Project Summary

Objectives

1. To be able to pass an unlimited number of ‘OR’ logic operators in the query
2. To be able to make complex queries involving nested ‘OR’ and ‘AND’ logic operators
3. To retrieve the tone of each particular article to help us generate our own tone chart and tonetimeline from a group of curated and personalised news articles
4. To be able to retrieve articles highly relevant to a search keyword or phrase

GDELT provides a number of different datasets made available for analysis.
In the end, the following datasets from GDELT were combined to achieve the objectives listed above:

1. webngrams: (https://blog.gdeltproject.org/announcing-the-new-web-news-ngrams-3-0-dataset/)
2. geg: (https://blog.gdeltproject.org/announcing-the-global-entity-graph-geg-and-a-new-11-billion-entity-dataset/)
3. gal: (https://blog.gdeltproject.org/announcing-the-gdelt-article-list-rss-feed/)
4. gsg: (https://blog.gdeltproject.org/announcing-the-global-similarity-graph/)

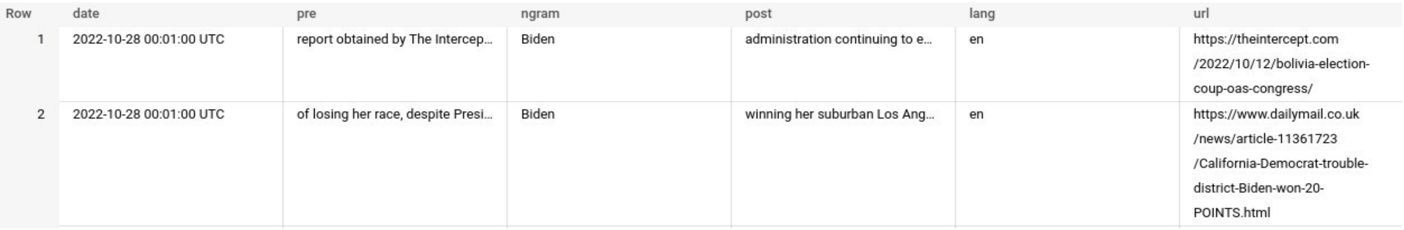
A description of how each dataset contributes to producing curated highly relevant articles is discussed below.

(a) **Webngrams : `gdelt-bq.gdeltv2.webngrams`**

Using a number of fields from the ngrams 3.0 dataset, we can retrieve the list of articles for a particular ngram
Assuming we want to extract news of Biden, we set the “ngram” column to “Biden”

```
SELECT `date`, `pre`,`ngram`, `post`, `lang`, `url` FROM `gdelt-bq.gdeltv2.webngrams`
WHERE `ngram`="Biden" AND `lang`= 'en' AND `date`BETWEEN TIMESTAMP("2022-10-28 00:00:00 UTC") AND TIMESTAMP("2022-10-28 23:59:59 UTC")
ORDER BY `date` ASC
```

The results are shown below



Row	date	pre	ngram	post	lang	url
1	2022-10-28 00:01:00 UTC	report obtained by The Intercep...	Biden	administration continuing to e...	en	https://theintercept.com /2022/10/12/bolivia-election- coup-oas-congress/
2	2022-10-28 00:01:00 UTC	of losing her race, despite Presi...	Biden	winning her suburban Los Ang...	en	https://www.dailymail.co.uk /news/article-11361723 /California-Democrat-trouble- district-Biden-won-20- POINTS.html

However, from the results you can notice that the values in the “pre” and “post” columns are the text snippets in the article which appeared before and after the “ngram” word which is “Biden”.

