02_AM_Transform_Vaccine_Tweets

July 13, 2021

[1]:
```
import pandas as pd
import numpy as np
```

1 Prepare dataset for hydration

Hydration describes the process of fetching tweet information via the twitter api. For that, we stripped the pre-hydrated dataset of everything so we can get fresh data.

[2]:
```
#raw_df = pd.read_csv("../data/raw/kaggle_dataset.csv")
```

[3]:
```
#raw_df = raw_df["id"].reset_index(drop=True)
```

[4]:
```
#raw_df.to_csv("../data/raw/tweet_ids.csv", index=False)
```

2 Transform Raw Vaccine Tweets

[5]:
```
raw_vaccine_tweets = pd.read_json("../data/raw/vaccine_tweets_hydrated.jsonl",␣
 ↪lines=True, encoding="iso-8859-1")
```

[6]:
```
raw_vaccine_tweets
```

[6]:
```
                       created_at                   id               id_str  \
0       2020-12-13 16:27:13+00:00  1338158543359250400  1338158543359250432
1       2020-12-12 19:22:45+00:00  1337840331522453500  1337840331522453504
2       2020-12-14 18:00:29+00:00  1338544403795882000  1338544403795881984
3       2020-12-12 12:26:34+00:00  1337735595704115200  1337735595704115200
4       2020-12-12 20:04:29+00:00  1337850832256176000  1337850832256176128
…                             …                    …                    …
119794 2021-06-23 13:13:28+00:00  1407688257177936000  1407688257177935872
119795 2021-06-23 13:57:27+00:00  1407699323035558000  1407699323035557888
119796 2021-06-23 13:59:29+00:00  1407699835856330800  1407699835856330752
119797 2021-06-23 12:50:59+00:00  1407682599515000800  1407682599515000832
119798 2021-06-23 13:45:00+00:00  1407696190578249700  1407696190578249728

                                                 full_text  truncated  \
0        While the world has been on the wrong side of …      False
1        @cnnbrk #COVID19 #CovidVaccine #vaccine #Coron…      False
```

```
2       The FDA Authorizes Emergency Use Of The Pfizer…      False
3       The #FDA finally issues #EUA now comes the pro…      False
4       There have not been many bright days in 2020 b…      False
…                                                     …        …
119794  #SputnikV Paid #Hyderabad https://t.co/oklatcuWLh      False
119795  The @WHO said its review of how #Russia produc…      False
119796  #WHO Finds Production Infringements at #Sputni…      False
119797  When was the #SputnikV\n\n1. Exploratory Stage…      False
119798  .@WHO raises concern on cross-contamination, i…      False

       display_text_range                                           entities  \
0                 [0, 275]  {'hashtags': [{'text': 'covid19', 'indices': […
1                 [8, 173]  {'hashtags': [{'text': 'COVID19', 'indices': […
2                 [0, 263]  {'hashtags': [{'text': 'PFE', 'indices': [79, …
3                 [0, 224]  {'hashtags': [{'text': 'FDA', 'indices': [4, 8…
4                 [0, 276]  {'hashtags': [{'text': 'BidenHarris', 'indices…
…                    …                                                  …
119794             [0, 25]  {'hashtags': [{'text': 'SputnikV', 'indices': …
119795            [0, 287]  {'hashtags': [{'text': 'Russia', 'indices': [3…
119796            [0, 133]  {'hashtags': [{'text': 'WHO', 'indices': [0, 4…
119797            [0, 282]  {'hashtags': [{'text': 'SputnikV', 'indices': …
119798            [0, 144]  {'hashtags': [{'text': 'SputnikV', 'indices': …

                                                  source  \
0       <a href="https://mobile.twitter.com" rel="nofo…
1       <a href="https://mobile.twitter.com" rel="nofo…
2       <a href="https://mobile.twitter.com" rel="nofo…
3       <a href="https://mobile.twitter.com" rel="nofo…
4       <a href="http://twitter.com/download/iphone" r…
…                                                     …
119794  <a href="http://twitter.com/download/android" …
119795  <a href="http://twitter.com/download/android" …
119796  <a href="http://twitter.com/download/iphone" r…
119797  <a href="http://twitter.com/download/android" …
119798  <a href="https://about.twitter.com/products/tw…

        in_reply_to_status_id  in_reply_to_status_id_str  …  favorited  \
0                         NaN                        NaN  …      False
1                1.337811e+18               1.337811e+18  …      False
2                         NaN                        NaN  …      False
3                         NaN                        NaN  …      False
4                         NaN                        NaN  …      False
…                       …                         …   …        …
119794                    NaN                        NaN  …      False
119795                    NaN                        NaN  …      False
119796                    NaN                        NaN  …      False
119797                    NaN                        NaN  …      False
```

```
119798                    NaN                        NaN  …      False

        retweeted possibly_sensitive lang  \
0           False                0.0   en
1           False                NaN   en
2           False                0.0   en
3           False                NaN   en
4           False                NaN   en
…           …                …  …
119794      False                0.0   en
119795      False                0.0   en
119796      False                0.0   en
119797      False                NaN   en
119798      False                0.0   en

                                        extended_entities quoted_status_id  \
0                                                     NaN              NaN
1                                                     NaN              NaN
2       {'media': [{'id': 1338544352956719000, 'id_str…             NaN
3                                                     NaN              NaN
4                                                     NaN              NaN
…                                                   …              …
119794  {'media': [{'id': 1407688245484216300, 'id_str…             NaN
119795                                                NaN              NaN
119796                                                NaN              NaN
119797                                                NaN              NaN
119798                                                NaN              NaN

       quoted_status_id_str  quoted_status_permalink  quoted_status  \
0                       NaN                      NaN            NaN
1                       NaN                      NaN            NaN
2                       NaN                      NaN            NaN
3                       NaN                      NaN            NaN
4                       NaN                      NaN            NaN
…                     …                      …            …
119794                  NaN                      NaN            NaN
119795                  NaN                      NaN            NaN
119796                  NaN                      NaN            NaN
119797                  NaN                      NaN            NaN
119798                  NaN                      NaN            NaN

        withheld_in_countries
0                         NaN
1                         NaN
2                         NaN
3                         NaN
4                         NaN
```

```
…                                 …
119794                              NaN
119795                              NaN
119796                              NaN
119797                              NaN
119798                              NaN

[119799 rows x 31 columns]
```

3 Data preparation

- Remove duplicate tweets
  - Drop if retweeted == true
  - Remove duplicate text (or tweet ids)
- Relevant columns = id, created_at, username, full_text, retweet, hashtags

3.1 Remove duplicate tweets

[7]:

```
raw_vaccine_tweets[raw_vaccine_tweets.retweeted == True]
```

[7]:

```
Empty DataFrame
Columns: [created_at, id, id_str, full_text, truncated, display_text_range,
entities, source, in_reply_to_status_id, in_reply_to_status_id_str,
in_reply_to_user_id, in_reply_to_user_id_str, in_reply_to_screen_name, user,
geo, coordinates, place, contributors, is_quote_status, retweet_count,
favorite_count, favorited, retweeted, possibly_sensitive, lang,
extended_entities, quoted_status_id, quoted_status_id_str,
quoted_status_permalink, quoted_status, withheld_in_countries]
Index: []

[0 rows x 31 columns]
```

[8]:

```
raw_vaccine_tweets.id.count()
```

[8]:

```
119799
```

[9]:

```
raw_vaccine_tweets = raw_vaccine_tweets.drop_duplicates(subset=["full_text"],␣
 ↪keep='first').reset_index(drop=True)
```

