Data validation between the data set in BigQuery and GraphDB

The objectives are to:

1. Ensure complete upload of data to BigQuery and GraphDB.
2. Check for data qualtiy issues.
3. Gain insights on the data.
4. Engineer features.

```
import pandas as pd
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```

```
# Mount to Google Drive to save results
from google.colab import drive
drive.mount('/content/drive')
%cd /content/drive/MyDrive/MSc/2020-21/Research\ Project/Colab/
%ls
```

```
Mounted at /content/drive
/content/drive/MyDrive/MSc/2020-21/Research Project/Colab
domain_count_df.csv  label_count_df.csv
```

```
# Connect to GCP Bucket
from google.colab import auth
auth.authenticate_user()
```

```
# Set GCP project ID and region to Europe West 2 - London
PROJECT = 'detect-fake-news-313201'
!gcloud config set project $PROJECT
REGION = 'europe-west2'
CLUSTER = '{}-cluster'.format(PROJECT)
!gcloud config set compute/region $REGION
!gcloud config set dataproc/region $REGION

!gcloud config list # show some information
```

```
Updated property [core/project].
Updated property [compute/region].
Updated property [dataproc/region].
[component_manager]
disable_update_check = True
[compute]
gce_metadata_read_timeout_sec = 0
region = europe-west2
[core]
account = aaron.altrock@gmail.com
project = detect-fake-news-313201
[dataproc]
region = europe-west2

Your active configuration is: [default]
```

Check the number of files in successive GCP cloud storage buckets

```
# Count the number of cleaned JSON files from the end of stage 1 in the pipel
!gsutil ls -l gs://fake_news_cleaned_json/*.json | wc -l
```

```
59733
```

```
# Count the number of parsed JSON and TTL files into triples at the end of st
!gsutil ls -l gs://fake_news_ttl_json/*.ttl | wc -l
!gsutil ls -l gs://fake_news_ttl_json/*.json | wc -l
```

```
27590
27590
```

The variance between the 59,733 cleaned files to 27,590 turtle documents would suggest this is due to the raw data containing duplicating records for the same news web page, when the turtles are indexed by the hash value of the URLs and therefore would overwrite leading to small number of samples.

```
# Based on https://cloud.google.com/bigquery/docs/quickstarts/quickstart-clie
from google.cloud import bigquery
client = bigquery.Client(PROJECT)

```

Profile the data in its original form held in BigQuery

```
# BIgQuery data row count
query_job = client.query(
    """
    SELECT COUNT(*) AS POPULATION_COUNT
    FROM `detect-fake-news-313201.fake_news_sql.src_fake_news`
    """
)

res_df = query_job.result().to_dataframe()  # Waits for job to complete.

res_df
```

	POPULATION_COUNT
0	27589

Therefore deviation by one record compared to the number of files in `gs://fake_news_ttl_json`.

```
# BIgQuery data row count
query_job = client.query(
    """
    WITH URL_LIST AS (
      SELECT
      URL
      , COUNT(*) AS URL_COUNT
      FROM `detect-fake-news-313201.fake_news_sql.src_fake_news`
      GROUP BY URL
    )
    SELECT * FROM URL_LIST WHERE URL_COUNT > 1
    """
)

res_df = query_job.result().to_dataframe()  # Waits for job to complete.

res_df
```

URL	URL_COUNT

Therefore no samples found to have duplicating URL in the BigQuery table, and all articles have unique URLs.

```
# BIgQuery data preview
query_job = client.query(
    """
    SELECT *
    FROM `detect-fake-news-313201.fake_news_sql.src_fake_news`
    LIMIT 10
    """
)

res_df = query_job.result().to_dataframe()  # Waits for job to complete.

res_df
```

	url	domain	domain_hash	
0	http://awm.com/woman-adopts-rescue-dog-starts-...	awm.com	domain_9725d0802cf38288710d8bff8f64dcba	title_761e

1	http://awm.com/grand-daughter-sings-favorite-s...	awm.com	domain_9725d0802cf38288710d8bff8f64dcba	title_2a30
2	http://awm.com/first-class-passenger-sees-a-so...	awm.com	domain_9725d0802cf38288710d8bff8f64dcba	title_b89b1
3	http://awm.com/model-who-boasts-she-transforme...	awm.com	domain_9725d0802cf38288710d8bff8f64dcba	title_986
4	http://awm.com/dad-puts-his-son-into-the-game-...	awm.com	domain_9725d0802cf38288710d8bff8f64dcba	title_08d
5	http://awm.com/little-girls-reaction-to-shooti...	awm.com	domain_9725d0802cf38288710d8bff8f64dcba	title_0ef2
6	http://awm.com/out-of-all-the-holiday-recipes-...	awm.com	domain_9725d0802cf38288710d8bff8f64dcba	title_0165
7	http://awm.com/a-starbucks-barista-couldnt-fin...	awm.com	domain_9725d0802cf38288710d8bff8f64dcba	title_55c1
8	http://awm.com/the-easily-offended-whine-when-...	awm.com	domain_9725d0802cf38288710d8bff8f64dcba	title_ab7f6

```
# BigQuery count by domain
query_job = client.query(
    """
    SELECT
    DOMAIN_HASH
    , LABEL
    COUNT(*) AS ARTICLES_COUNT
    FROM `detect-fake-news-313201.fake_news_sql.src_fake_news`
    GROUP BY DOMAIN_HASH, LABEL
    ORDER BY ARTICLES_COUNT DESC
    """
)

res_df = query_job.result().to_dataframe()

# Tally the domain hash to ensure each domain only has one label
domain_tally_ls = []
duplicate_domain_tally_ls = []
for i, row in res_df.iterrows():
  if row['DOMAIN_HASH'] in domain_tally_ls:
    duplicate_domain_tally_ls += [True]
  else:
    duplicate_domain_tally_ls += [False]

  # Add domain hash to the list of domains already reviwed
  domain_tally_ls += [row['DOMAIN_HASH']]

res_df['DOMAIN_HAS_MULTIPLE_LABEL'] = duplicate_domain_tally_ls

res_df
```

	DOMAIN_HASH	LABEL	ARTICLES_COUNT	DOMAIN_
0	domain_8f00b2b61ba335244231d632d390bf8d	fake	7046	
1	domain_9f5bbbcbad4a48edd86162fabbe90b6e	political	5441	
2	domain_7e058b16a00bca2c3620d5147881e34d	conspiracy	2819	
3	domain_98845529bda170844508657ae469197a	bias	1162	
4	domain_2bb13bea21c458421e33179105662bdd	political	1061	
5	domain_bd8db3098e8505e2d80c8b89a2eccdb8	clickbait	813	
6	domain_0d6ca907eef322628f81ae5a38735d44	bias	730	
7	domain_f573aed8d0dcf08ef6ce3efd5f5c48c0	political	729	
8	domain_213f2e7f6968dd534eb43ef113fce3e1	political	671	
9	domain_8fbb80fdc3a337fd46babecb01f203fa	junksci	615	
10	domain_799497bdca60cb2f1522d66166f15274	political	435	
11	domain_bfcde262dcc0a3fe874d75293764fb4c	political	415	
12	domain_69a49c3be87af6ae0c48f5c9b005a21b	None	413	
13	domain_f9f6f6f07772425d1522dd8a8ab25193	None	347	