Now according to the ngram docs, if we want to search for a bigram, eg “Joe Biden”. The “pre” column is assigned “Joe” and “ngram” column is assigned “Biden”. However if we were to run this in big query we would get no results because if we specify that the “pre” column should be “Joe”, then it means in the article the only text snippet which appears before the “ngram” ie. Biden, should be only “Joe”

Therefore the the following query below produces no results

```
SELECT `date`, `pre`,`ngram`, `post`, `lang`, `url` FROM `gdelt-bq.gdeltv2.webngrams`
WHERE `pre`="Joe" AND `ngram`="Biden" AND `lang`= 'en' AND `date`BETWEEN
TIMESTAMP("2022-10-28 00:00:00 UTC") AND TIMESTAMP("2022-10-28 23:59:59
UTC")
ORDER BY `date` ASC
```

According to this link (https://blog.gdeltproject.org/custom-entity-extraction-over-the-news-using-web-ngrams-3-0/) we can concatenate the pre, ngram and post fields together to run our own entity extraction.

But if we are extracting the entities, then why not use the Global Entity Graph then?

(b) **Gobal Entity Graph : `gdelt-bq.gdeltv2.geg_gcnlapi`**

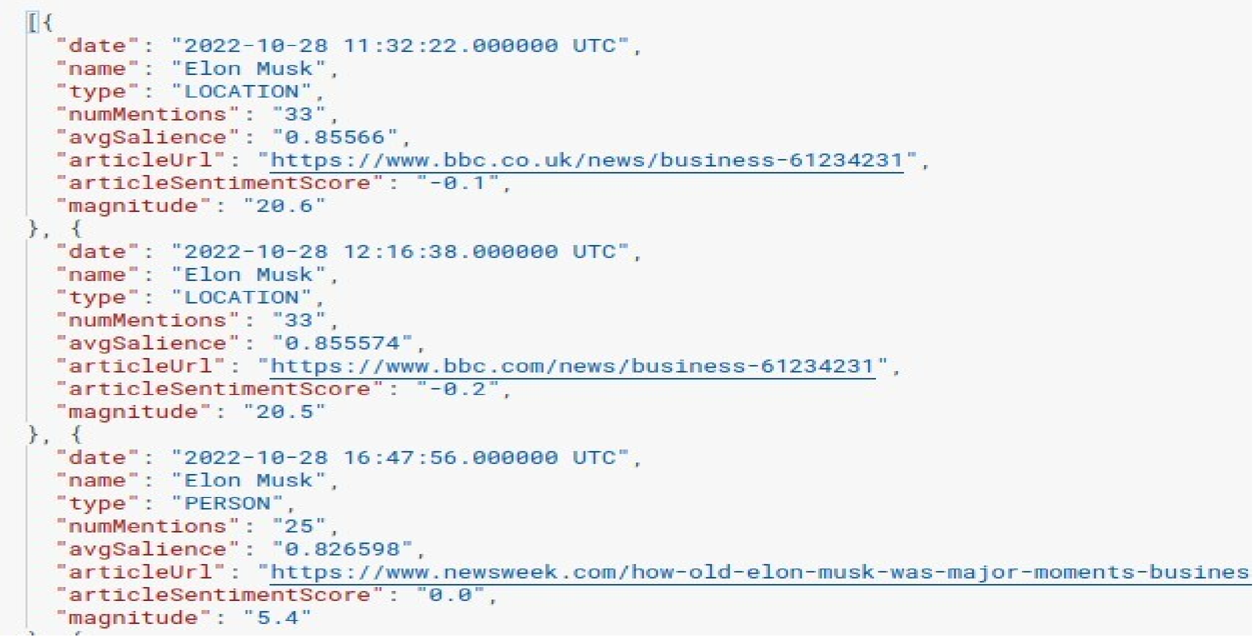
This dataset allows us to filter articles based on a particular ‘entity’. An entity can be a person eg. “Donald Trump”, organization eg. “republican party”, event eg. “replublican national convention”, location eg. “cleavland”.

The importance of this dataset is that it gives us the `salience score` which gives information on the importance or centrality of that entity to the entire document text. Using

this information we can filter out and eliminate articles which were returned only because they were mentioned once or twice in the article and really had nothing to do with the article’s main story.
Also, the dataset gives us the sentiment score of the article which can be used to describe the tone of the article.

