SIGMOID

CASE STUDY PRESENTATION

OLIVE CHAUDHURI
KALUVA PREETHI
MENTORED BY: PRATEEK SRIVASTAVA

PROBLEM STATEMENT

- The Twitter Producer should get data from Twitter based on some keywords and put them in any distributed messaging system.
- The Consumer component should get data from your messaging system.
- Insertion of the data into a distributed datastore after doing some basic transformations
- Write an API service which can be queried using any tool like Postman

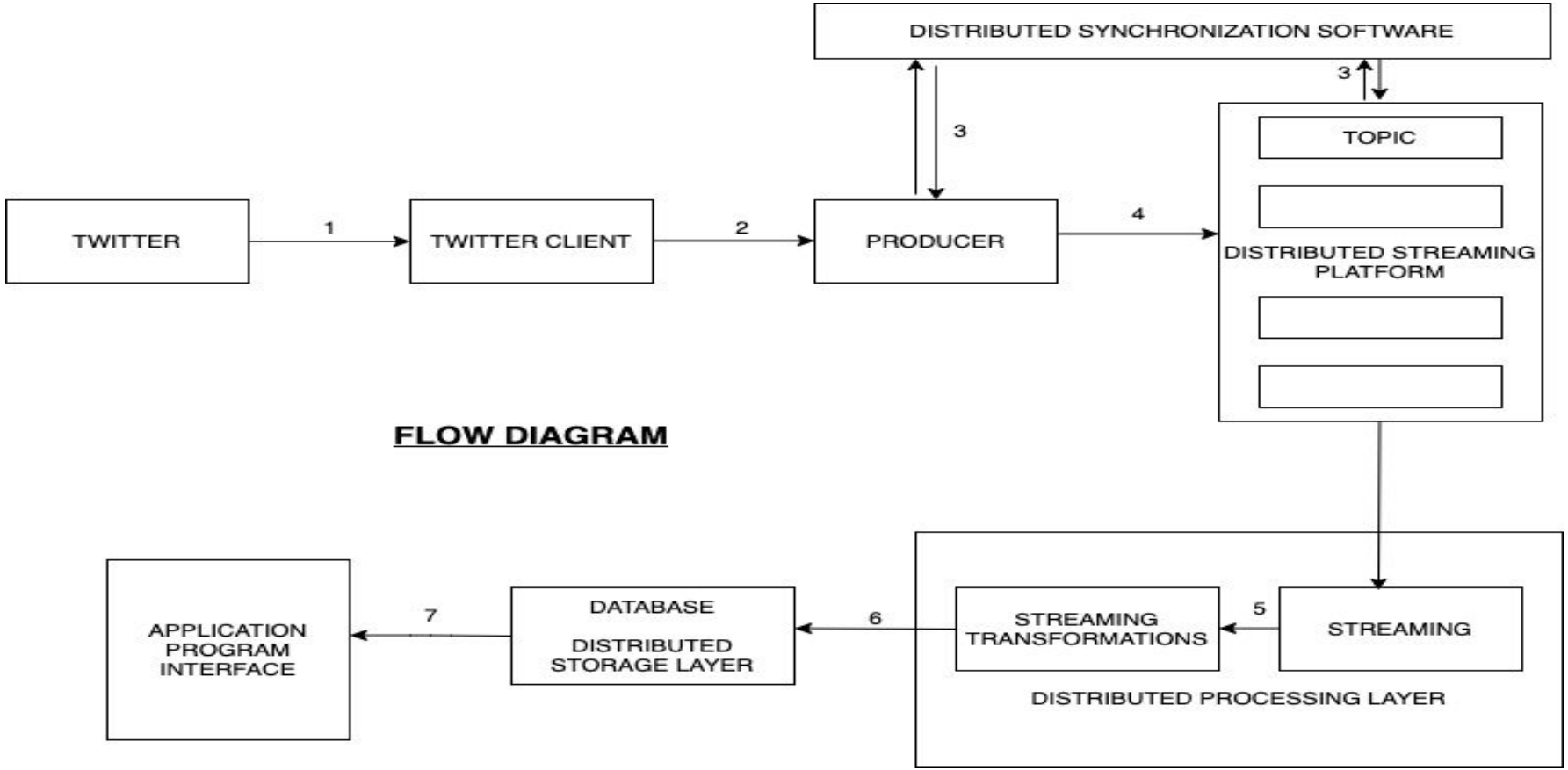
Implementation of following functionalities :

1. **Need to find overall number of tweets on coronavirus (keywords can be virus/covid-19/corona etc etc) per country in the last 3 months.**
2. **Next, need to find overall number of tweets per country on a daily basis.**
3. **Also need to find the top 100 words occurring on tweets involving coronavirus. (words should be nouns/verbs and not involving common ones like the, is, are, etc etc).**
4. **Find the top 100 words occurring on tweets involving coronavirus on a per country basis.**
5. **Total no. of donations received towards COVID 19 country wise, in all the affected countries.**

TECHNICAL REQUIREMENTS

1. Twitter developer account - For the twitter data.
2. Twitter Client - To fetch the data from twitter.
3. Streaming Producer - To store the data in the topic of distributed streaming platforms.
4. Distributed Streaming Service - To get the data, stream it, transform it in the distributed processing layer.
5. Distributed Datastore - to store all the data in a distributed storage layer.
6. REST API - to fetch the required data from the database (integration with UI)

FLOW DIAGRAM



TECHNOLOGY STACK

Twitter Streaming API

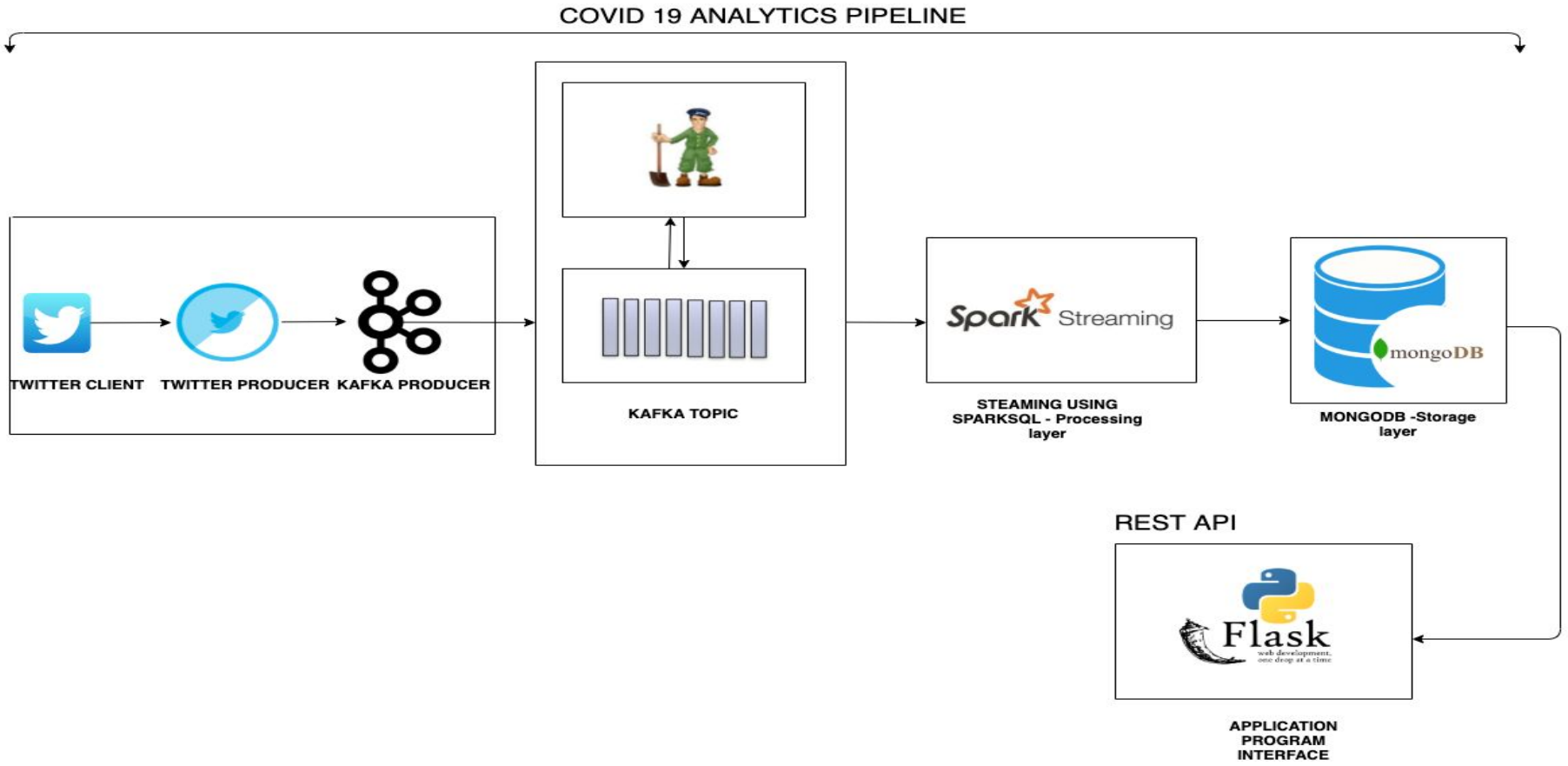
Apache Kafka

Apache Spark Stream

MongoDB

Flask

ARCHITECTURE DIAGRAM



APPROACH

Module 1 : Data Fetching

1. Creation of Twitter client
2. Creation of Kafka Producer

Module 2 : Streaming and Processing

1. Integration of spark shell and kafka topic using Spark Streaming
2. Setting format of the tweets.
3. Insertion of tweets into MongoDB collections

Module 3 : Rest API

1. Creation of a flask API in python
2. Performing queries on MongoDB

CORNER CASES HANDLED WHILE FORMATTING TWEETS

- ➔ Removal of invalid country names by performing a lookup in a set of predefined countries
- ➔ Extraction of tweeted text filtered using a set of stopwords
- ➔ Extraction of date formatted as mmm dd, yyyy
- ➔ In case of unavailability of the extracted date, timestamp(in ms) is used and converted to above date format

These three data attributes were inserted into 4 collections and an update of count is maintained. This is done to make the query faster.

