обстановка, что будет отображаться как более яркое/бледное закрашивание маркера



Задачи

01

DICT_COVID

Выкачать новости с коронавирусной обстановкой в мире и сделать из них словарь

04

DYNAMIC

Выявить положительную/отрицательную динамику в каждой из упомянутых в новостях стран/регионов/городов

02

NATASHA

Вытащить и лемматизировать географические топонимы из новостей

05

DATAFRAME

Записать все данные в DataFrame

03

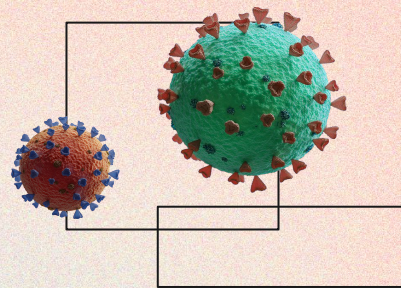
GEOPANDAS

Присвоить каждой локации координаты

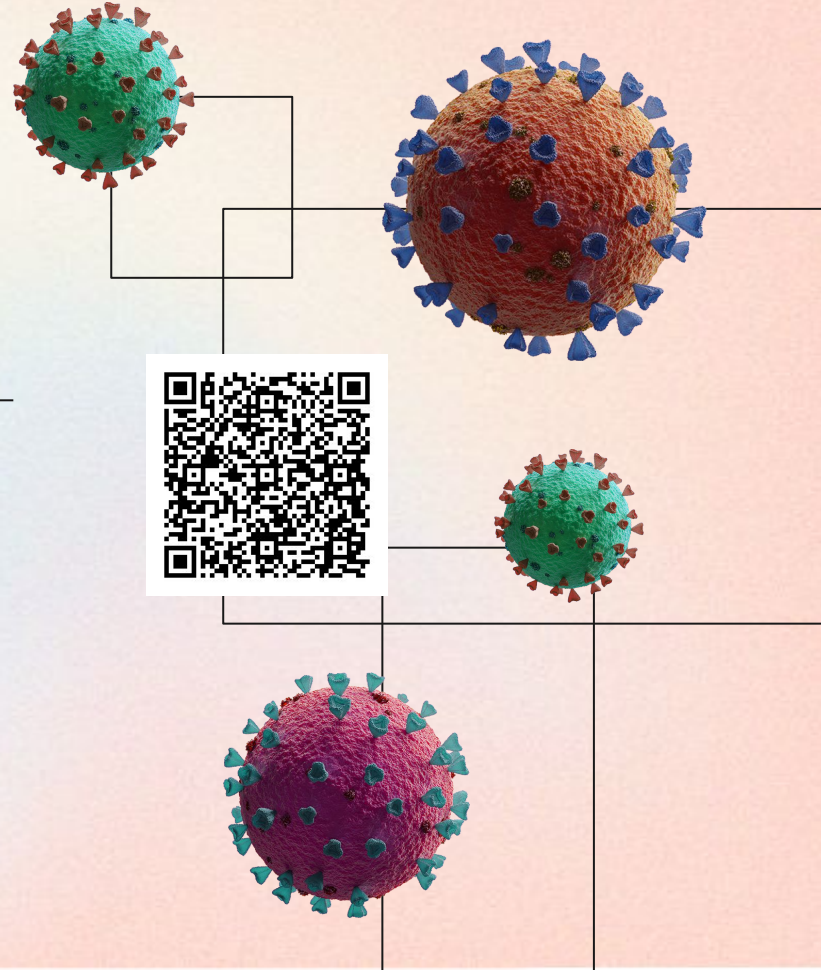
06

COVID MAP

Вывести все данные на интерактивной карте, отражающей ситуацию в период с января 2020 по февраль 2022 года



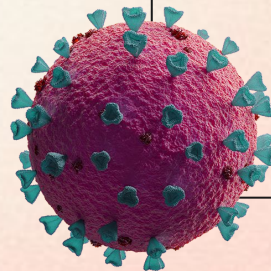
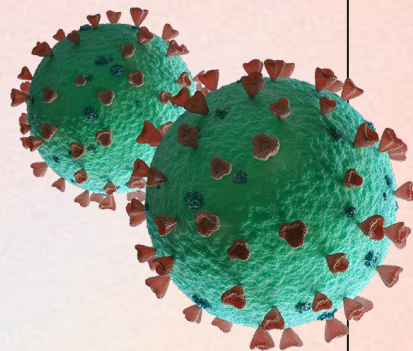
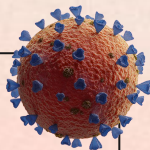
DICT_COVID