[10]:

```
raw_vaccine_tweets.id.count()
```

[10]:

```
118272
```

3.2 Select relevant columns

[11]:

```
raw_vaccine_tweets.columns
```

```
[11]: Index(['created_at', 'id', 'id_str', 'full_text', 'truncated',
              'display_text_range', 'entities', 'source', 'in_reply_to_status_id',
              'in_reply_to_status_id_str', 'in_reply_to_user_id',
              'in_reply_to_user_id_str', 'in_reply_to_screen_name', 'user', 'geo',
              'coordinates', 'place', 'contributors', 'is_quote_status',
              'retweet_count', 'favorite_count', 'favorited', 'retweeted',
              'possibly_sensitive', 'lang', 'extended_entities', 'quoted_status_id',
              'quoted_status_id_str', 'quoted_status_permalink', 'quoted_status',
              'withheld_in_countries'],
             dtype='object')
```

```
[12]: raw_vaccine_tweets =␣
       ↪raw_vaccine_tweets[["id_str","created_at","user","geo","full_text",␣
       ↪"entities"]]
```

3.3 Extracting user_ids

```
[13]: raw_vaccine_tweets["user_id"] = int

      for i in range(len(raw_vaccine_tweets)):
          raw_vaccine_tweets.user_id[i] = raw_vaccine_tweets.user[i]["id"]
```

```
<ipython-input-13-589c50f79ebf>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets.user_id[i] = raw_vaccine_tweets.user[i]["id"]
```

```
[14]: raw_vaccine_tweets.user_id
```

```
[14]: 0                     76052772
      1          1300382181605494800
      2          1164717209253552000
      3          1316036067754205200
      4          1110032180237852700
                         …
      118267     1263779139397382100
      118268                40623001
      118269                61611674
      118270               126591034
      118271               231692806
      Name: user_id, Length: 118272, dtype: object
```

3.4 Hashtags

renaming entities column to hashtags:

[15]:

```
raw_vaccine_tweets = raw_vaccine_tweets.rename(columns={'entities': 'hashtags',␣
 ↪'id_str':'id'})
```

Extracting hashtags which are stored in entities>hashtags>text:

[16]:

```
for i in range(len(raw_vaccine_tweets)):
    try:
        raw_vaccine_tweets.hashtags[i] = [value["text"] for value in␣
 ↪raw_vaccine_tweets.iloc[i]["hashtags"]["hashtags"]]
    except:
        print("failed: ",i)

```

```
<ipython-input-16-f542ec085c11>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets.hashtags[i] = [value["text"] for value in
raw_vaccine_tweets.iloc[i]["hashtags"]["hashtags"]]
```

Storing hashtags:count as key:value in a dict:

[17]:

```
hashtag_dict = {}
for i in range(len(raw_vaccine_tweets)):
    for hashtag in raw_vaccine_tweets["hashtags"][i]:
        if hashtag not in hashtag_dict:
            hashtag_dict[hashtag] = 1
        else:
            hashtag_dict[hashtag] += 1
```

Identifying relevant hashtags for each vaccine manufacturer:

[18]:

```
sorted(hashtag_dict.items(), key=lambda x: x[1], reverse=True)
```

[18]:

```
[('Moderna', 28868),
 ('Covaxin', 24394),
 ('COVID19', 19027),
 ('SputnikV', 17623),
 ('COVAXIN', 13822),
 ('vaccine', 12270),
 ('Pfizer', 9767),
 ('moderna', 7338),
 ('Sinovac', 7268),
 ('CovidVaccine', 6876),
 ('BBMP', 6821),
 ('PfizerBioNTech', 6569),
 ('Sinopharm', 6177),
 ('Covishield', 5386),
 ('AstraZeneca', 5106),
 ('coronavirus', 3869),
 ('vaccinated', 3645),
 ('COVID19Vaccine', 3558),
 ('covaxin', 3360),
 ('vaccination', 3056),
 ('Vaccine', 2769),
 ('vaccines', 2762),
 ('OxfordAstraZeneca', 2498),
 ('BharatBiotech', 2370),
 ('India', 2333),
 ('lockdown', 2316),
 ('Covid19', 2279),
 ('China', 2279),
 ('COVID', 2128),
 ('Russia', 1990),
 ('covid19', 1798),
 ('GetVaccinated', 1790),
 ('PfizerVaccine', 1772),
 ('COVIDVaccination', 1668),
 ('pfizer', 1585),
 ('oxfordastrazeneca', 1549),
 ('PfizerBiontech', 1525),
 ('covid', 1474),
 ('CoronaVaccine', 1444),
 ('COVID19Vaccination', 1404),
 ('COVISHIELD', 1213),
 ('Covid', 1207),
 ('Coronavirus', 1156),
 ('Covid_19', 1144),
 ('CovishieldVaccine', 1059),
 ('WHO', 1039),
 ('COVID19India', 984),
 ('mRNA', 947),
 ('Vaccination', 934),
 ('IndiaFightsCorona', 931),
 ('sinovac', 897),
 ('EU', 867),
 ('COVIDVaccine', 797),
 ('johnsonandjohnson', 795),
 ('MentalHealth', 766),
 ('modernavaccine', 752),
 ('VaccineForAll', 749),
 ('VaccinationDrive', 728),
 ('GurgaonCOVAXIN', 723),
 ('Sputnik', 712),
 ('VaccinesWork', 690),
 ('pandemic', 687),
 ('JohnsonandJohnson', 684),
 ('sputnikv', 676),
 ('MumbaiCOVAXIN', 664),
 ('BioNTech', 663),
 ('covidvaccine', 647),
 ('Corona', 629),
 ('covidvacccine', 629),
 ('Pakistan', 614),
 ('covishield', 614),
 ('LargestVaccineDrive', 612),
 ('VaccinesSaveLives', 576),
 ('SriLanka', 565),
 ('covid_19', 531),
 ('Vaccines', 523),
 ('Jesus', 518),
 ('DeltaVariant', 514),
 ('COVIDã\x83¼19', 510),
 ('MentalHealthMatters', 510),
 ('health', 507),
 ('FullyVaccinated', 504),
 ('astrazeneca', 498),
 ('FDA', 491),
 ('CoronavirusVaccine', 491),
 ('healthcare', 488),
 ('COVIDSecondWave', 483),
 ('UK', 482),
 ('Vaccinated', 468),
 ('NarendraModi', 463),
 ('AstraZenaca', 453),
 ('COVAX', 447),
 ('BREAKING', 440),
 ('Ocugen', 422),
 ('OCGN', 416),
 ('US', 411),
 ('Brazil', 410),
 ('news', 409),
 ('Delhi', 409),
 ('Canada', 389),
 ('PMModi', 376),
 ('sinopharm', 373),
 ('lka', 371),
 ('ocgn', 371),
 ('NHS', 362),
 ('USA', 361),
 ('CoWIN', 359),
 ('JohnsonAndJohnson', 357),
 ('UAE', 353),
 ('CoronavirusPandemic', 352),
 ('Chinese', 350),
 ('CDC', 347),
 ('CoronavirusIndia', 337),
 ('CovidVaccines', 335),
 ('SARSCoV2', 335),
 ('india', 324),
 ('ocugen', 324),
 ('mentalhealth', 320),
 ('cdnpoli', 319),
 ('Bengaluru', 316),
 ('pfizerbiontech', 314),
 ('AIIMS', 310),
 ('PfizerCovidVaccine', 307),
 ('COVIDEmergency', 303),
 ('PuneCOVAXIN', 301),
 ('Hyderabad', 300),
 ('Russian', 299),
 ('Pfizervaccine', 295),
 ('seconddose', 289),
 ('corona', 288),
 ('COVID19vaccine', 283),
 ('CoronaVac', 281),
 ('Novavax', 281),
 ('spring', 281),
 ('Health', 277),
 ('Putin', 274),
 ('WearAMask', 273),
 ('CovidIndia', 273),
 ('Germany', 268),
 ('investing', 268),
 ('Janssen', 261),
 ('IndiaFightsCOVID19', 260),
 ('BreakingNews', 259),
 ('VaccineShortage', 259),
 ('science', 254),
 ('Philippines', 250),
 ('FauciOuchie', 249),
 ('Thailand', 248),
 ('JohnsonAndJohnsonVaccine', 244),
 ('Covid19Vaccine', 242),
 ('CoronaVirusUpdates', 242),
 ('COVID19Vaccines', 240),
 ('Johnson', 239),
 ('Modi', 238),
 ('HongKong', 236),
 ('PfizerBioNtech', 235),
 ('COVIDvaccine', 235),
 ('vaccinationdone', 233),
 ('StaySafe', 225),
 ('vaccinate', 221),
 ('Mumbai', 220),
 ('MODERNA', 219),
 ('Europe', 218),
 ('Israel', 218),
 ('astrazenecavaccine', 218),
 ('ThaneCOVAXIN', 217),
 ('SputnikUpdates', 213),
 ('JandJ', 212),
 ('Biden', 210),
 ('RDIF', 209),
 ('DCGI', 207),
 ('Iran', 206),
 ('StocksToWatch', 203),
 ('CoronaSecondWave', 203),
 ('Argentina', 202),
 ('Remdesivir', 201),
 ('SinoVac', 200),
 ('CowinApp', 196),
 ('COVIDSHIELD', 195),
 ('EMA', 194),
 ('ICMR', 193),
 ('Cipla', 193),
 ('Fauci', 188),
 ('Slovakia', 187),
 ('Syria', 187),
 ('COVID19SL', 186),
 ('Unite2FightCorona', 186),
 ('BangaloreUrban', 186),
 ('Pandemic', 185),
 ('VACCINE', 184),
 ('VaccineMaitri', 184),
 ('covid19vaccine', 182),
 ('COVIDIOTS', 181),
 ('OCUGEN', 178),
 ('Science', 177),
 ('COVID19LK', 177),
 ('SouthAfrica', 176),
 ('Hungary', 175),
 ('BigPharma', 174),
 ('EuropeanUnion', 173),
 ('Italy', 173),
 ('Pune', 173),
 ('vaccinations', 171),
 ('Delta', 171),
 ('Oxford', 170),
 ('2nddose', 170),
 ('Bangladesh', 168),
 ('CoronaVaccination', 167),
 ('EUA', 166),
 ('Maharashtra', 164),
 ('LargestVaccinationDrive', 163),
 ('grateful', 162),
 ('Zimbabwe', 162),
 ('Biontech', 161),
 ('MaskUp', 161),
 ('markets', 161),
 ('Covidvax', 161),
 ('oxygen', 161),
 ('oxfordvaccine', 160),
 ('StockMarket', 160),
 ('Indonesia', 159),
 ('ontariolockdown', 158),
 ('china', 158),
 ('COVIDVaccines', 156),
 ('VaccineRegistration', 156),
 ('MadeInIndia', 155),
 ('Sputnikvaccine', 152),
 ('News', 151),
 ('MondayMotivation', 151),
 ('á´\xa0á´\x80á´\x84á´\x84ÉªÉ´á´\x87ssá´\x80á´\xa0á´\x87Ê\x9fÉªá´\xa0á´\x87s',
  151),
 ('COVIDEmergency2021', 150),
 ('VaccineFor18Plus', 150),
 ('CCP', 149),
 ('vaccineSideEffects', 149),
 ('Indian', 148),
 ('SINOVAC', 148),
 ('CoronaVirus', 147),
 ('bharatBiotech', 145),
 ('mumbai', 145),
 ('COVID19Vic', 144),
 ('Egypt', 143),
 ('ThisIsOurShot', 142),
 ('Turkey', 140),
 ('COVIDSecondWaveInIndia', 140),
 ('bharatbiotech', 139),
 ('Singapore', 138),
 ('sputnikV', 137),
 ('takeUsBackToChina', 137),
 ('COVID19vaccines', 136),
 ('ChineseVirus', 136),
 ('AatmanirbharBharat', 136),
 ('coronavirusvaccine', 134),
 ('AtmaNirbharBharat', 133),
 ('BJP', 133),
 ('mondaythoughts', 132),
 ('lockdown2021', 132),
 ('Astrazeneca', 131),
 ('Covid19UK', 130),
 ('Dubai', 130),
 ('getvaccinated', 129),
 ('covaxine', 128),
 ('VaccinePassports', 128),
 ('StayHome', 127),
 ('Mexico', 127),
 ('sideeffects', 126),
 ('medical', 126),
 ('Covid19vaccine', 125),
 ('MakeInIndia', 124),
 ('Sanofi', 122),
 ('variants', 122),
 ('Cambodia', 122),
 ('TamilNadu', 122),
 ('Comirnaty', 121),
 ('firstdose', 120),
 ('France', 120),
 ('Lockdown', 120),
 ('VaccinateIndia', 119),
 ('Congress', 118),
 ('Covax', 118),
 ('CanSino', 117),
 ('OxygenCylinders', 117),
 ('ModernaVaccine', 116),
 ('SerumInstituteofIndia', 116),
 ('Ukraine', 116),
 ('NovelCoronavirus', 116),
 ('AZ', 115),
 ('Serbia', 115),
 ('Breaking', 114),
 ('COVID19Ontario', 114),
 ('SecondDose', 114),
 ('stocks', 113),
 ('COVID19ON', 113),
 ('Africa', 112),
 ('unite2fightcorona', 112),
 ('sputnik', 111),
 ('AI', 110),
 ('trading', 110),
 ('StocksToBuy', 110),
 ('bloodclots', 110),
 ('largestvaccinedrive', 110),
 ('COVID19AB', 109),
 ('sensex', 108),
 ('Malaysia', 107),
 ('PFIZER', 107),
 ('healthforall', 107),
 ('Nepal', 107),
 ('nifty', 107),
 ('children', 107),
 ('COVID_19', 106),
 ('phizer', 106),
 ('Ontario', 105),
 ('BillGates', 105),
 ('virus', 105),
 ('VaccinationCovid', 105),
 ('VaccineDiplomacy', 105),
 ('StocksToTrade', 105),