SCHEMA OF MONGODB COLLECTIONS:

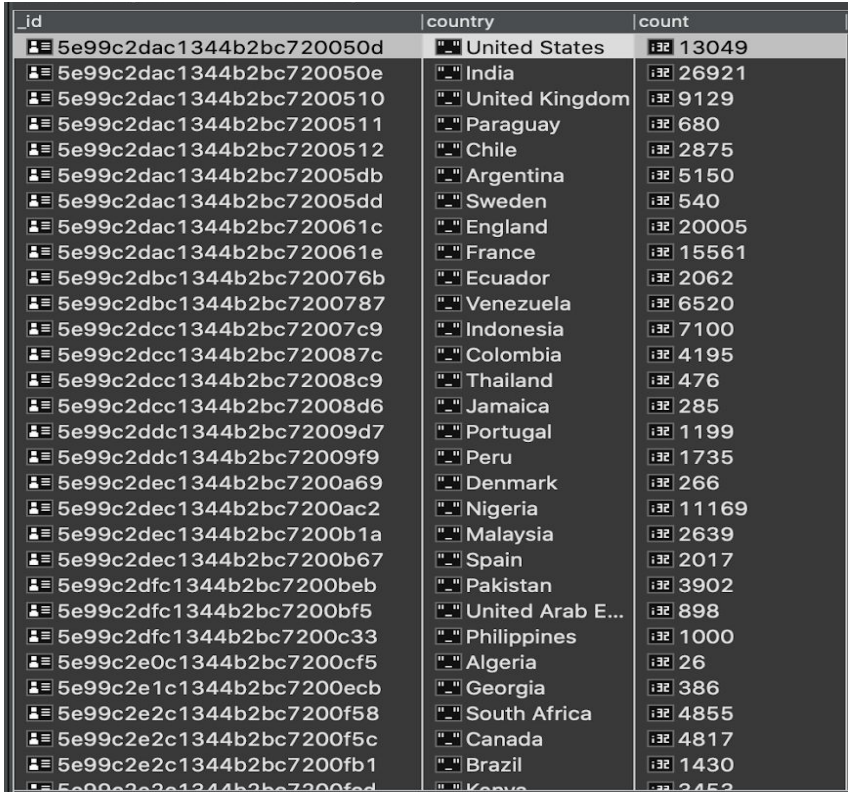
4 collections have been used to solve 5 queries

1st Collection:

_id : system generated unique object id for this row during creation

country : names of country

count : counter of the number of tweets updated for each country



_id	country	count
5e99c2dac1344b2bc720050d	United States	13049
5e99c2dac1344b2bc720050e	India	26921
5e99c2dac1344b2bc7200510	United Kingdom	9129
5e99c2dac1344b2bc7200511	Paraguay	680
5e99c2dac1344b2bc7200512	Chile	2875
5e99c2dac1344b2bc72005db	Argentina	5150
5e99c2dac1344b2bc72005dd	Sweden	540
5e99c2dac1344b2bc720061c	England	20005
5e99c2dac1344b2bc720061e	France	15561
5e99c2dbc1344b2bc720076b	Ecuador	2062
5e99c2dbc1344b2bc7200787	Venezuela	6520
5e99c2dcc1344b2bc72007c9	Indonesia	7100
5e99c2dcc1344b2bc720087c	Colombia	4195
5e99c2dcc1344b2bc72008c9	Thailand	476
5e99c2dcc1344b2bc72008d6	Jamaica	285
5e99c2ddc1344b2bc72009d7	Portugal	1199
5e99c2ddc1344b2bc72009f9	Peru	1735
5e99c2dec1344b2bc7200a69	Denmark	266
5e99c2dec1344b2bc7200ac2	Nigeria	11169
5e99c2dec1344b2bc7200b1a	Malaysia	2639
5e99c2dec1344b2bc7200b67	Spain	2017
5e99c2dfc1344b2bc7200beb	Pakistan	3902
5e99c2dfc1344b2bc7200bf5	United Arab E...	898
5e99c2dfc1344b2bc7200c33	Philippines	1000
5e99c2e0c1344b2bc7200cf5	Algeria	26
5e99c2e1c1344b2bc7200ecb	Georgia	386
5e99c2e2c1344b2bc7200f58	South Africa	4855
5e99c2e2c1344b2bc7200f5c	Canada	4817
5e99c2e2c1344b2bc7200fb1	Brazil	1430

2nd Collection:

_id : system generated unique object id for this row during creation

country : names of country

date : date of tweet

count : counter of the number of tweets updated for each country and date

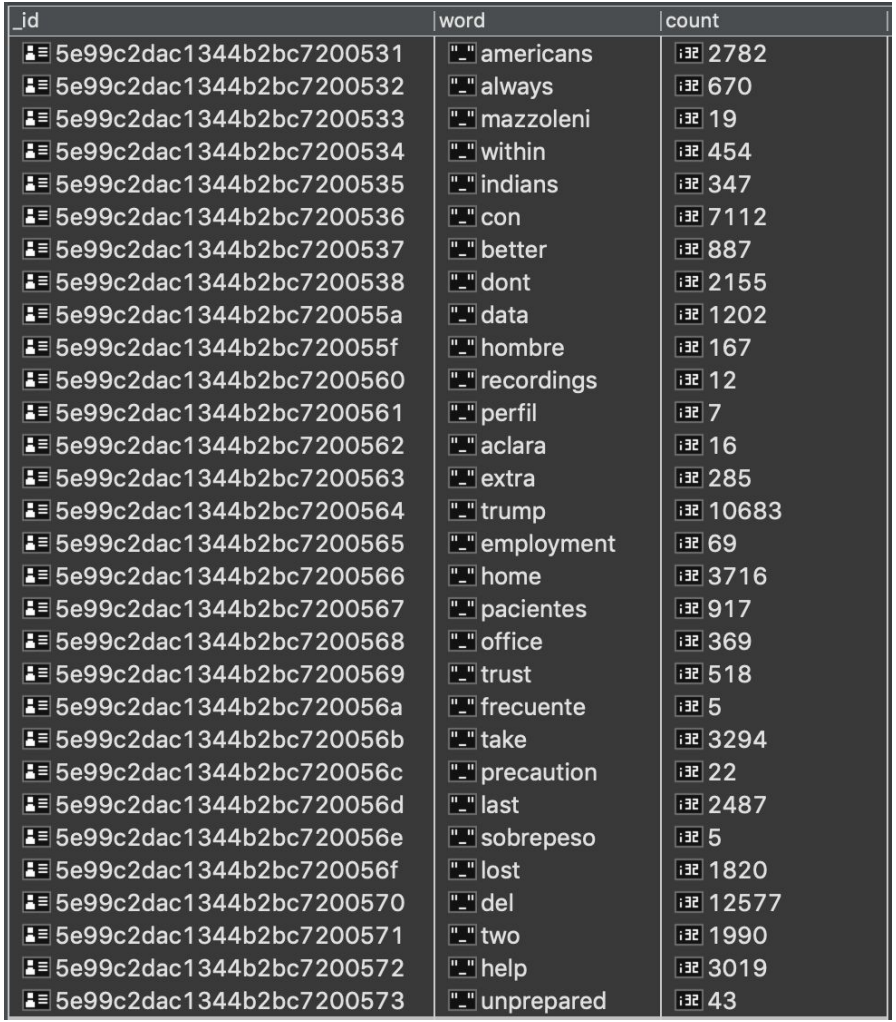
_id	country	date	count
5e99c2dac1344b2bc7200559	Chile	Apr 10, 2020	274
5e99c2dac1344b2bc720055b	India	Apr 10, 2020	1010
5e99c2dac1344b2bc720055c	United States	Apr 10, 2020	1070
5e99c2dac1344b2bc720055d	United Kingdom	Apr 10, 2020	366
5e99c2dac1344b2bc720055e	Paraguay	Apr 10, 2020	76
5e99c2dac1344b2bc72005e7	Argentina	Apr 10, 2020	364
5e99c2dac1344b2bc72005e9	Sweden	Apr 10, 2020	27
5e99c2dac1344b2bc7200625	England	Apr 10, 2020	895
5e99c2dac1344b2bc720062a	France	Apr 10, 2020	763
5e99c2dbc1344b2bc7200779	Ecuador	Apr 10, 2020	206
5e99c2dbc1344b2bc7200793	Venezuela	Apr 10, 2020	477
5e99c2dcc1344b2bc72007d5	Indonesia	Apr 10, 2020	135
5e99c2dcc1344b2bc7200887	Colombia	Apr 10, 2020	340
5e99c2dcc1344b2bc72008d1	Thailand	Apr 10, 2020	31
5e99c2dcc1344b2bc72008e1	Jamaica	Apr 10, 2020	17
5e99c2ddc1344b2bc72009e1	Portugal	Apr 10, 2020	39
5e99c2ddc1344b2bc7200a00	Peru	Apr 10, 2020	221
5e99c2dec1344b2bc7200a70	Denmark	Apr 10, 2020	16
5e99c2dec1344b2bc7200aca	Nigeria	Apr 10, 2020	454
5e99c2dec1344b2bc7200b27	Malaysia	Apr 10, 2020	147
5e99c2dec1344b2bc7200b72	Spain	Apr 10, 2020	95
5e99c2dfc1344b2bc7200bf3	Pakistan	Apr 10, 2020	115
5e99c2dfc1344b2bc7200c00	United Arab Emirates	Apr 10, 2020	29
5e99c2dfc1344b2bc7200c3c	Philippines	Apr 10, 2020	15
5e99c2e0c1344b2bc7200cff	Algeria	Apr 10, 2020	3
5e99c2e1c1344b2bc7200ed5	Georgia	Apr 10, 2020	49
5e99c2e2c1344b2bc7200f62	South Africa	Apr 10, 2020	250
5e99c2e2c1344b2bc7200f65	Canada	Apr 10, 2020	418
5e99c2e2c1344b2bc7200fb7	Brazil	Apr 10, 2020	73
5e99c2e2c1344b2bc7200fd9	Kenya	Apr 10, 2020	163