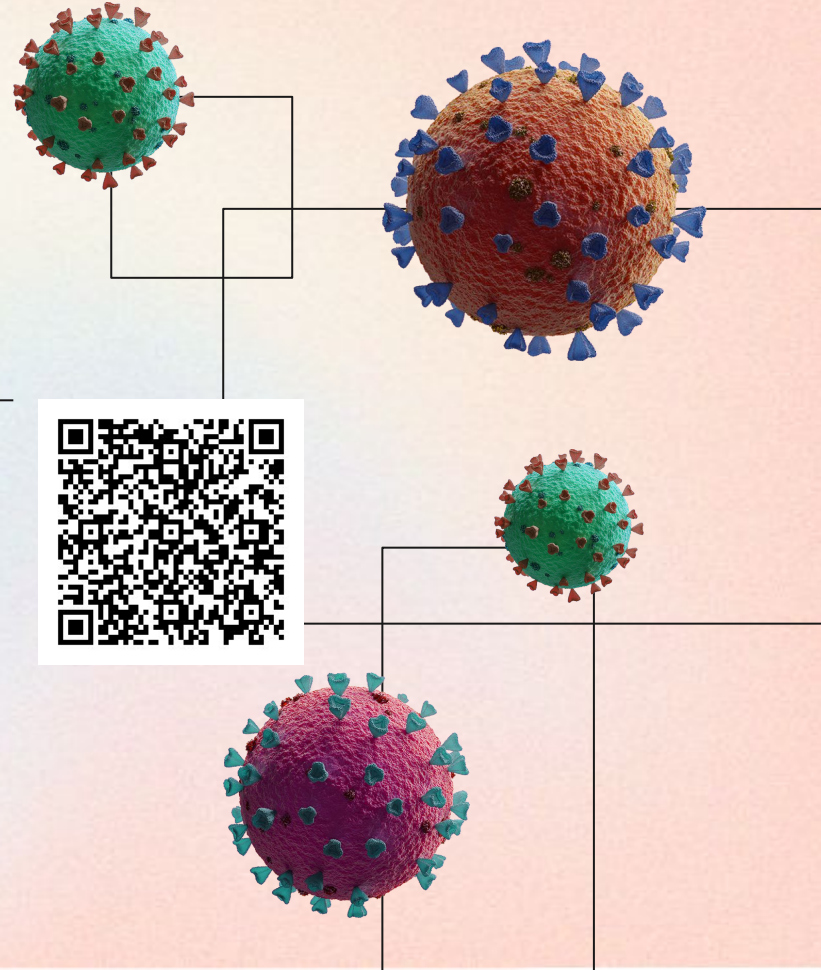
Для дальнейшей работы с данными мы создали словарь вида `{дата: новости}` и сложили его в `news_dict.json`. После было необходимо из всех данных отобрать только те, где содержалась информация про распространение коронавируса. Для этого мы воспользовались регулярками `[Cc]ovid[COVID]коронавирус[Пп]андем[a-я]`.

Почему именно такие?

При выборе телеграм канала мы пытались посмотреть как можно больше новостей про распространение коронавируса и чаще всего в них употреблялись именно эти слова, на основе которых мы и составили регулярки. После мы сложили полученный словарь про распространение коронавируса в отдельный файл, который назвали `Covid_dict.json`.

