Joel Duran, Èric Dalmases and Marc Aguilar

Report | Project Part 2

The project has been published on GitHub at the following URL and is available within this specific tag:

URL: https://github.com/ericdalmases/IRWA.git
Tag: IRWA-2023-part-2
(https://github.com/ericdalmases/IRWA/releases/tag/IRWA-2023-part-2)

Since this deliverable is a continuation of the previous one, some of the code of the first deliverable has been added in this delivery.

Indexing

Implementation of inverted index and TF-IDF

The primary objective of this project phase was to establish an indexing system to enhance the efficiency and effectiveness of information retrieval. Indexing plays a crucial role in improving the precision and ranking of queries, significantly impacting their effectiveness.

This project specifically requires the utilization of an inverted index, particularly beneficial when handling extensive document collections, such as the large volume of tweets in our case. The task involved implementing the *create_index_tfidf* function, provided to us, and customizing the code to align with the preprocessing phase. Creating the inverted index with our document collection of tweets took around 235 seconds.

We've incorporated functions, such as *search_tf_idf*, from Lab 2. This function aids in retrieving documents containing any term from the query. Additionally, we've utilized *rank_documents*, also derived from Lab 2, which evaluates the relevance of each document using TF-IDF, considering the query parameters.

Query processing and ground truth selection

To facilitate the posterior analysis of our queries, we decided to create two subclasses. The first one is TweetSimilarityPair, which just contains a tweet and a similarity score, a float. The second one is the class Query, which contains the query in text, in tokens, an id and a couple of sets which will store the relevant and non-relevant tweets for the query. Instead of containing the Tweet itself it has instances of TweetSimilairtyPair, since it will be useful to then compute all the metrics in the evaluation phase.

Then, we started introducing our own queries to then set the ground truth for each of them. We decided to use ["presidents visiting Kyiv", "conflict in Ukraine", "gas in Ukraine", "killing innocents", "donations for Ukraine"], which in the Annex you can find the ranking of the 10 best searches of each of the query.

In order to set the ground truth for each of them we considered building a custom algorithm that automatically assigned relevant and non-relevant Tweets to each of the queries. During the process we considered the two main constraints needed for our evaluation phase, that was that no tweet could be repeated among the 5 queries. So, all the relevant and non-relevant tweets for each query and between them must be unique.

Our algorithm basically uses the tokenized query and tokenized tweet to create a set for each of them. Then, it computes the intersection between those two sets. We perform this for each query and for each tweet. Then, we start by adding the top 10 tweets, which are the ones with larger intersection length, to our relevant tweets for the query. The same for the other 4 queries but checking that the Tweet is not relevant to any other query.

Up to this point we have already our 10 most relevant tweets for each of the queries. To select the non-relevant, we randomly select tweets for each of the queries and then we manually check that they were in fact not relevant. It worked as expected due to the large amount of possible tweets.

Evaluation

Upon completion of the indexing and ranking process, the necessity arose to verify the accuracy of the algorithms used and the relevance of the results for the search's intended purpose. To do so we implemented the class Evaluation which basically has a function for each metric which facilitates the process of evaluating the results.

The different measures we are asked to use:

- **Precision@K** measures the fraction of retrieved documents which is relevant to the searched query.
- **Recall@K** measures the fraction of the relevant documents which have been retrieved.
- **Average Precision@K** measures how much relevant documents are concentrated in the highest ranked predictions.
- **F1-Score@K** combines precision and recall measures to evaluate the relevance of the top-K documents in the ranked list for a given query.
- **Mean Average Precision** is the mean of all queries average precision.
- **Mean Reciprocal Rank** measures how well the system ranks the first relevant result.
- **Normalized Discounted Cumulative Gain** measures how well the ranked list of results corresponds to a list of items that are graded in relevance.

In order to conduct the evaluation for a single query we created a table (dataframe) which contained the following columns: Query_id which is used to identify the query, score which is the cosine similarity between the tweet and the query and label which is the binary ground truth of the query.

As required in the project statement, we evaluated each query with the 10 relevant and non-relevant tweets to the query, and the other non-relevant tweets of the other queries. Therefore, in the case of our queries we evaluated each query with 60 tweets: the 10 relevant and 10 non-relevant to the specific query, and the other 40 non-relevant from the other 4 queries (10 per query).