 ('Australia', 104),
 ('onpoli', 104),
 ('Telangana', 104),
 ('Covidvaccine', 103),
 ('stock', 103),
 ('gold', 103),
 ('Toronto', 102),
 ('nagpur', 102),
 ('California', 101),
 ('mask', 101),
 ('tuesdaymotivations', 101),
 ('Punjab', 101),
 ('Assam', 101),
 ('OxfordVaccine', 100),
 ('SPUTNIKV', 100),
 ('silver', 99),
 ('joebiden', 99),
 ('bitcoin', 98),
 ('DollyParton', 98),
 ('SinoPharm', 98),
 ('Trump', 97),
 ('NSTnation', 97),
 ('Bahrain', 96),
 ('auspol', 96),
 ('Chennai', 96),
 ('JoeBiden', 95),
 ('covid19vacccine', 95),
 ('maskup', 95),
 ('staysafe', 94),
 ('jab', 94),
 ('sundayvibes', 94),
 ('Regeneron', 94),
 ('EUL', 94),
 ('PfizerGang', 94),
 ('nhs', 93),
 ('SaudiArabia', 93),
 ('Moscow', 93),
 ('stockmarket', 93),
 ('cowinregistration', 93),
 ('russia', 92),
 ('B1617', 92),
 ('MedTwitter', 91),
 ('PublicHealth', 91),
 ('Cuba', 91),
 ('CoronaPandemic', 91),
 ('ArvindKejriwal', 91),
 ('Chile', 90),
 ('VAERS', 90),
 ('biotech', 89),
 ('coronavaccine', 89),
 ('myocarditis', 89),
 ('medicine', 88),
 ('Covisheild', 88),
 ('jesus', 88),
 ('VaccinePassport', 87),
 ('money', 87),
 ('wednesdaythought', 87),
 ('Gujarat', 87),
 ('XiJinping', 87),
 ('ImranKhan', 87),
 ('BBMPCOVAXIN', 86),
 ('travel', 85),
 ('who', 85),
 ('wearamask', 85),
 ('Karnataka', 85),
 ('2ndDose', 84),
 ('NCOC', 84),
 ('GetVaccinatedASAP', 84),
 ('mrna', 83),
 ('cowin', 83),
 ('modernagang', 83),
 ('publichealth', 82),
 ('hope', 81),
 ('premiabiotech', 81),
 ('UnitedStates', 80),
 ('America', 79),
 ('CovidVaccination', 79),
 ('research', 79),
 ('OxfordAstrazeneca', 79),
 ('vaccineshortage', 79),
 ('DrReddy', 79),
 ('Srilanka', 78),
 ('Astrazenaca', 78),
 ('novavax', 77),
 ('COVISHEILD', 77),
 ('today', 76),
 ('covidindia', 76),
 ('NIH', 75),
 ('IndiaPostUSA', 75),
 ('modi', 75),
 ('help', 75),
 ('pharma', 74),
 ('HerdImmunity', 74),
 ('clinicaltrials', 74),
 ('vaccini', 74),
 ('emergency', 74),
 ('Cowin', 74),
 ('Tech', 74),
 ('DRreddys', 74),
 ('Covid19IndiaHelp', 74),
 ('OxygenConcentrator', 74),
 ('Trending', 73),
 ('Vaccin', 73),
 ('LKA', 73),
 ('Trudeau', 73),
 ('FreeTibet', 73),
 ('DrReddys', 73),
 ('UttarPradesh', 73),
 ('Covidshield', 72),
 ('fullyvaccinated', 72),
 ('Lebanon', 71),
 ('cdc', 71),
 ('vaccinerollout', 71),
 ('Norway', 71),
 ('Austria', 71),
 ('StayHomeStaySafe', 71),
 ('ThirdWave', 71),
 ('Healthcare', 70),
 ('B117', 70),
 ('tuesdayvibe', 70),
 ('PMOIndia', 70),
 ('JohnsonJohnson', 70),
 ('Biotech', 70),
 ('oxygenPlants', 70),
 ('DeltaPlusVariant', 70),
 ('igottheshot', 69),
 ('BoycottSinovac', 69),
 ('BoycottSinopharm', 69),
 ('FAIL', 69),
 ('VaccinationForAll', 69),
 ('URBAN', 69),
 ('variant', 68),
 ('covidshield', 68),
 ('HoldChinaAccountable', 68),
 ('farmlaws', 68),
 ('powell', 68),
 ('uk', 67),
 ('England', 67),
 ('bcpoli', 67),
 ('Biology', 67),
 ('Coimbatore', 67),
 ('Bangalore', 67),
 ('JNJ', 66),
 ('daytrading', 66),
 ('Covid19India', 65),
 ('World', 65),
 ('SerumInstitute', 65),
 ('UN', 65),
 ('Taiwan', 65),
 ('SputnikLight', 65),
 ('AtmanirbharBharat', 65),
 ('NHSCovidVaccine', 65),
 ('FirstDose', 64),
 ('Bitcoin', 64),
 ('biontech', 64),
 ('Nifty', 64),
 ('J', 64),
 ('CoWin', 64),
 ('RahulGandhi', 64),
 ('delhi', 64),
 ('Appointments', 64),
 ('TikaUtsav', 64),
 ('EURO2020', 64),
 ('canada', 63),
 ('disease', 63),
 ('Japan', 63),
 ('HIV', 63),
 ('vaccineforall', 63),
 ('AstraZeneka', 63),
 ('Vietnam', 63),
 ('fullyvacinnated', 63),
 ('Easter', 63),
 ('SocialDistancing', 62),
 ('economy', 62),
 ('remdesivir', 62),
 ('Preclinical', 62),
 ('NYC', 61),
 ('Odisha', 61),
 ('Pharma', 61),
 ('Islamabad', 61),
 ('Seychelles', 61),
 ('GST', 61),
 ('CowinPortal', 61),
 ('Coronavac', 60),
 ('premiaholdings', 60),
 ('We4Vaccine', 60),
 ('DrugDiscovery', 60),
 ('CCPVirus', 59),
 ('Merkel', 59),
 ('AnthonyFauci', 59),
 ('Grateful', 59),
 ('janssen', 59),
 ('SputnikVaccinated', 59),
 ('vaxxed', 58),
 ('Kerala', 58),
 ('CoVaxin', 58),
 ('Ahmedabad', 58),
 ('VaccinateEveryIndian', 58),
 ('oxygencylinder', 58),
 ('UPDATE', 57),
 ('Indians', 57),
 ('lockdowns', 57),
 ('Navalny', 57),
 ('finance', 57),
 ('love', 57),
 ('yields', 57),
 ('BreakTheChain', 57),
 ('DrugDesign', 57),
 ('DNA', 56),
 ('efficacy', 56),
 ('dollyparton', 56),
 ('Duterte', 56),
 ('Chemistry', 56),
 ('Ramadan2021', 56),
 ('StockMarketNews', 56),
 ('VirtualEvent', 56),
 ('urgent', 56),
 ('PfizerCOVIDvaccine', 55),
 ('eu', 55),
 ('Vaccinate', 55),
 ('Vaccinatie', 55),
 ('Gaza', 55),
 ('COVID19Aus', 55),
 ('CureVac', 55),
 ('Bharat', 55),
 ('Venezuela', 55),
 ('stockmarketnews', 55),
 ('HBDMKStalin', 55),
 ('VIRTUALPDDP', 55),
 ('IndianVariant', 55),
 ('IndiaCovidCrisis', 55),
 ('pfizervaccine', 54),
 ('God', 54),
 ('malaysia', 54),
 ('RESBAKUNA', 54),
 ('SideEffects', 53),
 ('johnson', 53),
 ('VaccineStrategy', 53),
 ('Bhubaneswar', 53),
 ('premiameds', 53),
 ('DRREDDY', 53),
 ('icu', 53),
 ('ThisIsOurShotCA', 53),
 ('business', 52),
 ('StopTheSpread', 52),
 ('Ireland', 52),
 ('CoviShield', 52),
 ('BorisJohnson', 52),
 ('CoronaVaccineUpdates', 52),
 ('DrHarshVardhan', 52),
 ('SastaBhiKargarBhi', 52),
 ('pmmodi', 52),
 ('maskupindia', 52),
 ('Kenya', 52),
 ('bed', 52),
 ('vaccinebot', 52),
 ('CoronaUpdate', 52),
 ('Covifor', 52),
 ('FreeVaccineForAll', 52),
 ('RURAL', 52),
 ('OperationWarpSpeed', 51),
 ('FMTNews', 51),
 ('immunity', 51),
 ('DrFauci', 51),
 ('TakeYourShot', 51),
 ('CoronaPatients', 51),
 ('antibodies', 51),
 ('fauciouchie', 51),
 ('firstdosedone', 51),
 ('fullyvaxxed', 51),
 ('COWIN', 51),
 ('MHRA', 50),
 ('Wuhan', 50),
 ('GSK', 50),
 ('Phizer', 50),
 ('Algeria', 50),
 ('herdimmunity', 50),
 ('CoronaVaccineNews', 50),
 ('ChinaLiedPeopleDied', 50),
 ('banknifty', 50),
 ('SupremeCourt', 50),
 ('nonprofit', 50),
 ('FightAgainstCOVID19', 50),
 ('helpnagar', 50),
 ('icubed', 50),
 ('Britain', 49),
 ('doses', 49),
 ('Qatar', 49),
 ('5G', 49),
 ('ASTRAZENECA', 49),
 ('Peru', 49),
 ('Morocco', 49),
 ('technicalanalysis', 49),
 ('Beijing', 49),
 ('Pzifer', 49),
 ('ventilator', 49),
 ('usa', 48),
 ('covax', 48),
 ('Brexit', 48),
 ('NewsAlert', 48),
 ('politics', 48),
 ('BioNTechpfizer', 48),
 ('Lancet', 48),
 ('IBM', 48),
 ('Stimuluschecks', 48),
 ('nifty50', 48),
 ('SputnikVaccineInKenya', 48),
 ('CoronaCurfew', 48),
 ('YellowCard', 48),
 ('USFDA', 48),
 ('Alpha', 48),
 ('CovidVaccinesideeffects', 47),
 ('AbuDhabi', 47),
 ('thankyouscience', 47),
 ('Wales', 47),
 ('Vaccinations', 47),
 ('budget', 47),
 ('Chhattisgarh', 47),
 ('resbakuna', 47),
 ('technology', 46),
 ('InformedConsent', 46),
 ('thankful', 46),
 ('Denmark', 46),
 ('secondshot', 46),
 ('shot', 46),
 ('CovaxiUpdates', 46),
 ('Children', 46),
 ('SII', 46),
 ('ChineseVaccine', 46),
 ('airtel', 46),
 ('SputnikVaccine', 46),
 ('covid19india', 46),
 ('GenXZeneca', 46),
 ('CovidHelp', 46),
 ('VaccinationUpdate', 46),
 ('Tunisia', 45),
 ('Video', 45),
 ('Tesla', 45),
 ('vax', 45),
 ('NATO', 45),
 ('Bolivia', 45),
 ('MRNA', 45),
 ('Belarus', 45),
 ('getvaccienated', 45),
 ('congressmuktbharat', 45),
 ('DailyVoice', 45),
 ('scientists', 44),
 ('CNN', 44),
 ('frontlineworkers', 44),
 ('SouthKorea', 44),