The following command below gives us news articles on “Elon Musk” together with the article’s sentiment score and salience

```
SELECT `date`, entities.name, entities.type, entities.numMentions, entities.avgSalience,
`url` AS articleUrl, score AS articleSentimentScore, magnitude FROM `gdelt-bq.gdeltv2.geg_gcnlapi`,
UNNEST(entities) entities
WHERE LOWER(entities.name) LIKE LOWER("%elon musk%") AND DATE BETWEEN
"2022-10-28 00:00:00" AND "2022-10-28 23:59:59" ORDER BY entities.avgSalience
DESC LIMIT 10000
```



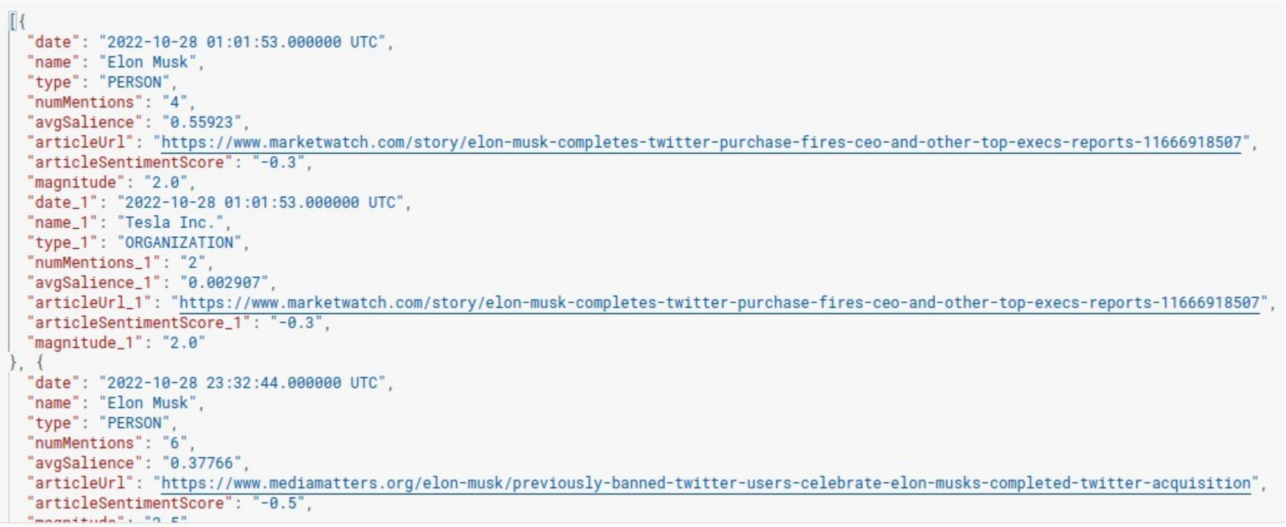
```
[{
  "date": "2022-10-28 11:32:22.000000 UTC",
  "name": "Elon Musk",
  "type": "LOCATION",
  "numMentions": "33",
  "avgSalience": "0.85566",
  "articleUrl": "https://www.bbc.co.uk/news/business-61234231",
  "articleSentimentScore": "-0.1",
  "magnitude": "20.6"
}, {
  "date": "2022-10-28 12:16:38.000000 UTC",
  "name": "Elon Musk",
  "type": "LOCATION",
  "numMentions": "33",
  "avgSalience": "0.855574",
  "articleUrl": "https://www.bbc.com/news/business-61234231",
  "articleSentimentScore": "-0.2",
  "magnitude": "20.5"
}, {
  "date": "2022-10-28 16:47:56.000000 UTC",
  "name": "Elon Musk",
  "type": "PERSON",
  "numMentions": "25",
  "avgSalience": "0.826598",
  "articleUrl": "https://www.newsweek.com/how-old-elon-musk-was-major-moments-busines
  "articleSentimentScore": "0.0",
  "magnitude": "5.4"
```

Example 2: Assuming we want to retrieve articles on Elon Musk and tesla. We would want to get the most relevant articles to ensure that we dont get for instance an article in our results which was mainly centered on elon musk and twitter but was retrieved because it had a mention of tesla.