```
14   domain_5dfa617b9d6dac8de770c521298c4e53     junksci   294
15   domain_3cbda5292ac5c0b572a8d764b0e52eec        bias   209
16   domain_fb4e14ee232e6a6acfda1de4f10a1fe3   political   209
17   domain_c29df259d72ab43dc7e2c43b1941a1c9   clickbait   178
18   domain_6a12243e8071c910a53039000309c901    reliable   168
19   domain_7cc6378eb1ac3bf4455eb536225df80b        bias   160
20   domain_f9e7e3cf456dda4da1f7ada4ba6ddc8c  conspiracy   153
21   domain_191e287683c85a13fa54c85c43281769  conspiracy   152
22   domain_3e2b4bc8d67cf01a6f156413acc03375   clickbait   144
23   domain_1682643b797f67fe89d09642a50174fe   political   120
24   domain_b55734eb636ef77462a42b666bc74c5f     unknown   114
25   domain_714bf72800dbdc780c8b895adc12ae3c  unreliable   108
26   domain_bfff154f86499943be9f3e65798b5d53   political   107
27   domain_d061e62c09e8c363752ddfa81c54b3ba        bias    99
28   domain_a66294ec7d2d0d8e3b22c293a88d14f4   political    97
29   domain_8776e5a68ce1577dd2809b859c3c6d82   political    90
30   domain_5014c2313218bc048eaef8f34771d2eb     unknown    84
31   domain_f03bcedc7a83b5568ff986bfec2c3413  conspiracy    77
32   domain_9725d0802cf38288710d8bff8f64dcba  unreliable    76
33   domain_fa702aa239d1e1c37cc9c5d75ab9b4f9   political    75
34   domain_f22938424edda66ec091ddd069854179  unreliable    74
35   domain_77b6aa19bdd7ac5f26e593472bd9a7d5        hate    71
36   domain_6afcd6f29c592f5495589242b12d1c25   political    67
37   domain_2c6ca1a760cdc4987f4283c9464dc639        None    63
38   domain_090b7aeff850c5dda848d5b9e75bb16d     unknown    58
39   domain_f047cbddff341a471b7b05bd5f4ee660       rumor    57
40   domain_daf12fe2bd3c219af940a4f8891f23f0        bias    57
41   domain_bc6e2fd29bb66b366afa86a3934609ce  conspiracy    54
42   domain_dbbb88dd5835637c6c4b2e3c54c2669e   political    50
```

43	domain_b04f082dca5817be6d561c1e25ea2c16	None	48
44	domain_b2f824ff829b40551a0f7e99a53745a0	satire	47
45	domain_e23a8807f4586e5ab98d2c0d335d998a	conspiracy	45
46	domain_9c34eb473390528f31c0303bda283c21	unreliable	44
47	domain_95e62162e78b589ec6362d5a6a19b3cb	bias	43
48	domain_b0c5e6bffbaa50cb2a380e69d80f0fde	bias	42
49	domain_c072fb2a857d0288f93a765a423629c1	conspiracy	42
50	domain_9a0d9e918e815c1424ffff912643042f	unreliable	41
51	domain_f3408542d52cb5a4c7f9d764ba80713b	bias	41
52	domain_e1e8ae12f228f5344ff99f14eb6926e8	None	33
53	domain_dc7dbd0a07a4d8dcf91c1d6de72611ea	unreliable	33
54	domain_1d3355bc85a66dfd13a09f3412c7f085	clickbait	33
55	domain_f446297823e3381ccbdcadd1033349f8	conspiracy	32
56	domain_ae885395fb95bb58ee77f8ae3925d64c	None	32
57	domain_1c74052b1eb7f1b234c3c19714e65a53	bias	31
58	domain_f9a3693da598b15b2c9235cb8c2c3e74	junksci	30
59	domain_2e59c440f7b5b4a04a68f652f5b5333b	political	28
60	domain_670f9a1790abd5f611073d167d0b8181	unknown	27
61	domain_d9b91044b82468a930dbba5a1be97efe	None	27
62	domain_e405e81e7af8517d01f31f4148fc71f8	bias	26
63	domain_e34675b4f28a906f84eeae9b500db072	bias	26
64	domain_0691d86f4f2449106e6b812635ae6dc2	conspiracy	24
65	domain_e0ed6141030ad9697efb6c0fbeaaf39a	bias	23
66	domain_7e6e1c6f7cf959d4b9bb48a6343ae80a	unknown	23
67	domain_902091121d014e622d75cbc7366f2ee6	unreliable	22
68	domain_bf1876a5bd135bede0c8d6dceee56287	bias	21
69	domain_ef1bc2898cde8803366bc549a5db7810	political	21
70	domain_80983712743801d65edf8f982efb226f	conspiracy	20
71	domain_77bb545e808a339883519ec87b37568b	conspiracy	20
72	domain_6dd68b0f98a01bc7b818f07015b03045	satire	19

73	domain_2ec5c58f61208a4c8c63d34ab78b0d84	junksci	17
74	domain_ebe3d7d9918e288972a59e1ca7fad839	political	17
75	domain_8e9372b5e0f099519daea57d17d06dc8	political	17
76	domain_725ab69c90f5c175022e9fff121b86b3	political	17
77	domain_8bcdc1ff44fe95e8fa1dbc98837e2d61	clickbait	16
78	domain_6516070744623d5b202c1aed9c8cd167	clickbait	16
79	domain_b3561e4c69fc2c3c42caff5d09c06797	clickbait	16
80	domain_9ec5ddf6e6b6f4394e9aa78d1e0b7df5	bias	16
81	domain_f6467a39846eb29840eeacd10a72fa18	bias	15
82	domain_cd6751faf82670aa1def1f49d631f5b5	political	15
83	domain_26f4e8623af4740617dbcd58ea5ea79e	political	15
84	domain_625aa2cfe2431342ba06d4b1091e0068	political	15
85	domain_ef1f79a3352040ff1e53c4792e0f82da	unknown	15
86	domain_b5562cb6e482ead8e7b41ad374075d6c	conspiracy	15
87	domain_9a9c6420af59e59d64f6d6efec7010a4	political	14
88	domain_ff30c35599b4b1c1cf2452fd9bc516ec	unknown	14
89	domain_be43d3d4c70a94cb184229a094d40ce6	junksci	13
90	domain_98329bfee9ee10e6193bc6e2f9d68326	bias	13
91	domain_333fbef8f20680b2f64c9bcdbf507eed	junksci	13
92	domain_25516cc33d26c336d56269cc3e470ef4	conspiracy	12
93	domain_0962e477ba5ef22cfa8f9f2a2bf1cf84	unreliable	11
94	domain_e3782a046734ce84f943b186a6b3b0d5	bias	11
95	domain_6ae10a38ca73549a2e745ac334073bbc	bias	10
96	domain_fedb6741f5466a1f2b94f3067260b45a	political	10
97	domain_283e8bea1b4bda6cfaef09fd756f583e	clickbait	10
98	domain_8d01354aadbf9b5bcbf69cb21e763e20	reliable	10
99	domain_8d36e2d1e8fd855649ffe49a1f6a6a64	satire	10
100	domain_9eccd3e363d514276c6d0ebd272f44d0	fake	9
101	domain_92398575509fbcdaf15f3300afa90f82	unreliable	9

102	domain_d229b60c4e7d85bb990a7c0114c56a59	bias	9
103	domain_bbdcc034a7e7286fa666aca650af364d	fake	8
104	domain_c2234ca891fbd2e5dd8b12f0f7ee9e57	satire	8
105	domain_4e2accee075ea7805bff7247a3b08308	satire	8