We decided to use K = 10 because we would be returning our 10 relevant labeled tweets, so it makes sense to evaluate over the 10 with higher score.

Evaluation of *evaluation* queries

‘Tank in Kharkiv’

Precision: 0.7

Recall: 0.7

Average Precision: 0.583

F1 Score: 0.7

Mean Reciprocal Rank: 1.0

Normalized Discounted Cumulative Gain: 0.7503

‘Nord Stream pipeline’

Precision: 1.0

Recall: 1.0

Average Precision: 1.0

F1 Score: 1.0

Mean Reciprocal Rank: 1.0

Normalized Discounted Cumulative Gain: 1.0

‘Annexation of territories by Russia’

Precision: 0.8

Recall: 0.8

Average Precision: 0.6484

F1 Score: 0.8

Mean Reciprocal Rank: 1.0

Normalized Discounted Cumulative Gain: 0.7917

MEAN AVERAGE PRECISION: 0.74394

Evaluation of *our* queries

As mentioned above, we defined the following queries:

A. ‘Presidents visiting Kyiv’
B. ‘Conflict in Ukraine’
C. ‘Gas in Ukraine’
D. ‘Killing innocents’
E. ‘Donations for Ukraine’

For each of these queries, we got the following results with k=10:

‘Presidents visiting Kyiv’

Precision: 1.0

Recall: 1.0

Average Precision: 1.0

F1 Score: 1.0

Mean Reciprocal Rank: 1.0

Normalized Discounted Cumulative Gain: 1.0

The evaluation results for this query are exceptional, with all metrics indicating a perfect performance in retrieving and ranking relevant documents.

‘Conflict in Ukraine’

Precision: 1.0

Recall: 1.0

Average Precision: 1.0

F1 Score: 1.0

Mean Reciprocal Rank: 1.0

Normalized Discounted Cumulative Gain: 1.0

The evaluation results for this query are exceptional, with all metrics indicating a perfect performance in retrieving and ranking relevant documents.

‘Gas in Ukraine’

Precision: 1.0

Recall: 1.0

Average Precision: 1.0

F1 Score: 1.0

Mean Reciprocal Rank: 1.0

Normalized Discounted Cumulative Gain: 1.0

The evaluation results for this query are exceptional, with all metrics indicating a perfect performance in retrieving and ranking relevant documents.

‘Killing innocents’

Precision: 1.0
Recall: 1.0
Average Precision: 1.0
F1 Score: 1.0
Mean Reciprocal Rank: 1.0
Normalized Discounted Cumulative Gain: 1.0

The evaluation results for this query are exceptional, with all metrics indicating a perfect performance in retrieving and ranking relevant documents.

‘Donations for Ukraine’
Precision: 0.6
Recall: 0.6
Average Precision: 0.6
F1 Score: 0.6
Mean Reciprocal Rank: 1.0
Normalized Discounted Cumulative Gain: 0.7273

In summary, the evaluation results for this query indicate a moderate performance. While the Mean Reciprocal Rank is perfect, the precision, recall, and average precision suggest that there is room for enhancement in the relevance and ranking of results. The F1 Score reflects the balance between precision and recall, and NDCG indicates the potential for improved ranking quality.

The mean average precision (MAP) of 0.92 for these queries indicates that, on average, the search engine is very effective at retrieving and ranking relevant documents. A MAP score of 0.92 is quite high, and it suggests that for the set of queries being evaluated, the search engine consistently presents relevant information at the top of the results, with minimal irrelevant or off-topic content. It is evident that the sole query not achieving a flawless average precision score is the one with the highly generic title "Donations for Ukraine." This query is likely to generate a considerable number of search results, making it more probable for tweets marked as "Non-relevant" to receive higher scores compared to the ones marked as "relevant." In contrast, for more specific queries like "Presidents visiting

Kyiv," it is expected to attain a perfect precision score due to the limited number of tweets related to that topic.

Word2Vec and TSNE

First of all we decided to train the word2vec model over both the evaluation queries and our own ones. So, the training data included the query and both the relevant and non-relevant tweets for all the queries of the project. After doing so, we first plotted the query and the relevant and non-relevant tweets for all the evaluation queries. The size of each datapoint is its score.

