УДК 004.89

АЛГОРИТМЫ DATA-SCIENCE КАК ФУНДАМЕНТАЛЬНАЯ ОСНОВА ВИЗУАЛИЗАЦИИ

Кудаева И. В., студентка группы ИТб-161, IV курс
Пылов П. А., студент группы ИТб-162, IV курс
Акилина М. В., студентка группы ИТб-162, IV курс
Протодьяконов А.В., [1] к.т.н., доцент
Кузбасский государственный технический университет имени Т.Ф. Горбачева
г. Кемерово

Вспышке инфекции коронавируса, начавшейся с обнаружения в первой декаде декабря 2019 года в городе Ухань[2] случаев пневмонии неизвестного происхождения у местного населения, было присвоено название COVID-19 Всемирной Организацией Здравоохранения[3]. Впоследствии вирус распространился по всем другим провинциям Китая и в более двадцати других стран Азии, Европы, Северной Америки и Океании[3]. Распространение вируса от человека к человеку подтверждено в Китае, Германии, Таиланде, Тайване, Японии и Соединенных Штатах[3]. Частный университет Джона Хопкинса запустил интерактивный дашборд[4] визуализации данных о заболевании, однако цель дашборда ограничена картой распространения, в ней отсутствует сортировка по городам и странам.

По состоянию на 29 февраля 2020 года было подтверждено 85409 случаев заражения, из которых 79251 были зарегистрированы в материковом Китае[5]. По состоянию на 1 февраля 2020 года за пределами Китая были люди, которые либо выехали из Ухани, либо были в непосредственном контакте с людьми, которые путешествовали из этого района.

В статье проводится аналитика и визуализация данных, предоставленных Всемирной Организацией Здравоохранения (ВОЗ), о коронавирусе COVID-19, которые были составлены в единый датасет, начиная с момента первых регистрируемых ежедневных отчётов ВОЗ, до момента написания статьи [6]. В представленной работе аналитика и визуализация данных не ограничивается мировой картой распространения вируса, дополнительными

---

[1] Научный руководитель
[2] Центральная часть Китая, Китайская Народная Республика
[3] Официальный сайт Всемирной Организацией Здравоохранения https://www.who.int/
[4] Dashboard of Johns Hopkins University https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
[5] Университет Джона Хопкинса
https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
[6] 01.03.2020

диаграммами представлены графики заболеваний по городам, штатам, провинциям и странам.

Платформой для визуализации данных был выбран язык программирования Python, так как с его помощью очень удобно представлять большое количество данных в виде интерактивных графиков и диаграмм.

Начальный этап виртуализации данных связан с их получением, поэтому загружаем в проект решения датасет исходных данных (метод загрузки представлен на рисунке 1)

```
dataCVirus = pd.read_csv('coronovirus_2019-20_clean_data.csv',
                         parse_dates=['Дата события'])
```

Рисунок 1. Загрузка данных.

Теперь выполним ячейку кода (рисунок 1) и проанализируем в качестве экземпляров первые 15 строк данных (рисунок 2)

	Провинция/ Штат	Страна/ Регион	Широта местонахождения	Долгота местонахождения	Дата события	Подтверждённое заражение	Летальный исход	Выздоровевшие
0	Anhui	Mainland China	31.82570	117.2264	2020-01-22	1	0	0
1	Beijing	Mainland China	40.18240	116.4142	2020-01-22	14	0	0
2	Chongqing	Mainland China	30.05720	107.8740	2020-01-22	6	0	0
3	Fujian	Mainland China	26.07890	117.9874	2020-01-22	1	0	0
4	Gansu	Mainland China	36.06110	103.8343	2020-01-22	0	0	0
5	Guangdong	Mainland China	23.34170	113.4244	2020-01-22	26	0	0
6	Guangxi	Mainland China	23.82980	108.7881	2020-01-22	2	0	0
7	Guizhou	Mainland China	26.81540	106.8748	2020-01-22	1	0	0
8	Hainan	Mainland China	19.19590	109.7453	2020-01-22	4	0	0
9	Hebei	Mainland China	38.04280	114.5149	2020-01-22	1	0	0
10	Heilongjiang	Mainland China	47.86200	127.7615	2020-01-22	0	0	0
11	Henan	Mainland China	33.88202	113.6140	2020-01-22	5	0	0
12	Hubei	Mainland China	30.97560	112.2707	2020-01-22	444	17	28
13	Hunan	Mainland China	27.61040	111.7088	2020-01-22	4	0	0
14	Inner Mongolia	Mainland China	44.09350	113.9448	2020-01-22	0	0	0

Рисунок 2. Датасет сформированных данных из отчётов ВОЗ.