 ('Palestine', 44),
 ('AMC', 44),
 ('PMCaresFund', 44),
 ('Karachi', 44),
 ('ThailandNews', 44),
 ('koolex', 44),
 ('PFIZERBIONTECH', 43),
 ('vaccin', 43),
 ('ChinaVirus', 43),
 ('CovidVaccineIndia', 43),
 ('FreeHK', 43),
 ('Coronil', 43),
 ('DelhiHighCourt', 43),
 ('Guwahati', 43),
 ('NSE', 43),
 ('vaccinationfor18plus', 43),
 ('safety', 42),
 ('nurses', 42),
 ('Virus', 42),
 ('Spain', 42),
 ('COVIDvaccines', 42),
 ('cryptocurrency', 42),
 ('Facebook', 42),
 ('chinesevaccine', 42),
 ('Iraq', 42),
 ('StocksInFocus', 42),
 ('wipro', 42),
 ('summer', 42),
 ('É¢á´\x87á´\x9bá´\xa0á´\x80á´\x84á´\x84ÉªÉ´á´\x80á´\x9bá´\x87á´\x85', 42),
 ('MaharashtraNeedsVaccine', 42),
 ('DelhiFightsCorona', 42),
 ('covaxinated', 42),
 ('ICYMI', 41),
 ('AngelaMerkel', 41),
 ('Immunity', 41),
 ('IndiaNarrative', 41),
 ('Impfung', 41),
 ('covidvaccines', 41),
 ('doctors', 41),
 ('covid19vaccines', 41),
 ('coronavirusindia', 41),
 ('Gurgaon', 41),
 ('MajorTeaser', 41),
 ('OxygenShortage', 41),
 ('Vivek', 41),
 ('WuhanVirus', 40),
 ('pfizervacine', 40),
 ('Texas', 40),
 ('Gamaleya', 40),
 ('icmr', 40),
 ('EpiVacCorona', 40),
 ('Pakistani', 40),
 ('StocksInNews', 40),
 ('slovakia', 40),
 ('PMNarendraModi', 40),
 ('Rajasthan', 40),
 ('Kolkata', 40),
 ('JaiHind', 40),
 ('Crypto', 40),
 ('ISIS', 40),
 ('BSE', 40),
 ('PanaceaBiotec', 40),
 ('stayhome', 39),
 ('coronavirusuk', 39),
 ('Government', 39),
 ('development', 39),
 ('worldnews', 39),
 ('socialdistancing', 39),
 ('comirnaty', 39),
 ('thursdayvibes', 39),
 ('CoronaCasesIndia', 39),
 ('family', 39),
 ('fridaymorning', 39),
 ('thursdaymorning', 39),
 ('covid19malaysia', 39),
 ('Bihar', 39),
 ('WestBengal', 39),
 ('Armenia', 39),
 ('ModiHaiTohMumkinHai', 39),
 ('Congo', 39),
 ('thesundaily', 39),
 ('n95', 39),
 ('art', 38),
 ('massacre', 38),
 ('B1351', 38),
 ('à¹\x82à¸\x84à¸§à¸´à¸\x9419', 38),
 ('MEDIA', 38),
 ('Book', 38),
 ('HealthMinistry', 38),
 ('FastForNation', 38),
 ('MKStalin', 38),
 ('firstshot', 38),
 ('ymedia', 38),
 ('Mangalore', 38),
 ('JammuAndKashmir', 38),
 ('NSTworld', 37),
 ('NurembergCode', 37),
 ('thankyou', 37),
 ('Florida', 37),
 ('thankyouNHS', 37),
 ('blessed', 37),
 ('WorldHealthOrganization', 37),
 ('dose', 37),
 ('Covid19Vaccines', 37),
 ('life', 37),
 ('SmartNews', 37),
 ('teammoderna', 37),
 ('Lahore', 37),
 ('Clairvoyant', 37),
 ('PSYCHIC', 37),
 ('Haryana', 37),
 ('TV9News', 37),
 ('JnJ', 37),
 ('Vaxzevria', 37),
 ('cipla', 37),
 ('ApolloHospital', 37),
 ('PfizerProud', 36),
 ('CoronaVirusUpdate', 36),
 ('healthcareworkers', 36),
 ('Thankful', 36),
 ('vaccinessavelives', 36),
 ('SARS_CoV_2', 36),
 ('London', 36),
 ('podcast', 36),
 ('putin', 36),
 ('VladimirPutin', 36),
 ('globalhealth', 36),
 ('media', 36),
 ('Senegal', 36),
 ('AmitShah', 36),
 ('GetTheShot', 36),
 ('jio', 36),
 ('UnlockOurCountry', 36),
 ('Protein', 36),
 ('Covidupdate', 36),
 ('COVIDEmergencyIndia', 36),
 ('MSNBC', 35),
 ('vaccineswork', 35),
 ('Business', 35),
 ('HealthCanada', 35),
 ('trustscience', 35),
 ('quarantine', 35),
 ('effective', 35),
 ('coronaviruspandemic', 35),
 ('1stDose', 35),
 ('Azerbaijan', 35),
 ('FarmersProstests', 35),
 ('premiamedical', 35),
 ('getyourshot', 35),
 ('VaccineEquity', 35),
 ('CHINA', 35),
 ('Kashmir', 35),
 ('telecom', 35),
 ('srilanka', 35),
 ('LockDown', 35),
 ('Mauritius', 35),
 ('banking', 35),
 ('DoctorsDay', 35),
 ('Nashik', 35),
 ('AssamCovidUpdate', 35),
 ('ontariovaccine', 35),
 ('covidhelpline', 35),
 ('fda', 34),
 ('IGotTheShot', 34),
 ('government', 34),
 ('education', 34),
 ('world', 34),
 ('Twitter', 34),
 ('COVID19PH', 34),
 ('TeamVaccine', 34),
 ('Covid19Lockdown', 34),
 ('hospitals', 34),
 ('chinavirus', 34),
 ('CVS', 34),
 ('1stdose', 34),
 ('Everlane', 34),
 ('CGBudget2021', 34),
 ('Resbakuna', 34),
 ('bseindia', 34),
 ('Apollo', 34),
 ('Vaccinateindia', 34),
 ('ApolloHospitals', 34),
 ('Amazon', 33),
 ('GreatReset', 33),
 ('StaySafeStayHealthy', 33),
 ('TikTok', 33),
 ('NewsUpdate', 33),
 ('SINOPHARM', 33),
 ('design', 33),
 ('IndianArmy', 33),
 ('mco2021', 33),
 ('investments', 33),
 ('NEET', 33),
 ('1stdosecovid19vaccine', 33),
 ('UNSPECIFIED', 33),
 ('DoctorsDay2021', 33),
 ('antivaxxers', 32),
 ('DGCI', 32),
 ('BioNtech', 32),
 ('vaccino', 32),
 ('Switzerland', 32),
 ('VAXXED', 32),
 ('Asia', 32),
 ('Antibodies', 32),
 ('ivermectin', 32),
 ('COVID19ireland', 32),
 ('Bosnia', 32),
 ('Greece', 32),
 ('pharmacy', 32),
 ('freedom', 32),
 ('Afghanistan', 32),
 ('doyourpart', 32),
 ('Pharmaceutical', 32),
 ('Investment', 32),
 ('Libya', 32),
 ('Ethereum', 32),
 ('Uruguay', 32),
 ('Colombo', 32),
 ('covaxinvaccine', 32),
 ('Growth', 32),
 ('MaskUpIndia', 32),
 ('CyberSecurity', 31),
 ('HealthcareHeroes', 31),
 ('NHSheroes', 31),
 ('safe', 31),
 ('Macron', 31),
 ('AstraZenecaVaccine', 31),
 ('uae', 31),
 ('approval', 31),
 ('data', 31),
 ('immunization', 31),
 ('people', 31),
 ('Covid19vaccines', 31),
 ('amazon', 31),
 ('seruminstituteofindia', 31),
 ('Scientist', 31),
 ('à¸§à¸±à¸\x84à¸\x8bà¸µà¸\x99à¹\x82à¸\x84à¸§à¸´à¸\x9419', 31),
 ('COVIDVACCINE', 31),
 ('kids', 31),
 ('ModeRNA', 31),
 ('WhatsApp', 31),
 ('à¤¬à¤¿à¤¹à¤¾à¤°_à¤¬à¥\x87à¤°à¥\x8bà¤\x9cà¤\x97à¤¾à¤°à¥\x80_à¤¦à¤¿à¤µà¤¸',
  31),
 ('NEWS', 31),
 ('Jaipur', 31),
 ('Cameroon', 31),
 ('thesun', 31),
 ('lockdownextension', 31),
 ('covid1948', 31),
 ('singapore', 30),
 ('FreeSpeech', 30),
 ('fakenews', 30),
 ('booster', 30),
 ('Sweden', 30),
 ('dogecoin', 30),
 ('growth', 30),
 ('SLnews', 30),
 ('Doctor', 30),
 ('snow', 30),
 ('CCPCHINA', 30),
 ('TrinidadandTobago', 30),
 ('GregPalast', 30),
 ('P1', 30),
 ('European', 30),
 ('FarmersProtest', 30),
 ('covisheild', 30),
 ('MotivationMonday', 30),
 ('cancelboardexams2021', 30),
 ('cancelboardexam2021', 30),
 ('AntibodySequencing', 30),
 ('ModiSpeech', 30),
 ('Deltavariant', 30),
 ('tourism', 29),
 ('ThankYouNHS', 29),
 ('Medical', 29),
 ('SaveLives', 29),
 ('VaccinesForAll', 29),
 ('doingmypart', 29),
 ('Research', 29),
 ('vaccinationCovid', 29),
 ('covid19news', 29),
 ('PrimeMinister', 29),
 ('production', 29),
 ('aztrazeneca', 29),
 ('wellness', 29),
 ('pakistan', 29),
 ('RedFort', 29),
 ('NovaScotia', 29),
 ('VaccinateNY', 29),
 ('Cansino', 29),
 ('Paraguay', 29),
 ('maga', 29),
 ('Gurugram', 29),
 ('Censorship', 29),
 ('fitness', 29),
 ('FarmLaws', 29),
 ('GetVaxxed', 29),
 ('Haffkine', 29),
 ('IndianLivesMatter', 29),
 ('SecondDoses', 29),
 ('magnetchallenge', 29),
 ('Clinical', 29),
 ('vaccinatedandhappy', 29),
 ('MyBMCVaccinationUpdate', 29),
 ('BLA', 29),
 ('DeltaPlus', 29),
 ('ARMY', 29),
 ('TorontoVaccineDay', 29),
 ('HappyDoctorsDay', 29),
 ('BC', 28),
 ('LongCovid', 28),
 ('UnitedKingdom', 28),
 ('nse', 28),
 ('clinicaltrial', 28),
 ('worthit', 28),
 ('INDIA', 28),
 ('novovax', 28),
 ('dose2', 28),
 ('cryptocurrencies', 28),
 ('Sindh', 28),
 ('covid19vaccination', 28),
 ('COMMUNISTCHINA', 28),
 ('Somalia', 28),
 ('shots', 28),
 ('IPL2021', 28),
 ('Motivation', 28),
 ('Fortis', 28),
 ('pune', 28),
 ('ModernaGang', 28),
 ('CAA', 28),
 ('justice', 28),
 ('Jammu', 28),
 ('vaccinehesitancy', 27),
 ('pregnant', 27),
 ('vaccinationdrive', 27),
 …]
```