3rd Collection:

_id : system generated unique object id for this row during creation

word : individual words extracted from filtered tweet

count : counter of the number of words updated for each record



_id	word	count
5e99c2dac1344b2bc7200531	americans	2782
5e99c2dac1344b2bc7200532	always	670
5e99c2dac1344b2bc7200533	mazzoleni	19
5e99c2dac1344b2bc7200534	within	454
5e99c2dac1344b2bc7200535	indians	347
5e99c2dac1344b2bc7200536	con	7112
5e99c2dac1344b2bc7200537	better	887
5e99c2dac1344b2bc7200538	dont	2155
5e99c2dac1344b2bc720055a	data	1202
5e99c2dac1344b2bc720055f	hombre	167
5e99c2dac1344b2bc7200560	recordings	12
5e99c2dac1344b2bc7200561	perfil	7
5e99c2dac1344b2bc7200562	aclara	16
5e99c2dac1344b2bc7200563	extra	285
5e99c2dac1344b2bc7200564	trump	10683
5e99c2dac1344b2bc7200565	employment	69
5e99c2dac1344b2bc7200566	home	3716
5e99c2dac1344b2bc7200567	pacientes	917
5e99c2dac1344b2bc7200568	office	369
5e99c2dac1344b2bc7200569	trust	518
5e99c2dac1344b2bc720056a	frecuente	5
5e99c2dac1344b2bc720056b	take	3294
5e99c2dac1344b2bc720056c	precaution	22
5e99c2dac1344b2bc720056d	last	2487
5e99c2dac1344b2bc720056e	sobrepeso	5
5e99c2dac1344b2bc720056f	lost	1820
5e99c2dac1344b2bc7200570	del	12577
5e99c2dac1344b2bc7200571	two	1990
5e99c2dac1344b2bc7200572	help	3019
5e99c2dac1344b2bc7200573	unprepared	43

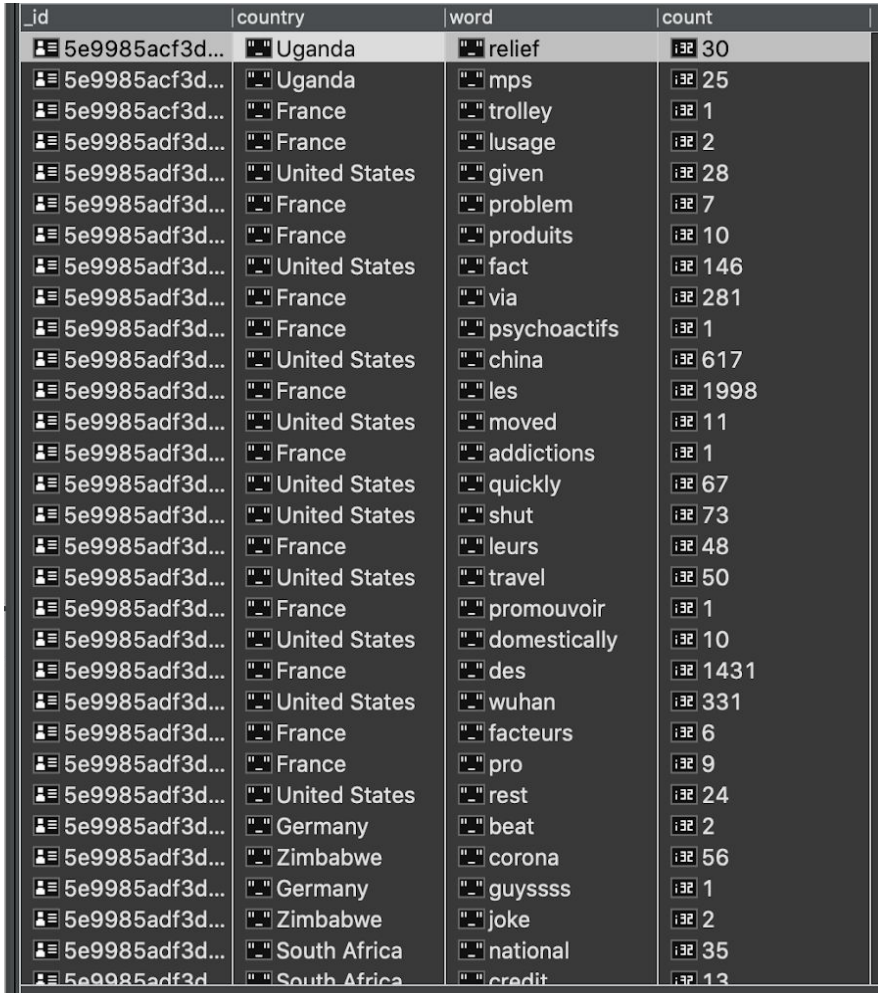
4th Collection:

_id : system generated unique object id for this row during creation

country : name of country

word : individual words extracted from filtered tweet

count : counter of the number of words for each country updated for each record



_id	country	word	count
5e9985acf3d...	Uganda	relief	30
5e9985acf3d...	Uganda	mps	25
5e9985acf3d...	France	trolley	1
5e9985adf3d...	France	lusage	2
5e9985adf3d...	United States	given	28
5e9985adf3d...	France	problem	7
5e9985adf3d...	France	produits	10
5e9985adf3d...	United States	fact	146
5e9985adf3d...	France	via	281
5e9985adf3d...	France	psychoactifs	1
5e9985adf3d...	United States	china	617
5e9985adf3d...	France	les	1998
5e9985adf3d...	United States	moved	11
5e9985adf3d...	France	addictions	1
5e9985adf3d...	United States	quickly	67
5e9985adf3d...	United States	shut	73
5e9985adf3d...	France	leurs	48
5e9985adf3d...	United States	travel	50
5e9985adf3d...	France	promouvoir	1
5e9985adf3d...	United States	domestically	10
5e9985adf3d...	France	des	1431
5e9985adf3d...	United States	wuhan	331
5e9985adf3d...	France	facteurs	6
5e9985adf3d...	France	pro	9
5e9985adf3d...	United States	rest	24
5e9985adf3d...	Germany	beat	2
5e9985adf3d...	Zimbabwe	corona	56
5e9985adf3d...	Germany	guyssss	1
5e9985adf3d...	Zimbabwe	joke	2
5e9985adf3d...	South Africa	national	35
5e9985adf3d...	South Africa	credit	13

CORNER CASES HANDLED IN THE API

- ➔ “Page Not found-404” error and “Database Connection Failure - 500” handled with an error handler that displays a default message.

- ➔ For the questions we have solved we have given an additional option to enter a country name as an input in the URL.