Принимая во внимание факт заражения людей на круизном лайнере «Diamond Princess cruise ship»[7], нужно отделить граждан стран, которые получили заболевание на корабле от самих стран (вирус мог отстутствовать на территории страны, подразумеваем территориальную компоненту, а не гражданство).

[7] Данные на портале https://www.princess.com/news/notices_and_advisories/notices/diamond-princess-update.html

Материковый Китай и Китай следует объединить в одну составляющую, так как разногласия в названиях вносят отрицательную лепту в статистику и превозносят дублирование идентичных строк.

Удобнее изучать данные, когда в них прослеживается четкая иерархия – охарактеризовать заболевания по городам, а потом включить все города в соответствущие им страны. Не лишним будет применить критерии «Зарегистрированное заражение», «Смертельные случаи», «Выздоровевшие» для осознавания ситуации в отдельном изучаемом городе/провинции/штате.

Дополнительным визуальным эффектом будет акцентирование внимание на эпицентре заражения, для этого выделим соответствующую строку красным оттенком цвета.

Результат всех вышеперечисленных тезисов представлен в виде части вывода на рисунке 3.

Страна/Регион	Провинция/Штат	Подтверждённое заражение	Летальный исход	Выздоровевшие
Afghanistan	NA	1	0	0
Algeria	NA	1	0	0
	From Diamond Princess	8	0	0
	New South Wales	4	0	4
Australia	Queensland	5	0	1
	South Australia	2	0	2
	Victoria	4	0	4
Austria	NA	3	0	0
Bahrain	NA	33	0	0
Belgium	NA	1	0	1
Brazil	NA	1	0	0
Cambodia	NA	1	0	1
	British Columbia	7	0	3
Canada	London, ON	1	0	1
	Toronto, ON	5	0	2
	Fujian	296	1	228
	Gansu	91	2	81
	Guangdong	1347	7	890
	Guangxi	252	2	161
	Guizhou	146	2	112
	Hainan	168	5	131
	Hebei	317	6	274
	Heilongjiang	480	13	270
	Henan	1272	20	1068
	Hubei	65596	2641	23383
	Hunan	1017	4	804
	Inner Mongolia	75	0	43
China	Jiangsu	631	0	498
	Jiangxi	934	1	754
	Jilin	93	1	67

Рисунок 3. Вывод переконфигурированных данных.

Изучение всей статистики будет неполным без совокупной строки всех случаев заражения (рисунок 4)

	Дата события	Подтверждённое заражение	Летальный исход	Выздоровевшие
38	2020-02-29 00:00:00	85308	2935	39772

Рисунок 4. Совокупная строка мирового распространения коронавируса.

После этого изучим ситуацию по странам, для этого исключим города из списка иерархии (рисунок 5), представленные на рисунке 3.

	Страна/Регион	Подтверждённое заражение	Летальный исход	Выздоровевшие
0	China	79251	2835	39279
1	South Korea	3150	16	27
2	Italy	1128	29	46
3	Iran	593	43	123
4	Japan	241	5	32
5	Singapore	102	0	72
6	France	100	2	12
7	Hong Kong	95	2	33
8	Germany	79	0	16
9	US	70	1	7
10	Kuwait	45	0	0
11	Spain	45	0	2
12	Thailand	42	0	28
13	Bahrain	41	0	0
14	Taiwan	39	1	9
15	Australia	25	0	11
16	Malaysia	25	0	18
17	UK	23	0	8
18	United Arab Emirates	21	0	5
19	Canada	20	0	6
20	Switzerland	18	0	0
21	Vietnam	16	0	16
22	Norway	15	0	0
23	Iraq	13	0	0

Рисунок 5. Заражение коронавирусам по странам Земли.