We perform two queries to see how the salience impacts our articles

Without specifying the salience for `tesla`

```
SELECT *
FROM
(
SELECT `date`, entities.name, entities.type, entities.numMentions, entities.avgSalience,
`url` as articleUrl, score as articleSentimentScore, magnitude FROM `gdelt-bq.gdeltv2.geg_gcnlapi`,
unnest(entities) entities
where lower(entities.name) like lower("%elon musk%") and entities.avgSalience>0.3 and
date between "2022-10-28 00:00:00" AND "2022-10-28 23:59:59"
) q1
INNER JOIN
(
SELECT `date`, entities.name, entities.type, entities.numMentions, entities.avgSalience,
`url` as articleUrl, score as articleSentimentScore, magnitude FROM `gdelt-bq.gdeltv2.geg_gcnlapi`,
unnest(entities) entities
where lower(entities.name) like lower("%tesla%") and date between "2022-10-28
00:00:00" AND "2022-10-28 23:59:59"
) q2
ON
q1.articleUrl = q2.articleUrl
```



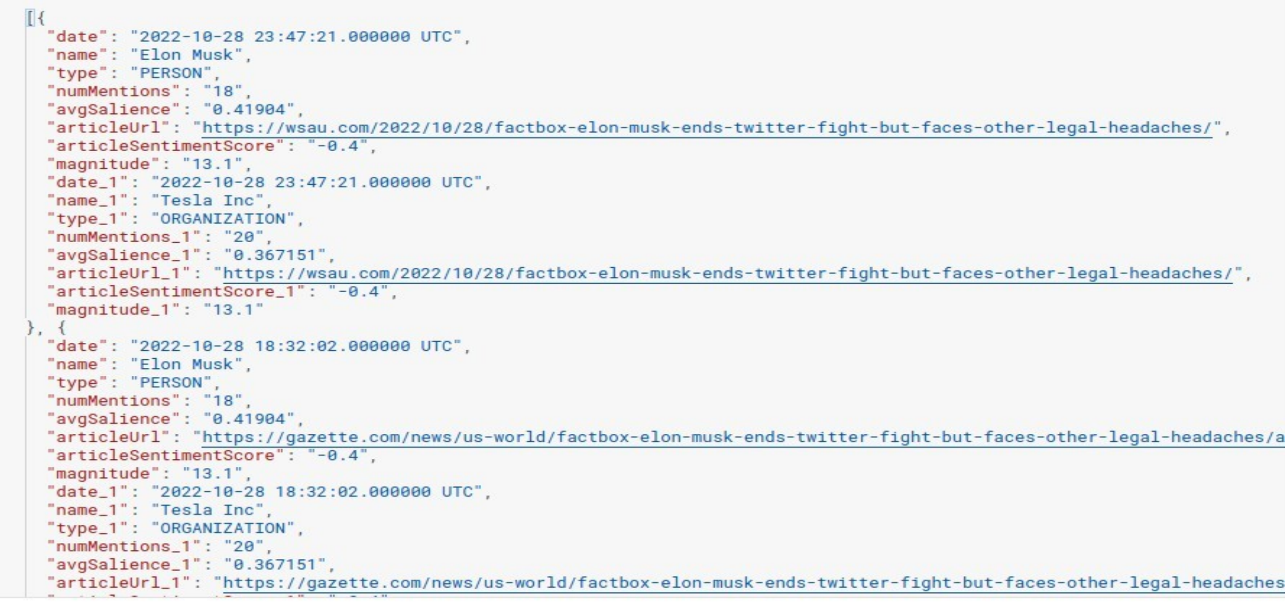
```
[{
  "date": "2022-10-28 01:01:53.000000 UTC",
  "name": "Elon Musk",
  "type": "PERSON",
  "numMentions": "4",
  "avgSalience": "0.55923",
  "articleUrl": "https://www.marketwatch.com/story/elon-musk-completes-twitter-purchase-fires-ceo-and-other-top-execs-reports-11666918507",
  "articleSentimentScore": "-0.3",
  "magnitude": "2.0",
  "date_1": "2022-10-28 01:01:53.000000 UTC",
  "name_1": "Tesla Inc.",
  "type_1": "ORGANIZATION",
  "numMentions_1": "2",
  "avgSalience_1": "0.002907",
  "articleUrl_1": "https://www.marketwatch.com/story/elon-musk-completes-twitter-purchase-fires-ceo-and-other-top-execs-reports-11666918507",
  "articleSentimentScore_1": "-0.3",
  "magnitude_1": "2.0"
}, {
  "date": "2022-10-28 23:32:44.000000 UTC",
  "name": "Elon Musk",
  "type": "PERSON",
  "numMentions": "6",
  "avgSalience": "0.37766",
  "articleUrl": "https://www.mediamatters.org/elon-musk/previously-banned-twitter-users-celebrate-elon-musks-completed-twitter-acquisition",
  "articleSentimentScore": "-0.5",
  "magnitude": "2.5"
```