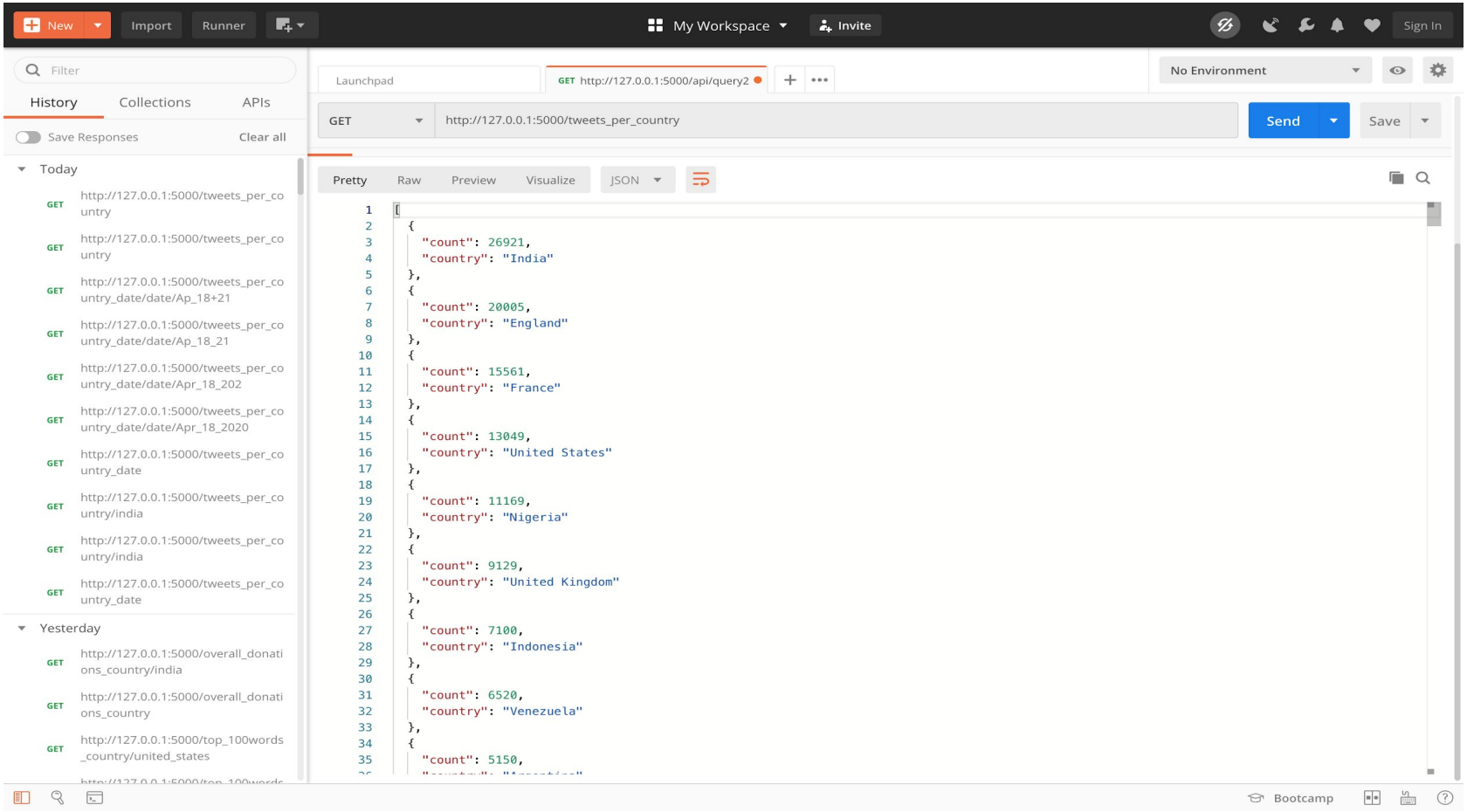
- ➔ Country name can be given in lowercase or uppercase with an underscore in between two words.

- ➔ Date can also be entered with underscores between the month, date and year in the same order.

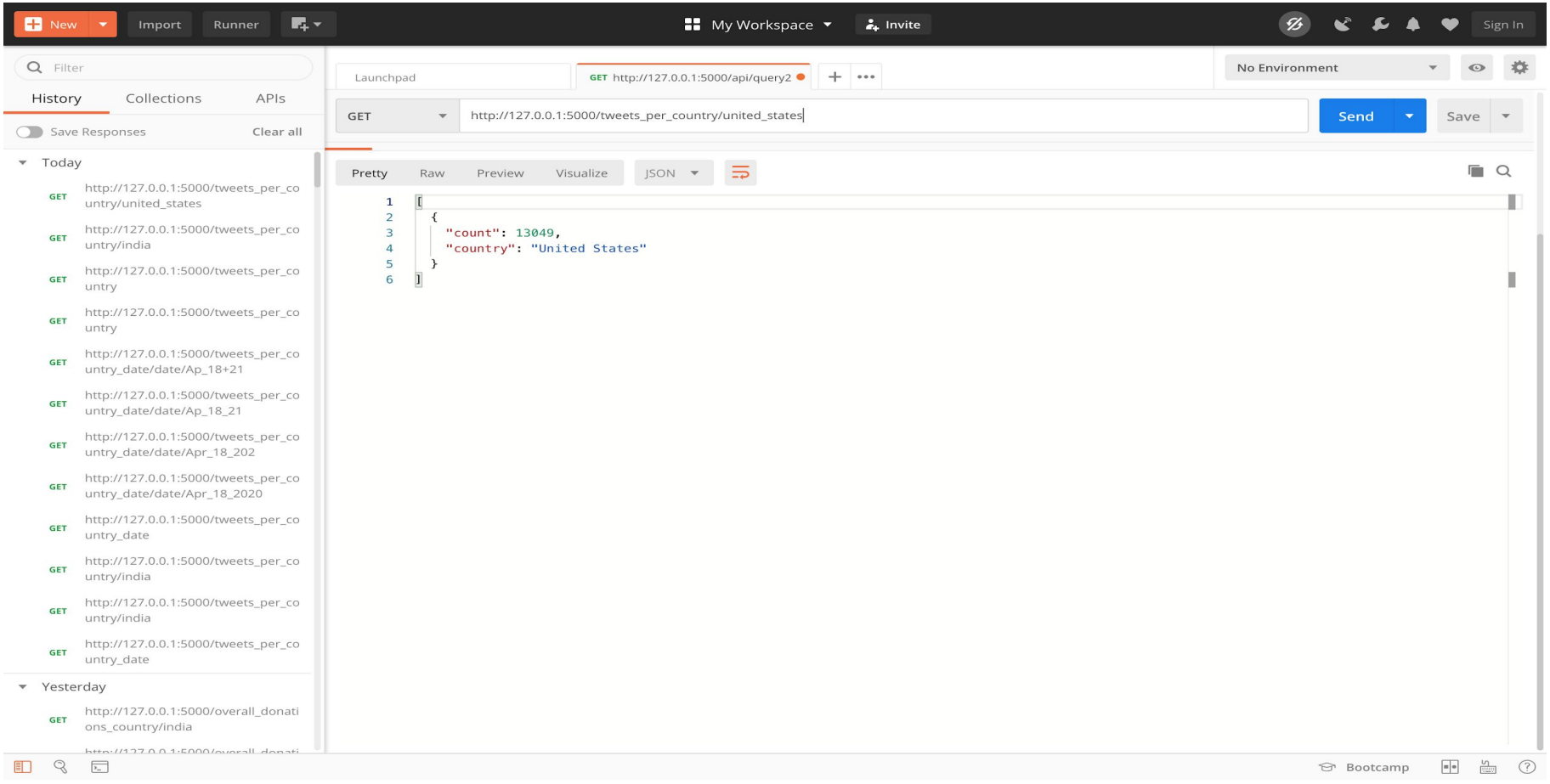
API SCHEMA

- ➔ localhost:5000/tweets_per_country
- ➔ localhost:5000/tweets_per_country/<country_name>
- ➔ localhost:5000/tweets_per_country_date
- ➔ localhost:5000/tweets_per_country_date/<country_name>
- ➔ localhost:5000/tweets_per_country_date/date/<date>
- ➔ localhost:5000/tweets_per_country_date/date_country/<country_name>/<date>
- ➔ localhost:5000/overall_top_100words
- ➔ localhost:5000/top_100words_country
- ➔ localhost:5000/top_100words_country/<country_name>
- ➔ localhost:5000/overall_donations_country
- ➔ localhost:5000/overall_donations_country/<country_name>