[19]:
```
pfizer_biontech_vax = ["Pfizer", "PFIZER", "PfizerBioNTech", "PfizerVaccine",␣
 ↪"pfizer", "PfizerBiontech", "BioNTech", "pfizerbiontech", "PfizerBioNtech",␣
 ↪"Biontech", "biontech", "PFIZERBIONTECH", "BioNTechpfizer"]
sputnik_vax = ["SputnikV", "Sputnik", "Sputnikv", "sputnikv", "SputnikUpdates",␣
 ↪"Sputnikvaccine", "sputnikV", "sputnik", "SPUTNIKV", "SputnikVaccinated",␣
 ↪"SputnikLight", "SputnikVaccineInKenya", "SputnikVaccine"]
sinopharm_vax = ["Sinopharm", "sinopharm", "SinoPharm", "BoycottSinopharm",␣
 ↪"SINOPHARM"]
sinovac_vax = ["Sinovac", "sinovac", "SinoVac", "SINOVAC", "BoycottSinovac"]
moderna_vax = ["Moderna", "moderna", "modernavaccine", "MODERNA",␣
 ↪"ModernaVaccine", "modernagang", "teammoderna", "modeRNA", "ModernaGang"]
oxford_az_vax = ["OxfordAstraZeneca", "oxfordastrazeneca", "Oxford",␣
 ↪"oxfordvaccine", "OxfordVaccine", "OxfordAstrazeneca", "AstraZeneca",␣
 ↪"astrazeneca", "AstraZenaca", "astrazenecavaccine", "Astrazeneca",␣
 ↪"AstraZeneka", "Astrazenaca", "ASTRAZENECA", "AstraZenecaVaccine"]
covaxin_vax = ["Covaxin", "COVAXIN", "covaxin", "GurgaonCOVAXIN",␣
 ↪"MumbaiCOVAXIN", "covaxine", "BBMPCOVAXIN", "PuneCOVAXIN", "CoVaxin",␣
 ↪"ThaneCOVAXIN", "covaxinated", "covaxinvaccine", "BharatBiotech",␣
 ↪"AatmanirbharBharat", "AtmaNirbharBharat", "bharatbiotech", "bharatBiotech",␣
 ↪"AtmanirbharBharat", "Bharat", "congressmuktbharat", "atmanirbharbharat"]
jandj_vax = ["johnsonandjohnson", "JohnsonandJohnson", "JohnsonAndJohnson",␣
 ↪"JohnsonAndJohnsonVaccine", "Johnson", "JandJ", "JohnsonJohnson", "johnson",␣
 ↪"JJ"]
```