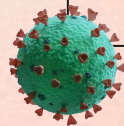
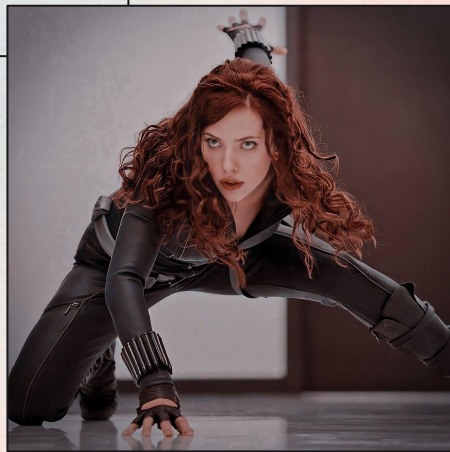
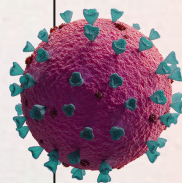


LOCATIONS



NATASHA

Мы нашли библиотеку, которая называется [Natasha](#). Эта библиотека может извлекать именованные сущности, а также лемматизировать слова. Если соответственно у слова был тип “[LOC](#)”, то мы его добавляли в список, который располагается в начале каждого сообщения. Таким образом у нас получился файл [Country_and_message.json](#), где к каждой дате в начале идёт список стран, а потом сообщение/новость.

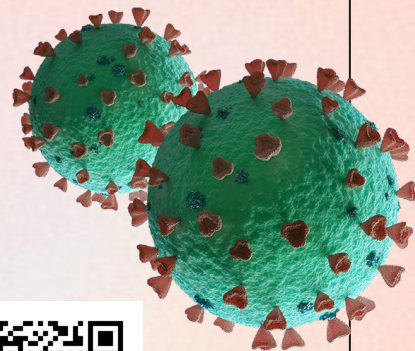
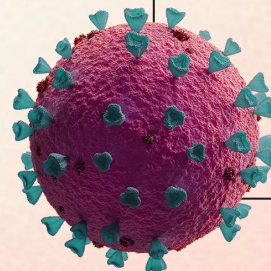
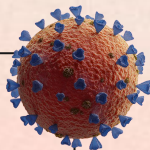


DEEP_TRANSLATOR

Воспользовавшись модулем `deep_translator` мы перевели все топонимы на английский язык для дальнейшей работы с картой.

Россия → Russia

США → USA

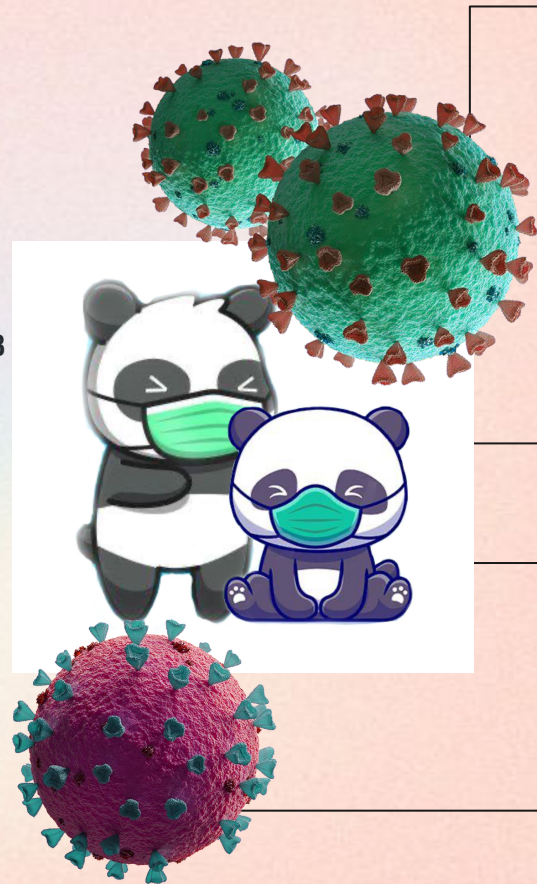
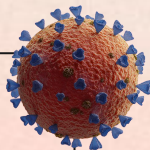