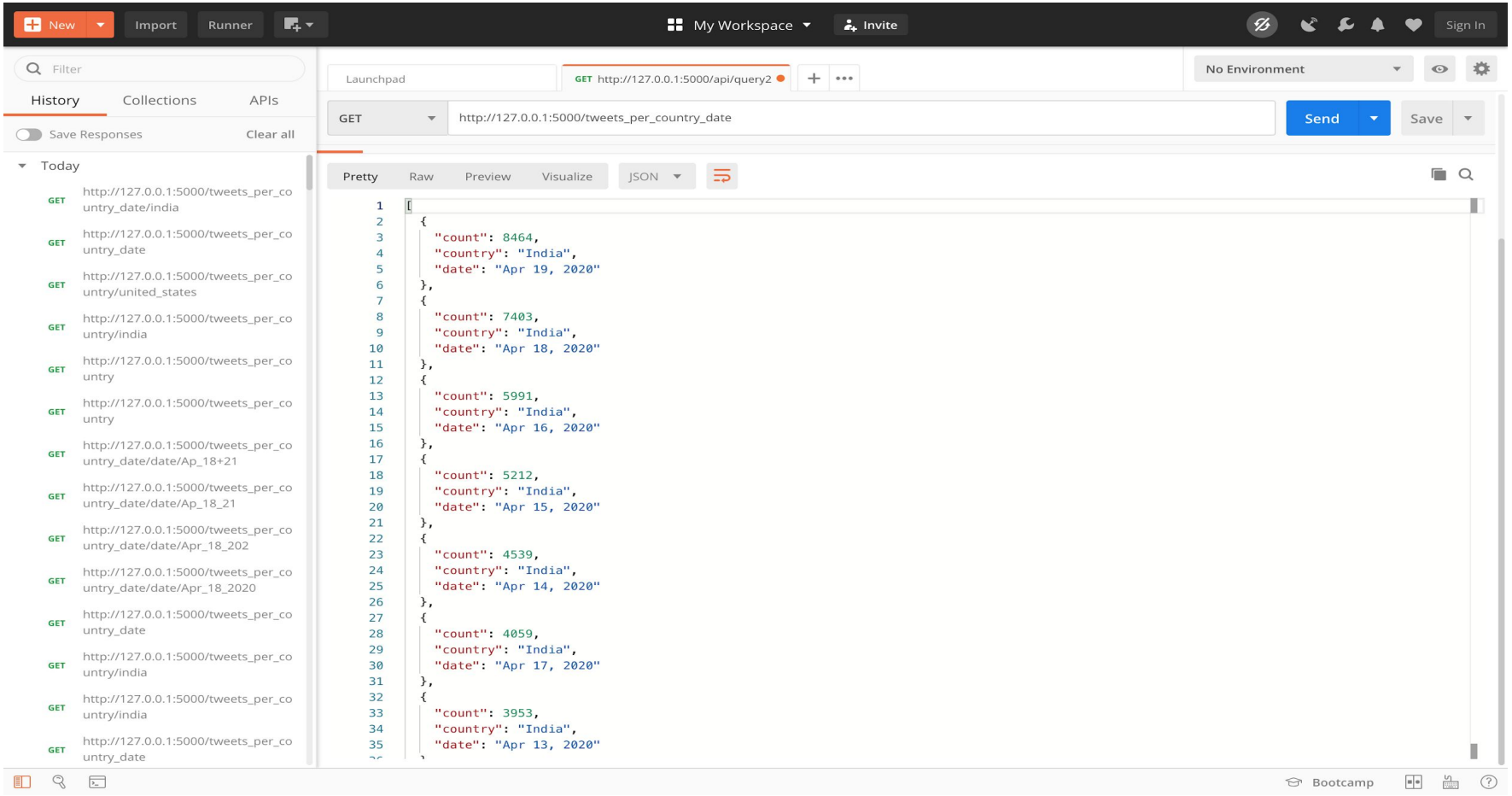
localhost:5000/tweets_per_country : Finding count of overall tweets per country.



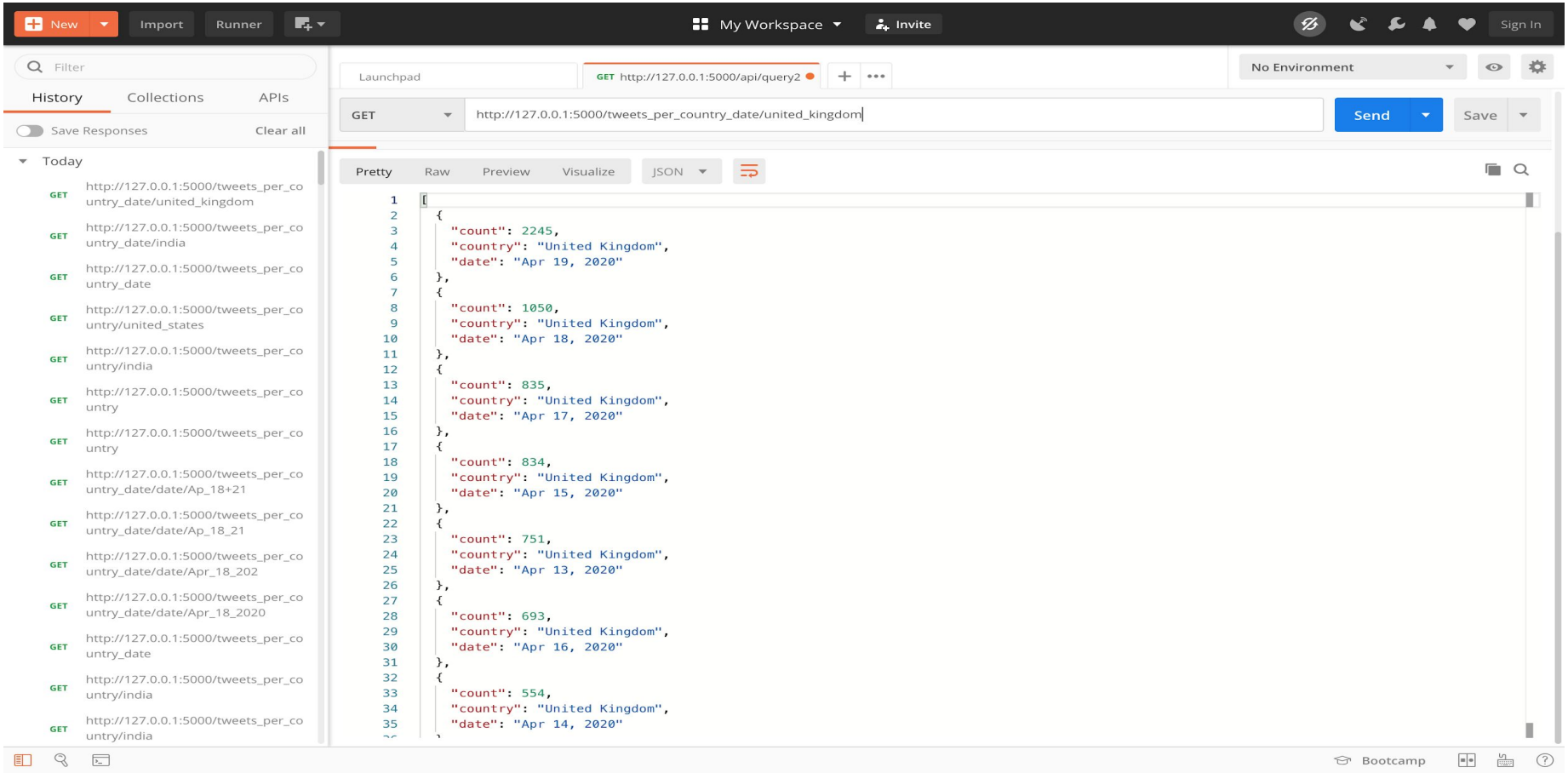
localhost:5000/tweets_per_country/<country_name> : Finding count of overall tweets per country for a particular searched country.



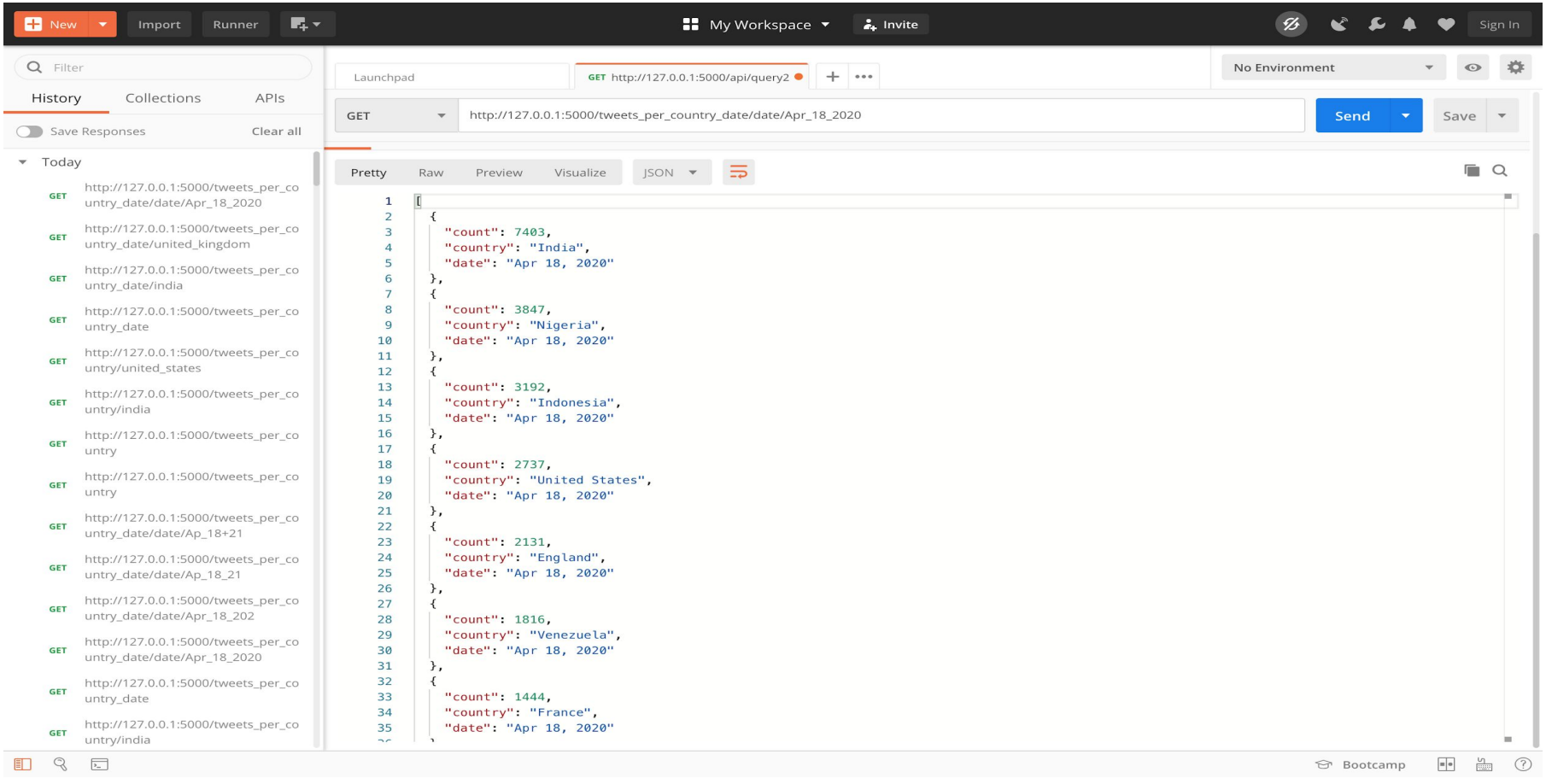
localhost:5000/tweets_per_country_date : Finding overall count of tweets per country on a daily basis.