The first url in the results for instance is mainly centred on twitter however it was returned because in the article, Musk was described as the “Tesla CEO”

To retrieve articles relevant to our query, we specify the salience for each keyword

Specifying the salience for `tesla`

```
SELECT *  FROM
(
SELECT `date`, entities.name, entities.type, entities.numMentions, entities.avgSalience, `url` as articleUrl, score as articleSentimentScore, magnitude FROM `gdelt-bq.gdeltv2.geg_gcnlapi`,
unnest(entities) entities
where lower(entities.name) like lower("%elon musk%") and entities.avgSalience>0.3 and date between "2022-10-28 00:00:00" AND "2022-10-28 23:59:59"
) q1
INNER JOIN
(
SELECT `date`, entities.name, entities.type, entities.numMentions, entities.avgSalience, `url` as articleUrl, score as articleSentimentScore, magnitude FROM `gdelt-bq.gdeltv2.geg_gcnlapi`,
unnest(entities) entities
where lower(entities.name) like lower("%tesla%") and entities.avgSalience>0.3 and date between "2022-10-28 00:00:00" AND "2022-10-28 23:59:59"
) q2
ON
q1.articleUrl = q2.articleUrl
```



```
[{
  "date": "2022-10-28 23:47:21.000000 UTC",
  "name": "Elon Musk",
  "type": "PERSON",
  "numMentions": "18",
  "avgSalience": "0.41904",
  "articleUrl": "https://wsau.com/2022/10/28/factbox-elon-musk-ends-twitter-fight-but-faces-other-legal-headaches/",
  "articleSentimentScore": "-0.4",
  "magnitude": "13.1",
  "date_1": "2022-10-28 23:47:21.000000 UTC",
  "name_1": "Tesla Inc",
  "type_1": "ORGANIZATION",
  "numMentions_1": "20",
  "avgSalience_1": "0.367151",
  "articleUrl_1": "https://wsau.com/2022/10/28/factbox-elon-musk-ends-twitter-fight-but-faces-other-legal-headaches/",
  "articleSentimentScore_1": "-0.4",
  "magnitude_1": "13.1"
}, {
  "date": "2022-10-28 18:32:02.000000 UTC",
  "name": "Elon Musk",
  "type": "PERSON",
  "numMentions": "18",
  "avgSalience": "0.41904",
  "articleUrl": "https://gazette.com/news/us-world/factbox-elon-musk-ends-twitter-fight-but-faces-other-legal-headaches/a
  "articleSentimentScore": "-0.4",
  "magnitude": "13.1",
  "date_1": "2022-10-28 18:32:02.000000 UTC",
  "name_1": "Tesla Inc",
  "type_1": "ORGANIZATION",
  "numMentions_1": "20",
  "avgSalience_1": "0.367151",
  "articleUrl_1": "https://gazette.com/news/us-world/factbox-elon-musk-ends-twitter-fight-but-faces-other-legal-headaches
```