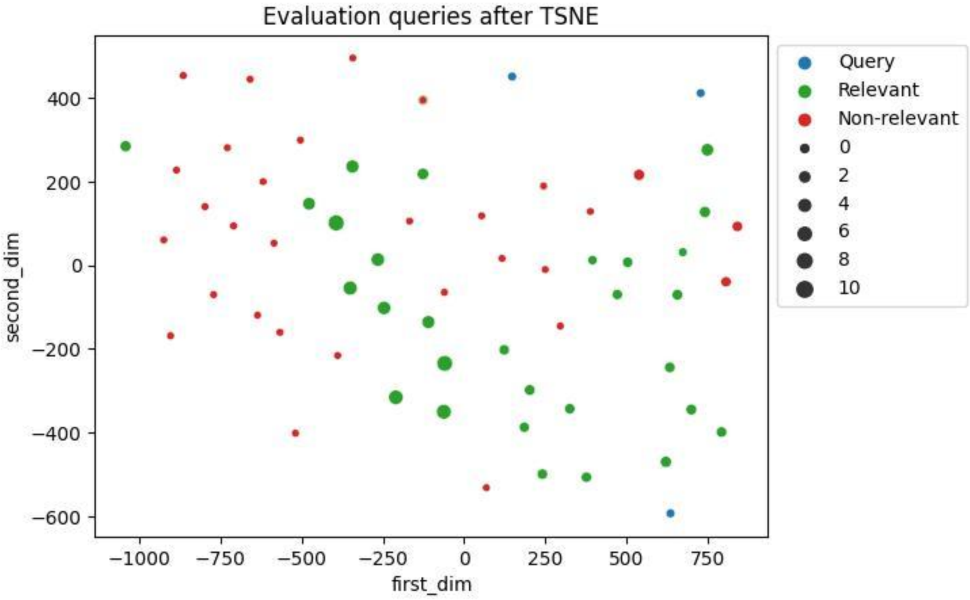


Figure 1

As it can be observed on Figure 1, the relevant queries are quite close to the queries except for the mid upper one. Therefore, the ground truth relevant tweets are close in the semantic space after performing the dimensionality reduction.

We obtain similar results for our own queries but with a higher relationship with queries and relevant documents. Almost all relevant documents are very close to the queries, or at least they are not close to non-relevant documents.

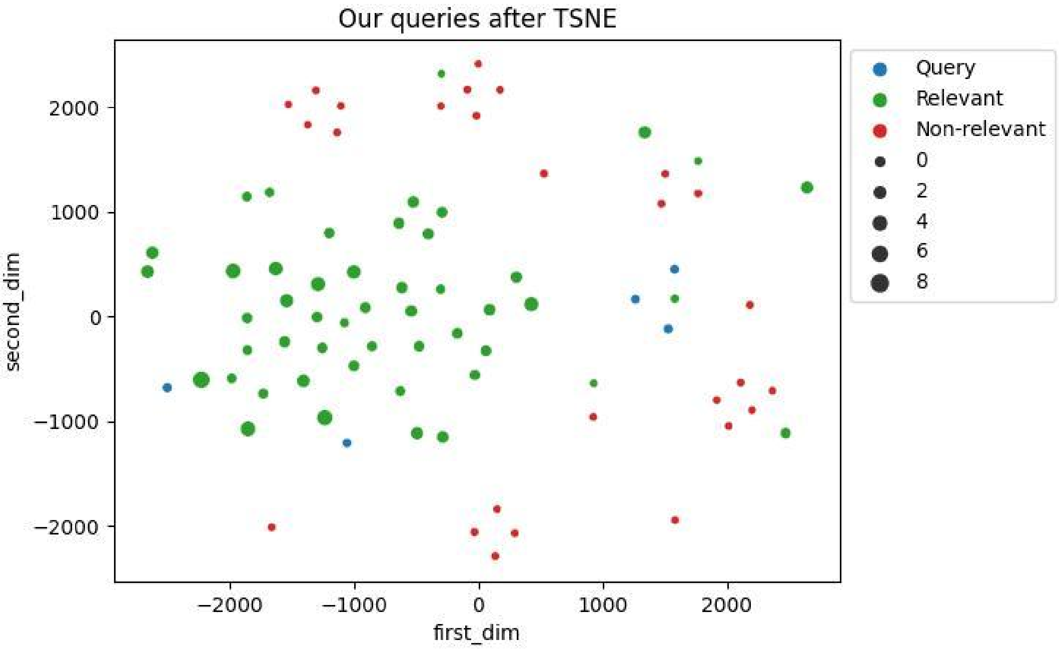


Figure 2

Then we did the same for a couple of single queries, to filter better which are the tweets that correspond to each query and the results are also very similar to the previous ones.