normalizing hashtags to all lowercase:

[20]:
```
for i in range(0, len(raw_vaccine_tweets)):
    for j in range(len(raw_vaccine_tweets["hashtags"][i])):
        review = raw_vaccine_tweets["hashtags"][i][j]
        review = review.lower()
        raw_vaccine_tweets["hashtags"][i][j] = review
```

adding columns for each vaccine manufacturer, based on the hashtags of a tweet:

[21]:
```
raw_vaccine_tweets["PfizerBiontech"] = 0
raw_vaccine_tweets["SputnikV"] = 0
raw_vaccine_tweets["Sinopharm"] = 0
raw_vaccine_tweets["Sinovac"] = 0
raw_vaccine_tweets["Moderna"] = 0
raw_vaccine_tweets["AstraZeneca"] = 0
raw_vaccine_tweets["Covaxin"] = 0
raw_vaccine_tweets["JandJ"] = 0
```

[22]:

```
for i in range(len(raw_vaccine_tweets)):
    for hashtag in raw_vaccine_tweets["hashtags"][i]:
        if hashtag in pfizer_biontech_vax:
            raw_vaccine_tweets["PfizerBiontech"][i] = 1
        if hashtag in sputnik_vax:
            raw_vaccine_tweets["SputnikV"][i] = 1
        if hashtag in sinopharm_vax:
            raw_vaccine_tweets["Sinopharm"][i] = 1
        if hashtag in sinovac_vax:
            raw_vaccine_tweets["Sinovac"][i] = 1
        if hashtag in moderna_vax:
            raw_vaccine_tweets["Moderna"][i] = 1
        if hashtag in oxford_az_vax:
            raw_vaccine_tweets["AstraZeneca"][i] = 1
        if hashtag in covaxin_vax:
            raw_vaccine_tweets["Covaxin"][i] = 1
        if hashtag in jandj_vax:
            raw_vaccine_tweets["JandJ"][i] = 1
```

```
<ipython-input-22-0e6b7adc50b2>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets["PfizerBiontech"][i] = 1
<ipython-input-22-0e6b7adc50b2>:12: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets["Moderna"][i] = 1
<ipython-input-22-0e6b7adc50b2>:14: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets["AstraZeneca"][i] = 1
<ipython-input-22-0e6b7adc50b2>:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets["Sinopharm"][i] = 1
<ipython-input-22-0e6b7adc50b2>:16: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
  raw_vaccine_tweets["Covaxin"][i] = 1
<ipython-input-22-0e6b7adc50b2>:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets["SputnikV"][i] = 1
<ipython-input-22-0e6b7adc50b2>:10: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets["Sinovac"][i] = 1
<ipython-input-22-0e6b7adc50b2>:18: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets["JandJ"][i] = 1
```

---

[23]:

```
from geopy import Nominatim
import reverse_geocode
```

3.5 Geo location

- Fetching location from tweet geo data
- Engineer location from user profile
- transform both to standardized country names (needed for pyplot)

Fetching location specified in user profile:

[24]:

```
raw_vaccine_tweets["user_location"] = None
for i in range(len(raw_vaccine_tweets["user"])):
    raw_vaccine_tweets["user_location"][i] =␣
 ↪raw_vaccine_tweets["user"][i]["location"]
```

```
<ipython-input-24-bf0c79a898dc>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets["user_location"][i] =
raw_vaccine_tweets["user"][i]["location"]
```

[25]:

```
countries = ['argentina', 'australia', 'austria', 'belgium', 'brazil',␣
 ↪'canada', 'france', 'germany', 'india', 'israel', 'italy', 'japan',␣
 ↪'mexico', 'pakistan', 'russia', 'spain', 'uae', 'uk', 'usa']
```

Extract location from user profile and standardize country name:

```
[26]: #set country for locations that contain that country
      for country in countries :
          raw_vaccine_tweets["user_location"][raw_vaccine_tweets["user_location"].str.
       ↪lower().str.contains(country)] = country

      #remove any location that isn't in country-list
      for i in range(len(raw_vaccine_tweets)):
          if raw_vaccine_tweets["user_location"][i] not in countries:
              raw_vaccine_tweets["user_location"][i] = None
```

```
<ipython-input-26-2f75fae531ee>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets["user_location"][raw_vaccine_tweets["user_location"].str.lo
wer().str.contains(country)] = country
<ipython-input-26-2f75fae531ee>:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets["user_location"][i] = None
```