Одним из самых востребованных инструментов визуализации является гистограмма распределения заражений по странам и состоянию больных. Визуализация представлена на рисунках 6 – 8

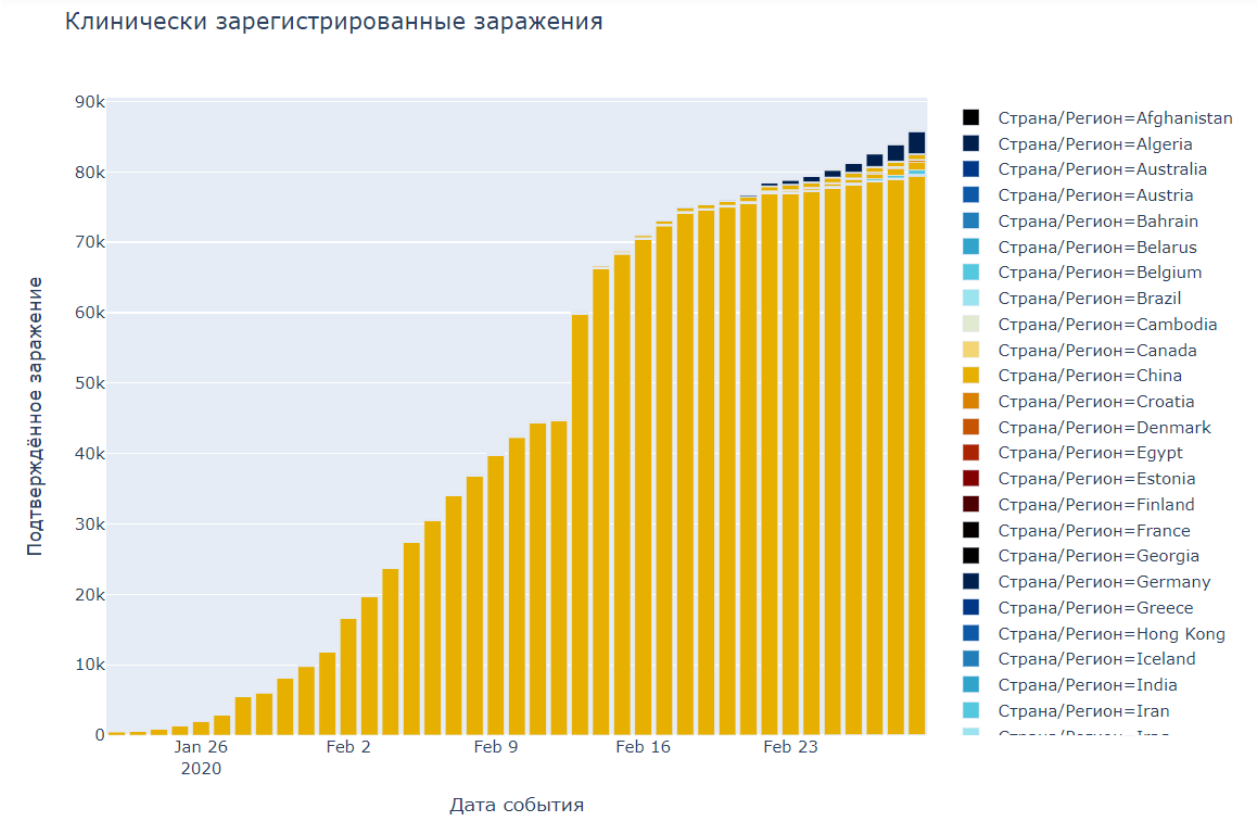

Рисунок 6. Визуализация клинически зарегистрированных случаев заражения.