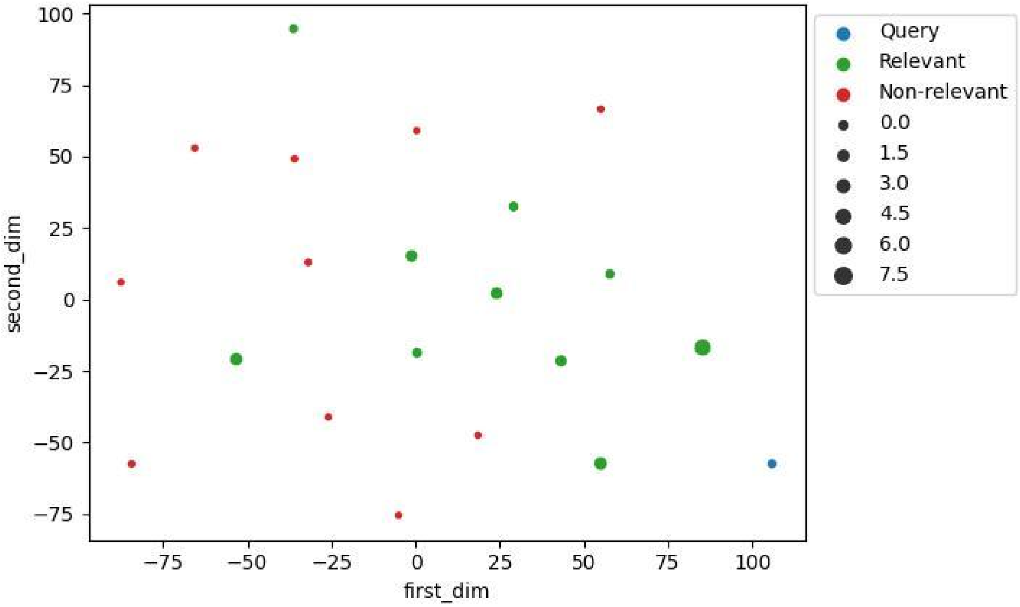


Figure 3

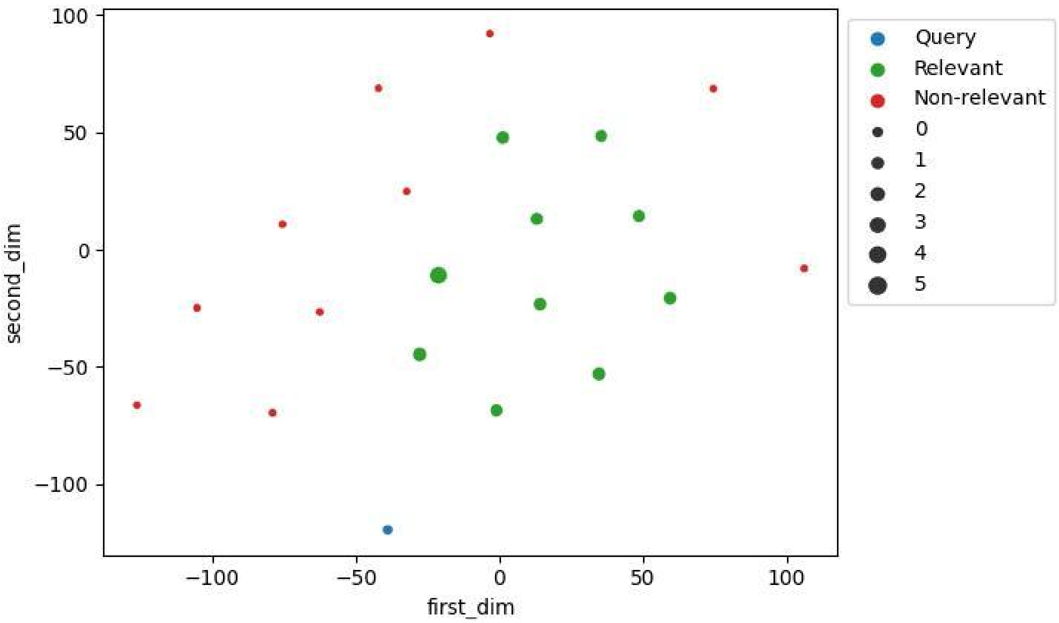
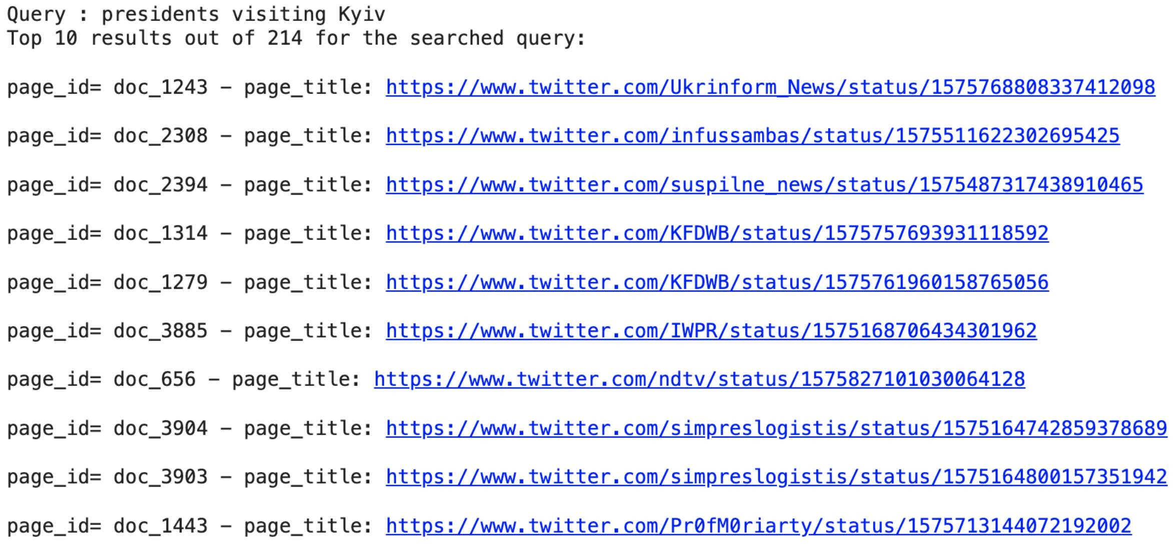


Figure 4

Therefore, we can conclude that there is a clear relationship between the