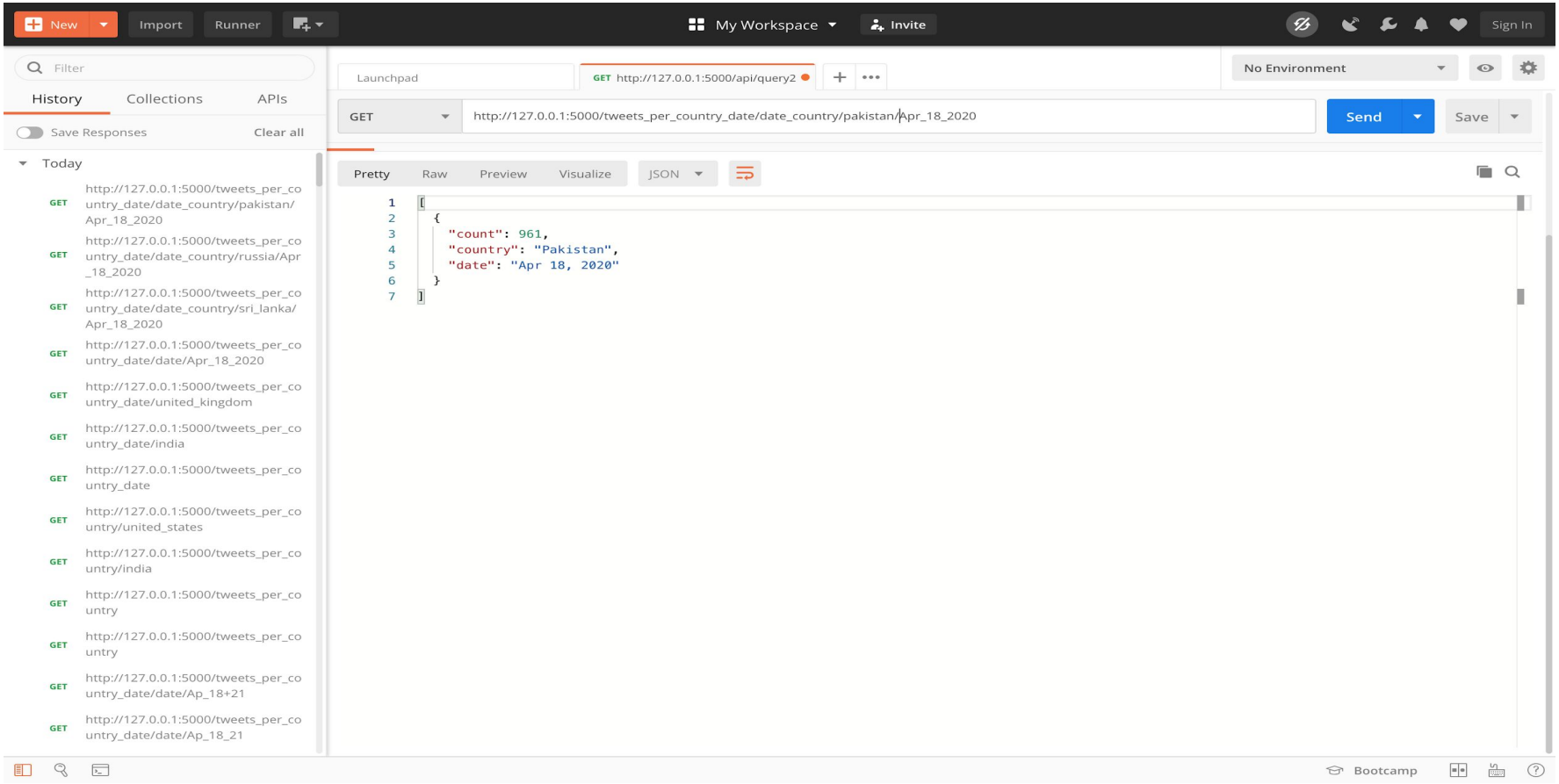
localhost:5000/tweets_per_country_date/<country_name> : Finding count of overall tweets per country on a daily basis for a particular searched country.



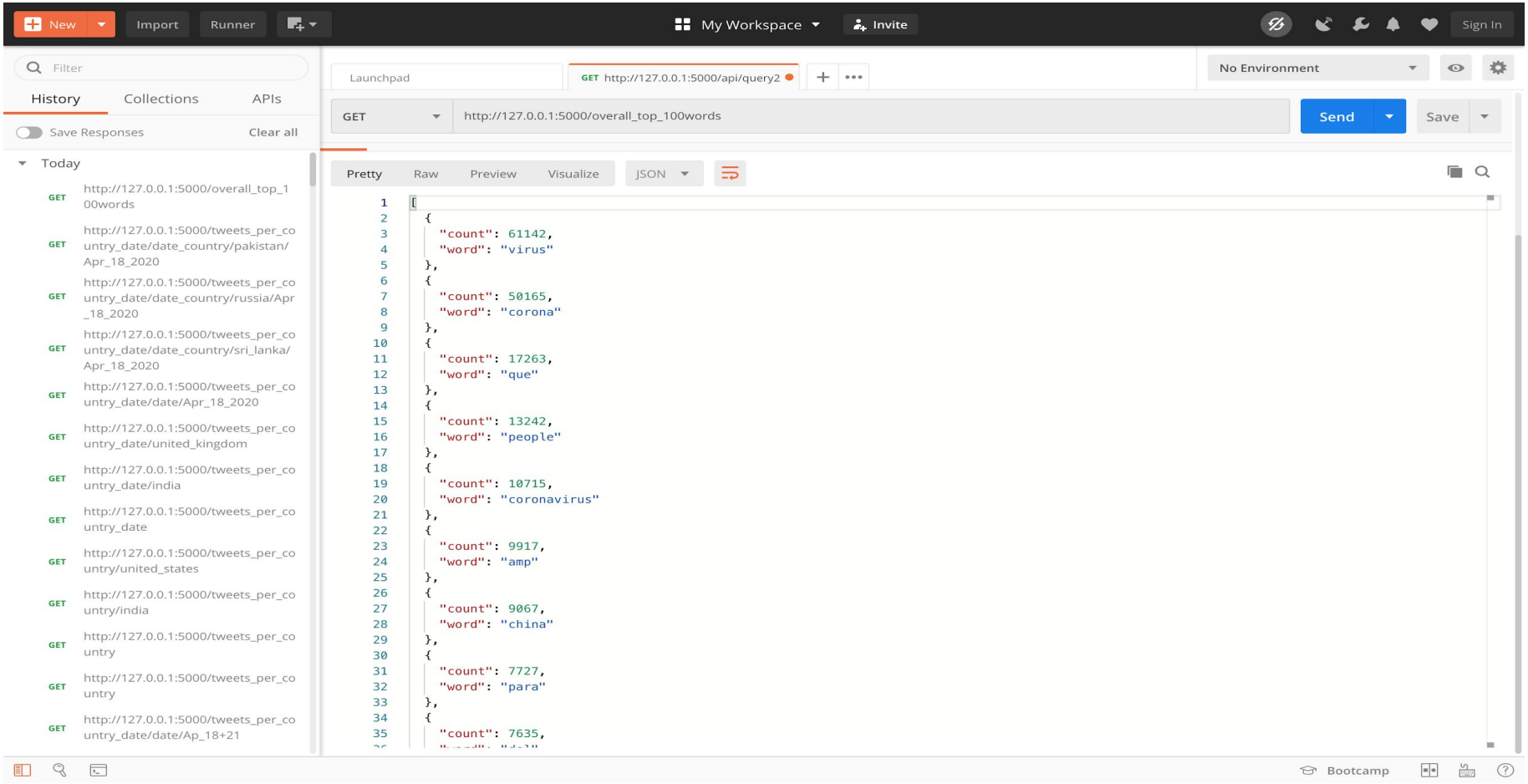
localhost:5000/tweets_per_country_date/date/<date> : Finding count of overall tweets per country on a daily basis for a particular searched date.