106	domain_1b114e669a32fd5523ab9a0169c5526f	conspiracy	7
107	domain_71a2bece9250fc73754674151d08ff0d	None	7
108	domain_39897a1f691f4ab9d29e150bf51ca4e3	clickbait	7
109	domain_d9aafe4c74d4dc82bcc371c7b1b906aa	unreliable	7
110	domain_5e2d02afb13db09442866e35b80beede	political	7
111	domain_81df9d37c2cc747e36887f85f8f12291	bias	6
112	domain_4640929e97eeb7aa066e21c89e25820b	conspiracy	6
113	domain_2b1b045d2f61d3b09d3d9de9ca54d67a	bias	6
114	domain_be827dfcbca489825ce5ed5091b87927	junksci	6
115	domain_d0a759558f0e28181573400f92ac533b	conspiracy	6
116	domain_a214f72bae25b351d8634b42efe7c961	political	6
117	domain_0abb7ad0f22beeddeaa66923c9f569b9	None	6
118	domain_9b94de25cdebeb77fe13e9c1bd4b9fb5	bias	6
119	domain_8371c7af10f4ca25e0f4fc8dbfa34086	fake	6
120	domain_aa4214b2bb89178bf26ccb85c29e87f6	unreliable	6
121	domain_20f4995b3b26bfb3fd3452488b508456	bias	5
122	domain_32941baccecb0e4772a4a00a9ddc2032	political	5
123	domain_9e8e2939ae6fc4ddf7d3c76dd7cf36c7	conspiracy	5
124	domain_228a30c466f1d9b6f15b783ca6d4d373	conspiracy	4
125	domain_0414276445efcdda835ae60c21f1257e	bias	4
126	domain_4b2772cf50d1497e8e4542710501bd5f	unknown	4
127	domain_71cde77da2e380141d923d8e192664bd	political	4
128	domain_c97825129fe35bfe256d3d3bfd74a68e	unknown	4
129	domain_fa20655a74bbf3d55e6093dc3f7b0b4b	unreliable	4
130	domain_35826b2bd227588ccf7f5d2d95cb8ead	fake	4
131	domain_a915d10a9cfcb1aaa463065abb64d0d1	clickbait	4

132	domain_09820ddecb343a96e4652ee94f3428ca	bias	4
133	domain_d2ab080f3f6c3ab5324c3171d904fb60	conspiracy	4
134	domain_7c80f485502f774206c4a955c73ec895	hate	3
135	domain_6dad3160ea412c366c718565e5ed50c7	unknown	3
136	domain_7439469439bb49107e23976d674fe054	hate	3
137	domain_2cd6c7e5b52562e349b54f8be0ad1eeb	None	3
138	domain_e7e9fcbd693e97a538800e8f8925833d	hate	3
139	domain_19727e7dfa4b40d50e6106d3cdf857fe	unknown	3
140	domain_f78d3e640568317fdc197922a83cc72d	bias	3
141	domain_7fe8b5ec02f3695cac07561cb3f06677	fake	3
142	domain_ed8a6546f82a5cd9f8663ea4bde6e638	bias	3
143	domain_ae5b067247d663e6a82ad6155a8950dd	conspiracy	3
144	domain_1eed9d343826bffb2d74bbd4ab92d725	fake	3
145	domain_5cb44e1c01a3e6659607671d7617c71c	unknown	2
146	domain_c07e15a699d0b7a3a8a85ceb082d03fe	political	2
147	domain_9a199d5872634a001effaf9e77ce0bbd	None	2
148	domain_b5f37665109c5d1055995893ffdedf83	conspiracy	2
149	domain_36de50180e84c35fe490bb60ccfa067e	bias	2
150	domain_b3af9c51233e0ee4e3bcb4eb429b56f3	political	2
151	domain_a2c5ba2a6834dcb8746a36f47e5fb9b9	unreliable	2
152	domain_66db98138a7da5d72c8becf3de0e4a31	conspiracy	2
153	domain_2e1c4e22c4feed37c72507c50d0232b1	bias	2
154	domain_069c5397b1af30b2ab38464644711aee	bias	2
155	domain_78868c04b77593ab9ae127dda5efd135	satire	2
156	domain_3e839bfdee6b0a2eb70e21686a474dfe	political	2
157	domain_00ed2a703fcb79060875666b27504053	political	2
158	domain_a5857e1ce2f61a5b22f70494134eaac0	political	2
159	domain_daf3443ff02927463d91818cfe02b50d	None	2
160	domain_185873d1f902490c0e61f488ea3a1a38	political	2
161	domain_71330e38a062718d01bf5cc906f58672	junksci	2

161	[illegible]	junksci	2
162	domain_d4d6e1ae6cdb4845995e37a870d29e02	satire	2
163	domain_4edff1b5750068279fb2e0230c3e5380	political	1
164	domain_2be28cabf750b41cfa89597f6e109e09	political	1
165	domain_75682947d32750e21851aa6bd371c829	conspiracy	1
166	domain_bbb95b551f642db5d9c3989002c1fa33	None	1
167	domain_ff5fe3728818a1b918e6d1a0221716ab	hate	1
168	domain_e60f2f982aeb75e173c158dd80ba5af2	clickbait	1
169	domain_3954d5639a2834fc2c3824caec12ebcd	conspiracy	1
170	domain_2b07b24e5cb6e6ee114747a0176d17a0	political	1
171	domain_8b500e62a848954dd1df3752a8da5919	political	1
172	domain_be6ccf3f58b38f629fa375fc5df60420	bias	1
173	domain_edd5909690ceb130b58fbf2abf882375	conspiracy	1
174	domain_341c80afad6e3b47016255d47078e0a6	conspiracy	1
175	domain_06d7d51d4c0251a3b42ebbfe7a26f5b4	bias	1
176	domain_9115aacafa05635ab12db50b331b2e91	bias	1
177	domain_d978280431610509cd3c1ffc63b5fad8	None	1
178	domain_32fb6e587463e29e5c17ef19a723f4f5	political	1
179	domain_2898ce92e0f40e6ae0fcf2172a57d962	bias	1
180	domain_dba255fee48039bec885525f34d93e3e	bias	1
181	domain_6ac3c748825f6ec75a1d590a4a0705f8	satire	1
182	domain_6d6083851f81d52d437a7c1c0bfebb0f	None	1
183	domain_315c9d54a7b67cd86a25cab716293e8b	None	1
184	domain_55c086c1f40ba2288dcb5bed04be39c2	conspiracy	1
185	domain_6a0258f088f96492673c324f167e5c55	junksci	1
186	domain_d223fe1505a39aab5565c357e417496e	conspiracy	1
187	domain_62f88633026e6e8f7cb86822defbc42c	fake	1
188	domain_5915307cbdc2d2909020ed5faa694dd7	unknown	1
189	domain_f73519c65f641722420e8f4cb2f4558b	satire	1
190	domain_d9fd2e6eb72f086ac4c4dda5f07b1af6	None	1

191	domain_43e820fdaa1f382bb0f2ce99866f09dd	bias	1
192	domain_8f7fb8558dbdcec0671e88504feccff4	None	1
193	domain_d2c1ed6780580239ade03edf73667f3f	unknown	1
194	domain_c888f140f0c2cd64fd7ba563b93e8097	conspiracy	1
195	domain_53cf7b3ea378b82f5dd0f3d7804f2d2b	satire	1
196	domain_7278328ff5bc42b1a29d9b51e9d3b946	fake	1
197	domain_9eb562d7fa2e4a32de9de63ca385db72	junksci	1
198	domain_1a482c97f790497c9809e63070e977dc	unknown	1
199	domain_ca14a1a10b6d95a384476925f6fd8898	junksci	1
200	domain_a715dbd3cfdbec1f4775714657dd3eb4	conspiracy	1
201	domain_5e30b0d401dc7f00af5f509e4aee8f22	None	1
202	domain_02cb0e25d8e2a8950c674f2d77e05712	political	1