GEOPANDAS

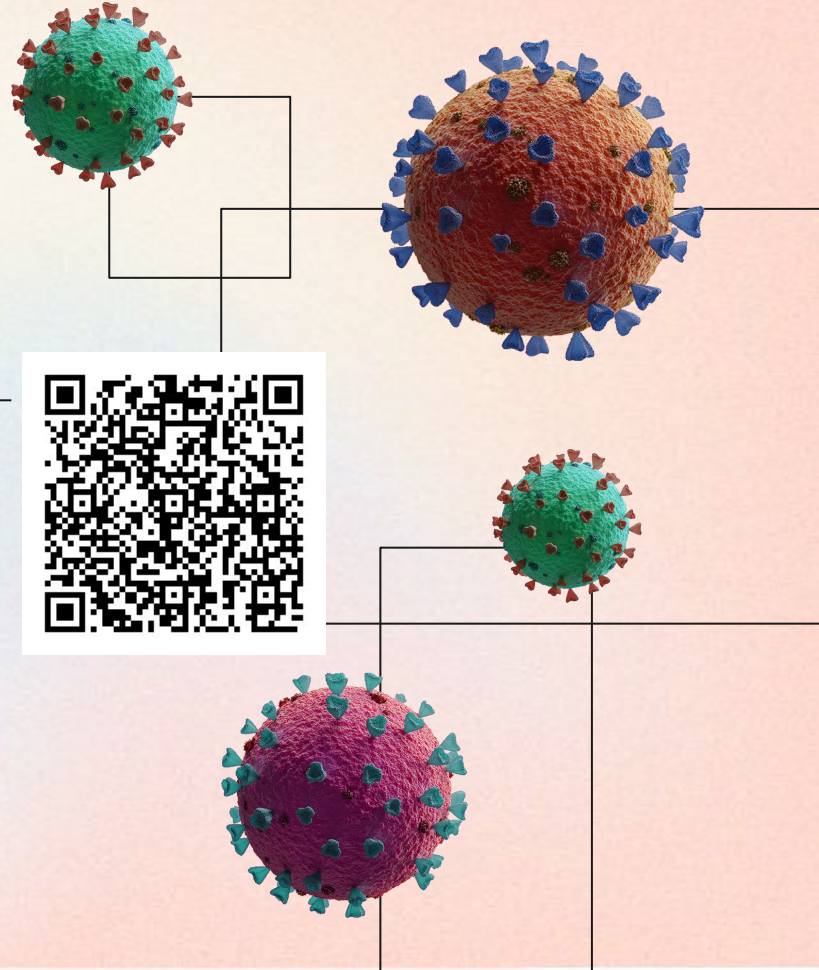
С помощью [geopandas](#) мы присваиваем координаты каждой из локаций, чтобы потом записать их в датасет и отметить на карте. Все результаты записывались в промежуточные файлы.

```
"China": {"lon": 104.999927, "lat": 35.000074}
```

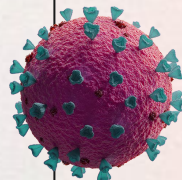
```
"Germany": {"lon": 10.4234469, "lat": 51.0834196}
```



DYNAMIC



DYNAMIC



После того, как каждой стране были присвоены координаты, было необходимо изучить динамику распространения вируса, пройдясь по тексту новости регулярками, присваивая каждой стране «+» или «-» в зависимости от содержания сообщения.

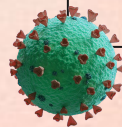
Для **положительной динамики** были использованы следующая регулярка:

```
[Мм]иновал|[Оо]слабл[а-я]+|[Сс]нят[а-я]+|[Уу]пад[а-я]+|[Сс]ниж[а-я]+|[Вв]ыходит|[Сс]мягч[а-я]+|[Пп]ад[а-я]*|[Зз]амедл[а-я]+|[Уу]был[а-я]+|[Сс]нима[а-я]+
```