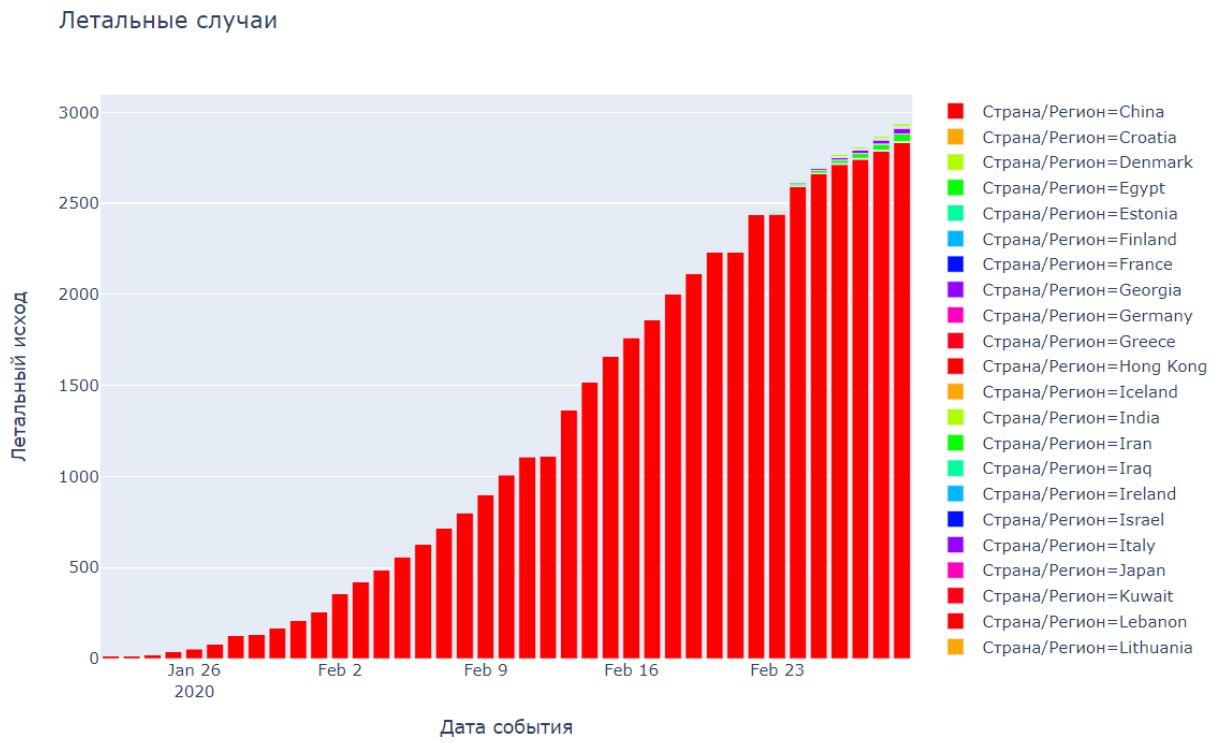

Рисунок 7. Визуализация летальных исходов по странам.

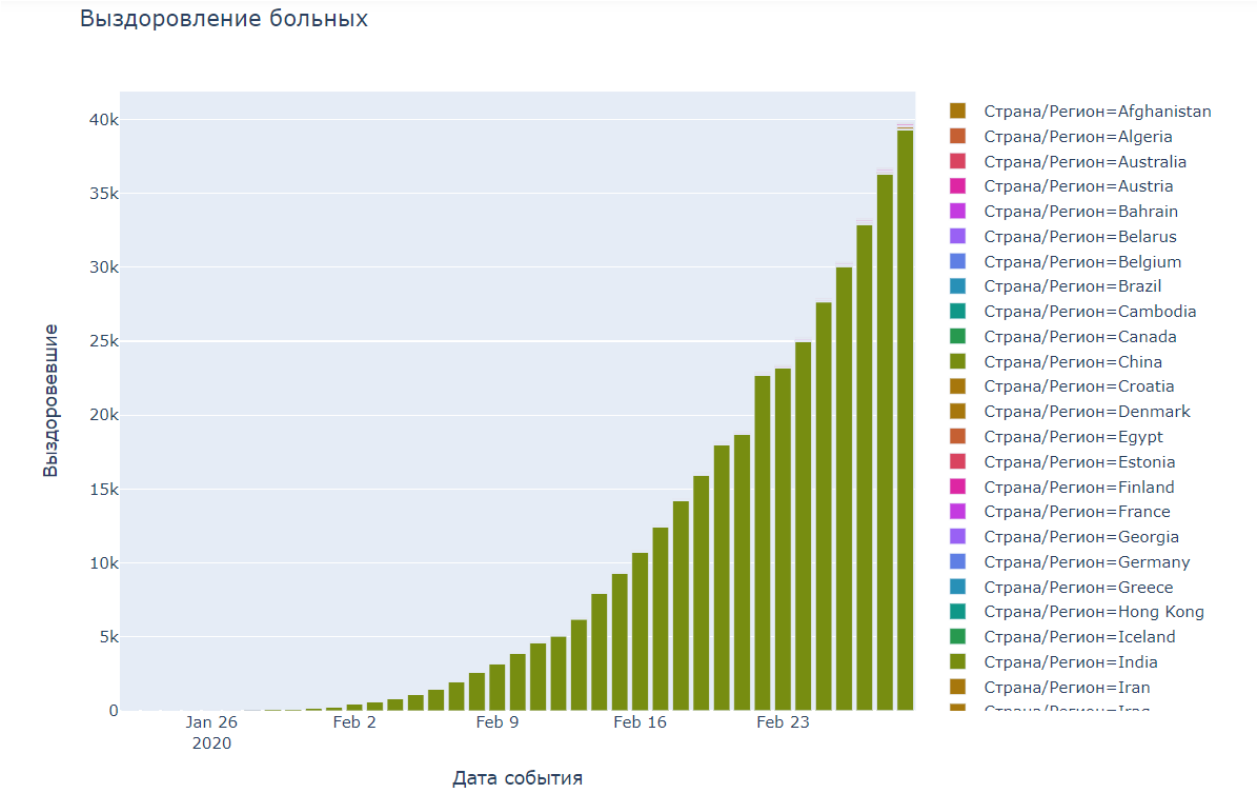

Рисунок 8. Визуализация выздоровевших больных по странам мира.