```
# BIgQuery count by label
query_job = client.query(
    """
    SELECT
    LABEL
    , COUNT(*) AS LABEL_COUNT
    FROM `detect-fake-news-313201.fake_news_sql.src_fake_news`
    GROUP BY LABEL
    ORDER BY LABEL_COUNT DESC
    """
)

res_df = query_job.result().to_dataframe()

print('Total: {}'.format(res_df['LABEL_COUNT'].sum()))

res_df
```

```
Total: 27589
```

	LABEL	LABEL_COUNT
0	political	9776
1	fake	7081
2	conspiracy	3512
3	bias	2793
4	clickbait	1238
5	junksci	993
6	None	990
7	unreliable	437
8	unknown	354
9	reliable	178
10	satire	99
11	hate	81
12	rumor	57

```
# BIgQuery list of URLs
query_job = client.query(
    """
    SELECT DISTINCT
    URL_HASH
    FROM `detect-fake-news-313201.fake_news_sql.src_fake_news`
    """
)

res_df = query_job.result().to_dataframe()

print('Total: {}'.format(res_df.shape[0]))

res_df.head()
```

```
Total: 27589
```

	URL_HASH
0	00d76dc2a2b9c02526ab6aba99cea68d
1	022b655e21ca0d38a1fc33f2a37165a8
2	0405cdbffb7b0f0aef94eafaba0863ff
3	09f269f0d217228587029f2012f875c9
4	0c739d0432f31e7805c0d62f6eff0849

Noted that there were no classification for 990 samples, and further 354 with unknown classifications.

Profile the data in GraphDB

```
# Install the wrapper package
# Source: https://github.com/RDFLib/sparqlwrapper
!pip install sparqlwrapper
```

```
Collecting sparqlwrapper
  Downloading SPARQLWrapper-1.8.5-py3-none-any.whl (26 kB)
Collecting rdflib>=4.0
  Downloading rdflib-6.0.2-py3-none-any.whl (407 kB)
     |████████████████████████████████| 407 kB 6.5 MB/s
Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-
Requirement already satisfied: pyparsing in /usr/local/lib/python3.7/dist-p
Collecting isodate
  Downloading isodate-0.6.0-py2.py3-none-any.whl (45 kB)
     |████████████████████████████████| 45 kB 3.2 MB/s
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-package
Installing collected packages: isodate, rdflib, sparqlwrapper
Successfully installed isodate-0.6.0 rdflib-6.0.2 sparqlwrapper-1.8.5
```

```
# Code based on: https://sparqlwrapper.readthedocs.io/en/latest/main.html
from SPARQLWrapper import SPARQLWrapper, JSON

queryString = """
PREFIX aa: <http://www.city.ac.uk/ds/inm363/aaron_altrock#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select (count(?url_hash) as ?url_count) where {
  ?url_hash rdf:type aa:urlHash .
}
"""


sparql = SPARQLWrapper("http://35.246.120.165:7200/repositories/src_fake_news
sparql.setReturnFormat(JSON)
sparql.setQuery(queryString)

try :
   res_dct = sparql.query().convert()
   print('OK')

except Exception as e:
   print('ERROR: {}'.format(e))

```

```
OK
```

```
# No. of URL hash in GraphDB
res_dct.get('results').get('bindings')[0].get('url_count').get('value')
```

```
'27589'
```

Therefore noted that the number of news articles as URL hashes were completely uploaded when compared to BigQury count given both have the same number of articles `27598`.

```
res_dct
```

```
{'head': {'vars': ['url_count']},
 'results': {'bindings': [{'url_count': {'datatype': 'http://www.w3.org/200
     'type': 'literal',
     'value': '27589'}}]}}
```

No. of URL hashes

```
# Code based on: https://sparqlwrapper.readthedocs.io/en/latest/main.html

queryString = """
PREFIX aa: <http://www.city.ac.uk/ds/inm363/aaron_altrock#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?domain_hash ?label (count(?url_hash) as ?url_count) where {
  ?domain_hash rdf:type aa:domainHash .
  ?url_hash rdf:type aa:urlHash .
  ?label rdf:type aa:newsLabel .
  ?url_hash aa:has_domain_hash ?domain_hash .
  ?url_hash aa:has_news_label ?label .
}
GROUP BY ?domain_hash ?label
ORDER BY ?url_count
"""


sparql = SPARQLWrapper("http://35.246.120.165:7200/repositories/src_fake_news
sparql.setReturnFormat(JSON)
sparql.setQuery(queryString)

try :
   res_dct = sparql.query().convert()
   print('OK')

except Exception as e:
   print('ERROR: {}'.format(e))

```

```
OK
```

```
res_dct
```

```
      'value': '33'}},
    {'domain_hash': {'type': 'uri',
      'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#domain_dc7dbd0
     'label': {'type': 'uri',
      'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#unreliable'},
     'url_count': {'datatype': 'http://www.w3.org/2001/XMLSchema#integer',
      'type': 'literal',
      'value': '33'}},
    {'domain_hash': {'type': 'uri',
      'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#domain_e1e8ae1
     'label': {'type': 'uri',
      'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#'},
     'url_count': {'datatype': 'http://www.w3.org/2001/XMLSchema#integer',
      'type': 'literal',
          'value': '33'}},
        {'domain_hash': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#domain_9a0d9e9
         'label': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#unreliable'},
         'url_count': {'datatype': 'http://www.w3.org/2001/XMLSchema#integer',
          'type': 'literal',
          'value': '41'}},
        {'domain_hash': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#domain_f340854
         'label': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#bias'},
         'url_count': {'datatype': 'http://www.w3.org/2001/XMLSchema#integer',
          'type': 'literal',
          'value': '41'}},
        {'domain_hash': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#domain_b0c5e6b
         'label': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#bias'},
         'url_count': {'datatype': 'http://www.w3.org/2001/XMLSchema#integer',
          'type': 'literal',
          'value': '42'}},
        {'domain_hash': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#domain_c072fb2
         'label': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#conspiracy'},
         'url_count': {'datatype': 'http://www.w3.org/2001/XMLSchema#integer',
          'type': 'literal',
          'value': '42'}},
        {'domain_hash': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#domain_95e6216
         'label': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#bias'},
         'url_count': {'datatype': 'http://www.w3.org/2001/XMLSchema#integer',
          'type': 'literal',
          'value': '43'}},
        {'domain_hash': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#domain_9c34eb4
         'label': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#unreliable'},
         'url_count': {'datatype': 'http://www.w3.org/2001/XMLSchema#integer',
          'type': 'literal',
          'value': '44'}},
        {'domain_hash': {'type': 'uri',
          'value': 'http://www.city.ac.uk/ds/inm363/aaron_altrock#domain_e23a880
         'label': {'type': 'uri',
```

```
import re
res_ls = res_dct.get('results').get('bindings')

# Helper func to transform SPARQLWrapper query result output (dict) to Pandas
def parse_to_dataframe(res_ls):

  # If query result has content then parse to data frame else return None
  if len(res_ls) > 0:
    # Get column names
    col_nm_ls = list(res_ls[0].keys())

    parsed_res_ls = []

    for res_dct in res_ls:
      __res_ls = []
      for k, v in res_dct.items():
        __res_ls += [v.get('value')]

      # __res_ls = [res_dct]
      parsed_res_ls += [__res_ls]

    res_df = pd.DataFrame.from_dict(parsed_res_ls)
    res_df.columns = col_nm_ls
    res_df.reset_index(inplace=True, drop=True)

    return res_df

  else:
    return None

# Parse dict output from SPARQL to Pandas data frame
domain_url_count_df = parse_to_dataframe(res_ls)

# Remove name space prefix
domain_url_count_df['domain_hash'] = domain_url_count_df['domain_hash'].map(l
domain_url_count_df['label'] = domain_url_count_df['label'].map(lambda str: s