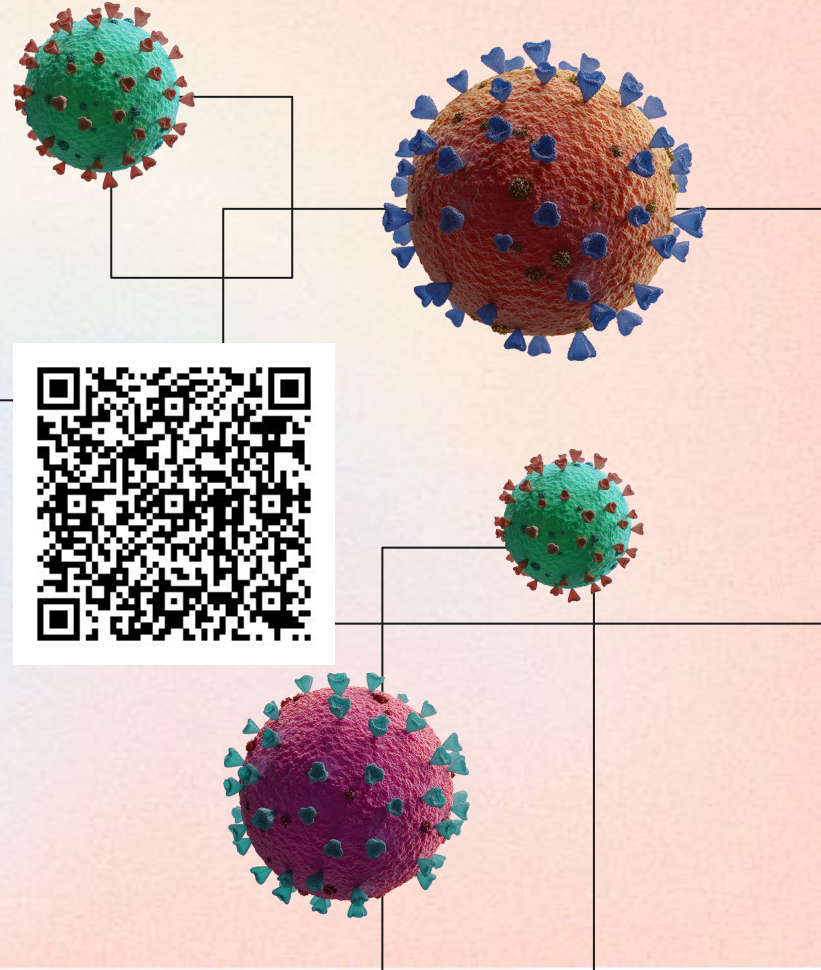
Для **отрицательной динамики** были использованы следующая регулярка:

```
[Пп]ик[а-я]|[Вв]спышк[а-я]|[Пп]ревы[а-я]+|[Уу]велич[а-я]+|[А-Яа-я]+?рекорд[а-я]+|[Уу]худш[а-я]+|[Р-р][ао]ст[а-я]+|[Зз]акры[а-я]+|[Вв]в[ео]д[а-я]т([а-я]+)?|[Мм]аксим[а-я]+|[Вв]ы?рос[тл][а-я]+|[Пп]рирост[а-я]|[Сс]кач[а-я]+|более|снова|[Уу]сил[а-я]+
```

После полученные данные были записаны в [Country_and_coord_and_dynFULL.json](#).

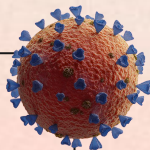
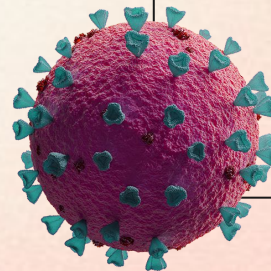
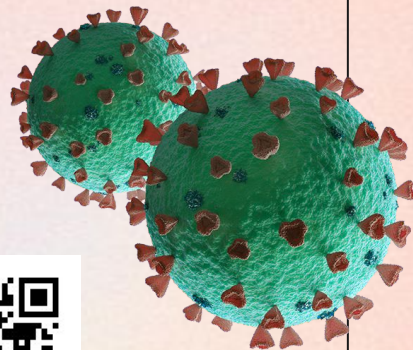


DATAFRAME

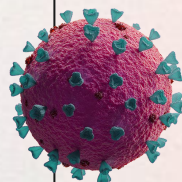


DELETE_TEXT

К тому моменту работы мы уже взяли из самого текста сообщений всё, что могли, поэтому следующим шагом стало удаление текста сообщений из словаря.

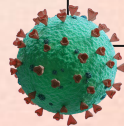


DATAFRAME



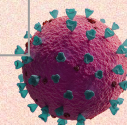
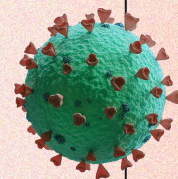
Мы создали новый словарь, в котором ключами были даты, а значениями — словари по каждой локации на этот момент. Каждой локации соответствовала накопленная ею динамика: встретив сообщение с динамикой "-" мы отнимали единицу, с динамикой "+" — соответственно прибавляли. Для координат мы создали отдельный словарь.