Get coordinates for every country:

```
[27]: geolocator = Nominatim(user_agent="CovaxAnalytica")
      location_coordinates = {}

      for country in countries:
          location = geolocator.geocode(country)
          try:
              location_coordinates[country] = [location.latitude, location.longitude]
          except:
              location_coordinates[country] = None
```

```
[28]: location_coordinates
```

```
[28]: {'argentina': [-34.9964963, -64.9672817],
       'australia': [-24.7761086, 134.755],
       'austria': [47.2, 13.2],
       'belgium': [50.6402809, 4.6667145],
       'brazil': [-10.3333333, -53.2],
       'canada': [61.0666922, -107.991707],
       'france': [46.603354, 1.8883335],
       'germany': [51.0834196, 10.4234469],
       'india': [22.3511148, 78.6677428],
       'israel': [31.5313113, 34.8667654],
       'italy': [42.6384261, 12.674297],
       'japan': [36.5748441, 139.2394179],
       'mexico': [22.5000485, -100.0000375],
       'pakistan': [30.3308401, 71.247499],
       'russia': [64.6863136, 97.7453061],
       'spain': [39.3260685, -4.8379791],
       'uae': [49.4871968, 31.2718321],
       'uk': [54.7023545, -3.2765753],
       'usa': [39.7837304, -100.4458825]}
```

Map country name to coordinates:

[29]:
```
raw_vaccine_tweets["coordinates"] = None
for i in range(len(raw_vaccine_tweets)):
     if raw_vaccine_tweets["user_location"][i] in countries:
        raw_vaccine_tweets["coordinates"][i] =␣
 ↪location_coordinates[raw_vaccine_tweets["user_location"][i].lower()]
```

```
<ipython-input-29-54fe49d17fbe>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets["coordinates"][i] =
location_coordinates[raw_vaccine_tweets["user_location"][i].lower()]
```

[30]:
```
raw_vaccine_tweets
```

[30]:
```
                         id                 created_at  \
0       1338158543359250432 2020-12-13 16:27:13+00:00
1       1337840331522453504 2020-12-12 19:22:45+00:00
2       1338544403795881984 2020-12-14 18:00:29+00:00
3       1337735595704115200 2020-12-12 12:26:34+00:00
4       1337850832256176128 2020-12-12 20:04:29+00:00
…                       …                         …
118267  1407688257177935872 2021-06-23 13:13:28+00:00
118268  1407699323035557888 2021-06-23 13:57:27+00:00
118269  1407699835856330752 2021-06-23 13:59:29+00:00
118270  1407682599515000832 2021-06-23 12:50:59+00:00
118271  1407696190578249728 2021-06-23 13:45:00+00:00

                                                    user   geo   \
0       {'id': 76052772, 'id_str': '76052772', 'name':…  None
1       {'id': 1300382181605494800, 'id_str': '1300382…  None
2       {'id': 1164717209253552000, 'id_str': '1164717…  None
3       {'id': 1316036067754205200, 'id_str': '1316036…  None
```

```
4       {'id': 1110032180237852700, 'id_str': '1110032…  None
…                                                   …   …
118267  {'id': 1263779139397382100, 'id_str': '1263779…  None
118268  {'id': 40623001, 'id_str': '40623001', 'name':…  None
118269  {'id': 61611674, 'id_str': '61611674', 'name':…  None
118270  {'id': 126591034, 'id_str': '126591034', 'name…  None
118271  {'id': 231692806, 'id_str': '231692806', 'name…  None

                                                full_text  \
0       While the world has been on the wrong side of …
1       @cnnbrk #COVID19 #CovidVaccine #vaccine #Coron…
2       The FDA Authorizes Emergency Use Of The Pfizer…
3       The #FDA finally issues #EUA now comes the pro…
4       There have not been many bright days in 2020 b…
…                                                   …
118267  #SputnikV Paid #Hyderabad https://t.co/oklatcuWLh
118268  The @WHO said its review of how #Russia produc…
118269  #WHO Finds Production Infringements at #Sputni…
118270  When was the #SputnikV\n\n1. Exploratory Stage…
118271  .@WHO raises concern on cross-contamination, i…

                                                 hashtags  \
0       [covid19, supplychain, logistics, vaccine, uni…
1       [covid19, covidvaccine, vaccine, corona, pfize…
2       [pfe, pfizer, pfizervaccine, pfizerbiontech, f…
3                  [fda, eua, pfizerbiontech, vaccinated]
4       [bidenharris, election2020, pfizerbiontech, co…
…                                                   …
118267                              [sputnikv, hyderabad]
118268                    [russia, sputnikv, coronavirus]
118269  [who, sputnikv, russia, covid19, corona, impfs…
118270                                         [sputnikv]
118271                                         [sputnikv]

                    user_id  PfizerBiontech  SputnikV  Sinopharm  Sinovac  \
0                  76052772               1         0          0        0
1       1300382181605494800               1         0          0        0
2       1164717209253552000               1         0          0        0
3       1316036067754205200               1         0          0        0
4       1110032180237852700               1         0          0        0
…                     …             …       …        …      …
118267  1263779139397382100               0         1          0        0
118268             40623001               0         1          0        0
118269             61611674               0         1          0        0
118270            126591034               0         1          0        0
118271            231692806               0         1          0        0
```

```
        Moderna  AstraZeneca  Covaxin  JandJ user_location  \
0             0            0        0      0          None
1             0            0        0      0          None
2             0            0        0      0          None
3             0            0        0      0          None
4             0            0        0      0          None
…         …          …      …     …           …
118267        0            0        0      0         india
118268        0            0        0      0          None
118269        0            0        0      0          None
118270        0            0        0      0          None
118271        0            0        0      0         india

                     coordinates
0                           None
1                           None
2                           None
3                           None
4                           None
…                          …
118267  [22.3511148, 78.6677428]
118268                      None
118269                      None
118270                      None
118271  [22.3511148, 78.6677428]

[118272 rows x 17 columns]
```

add the coordinates from the geo column:

[31]:
```
for i in range(len(raw_vaccine_tweets)):
    if raw_vaccine_tweets.geo[i] != None:
        raw_vaccine_tweets.coordinates[i] = raw_vaccine_tweets.
 ↪geo[i]["coordinates"]
```

```
<ipython-input-31-35795c746503>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets.coordinates[i] = raw_vaccine_tweets.geo[i]["coordinates"]
```

renaming all countries in standardized way using “reverse_geocode”:

[32]:
```
for i in range(len(raw_vaccine_tweets)):
    if raw_vaccine_tweets.coordinates[i] != None:
        raw_vaccine_tweets.user_location[i] = reverse_geocode.
 ↪search(tuple([raw_vaccine_tweets.coordinates[i],(1,1)]))[0]["country"]
```

```
<ipython-input-32-d848799ac007>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  raw_vaccine_tweets.user_location[i] = reverse_geocode.search(tuple([raw_vaccin
e_tweets.coordinates[i],(1,1)]))[0]["country"]
```

```
[33]: raw_vaccine_tweets["user_location"].unique()
```

```
[33]: array([None, 'Canada', 'Palestinian Territory', 'India', 'Germany',
             'United States', 'Italy', 'United Kingdom', 'France',
             'Russian Federation', 'Mexico', 'Belgium', 'Spain', 'Australia',
             'Pakistan', 'Ukraine', 'Argentina', 'Austria',
             'Virgin Islands, U.S.', 'Malaysia', 'Japan', 'Brazil',
             'United Arab Emirates', 'Jersey', 'Philippines', 'Chile',
             'Indonesia', 'Hong Kong', 'Qatar', 'Netherlands', 'China',
             'Saudi Arabia', 'Guyana', 'Thailand', 'Singapore', 'Croatia',
             'Switzerland', 'Trinidad and Tobago', 'Greece', 'Isle of Man',
             'Sweden'], dtype=object)
```

4 Exporting dataset

```
[34]: raw_vaccine_tweets.to_csv("../data/interim/cleaned_vaccine_tweets.csv")
```