# Convert url_count to integer
domain_url_count_df['url_count'] = domain_url_count_df['url_count'].map(int)

# Count percentage
domain_url_count_df['url_count_pct'] = domain_url_count_df['url_count'].map(l

domain_url_count_df.sort_values(by=['url_count'], ascending=False, inplace=Tr

domain_url_count_df
```

	domain_hash	label	url_count	url_count_pc
202	domain_8f00b2b61ba335244231d632d390bf8d	fake	7046	25.53916
201	domain_9f5bbbcbad4a48edd86162fabbe90b6e	political	5441	19.72162
200	domain_7e058b16a00bca2c3620d5147881e34d	conspiracy	2819	10.21784
199	domain_98845529bda170844508657ae469197a	bias	1162	4.21182

[illegible]	[illegible]	bias	[illegible]	[illegible]
198	domain_2bb13bea21c458421e33179105662bdd	political	1061	3.84573
197	domain_bd8db3098e8505e2d80c8b89a2eccdb8	clickbait	813	2.94682
196	domain_0d6ca907eef322628f81ae5a38735d44	bias	730	2.64598
195	domain_f573aed8d0dcf08ef6ce3efd5f5c48c0	political	729	2.64235
194	domain_213f2e7f6968dd534eb43ef113fce3e1	political	671	2.43212
193	domain_8fbb80fdc3a337fd46babecb01f203fa	junksci	615	2.22914
192	domain_799497bdca60cb2f1522d66166f15274	political	435	1.57671
191	domain_bfcde262dcc0a3fe874d75293764fb4c	political	415	1.50422
190	domain_69a49c3be87af6ae0c48f5c9b005a21b		413	1.49697
189	domain_f9f6f6f07772425d1522dd8a8ab25193		347	1.25774
188	domain_5dfa617b9d6dac8de770c521298c4e53	junksci	294	1.06564
187	domain_fb4e14ee232e6a6acfda1de4f10a1fe3	political	209	0.75754
186	domain_3cbda5292ac5c0b572a8d764b0e52eec	bias	209	0.75754
185	domain_c29df259d72ab43dc7e2c43b1941a1c9	clickbait	178	0.64518
184	domain_6a12243e8071c910a53039000309c901	reliable	168	0.60893
183	domain_7cc6378eb1ac3bf4455eb536225df80b	bias	160	0.57994
182	domain_f9e7e3cf456dda4da1f7ada4ba6ddc8c	conspiracy	153	0.55456
181	domain_191e287683c85a13fa54c85c43281769	conspiracy	152	0.55094
180	domain_3e2b4bc8d67cf01a6f156413acc03375	clickbait	144	0.52194
179	domain_1682643b797f67fe89d09642a50174fe	political	120	0.43495
178	domain_b55734eb636ef77462a42b666bc74c5f	unknown	114	0.41320
177	domain_714bf72800dbdc780c8b895adc12ae3c	unreliable	108	0.39146
176	domain_bfff154f86499943be9f3e65798b5d53	political	107	0.38783
175	domain_d061e62c09e8c363752ddfa81c54b3ba	bias	99	0.35883
174	domain_a66294ec7d2d0d8e3b22c293a88d14f4	political	97	0.35158
173	domain_8776e5a68ce1577dd2809b859c3c6d82	political	90	0.32621
172	domain_5014c2313218bc048eaef8f34771d2eb	unknown	84	0.30446
171	domain_f03bcedc7a83b5568ff986bfec2c3413	conspiracy	77	0.27909
170	domain_9725d0802cf38288710d8bff8f64dcba	unreliable	76	0.27547

169	domain_fa702aa239d1e1c37cc9c5d75ab9b4f9	political	75	0.27184
168	domain_f22938424edda66ec091ddd069854179	unreliable	74	0.26822
167	domain_77b6aa19bdd7ac5f26e593472bd9a7d5	hate	71	0.25734
166	domain_6afcd6f29c592f5495589242b12d1c25	political	67	0.24285
165	domain_2c6ca1a760cdc4987f4283c9464dc639		63	0.22835
164	domain_090b7aeff850c5dda848d5b9e75bb16d	unknown	58	0.21022
163	domain_f047cbddff341a471b7b05bd5f4ee660	rumor	57	0.20660
162	domain_daf12fe2bd3c219af940a4f8891f23f0	bias	57	0.20660
161	domain_bc6e2fd29bb66b366afa86a3934609ce	conspiracy	54	0.19573
160	domain_dbbb88dd5835637c6c4b2e3c54c2669e	political	50	0.18123
159	domain_b04f082dca5817be6d561c1e25ea2c16		48	0.17398
158	domain_b2f824ff829b40551a0f7e99a53745a0	satire	47	0.17035
157	domain_e23a8807f4586e5ab98d2c0d335d998a	conspiracy	45	0.16310
156	domain_9c34eb473390528f31c0303bda283c21	unreliable	44	0.15948
155	domain_95e62162e78b589ec6362d5a6a19b3cb	bias	43	0.15585
154	domain_c072fb2a857d0288f93a765a423629c1	conspiracy	42	0.15223
153	domain_b0c5e6bffbaa50cb2a380e69d80f0fde	bias	42	0.15223
152	domain_f3408542d52cb5a4c7f9d764ba80713b	bias	41	0.14861
151	domain_9a0d9e918e815c1424ffff912643042f	unreliable	41	0.14861
150	domain_e1e8ae12f228f5344ff99f14eb6926e8		33	0.11961
149	domain_dc7dbd0a07a4d8dcf91c1d6de72611ea	unreliable	33	0.11961
148	domain_1d3355bc85a66dfd13a09f3412c7f085	clickbait	33	0.11961
147	domain_f446297823e3381ccbdcadd1033349f8	conspiracy	32	0.11598
146	domain_ae885395fb95bb58ee77f8ae3925d64c		32	0.11598
145	domain_1c74052b1eb7f1b234c3c19714e65a53	bias	31	0.11236
144	domain_f9a3693da598b15b2c9235cb8c2c3e74	junksci	30	0.10873
143	domain_2e59c440f7b5b4a04a68f652f5b5333b	political	28	0.10149
141	domain_670f9a1790abd5f611073d167d0b8181	unknown	27	0.09786
142	domain_d9b91044b82468a930dbba5a1be97efe		27	0.09786
140	domain_e405e81e7af8517d01f31f4148fc71f8	bias	26	0.09424

139	domain_e34675b4f28a906f84eeae9b500db072	bias	26	0.09424
138	domain_0691d86f4f2449106e6b812635ae6dc2	conspiracy	24	0.08699
137	domain_e0ed6141030ad9697efb6c0fbeaaf39a	bias	23	0.08336
136	domain_7e6e1c6f7cf959d4b9bb48a6343ae80a	unknown	23	0.08336
135	domain_902091121d014e622d75cbc7366f2ee6	unreliable	22	0.07974
134	domain_ef1bc2898cde8803366bc549a5db7810	political	21	0.07611
133	domain_bf1876a5bd135bede0c8d6dceee56287	bias	21	0.07611
131	domain_77bb545e808a339883519ec87b37568b	conspiracy	20	0.07249
132	domain_80983712743801d65edf8f982efb226f	conspiracy	20	0.07249
130	domain_6dd68b0f98a01bc7b818f07015b03045	satire	19	0.06886
126	domain_2ec5c58f61208a4c8c63d34ab78b0d84	junksci	17	0.06161
127	domain_725ab69c90f5c175022e9fff121b86b3	political	17	0.06161
128	domain_8e9372b5e0f099519daea57d17d06dc8	political	17	0.06161
129	domain_ebe3d7d9918e288972a59e1ca7fad839	political	17	0.06161
125	domain_b3561e4c69fc2c3c42caff5d09c06797	clickbait	16	0.05799
124	domain_9ec5ddf6e6b6f4394e9aa78d1e0b7df5	bias	16	0.05799
123	domain_8bcdc1ff44fe95e8fa1dbc98837e2d61	clickbait	16	0.05799
122	domain_6516070744623d5b202c1aed9c8cd167	clickbait	16	0.05799
119	domain_cd6751faf82670aa1def1f49d631f5b5	political	15	0.05436
116	domain_26f4e8623af4740617dbcd58ea5ea79e	political	15	0.05436
118	domain_b5562cb6e482ead8e7b41ad374075d6c	conspiracy	15	0.05436
117	domain_625aa2cfe2431342ba06d4b1091e0068	political	15	0.05436
120	domain_ef1f79a3352040ff1e53c4792e0f82da	unknown	15	0.05436
121	domain_f6467a39846eb29840eeacd10a72fa18	bias	15	0.05436
115	domain_ff30c35599b4b1c1cf2452fd9bc516ec	unknown	14	0.05074
114	domain_9a9c6420af59e59d64f6d6efec7010a4	political	14	0.05074
113	domain_be43d3d4c70a94cb184229a094d40ce6	junksci	13	0.04712
112	domain_98329bfee9ee10e6193bc6e2f9d68326	bias	13	0.04712
111	domain_333fbef8f20680b2f64c9bcdbf507eed	junksci	13	0.04712

110	domain_25516cc33d26c336d56269cc3e470ef4	conspiracy	12	0.04349
109	domain_e3782a046734ce84f943b186a6b3b0d5	bias	11	0.03987
108	domain_0962e477ba5ef22cfa8f9f2a2bf1cf84	unreliable	11	0.03987
105	domain_8d01354aadbf9b5bcbf69cb21e763e20	reliable	10	0.03624