relevant tweets and the queries in the semantic space.

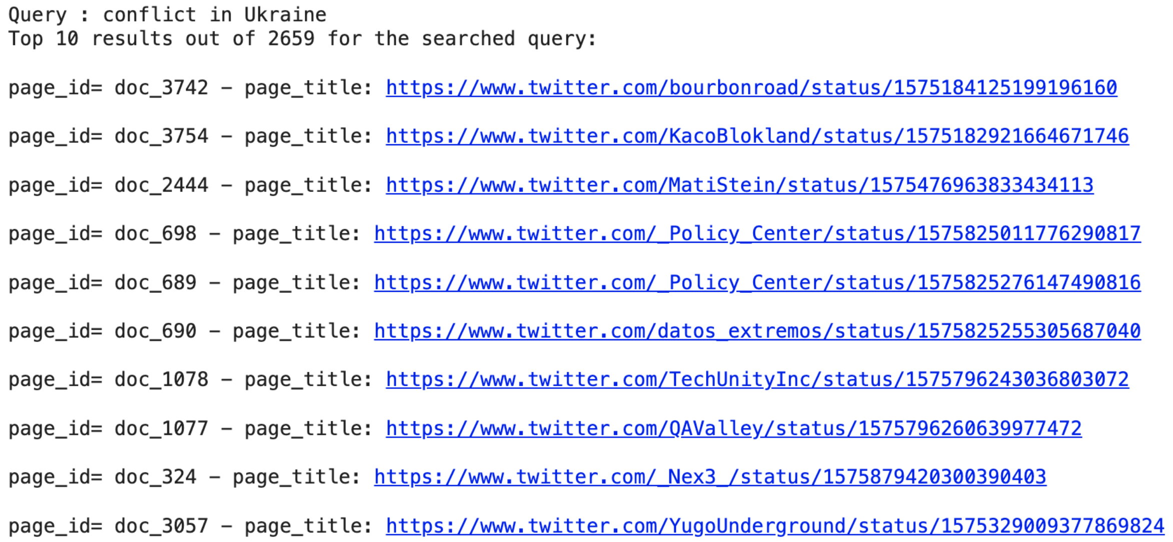
Annex

After performing the ranking of the tweets for our queries explained above, we obtained the following results:



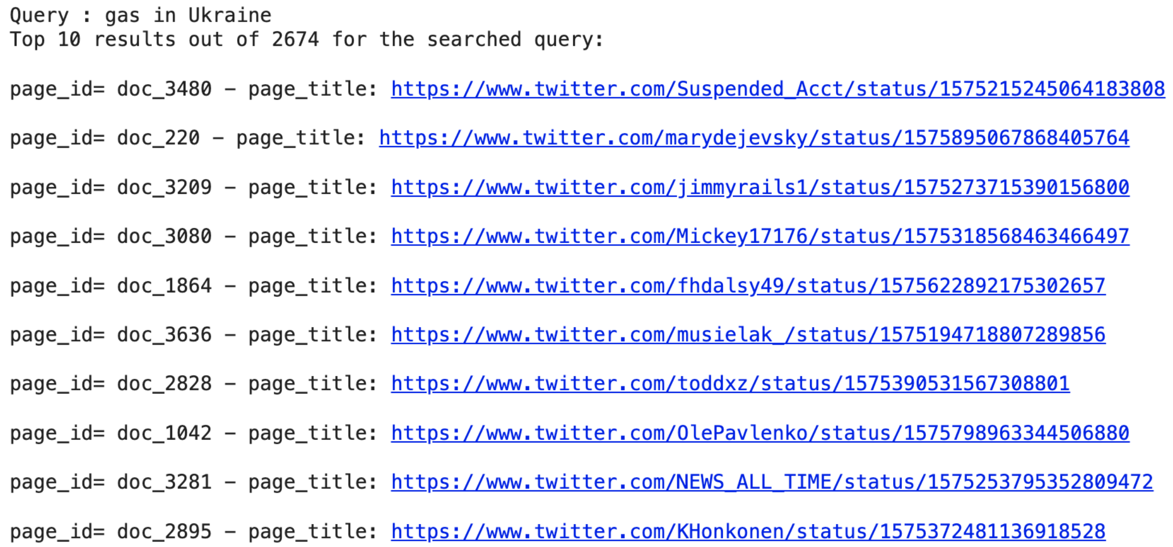
```
Query : presidents visiting Kyiv
Top 10 results out of 214 for the searched query:

page_id= doc_1243 - page_title: https://www.twitter.com/Ukrinform_News/status/1575768808337412098

page_id= doc_2308 - page_title: https://www.twitter.com/infussambas/status/1575511622302695425

page_id= doc_2394 - page_title: https://www.twitter.com/suspilne_news/status/1575487317438910465

page_id= doc_1314 - page_title: https://www.twitter.com/KFDWB/status/1575757693931118592

page_id= doc_1279 - page_title: https://www.twitter.com/KFDWB/status/1575761960158765056

page_id= doc_3885 - page_title: https://www.twitter.com/IWPR/status/1575168706434301962

page_id= doc_656 - page_title: https://www.twitter.com/ndtv/status/1575827101030064128

page_id= doc_3904 - page_title: https://www.twitter.com/simpreslogistis/status/1575164742859378689

page_id= doc_3903 - page_title: https://www.twitter.com/simpreslogistis/status/1575164800157351942

page_id= doc_1443 - page_title: https://www.twitter.com/Pr0fM0riarty/status/1575713144072192002
```