Наконец, представим генерализацию мировой карты появления корона-вируса по состоянию данных 29 февраля 2020 года по Гавайско-Алеутскому

времени[8], предварительно разделив события на заражения внутри КНР[9] (рисунок 9) и за её пределами, которая представлена на рисунке 10.

Рисунок 9. Заражение коронавирусом внутри Китая.

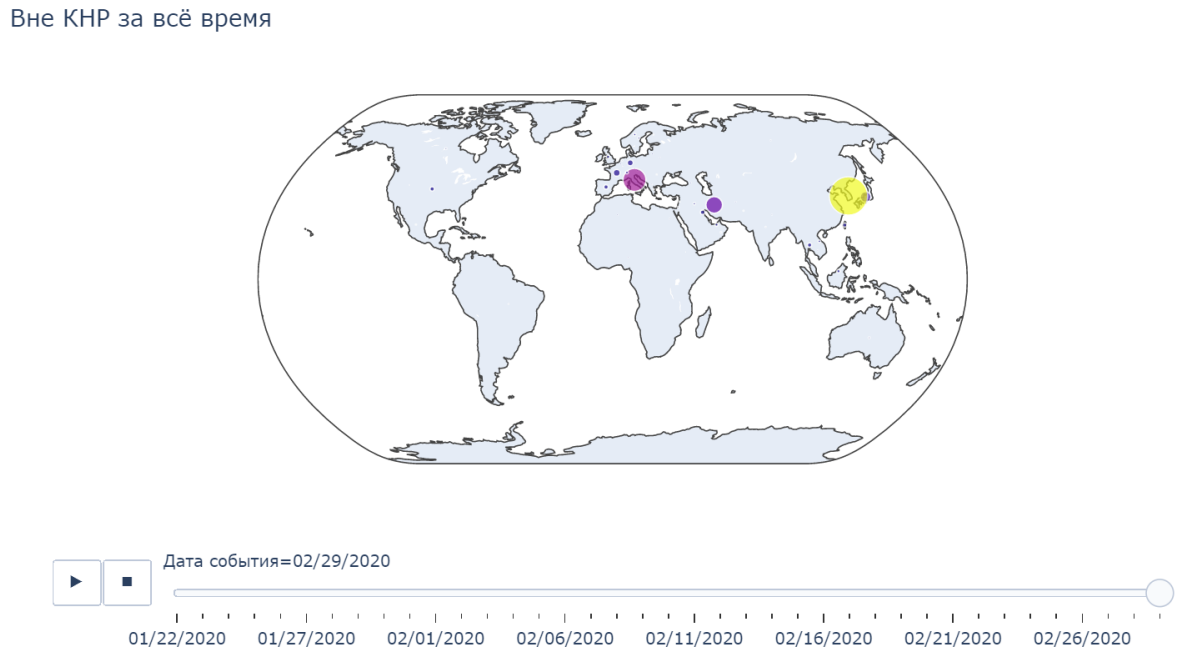

Рисунок 10. Распространение коронавируса за пределами Китая.

Эта работа не противопоставляется дашбордам университета Джона Хопкинса в виде конкурирующего способа визуализации. Изложенный проект представляет диаграммы другого типа, наиболее подходящие исследователям

[8] UTC -10, самое позднее время из всех часовых поясов Земли
[9] КНР – Китайская Народная Республика

в области медицины и специалистам data-science сообщества, поскольку такие графики незаменимы при построение моделей машинного обучения и систем области искусственного интеллекта.

Список литературы:

1. Christopher Bishop. Pattern Recognition and Machine Learning, Plenum press, New York – London, 1971
2. J. Knupp. Writing Idiomatic Python, Atlantic City, 2013
3. В.В. Мещеряков. Моделирование и визуализация случайных данных на языке Python, Интерактивное учебное пособие, Москва, 2015
4. В.В. Мещеряков. Эмпирическое распределение случайных данных. Язык Python, Учебное пособие, Москва, 2016
5. M. Osvaldo. Bayesian Analysis with Python, Packt, 2016