104	domain_6ae10a38ca73549a2e745ac334073bbc	bias	10	0.03624
103	domain_283e8bea1b4bda6cfaef09fd756f583e	clickbait	10	0.03624
106	domain_8d36e2d1e8fd855649ffe49a1f6a6a64	satire	10	0.03624
107	domain_fedb6741f5466a1f2b94f3067260b45a	political	10	0.03624
102	domain_d229b60c4e7d85bb990a7c0114c56a59	bias	9	0.03262
101	domain_9eccd3e363d514276c6d0ebd272f44d0	fake	9	0.03262
100	domain_92398575509fbcdaf15f3300afa90f82	unreliable	9	0.03262
99	domain_c2234ca891fbd2e5dd8b12f0f7ee9e57	satire	8	0.02899
98	domain_bbdcc034a7e7286fa666aca650af364d	fake	8	0.02899
97	domain_4e2accee075ea7805bff7247a3b08308	satire	8	0.02899
96	domain_d9aafe4c74d4dc82bcc371c7b1b906aa	unreliable	7	0.02537
95	domain_71a2bece9250fc73754674151d08ff0d		7	0.02537
94	domain_5e2d02afb13db09442866e35b80beede	political	7	0.02537
93	domain_39897a1f691f4ab9d29e150bf51ca4e3	clickbait	7	0.02537
92	domain_1b114e669a32fd5523ab9a0169c5526f	conspiracy	7	0.02537
87	domain_9b94de25cdebeb77fe13e9c1bd4b9fb5	bias	6	0.02174
82	domain_0abb7ad0f22beeddeaa66923c9f569b9		6	0.02174
84	domain_4640929e97eeb7aa066e21c89e25820b	conspiracy	6	0.02174
85	domain_81df9d37c2cc747e36887f85f8f12291	bias	6	0.02174
86	domain_8371c7af10f4ca25e0f4fc8dbfa34086	fake	6	0.02174
83	domain_2b1b045d2f61d3b09d3d9de9ca54d67a	bias	6	0.02174
89	domain_aa4214b2bb89178bf26ccb85c29e87f6	unreliable	6	0.02174
90	domain_be827dfcbca489825ce5ed5091b87927	junksci	6	0.02174
91	domain_d0a759558f0e28181573400f92ac533b	conspiracy	6	0.02174
88	domain_a214f72bae25b351d8634b42efe7c961	political	6	0.02174
81	domain_9e8e2939ae6fc4ddf7d3c76dd7cf36c7	conspiracy	5	0.01812

80	domain_32941baccecb0e4772a4a00a9ddc2032	political	5	0.01812
79	domain_20f4995b3b26bfb3fd3452488b508456	bias	5	0.01812
73	domain_4b2772cf50d1497e8e4542710501bd5f	unknown	4	0.01449
69	domain_0414276445efcdda835ae60c21f1257e	bias	4	0.01449
70	domain_09820ddecb343a96e4652ee94f3428ca	bias	4	0.01449
71	domain_228a30c466f1d9b6f15b783ca6d4d373	conspiracy	4	0.01449
72	domain_35826b2bd227588ccf7f5d2d95cb8ead	fake	4	0.01449
75	domain_a915d10a9cfcb1aaa463065abb64d0d1	clickbait	4	0.01449
74	domain_71cde77da2e380141d923d8e192664bd	political	4	0.01449
76	domain_c97825129fe35bfe256d3d3bfd74a68e	unknown	4	0.01449
77	domain_d2ab080f3f6c3ab5324c3171d904fb60	conspiracy	4	0.01449
78	domain_fa20655a74bbf3d55e6093dc3f7b0b4b	unreliable	4	0.01449
63	domain_7c80f485502f774206c4a955c73ec895	hate	3	0.01087
58	domain_19727e7dfa4b40d50e6106d3cdf857fe	unknown	3	0.01087
59	domain_1eed9d343826bffb2d74bbd4ab92d725	fake	3	0.01087
60	domain_2cd6c7e5b52562e349b54f8be0ad1eeb		3	0.01087
62	domain_7439469439bb49107e23976d674fe054	hate	3	0.01087
61	domain_6dad3160ea412c366c718565e5ed50c7	unknown	3	0.01087
64	domain_7fe8b5ec02f3695cac07561cb3f06677	fake	3	0.01087
65	domain_ae5b067247d663e6a82ad6155a8950dd	conspiracy	3	0.01087
66	domain_e7e9fcbd693e97a538800e8f8925833d	hate	3	0.01087
67	domain_ed8a6546f82a5cd9f8663ea4bde6e638	bias	3	0.01087
68	domain_f78d3e640568317fdc197922a83cc72d	bias	3	0.01087
40	domain_00ed2a703fcb79060875666b27504053	political	2	0.00724
41	domain_069c5397b1af30b2ab38464644711aee	bias	2	0.00724
42	domain_185873d1f902490c0e61f488ea3a1a38	political	2	0.00724
43	domain_2e1c4e22c4feed37c72507c50d0232b1	bias	2	0.00724
44	domain_36de50180e84c35fe490bb60ccfa067e	bias	2	0.00724
45	domain_3e839bfdee6b0a2eb70e21686a474dfe	political	2	0.00724
46	domain_5cb44e1c01a3e6659607671d7617e71e	unknown	2	0.00724

46	domain_5cb44e1c01a3e6659607671d7617c71c	unknown	2	0.00724
48	domain_71330e38a062718d01bf5cc906f58672	junksci	2	0.00724
47	domain_66db98138a7da5d72c8becf3de0e4a31	conspiracy	2	0.00724
51	domain_a2c5ba2a6834dcb8746a36f47e5fb9b9	unreliable	2	0.00724
49	domain_78868c04b77593ab9ae127dda5efd135	satire	2	0.00724
52	domain_a5857e1ce2f61a5b22f70494134eaac0	political	2	0.00724
53	domain_b3af9c51233e0ee4e3bcb4eb429b56f3	political	2	0.00724
54	domain_b5f37665109c5d1055995893ffdedf83	conspiracy	2	0.00724
55	domain_c07e15a699d0b7a3a8a85ceb082d03fe	political	2	0.00724
56	domain_d4d6e1ae6cdb4845995e37a870d29e02	satire	2	0.00724
57	domain_daf3443ff02927463d91818cfe02b50d		2	0.00724
50	domain_9a199d5872634a001effaf9e77ce0bbd		2	0.00724
14	domain_5915307cbdc2d2909020ed5faa694dd7	unknown	1	0.00362
12	domain_53cf7b3ea378b82f5dd0f3d7804f2d2b	satire	1	0.00362
13	domain_55c086c1f40ba2288dcb5bed04be39c2	conspiracy	1	0.00362
18	domain_6ac3c748825f6ec75a1d590a4a0705f8	satire	1	0.00362
15	domain_5e30b0d401dc7f00af5f509e4aee8f22		1	0.00362
16	domain_62f88633026e6e8f7cb86822defbc42c	fake	1	0.00362
17	domain_6a0258f088f96492673c324f167e5c55	junksci	1	0.00362
10	domain_43e820fdaa1f382bb0f2ce99866f09dd	bias	1	0.00362
11	domain_4edff1b5750068279fb2e0230c3e5380	political	1	0.00362
6	domain_315c9d54a7b67cd86a25cab716293e8b		1	0.00362
9	domain_3954d5639a2834fc2c3824caec12ebcd	conspiracy	1	0.00362
8	domain_341c80afad6e3b47016255d47078e0a6	conspiracy	1	0.00362
7	domain_32fb6e587463e29e5c17ef19a723f4f5	political	1	0.00362
5	domain_2be28cabf750b41cfa89597f6e109e09	political	1	0.00362
4	domain_2b07b24e5cb6e6ee114747a0176d17a0	political	1	0.00362
3	domain_2898ce92e0f40e6ae0fcf2172a57d962	bias	1	0.00362
2	domain_1a482c97f790497c9809e63070e977dc	unknown	1	0.00362
20	domain_7278328ff5bc42b1a29d9b51e9d3b946	fake	1	0.00362

19	domain_6d6083851f81d52d437a7c1c0bfebb0f		1	0.00362
1	domain_06d7d51d4c0251a3b42ebbfe7a26f5b4	bias	1	0.00362
21	domain_75682947d32750e21851aa6bd371c829	conspiracy	1	0.00362
22	domain_8b500e62a848954dd1df3752a8da5919	political	1	0.00362
39	domain_ff5fe3728818a1b918e6d1a0221716ab	hate	1	0.00362
38	domain_f73519c65f641722420e8f4cb2f4558b	satire	1	0.00362
37	domain_edd5909690ceb130b58fbf2abf882375	conspiracy	1	0.00362
36	domain_e60f2f982aeb75e173c158dd80ba5af2	clickbait	1	0.00362
35	domain_dba255fee48039bec885525f34d93e3e	bias	1	0.00362
34	domain_d9fd2e6eb72f086ac4c4dda5f07b1af6		1	0.00362
33	domain_d978280431610509cd3c1ffc63b5fad8		1	0.00362
32	domain_d2c1ed6780580239ade03edf73667f3f	unknown	1	0.00362
31	domain_d223fe1505a39aab5565c357e417496e	conspiracy	1	0.00362
30	domain_ca14a1a10b6d95a384476925f6fd8898	junksci	1	0.00362
29	domain_c888f140f0c2cd64fd7ba563b93e8097	conspiracy	1	0.00362
28	domain_be6ccf3f58b38f629fa375fc5df60420	bias	1	0.00362
27	domain_bbb95b551f642db5d9c3989002c1fa33		1	0.00362
26	domain_a715dbd3cfdbec1f4775714657dd3eb4	conspiracy	1	0.00362
25	domain_9eb562d7fa2e4a32de9de63ca385db72	junksci	1	0.00362
24	domain_9115aacafa05635ab12db50b331b2e91	bias	1	0.00362
23	domain_8f7fb8558dbdcec0671e88504feccff4		1	0.00362
0	domain_02cb0e25d8e2a8950c674f2d77e05712	political	1	0.00362