This time, the query returns a small number of rows however, each of these articles is centred mainly on Elon Musk and Tesla since we set a minimum value for the salience score of Tesla.
With this we are able to retrieve highly relevant articles as well as the sentiment/tone for each article which tackles Objective 3 and Objective 4

Using sql syntax in big query also gives us the freedom to combine OR and AND logical operators in complex queries and longer lengths than provided by the GDELT summary api.

Assuming we want to search for news on “Elon musk” (tesla or twitter) OR Biden (Russia OR “energy
crisis”).
The query when used in the GDELT summary api returns an invalid query however GDELT’s database allows us to query this statement through big query

```
SELECT `ngram`, `url` from `gdelt-bq.gdeltv2.webngrams` where `ngram` = "elon musk" and `date` between TIMESTAMP("2022-10-28 00:00:00 UTC") and TIMESTAMP("2022-10-28 00:15:00 UTC") and `url` in (
SELECT `url` from `gdelt-bq.gdeltv2.webngrams` where `ngram` = "tesla" or `ngram`= "twitter" and `date` between TIMESTAMP("2022-10-28 00:00:00 UTC") and TIMESTAMP("2022-10-28 00:15:00 UTC")
)
UNION ALL
SELECT `ngram`, `url` from `gdelt-bq.gdeltv2.webngrams` where `ngram` = "Biden" and `date` between TIMESTAMP("2022-10-28 00:00:00 UTC") and TIMESTAMP("2022-10-28 00:15:00 UTC") and `url` in (
SELECT `url` from `gdelt-bq.gdeltv2.webngrams` where `ngram` = "Russia" or `ngram`= "energy crisis" and `date` between TIMESTAMP("2022-10-28 00:00:00 UTC") and TIMESTAMP("2022-10-28 00:15:00 UTC")
)
```

Though the syntax length is limited in big query, it is enough to cater for almost all the queries you might want to make. Hence it is adequate enough to tackle Objective 1 and Objective 2

More on the queries and limitations can be found here (https://cloud.google.com/bigquery/quotas)

Miscellaneous

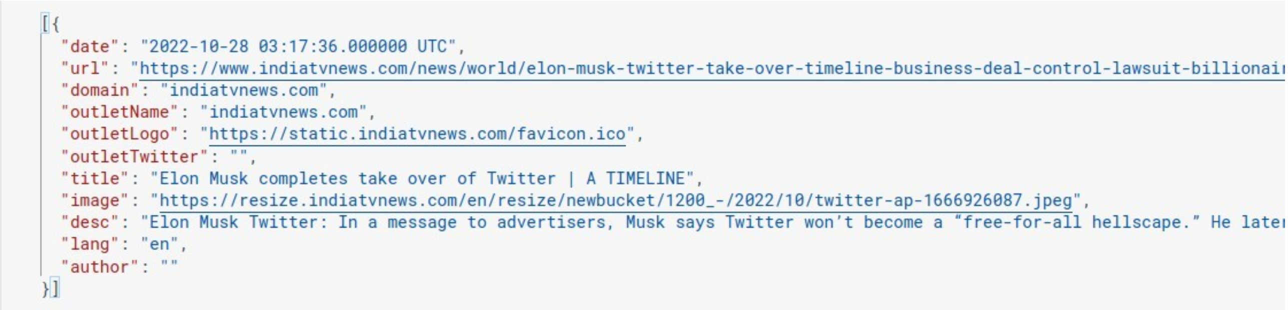
According to the docs on webngrams, we can use the ngrams 3.0 dataset (webngrams) together with the gdelt article list (gal), global entity graph (geg) and global similarity graph (gsg) datasets to filter the most relevant news. The information below talks about how the gal and gsg also help to improve the searches.