Figure 5 - 10 best searches of presidents visiting Kyiv



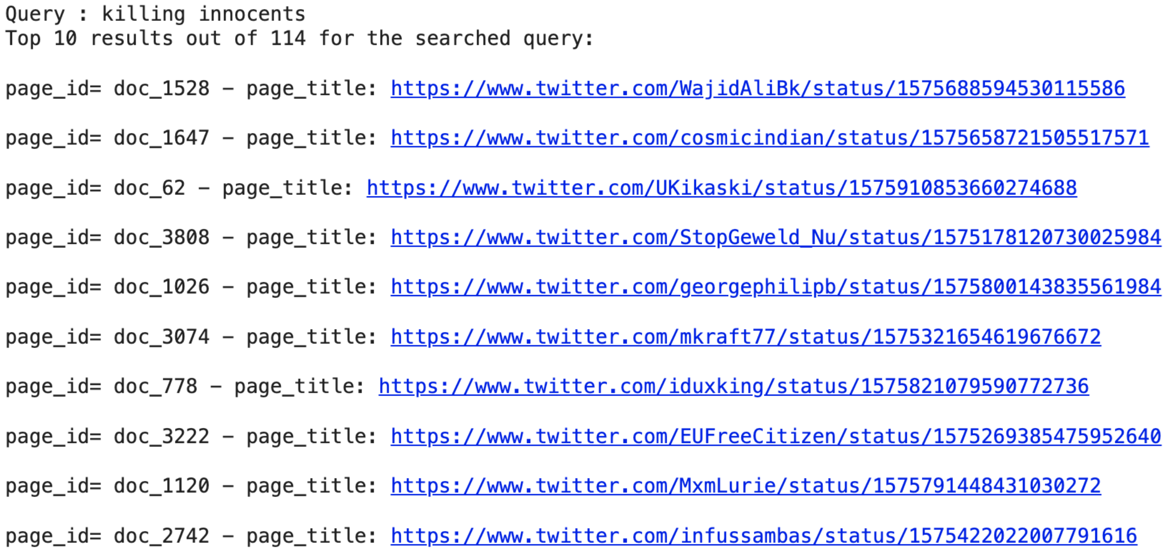
```
Query : conflict in Ukraine
Top 10 results out of 2659 for the searched query:

page_id= doc_3742 - page_title: https://www.twitter.com/bourbonroad/status/1575184125199196160

page_id= doc_3754 - page_title: https://www.twitter.com/KacoBlokland/status/1575182921664671746

page_id= doc_2444 - page_title: https://www.twitter.com/MatiStein/status/1575476963833434113

page_id= doc_698 - page_title: https://www.twitter.com/_Policy_Center/status/1575825011776290817

page_id= doc_689 - page_title: https://www.twitter.com/_Policy_Center/status/1575825276147490816

page_id= doc_690 - page_title: https://www.twitter.com/datos_extremos/status/1575825255305687040

page_id= doc_1078 - page_title: https://www.twitter.com/TechUnityInc/status/1575796243036803072

page_id= doc_1077 - page_title: https://www.twitter.com/QAValley/status/1575796260639977472

page_id= doc_324 - page_title: https://www.twitter.com/_Nex3_/status/1575879420300390403

page_id= doc_3057 - page_title: https://www.twitter.com/YugoUnderground/status/1575329009377869824
```