```
# Summary of the classification by count of articles
label_count_df = domain_url_count_df[['label', 'url_count']].groupby('label')
total_rec_count = sum(label_count_df['url_count'].to_list())
print('Total number of articles: {}'.format(total_rec_count))
label_count_df['url_count_pct'] = label_count_df['url_count'].map(lambda val:
label_count_df
```

```
Total number of articles: 27589
```

	label	url_count	url_count_pct
0	political	9776	35.434412
1	fake	7081	25.666026
2	conspiracy	3512	12.729711
3	bias	2793	10.123600
4	clickbait	1238	4.487296
5	junksci	993	3.599261
6		990	3.588387
7	unreliable	437	1.583965
8	unknown	354	1.283120
9	reliable	178	0.645185
10	satire	99	0.358839
11	hate	81	0.293595
12	rumor	57	0.206604

Per the https://github.com/several27/FakeNewsCorpus details on each news article classification noted no such classifications as `unknown` nor a classification of `None`. These may be added by the author of the data set who performed scrapping to backfill classifications that could not be found.

```
# Save summary to CSV
label_count_df.to_csv(r'label_count_df.csv', index=False)
```

No. of domains by classification

```
print('No. of domains in total: {}'.format(domain_url_count_df['domain_hash']
domain_count_df = domain_url_count_df[['domain_hash', 'label']].groupby(by='l
domain_count_df.rename(columns={'domain_hash': 'domain_count'}, inplace=True)
domain_count_df['domain_count_pct'] = domain_count_df['domain_count'].map(lam
domain_count_df.to_csv('domain_count_df.csv', index=False)
domain_count_df
```

```
No. of domains in total: 203
```

	label	domain_count	domain_count_pct
0	political	40	19.704433
1	bias	37	18.226601
2	conspiracy	30	14.778325
3		19	9.359606
4	unknown	15	7.389163
5	unreliable	13	6.403941
6	clickbait	11	5.418719
7	junksci	11	5.418719
8	satire	10	4.926108
9	fake	9	4.433498
10	hate	5	2.463054
11	reliable	2	0.985222
12	rumor	1	0.492611

Repeated publication of the same news articles by URL, domain and classifications

```
# Code based on: https://sparqlwrapper.readthedocs.io/en/latest/main.html

queryString = """
PREFIX aa: <http://www.city.ac.uk/ds/inm363/aaron_altrock#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?domain_hash ?url_hash ?body_hash ?label where {
  ?domain_hash rdf:type aa:domainHash .
  ?url_hash rdf:type aa:urlHash .
  ?body_hash rdf:type aa:bodyHash .
  ?label rdf:type aa:newsLabel .
  ?url_hash aa:has_domain_hash ?domain_hash .
  ?url_hash aa:has_news_label ?label .
  ?url_hash aa:has_body_hash ?body_hash .
}
ORDER BY ?url_count
"""


sparql = SPARQLWrapper("http://35.246.120.165:7200/repositories/src_fake_news
sparql.setReturnFormat(JSON)
sparql.setQuery(queryString)

try :
   res_dct = sparql.query().convert()
   print('OK')

except Exception as e:
   print('ERROR: {}'.format(e))


# Parse dict output from SPARQL to Pandas data frame
res_ls = res_dct.get('results').get('bindings')
domain_url_body_df = parse_to_dataframe(res_ls)

# Remove name space prefix
domain_url_body_df['domain_hash'] = domain_url_body_df['domain_hash'].map(lam
domain_url_body_df['url_hash'] = domain_url_body_df['url_hash'].map(lambda st
domain_url_body_df['body_hash'] = domain_url_body_df['body_hash'].map(lambda 
domain_url_body_df['label'] = domain_url_body_df['label'].map(lambda str: str

domain_url_body_df.head()
```

```
OK
```

	domain_hash	url_hash	
0	domain_8f00b2b61ba335244231d632d390bf8d	e681402ef05d57310574314a7918aa6c	bo
1	domain_8f00b2b61ba335244231d632d390bf8d	3e3b0220abefd45f00916f2b69d13051	body_
2	domain_8f00b2b61ba335244231d632d390bf8d	7c6de65487ed9706801ee482e5f4343f	body
3	domain_3e2b4bc8d67cf01a6f156413acc03375	4a883a995ada072079326ffb6d55f992	body
4	domain_9c34eb473390528f31c0303bda283c21	a4d57e74eaf6511fdcb675e1c7b55a15	boc

```
# Summarise the number of URLs with the same text body in news articles
reuse_content_url_df = domain_url_body_df[['body_hash', 'label', 'url_hash']]
reuse_content_url_df.to_csv('reuse_content_url_df.csv', index=False)
print('No. of distinct text corpora: {}'.format(reuse_content_df.shape[0]))
reuse_content_url_df.head()
```

```
No. of distinct text corpora: 50
```

	body_hash	label	url_hash
15548	body_a91766cb7db645f57aa52fca4b5a77e0	bias	410
12169	body_845af446af521c88c801695145743815	junksci	325
11210	body_7a03935701b11bf99ae50445a0e67793	fake	82
16030	body_ae1d6ee9a8d8345cd74e6983178a7a45	political	52
20048	body_d9ad78c5666f8bcb7c2b954427cdaa2b	political	43

```
print('No. of text bodies re-used: {}'.format(reuse_content_url_df[reuse_cont
```

```
No. of text bodies re-used: 50
```

Therefore noted all articles were re-used up to 410 times.