Gdelt article list : `gdelt-bq.gdeltv2.gal`

This dataset is based on the gemg dataset and gives us a standardized and minimized basic set of metadata for every article that GDELT monitors which includes the article’s image, full outlet name, outlet logo, one-sentence summary description and author of the news article. So lets assume we searched for “Elon Musk” in the webngrams dataset or entity dataset, we can take any record for that results and then for the article url of that record, we can look it up in the `gal` dataset to retrieve its metadata which contains more info on that article.

This query below uses the `gal` dataset to fetch more info about a given article url on “Elon Musk”

```
SELECT * from `gdelt-bq.gdeltv2.gal`
where `url`="https://www.indiatvnews.com/news/world/elon-musk-twitter-take-over-timeline-business-deal-control-lawsuit-billionaire-tesla-ceo-social-media-platform-owner-latest-updates-2022-10-28-819380" and `date` >= TIMESTAMP("2022-10-28 00:00:00 UTC") and `date` <= TIMESTAMP("2022-10-28 23:59:59 UTC")
order by `date` asc
limit 10
```



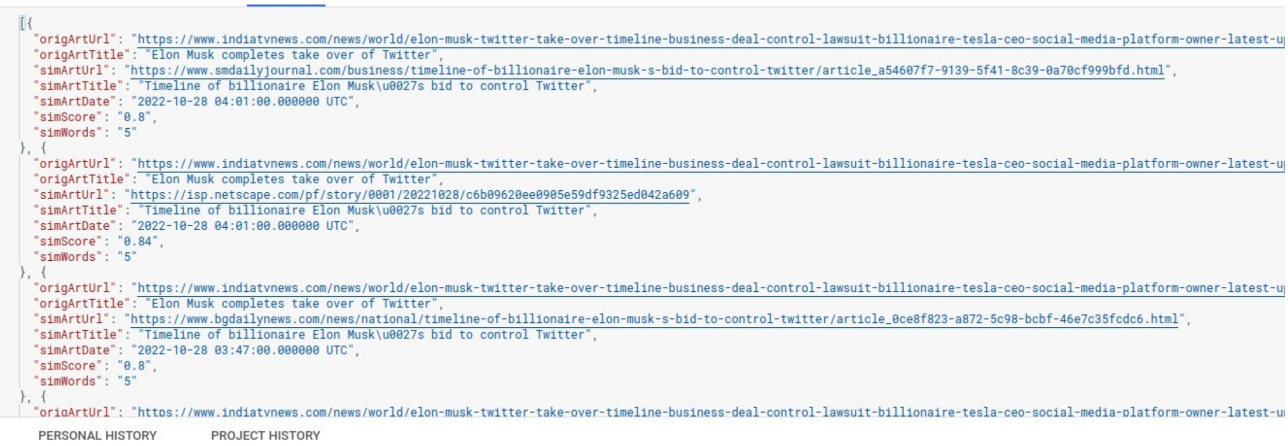
```
[{
  "date": "2022-10-28 03:17:36.000000 UTC",
  "url": "https://www.indiatvnews.com/news/world/elon-musk-twitter-take-over-timeline-business-deal-control-lawsuit-billionai
  "domain": "indiatvnews.com",
  "outletName": "indiatvnews.com",
  "outletLogo": "https://static.indiatvnews.com/favicon.ico",
  "outletTwitter": "",
  "title": "Elon Musk completes take over of Twitter | A TIMELINE",
  "image": "https://resize.indiatvnews.com/en/resize/newbucket/1200_-/2022/10/twitter-ap-1666926087.jpeg",
  "desc": "Elon Musk Twitter: In a message to advertisers, Musk says Twitter won’t become a “free-for-all hellscape.” He late
  "lang": "en",
  "author": ""
}]
```