Figure 6 - 10 best searches of conflict in Ukraine



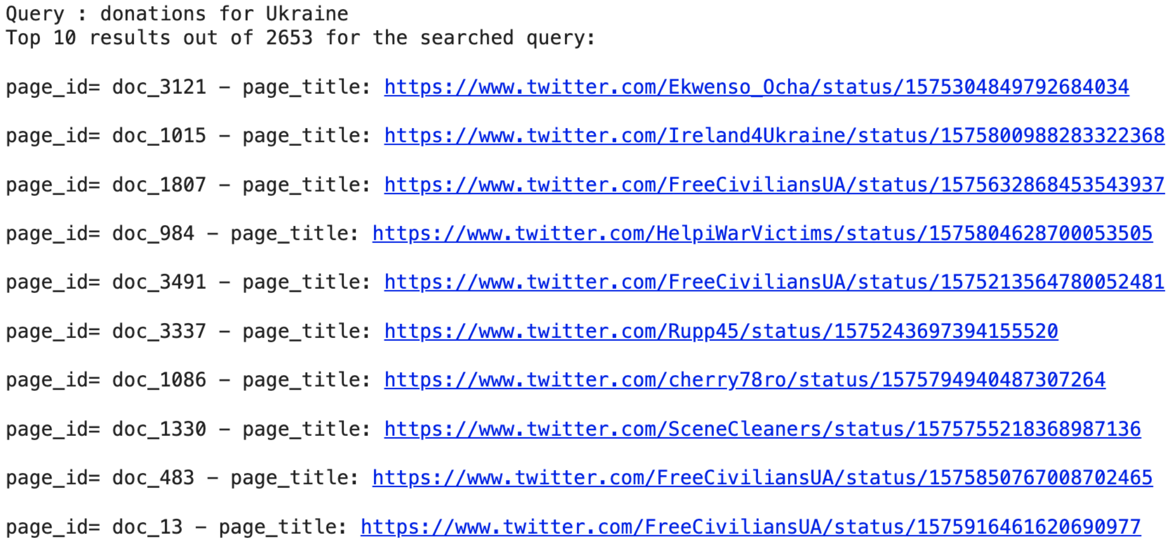
```
Query : gas in Ukraine
Top 10 results out of 2674 for the searched query:

page_id= doc_3480 - page_title: https://www.twitter.com/Suspended_Acct/status/1575215245064183808

page_id= doc_220 - page_title: https://www.twitter.com/marydejevsky/status/1575895067868405764

page_id= doc_3209 - page_title: https://www.twitter.com/jimmyrails1/status/1575273715390156800

page_id= doc_3080 - page_title: https://www.twitter.com/Mickey17176/status/1575318568463466497

page_id= doc_1864 - page_title: https://www.twitter.com/fhdalsy49/status/1575622892175302657

page_id= doc_3636 - page_title: https://www.twitter.com/musielak_/status/1575194718807289856

page_id= doc_2828 - page_title: https://www.twitter.com/toddxz/status/1575390531567308801

page_id= doc_1042 - page_title: https://www.twitter.com/OlePavlenko/status/1575798963344506880

page_id= doc_3281 - page_title: https://www.twitter.com/NEWS_ALL_TIME/status/1575253795352809472

page_id= doc_2895 - page_title: https://www.twitter.com/KHonkonen/status/1575372481136918528
```

Figure 7 - 10 best searches of gas in Ukraine



```
Query : killing innocents
Top 10 results out of 114 for the searched query:

page_id= doc_1528 – page_title: https://www.twitter.com/WajidAliBk/status/1575688594530115586

page_id= doc_1647 – page_title: https://www.twitter.com/cosmicindian/status/1575658721505517571

page_id= doc_62 – page_title: https://www.twitter.com/UKikaski/status/1575910853660274688

page_id= doc_3808 – page_title: https://www.twitter.com/StopGeweld_Nu/status/1575178120730025984

page_id= doc_1026 – page_title: https://www.twitter.com/georgephilipb/status/1575800143835561984

page_id= doc_3074 – page_title: https://www.twitter.com/mkraft77/status/1575321654619676672

page_id= doc_778 – page_title: https://www.twitter.com/iduxking/status/1575821079590772736

page_id= doc_3222 – page_title: https://www.twitter.com/EUFreeCitizen/status/1575269385475952640

page_id= doc_1120 – page_title: https://www.twitter.com/MxmLurie/status/1575791448431030272

page_id= doc_2742 – page_title: https://www.twitter.com/infussambas/status/1575422022007791616
```

Figure 8 - 10 best searches of killing innocents



```
Query : donations for Ukraine
Top 10 results out of 2653 for the searched query:

page_id= doc_3121 – page_title: https://www.twitter.com/Ekwenso_Ocha/status/1575304849792684034

page_id= doc_1015 – page_title: https://www.twitter.com/Ireland4Ukraine/status/1575800988283322368

page_id= doc_1807 – page_title: https://www.twitter.com/FreeCiviliansUA/status/1575632868453543937

page_id= doc_984 – page_title: https://www.twitter.com/HelpiWarVictims/status/1575804628700053505

page_id= doc_3491 – page_title: https://www.twitter.com/FreeCiviliansUA/status/1575213564780052481

page_id= doc_3337 – page_title: https://www.twitter.com/Rupp45/status/1575243697394155520

page_id= doc_1086 – page_title: https://www.twitter.com/cherry78ro/status/1575794940487307264

page_id= doc_1330 – page_title: https://www.twitter.com/SceneCleaners/status/1575755218368987136

page_id= doc_483 – page_title: https://www.twitter.com/FreeCiviliansUA/status/1575850767008702465

page_id= doc_13 – page_title: https://www.twitter.com/FreeCiviliansUA/status/1575916461620690977
```

Figure 9 - 10 best searches of donations for Ukraine