```
# Summarise the number of domains with the same text body in news articles
reuse_content_domain_df = domain_url_body_df[['body_hash', 'label', 'domain_h
reuse_content_domain_df.to_csv('reuse_content_domain_df.csv', index=False)
print('No. of distinct text corpora: {}'.format(reuse_content_df.shape[0]))
reuse_content_domain_df.head()
```

```
No. of distinct text corpora: 50
```

	body_hash	domain_ha
13088	body_8dda333b27153fe5e09a6edc4e243192	domain_ff5fe3728818a1b918e6d1a022171
6638	body_4838596bacba67022cc422a5c7533e80	domain_ff30c35599b4b1c1cf2452fd9bc51
6429	body_45d0af00418769088d34b1bdd22793ad	domain_ff30c35599b4b1c1cf2452fd9bc51
4184	body_2c5471fb74e36f68f3f518e9cf28cb91	domain_ff30c35599b4b1c1cf2452fd9bc51
15629	body_aa111236c97009e280bf260d504e8af5	domain_ff30c35599b4b1c1cf2452fd9bc51

```
# Find text corpora referenced by more than one domains
reuse_content_domain_df[reuse_content_domain_df['label'] > 1]
```

body_hash	domain_hash	label

Noted therefore whilst there were repeated republication of the same news text corpora as the bodies they were confined to within the same web domains.

```
# BigQuery count by domain
query_job = client.query(
    """
    SELECT
    *
    FROM `detect-fake-news-313201.fake_news_sql.src_fake_news`
    WHERE BODY_HASH = 'body_7a03935701b11bf99ae50445a0e67793'
    """
)

res_df = query_job.result().to_dataframe()
```

```
res_df.head()
```

	url	domain	
0	http://beforeitsnews.com/earthquakes/2017/10/m...	beforeitsnews.com	domain_8f00b2b61ba3
1	http://beforeitsnews.com/earthquakes/2018/01/m...	beforeitsnews.com	domain_8f00b2b61ba3
2	http://beforeitsnews.com/earthquakes/2017/12/m...	beforeitsnews.com	domain_8f00b2b61ba3
3	http://beforeitsnews.com/earthquakes/2018/01/m...	beforeitsnews.com	domain_8f00b2b61ba3
4	http://beforeitsnews.com/earthquakes/2018/01/v...	beforeitsnews.com	domain_8f00b2b61ba3

```
res_df['body'].iloc[0]
```

```
'I felt the shaking * Now Today Earlier\n\nCountry where you felt the eart
hquake * Afghanistan aland Islands Albania Algeria American Samoa Andorra
Angola Anguilla Antarctica Antigua and Barbuda Argentina Armenia Aruba Aus
tralia Austria Azerbaijan Bahamas Bahrain Bangladesh Barbados Belarus Belg
ium Belize Benin Bermuda Bhutan Bolivia Bosnia and Herzegovina Botswana Bo
uvet Island Brazil British Antarctic Territory British Indian Ocean Territ
ory British Virgin Islands Brunei Bulgaria Burkina Faso Burundi Cambodia C
ameroon Canada Canton and Enderbury Islands Cape Verde Cayman Islands Cent
ral African Republic Chad Chile China Christmas Island Cocos [Keeling] Isl
ands Colombia Comoros Congo - Brazzaville Congo - Kinshasa Cook Islands Co
sta Rica Côte d’Ivoire Croatia Cuba Cyprus Czech Republic Denmark Djibouti
Dominica Dominican Republic Dronning Maud Land East Germany Ecuador Egypt
El Salvador Equatorial Guinea Eritrea Estonia Ethiopia Falkland Islands Fa
roe Islands Fiji Finland France French Guiana French Polynesia French Sout
hern and Antarctic Territories French Southern Territories Gabon Gambia Ge
orgia Germany Ghana Gibraltar Greece Greenland Grenada Guadeloupe Guam Gua
temala Guernsey Guinea Guinea-Bissau Guyana Haiti Heard Island and McDonal
d Islands Honduras Hong Kong SAR China Hungary Iceland India Indonesia Ira
n Iraq Ireland Isle of Man Israel Italy Jamaica Japan Jersey Johnston Isla
nd Jordan Kazakhstan Kenya Kiribati Kuwait Kyrgyzstan Laos Latvia Lebanon
Lesotho Liberia Libya Liechtenstein Lithuania Luxembourg Macau SAR China M
acedonia Madagascar Malawi Malaysia Maldives Mali Malta Marshall Islands M
artinique Mauritania Mauritius Mayotte Metropolitan France Mexico Micrones
ia Midway Islands Moldova Monaco Mongolia Montenegro Montserrat Morocco Mo
zambique Myanmar [Burma] Namibia Nauru Nepal Netherlands Netherlands Antil
les Neutral Zone New Caledonia New Zealand Nicaragua Niger Nigeria Niue No
rfolk Island North Korea North Vietnam Northern Mariana Islands Norway Oma
n Pacific Islands Trust Territory Pakistan Palau Palestinian Territories P
anama Panama Canal Zone Papua New Guinea Paraguay People's Democratic Repu
blic of Yemen Peru Philippines Pitcairn Islands Poland Portugal Puerto Ric
o Qatar Réunion Romania Russia Rwanda Saint Barthélemy Saint Helena Saint
Kitts and Nevis Saint Lucia Saint Martin Saint Pierre and Miquelon Saint V
incent and the Grenadines Samoa San Marino São Tomé and Príncipe Saudi Ara
bia Senegal Serbia Serbia and Montenegro Seychelles Sierra Leone Singapore
Slovakia Slovenia Solomon Islands Somalia South Africa South Georgia and t
he South Sandwich Islands South Korea Spain Sri Lanka Sudan Suriname Svalb
ard and Jan Mayen Swaziland Sweden Switzerland Syria Taiwan Tajikistan Tan
zania Thailand Timor-Leste Togo Tokelau Tonga Trinidad and Tobago Tunisia
Turkey Turkmenistan Turks and Caicos Islands Tuvalu U.S. Minor Outlying Is
lands U.S. Miscellaneous Pacific Islands U.S. Virgin Islands Uganda Ukrain
e Union of Soviet Socialist Republics United Arab Emirates United Kingdom
United States Unknown or Invalid Region Uruguay Uzbekistan Vanuatu Vatican
City Venezuela Vietnam Wake Island Wallis and Futuna Western Sahara Yemen
Zambia Zimbabwe\n\nCity/Village where you felt the earthquake *\n\nStreet
or suburb (area) where you felt the earthquake\n\nLatitude (area) where yo
u felt the earthquake\n\nLongitude (area) where you felt the earthquake\n\
```

● ✕