Global similarity graph: `gdelt-bq.gdeltv2.gsg`

This dataset returns similar articles based on a given article url. The fields of this dataset can be adjusted to return more preferred articles. Eg. using the similarity score field

We find similar articles from the url in the previous command.

```
SELECT `fromUrl` as origArtUrl, `fromTitle` as origArtTitle, `toUrl` as simArtUrl, `toTitle` as
simArtTitle,`toDate` as simArtDate, `simScore`,`simWords` from `gdelt-bq.gdeltv2.gsg`
where `fromUrl`="https://www.indiatvnews.com/news/world/elon-musk-twitter-take-over-timeline-business-deal-control-lawsuit-billionaire-tesla-ceo-social-media-platform-owner-latest-updates-2022-10-28-819380" and `fromDate` >= TIMESTAMP("2022-10-28
00:00:00 UTC") and `fromDate` <= TIMESTAMP("2022-10-28 23:59:59 UTC") and
`simScore` >0.5
order by `fromDate` asc
limit 10
```



```
[{
  "origArtUrl": "https://www.indiatvnews.com/news/world/elon-musk-twitter-take-over-timeline-business-deal-control-lawsuit-billionaire-tesla-ceo-social-media-platform-owner-latest-u
  "origArtTitle": "Elon Musk completes take over of Twitter",
  "simArtUrl": "https://www.smdailyjournal.com/business/timeline-of-billionaire-elon-musk-s-bid-to-control-twitter/article_a54607f7-9139-5f41-8c39-0a70cf999bfd.html",
  "simArtTitle": "Timeline of billionaire Elon Musk\u0027s bid to control Twitter",
  "simArtDate": "2022-10-28 04:01:00.000000 UTC",
  "simScore": "0.8",
  "simWords": "5"
}, {
  "origArtUrl": "https://www.indiatvnews.com/news/world/elon-musk-twitter-take-over-timeline-business-deal-control-lawsuit-billionaire-tesla-ceo-social-media-platform-owner-latest-u
  "origArtTitle": "Elon Musk completes take over of Twitter",
  "simArtUrl": "https://isp.netscape.com/pf/story/0001/20221028/c6b09620ee0905e59df9325ed042a609",
  "simArtTitle": "Timeline of billionaire Elon Musk\u0027s bid to control Twitter",
  "simArtDate": "2022-10-28 04:01:00.000000 UTC",
  "simScore": "0.84",
  "simWords": "5"
}, {
  "origArtUrl": "https://www.indiatvnews.com/news/world/elon-musk-twitter-take-over-timeline-business-deal-control-lawsuit-billionaire-tesla-ceo-social-media-platform-owner-latest-u
  "origArtTitle": "Elon Musk completes take over of Twitter",
  "simArtUrl": "https://www.bgdailynews.com/news/national/timeline-of-billionaire-elon-musk-s-bid-to-control-twitter/article_0ce8f823-a872-5c98-bcbf-46e7c35fcdc6.html",
  "simArtTitle": "Timeline of billionaire Elon Musk\u0027s bid to control Twitter",
  "simArtDate": "2022-10-28 03:47:00.000000 UTC",
  "simScore": "0.8",
  "simWords": "5"
}, {
  "origArtUrl": "https://www.indiatvnews.com/news/world/elon-musk-twitter-take-over-timeline-business-deal-control-lawsuit-billionaire-tesla-ceo-social-media-platform-owner-latest-u
```

Limitations of the proposed solution above

1. The sql queries used above consume a large quota of data and cost. Solution to increasing query performance and reducing cost on big query can be found here. (https://cloud.google.com/bigquery/docs/best-practices-performance-communication)