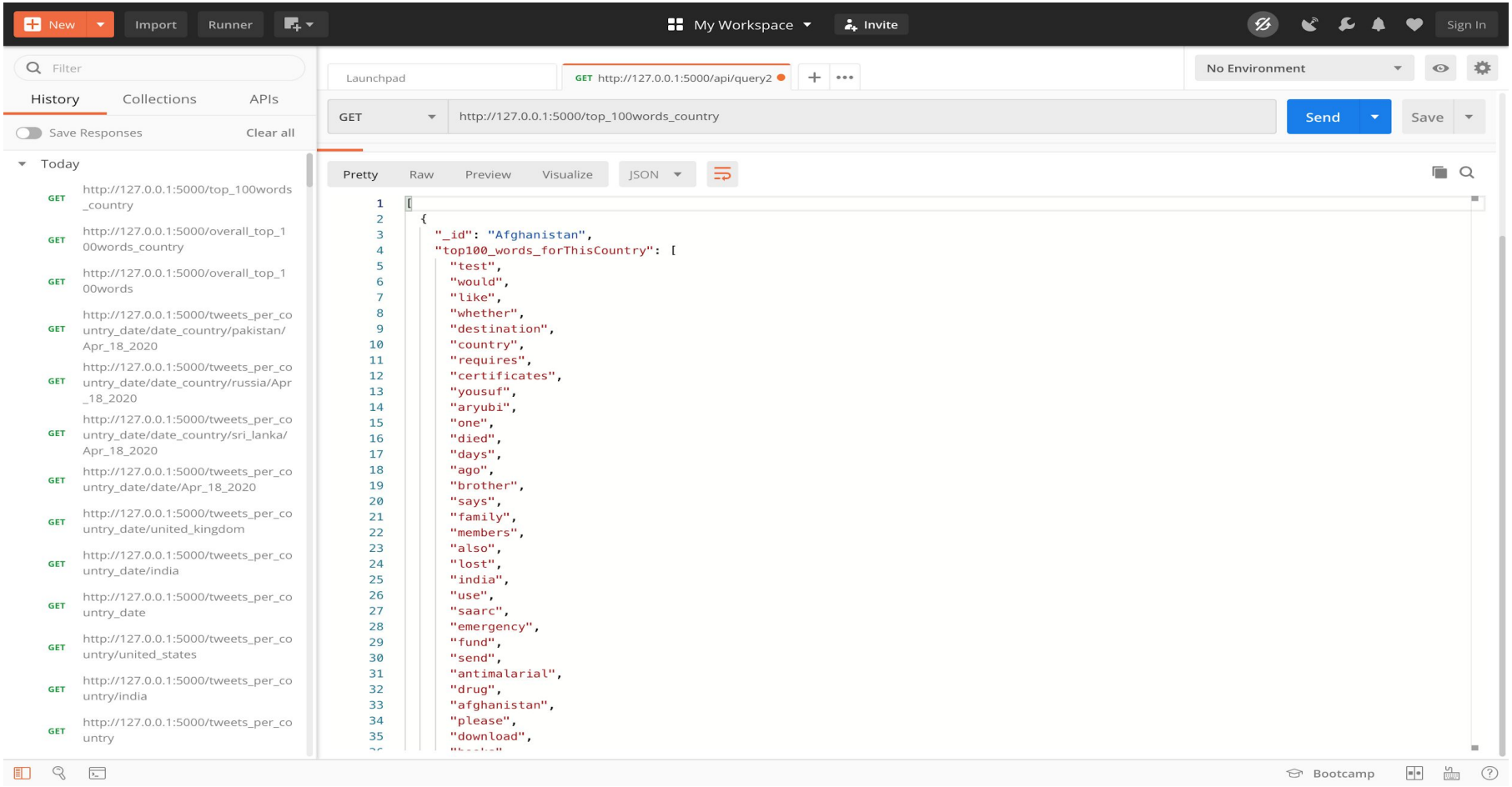
localhost:5000/tweets_per_country_date/date_country/<country_name><date> : Finding count of overall tweets per country on a daily basis for a particular searched date and country.



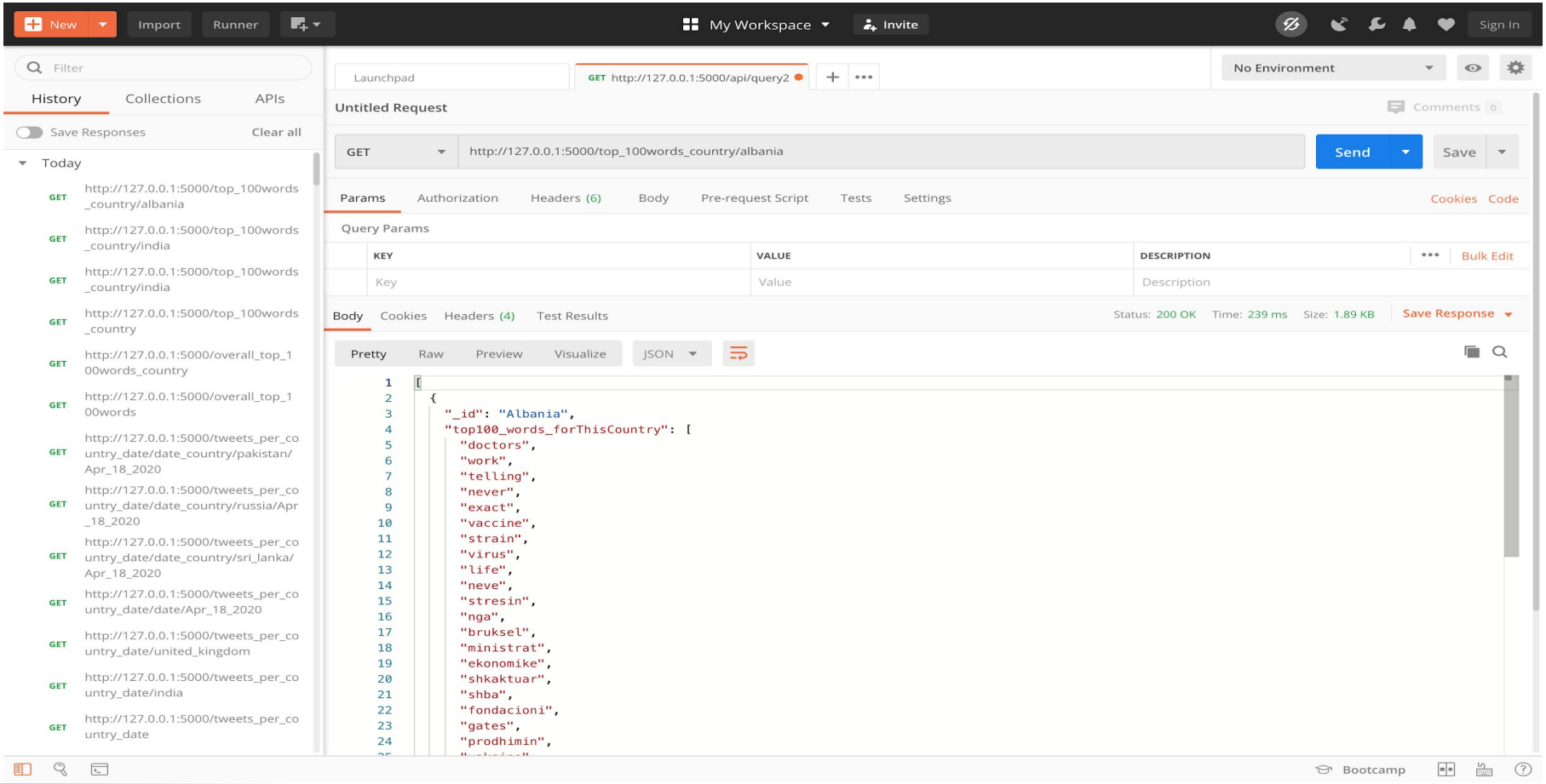
localhost:5000/overall_top_100words : Finding the top 100 words occurring on tweets.