Чтобы на основе датафрейма получилась анимация, недостаточно было просто иметь в нем строками даты, а столбцами - страны. Поэтому датафрейм был создан такого вида: страны повторялись в строках столько раз, сколько было дат. В столбце "динамика" для каждой страны на каждый момент времени было соответствующее число. Последним шагом мы добавили каждой стране ее координаты.

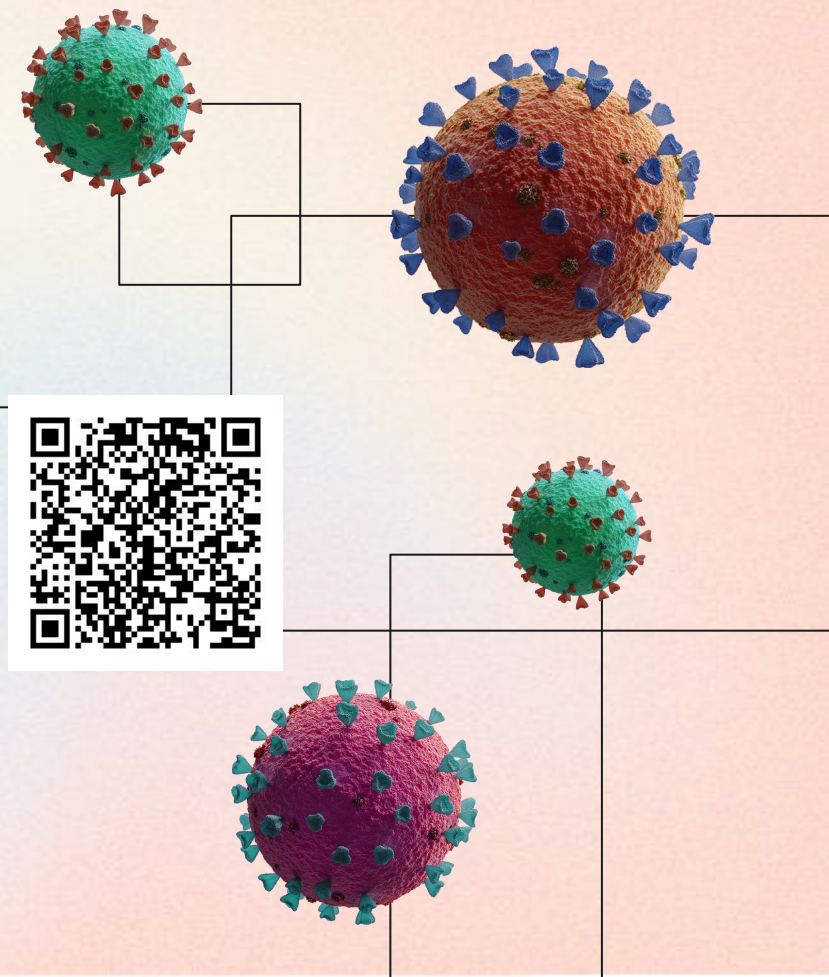


DATAFRAME

	Место	Дата	Динамика	lats	lons
1	China	2020-01-22T14:33:33	-1	35.00000	103.00000
2	China	2020-01-23T16:57:05	-2	35.00000	103.00000
3	China	2020-01-24T09:15:37	-3	35.00000	103.00000
...
160627	Bali	2022-02-03T12:03:05	0	48.85889	2.320041
160628	Bali	2022-02-05T16:17:10	0	48.85889	2.320041
160629	Bali	2022-02-07T20:02:01	-1	48.85889	2.320041