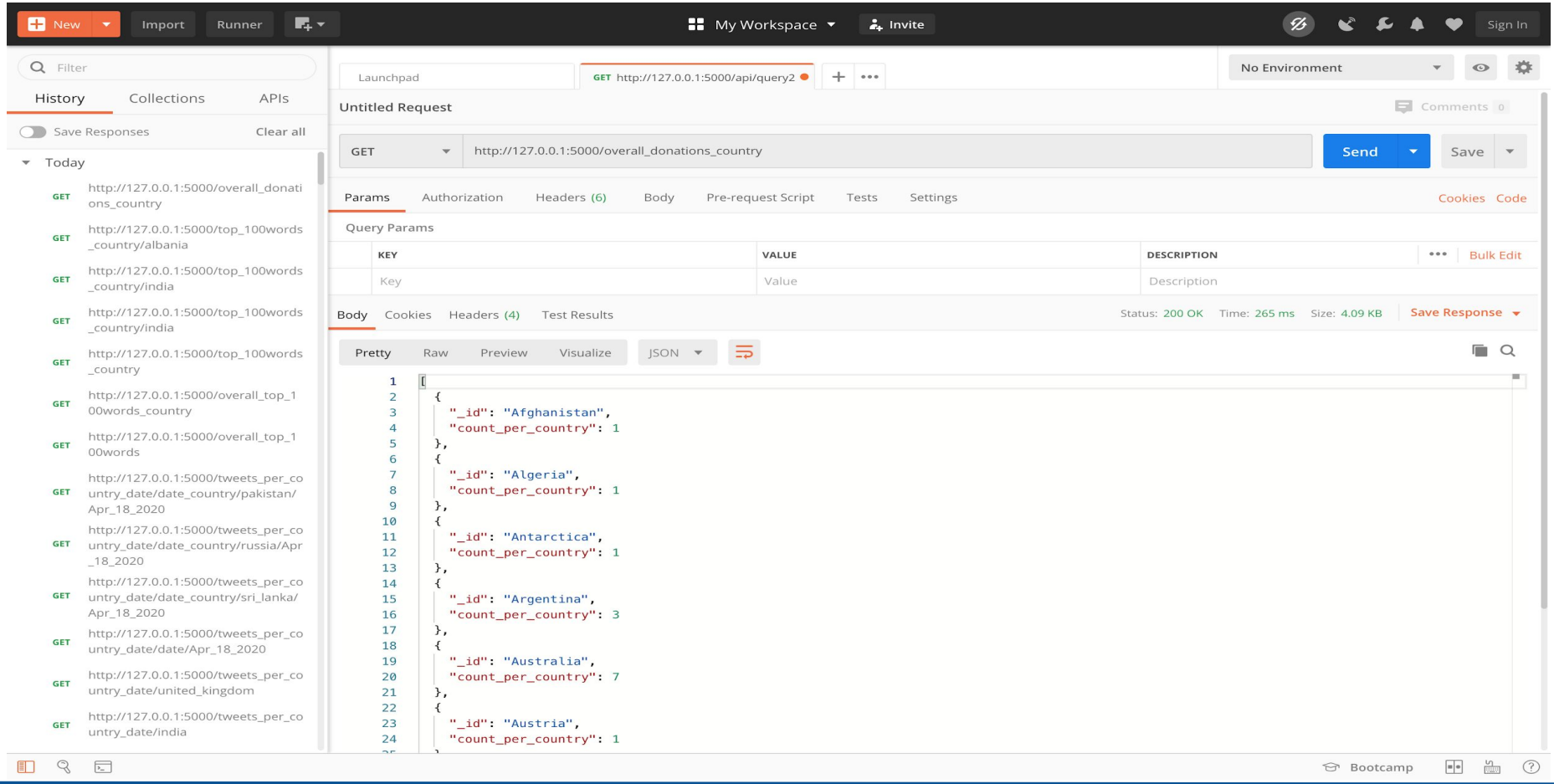
localhost:5000/top_100words_country: Finding top 100 words per country.



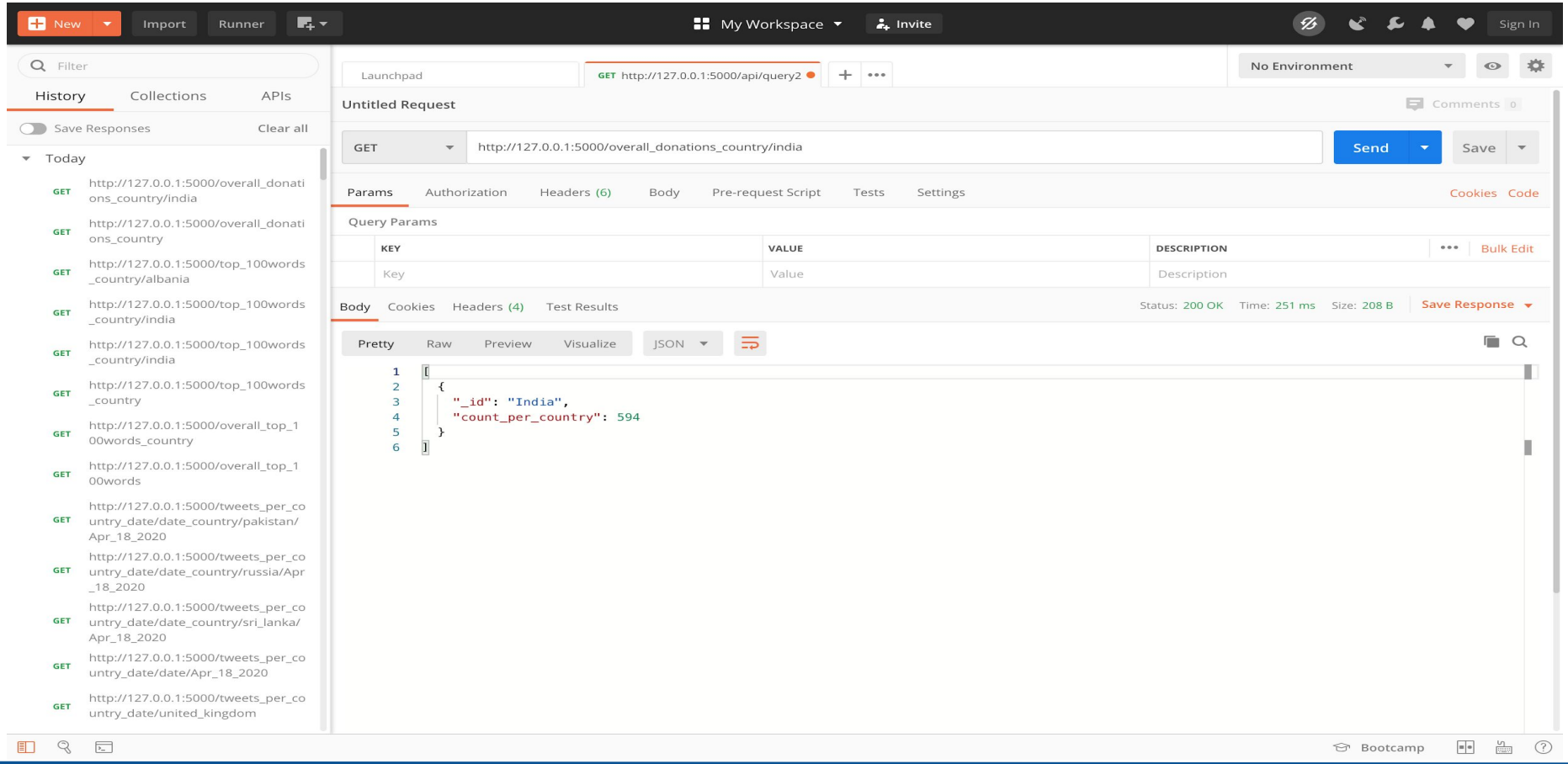
localhost:5000/top_100words_country/<country_name>: Finding top 100 words per country for a particular searched country.



localhost:5000/overall_donations_country: Finding total number of donations per country.



localhost:5000/overall_donations_country/<country_name>: Finding total number of donations per country for a particular searched country.



TEST CASES

- Twitter producer
  - Authentication of credentials.

- Spark Streaming
  - Testing the business logic involving formatting of tweets.

- Flask API
  - Verification of API routes.

FURTHER IMPROVEMENT

1. Communication Security: Network Protocols like Kerberos can be added to Kafka to provide strong authentication for client/server

2. Security of REST API: Configuration of APIs with SSL can help to keep sensitive information sent across the internet encrypted when the server is moved away from localhost.

3. Integration testing : Advanced mocking techniques like mockito can be used to mock the client connection.

THANK YOU !

APPENDIX

SLIDE 4: FLOW DIAGRAM

1. TWITTER CLIENT GETTING THE INFORMATION (TWEETS)
2. FETCHING TWITTER DATA
3. DISTRIBUTED SYNCHRONISATION SOFTWARE RUNNING IN THE BACKGROUND (ZOOKEEPER IN THIS CASE)
4. WRITING DATA TO A DISTRIBUTED STREAMING PLATFORM (KAFKA IN THIS CASE)
5. STREAM AND TRANSFORMATIONS (SPARK SQL IN THIS CASE)
6. STORING DATA IN THE DATABASE (MONGODB IN THIS CASE)
7. TO ACCESS THE DATABASE AND GET THE DESIRED OUTPUT USING API