PLOTLY MAP

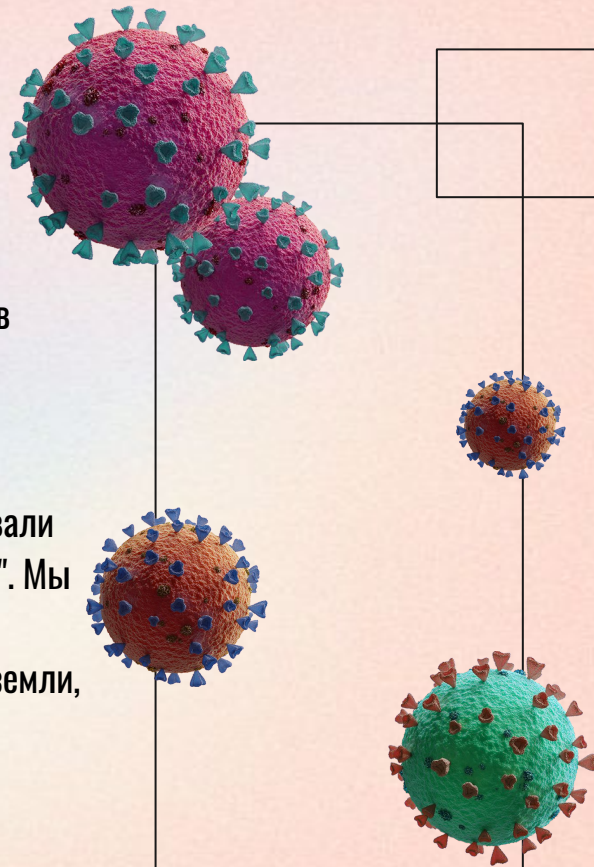


PLOTLY MAP

Функция [Plotly Express](#), которую мы решили использовать — [geoscatteer](#). Она в первую требует в качестве аргументов:

- 1) фреймы анимации, в нашем случае это столбец "дата";
- 2) ширину и долготу — соответствующие столбцы в датафрейме

Отразить динамику мы решили с помощью тепловой шкалы, которой окрашивали точки на карте. Параметру "цвет" мы присвоили значения столбца "Динамика". Мы настроили шкалу так, чтобы она была чувствительна к изменениям, а также настроили некоторые внешние параметры карты (скорость анимации, цвета земли, цвета границ и так далее).



COVID MAP

У карты есть две опции: перемещаться по временной шкале с помощью ползунка и смотреть полную анимацию с помощью кнопки **play**.

Можно также останавливать на любом моменте.

Карта представлена файлом [html](#).

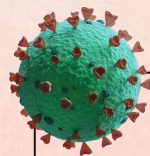
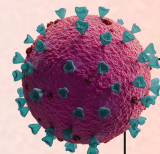
UNITED STATES

Дата=2020-01-22T14:33:33


lats=39.78373


lons=-100.4459


Динамика=0




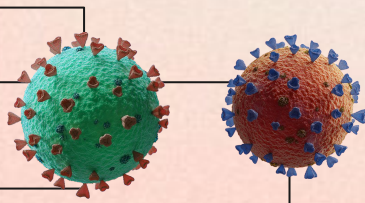
DIFFICULTIES & PERSPECTIVES

 Сложно составить регулярку, которая бы охватила всю необходимую информацию

 Сложно найти способ полностью автоматизировать работу: оставались вещи, которые приходилось форматировать вручную

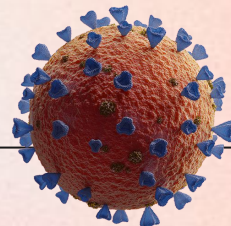
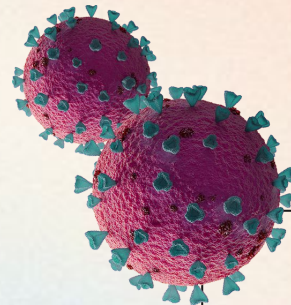
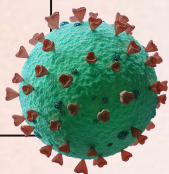
 Создание такой же интерактивной карты на основе одного или нескольких новостных каналов для сравнения подлинности и полноты предоставляемой ими информации

 Реализовать функцию, которая даст возможность посмотреть информацию (количество заболевших/умерших) в конкретной стране при наведении на неё курсора



РАСПРЕДЕЛЕНИЕ ОБЯЗАННОСТЕЙ

- ❑ **Работа с регулярными выражениями** – Чан Тхюи Хуен
- ❑ **Работа с модулем `natasha`** – Харская Стефания
- ❑ **Работа с `DataFrame` и визуализацией карты** – Криволап Мария
- ❑ **Работ с форматированием данных** – Кузьмина Александра
- ❑ **Подбор материала** – вся группа



GITHUB:



map_COVID

СПАСИБО ЗА ВНИМАНИЕ !

