Gerard Morales - 242781
Patricia Garay - 229260
Maren Clapers - 243397
GROUP 102_6

Universitat Pompeu Fabra
Barcelona

Part 2: Indexing and Evaluation

Information Retrieval and Web Analysis

Introduction

In this second part of the project, the goal is to build an inverted index to have a better control over the words in the tweets and evaluate them through queries and rankings. In this report, it is explained why we chose each query, the explanation of the ranking results and the analysis of the evaluation of Russia-Ukraine war tweets dataset.

Indexing

1. Building the inverted index

We used the functions create_index_tfidf, rank_documents and search_tf_idf implemented in the first laboratory to build the inverted index and be able to find any information within the tweets.

2. Test queries proposal

Given the keyword rank frequency obtained from the first part of this project (in the Exploratory Data Analysis), we defined this five queries that will be used to evaluate our search engine:

- **"Presidents visiting Kyiv":** This query is relevant to see which president visited the capital. Can be used to test which countries support Ukraine.
- **"Countries supporting Ukraine"**: This query includes the terms "Countries" and "Ukraine" and is relevant to the situation involving Ukraine support from other countries.
- **"Humanitarian aid in Ukraine":** Humanitarian aid is extremely important in a war and this query can show the countries providing this help and how it is being handled.
- **"Citizens fleeing Ukraine"**: It might be useful to see which destinations these citizens are fleeing and the volume of war refugees the war is causing.
- **"Putin and Zelensky peace talks"**: This query is focusing on the term “peace” between the two presidents. It can be useful to see the proportion of tweets talking about peace between these countries.

Gerard Morales - 242781
Patricia Garay - 229260
Maren Clapers - 243397
GROUP 102_6

Universitat Pompeu Fabra Barcelona

3. **Results ranking with tf-idf**

We used the ranking functions to obtain the top 10 most relevant tweets related with each of the proposed queries and these are the highlights of each one:

- **"Presidents visiting Kyiv":** There is one tweet about the British defense minister paying a secret visit to Kyiv, president Putin calling on Kyiv to cease military action, President Zelensky request to join NATO, and other tweets that relate with the query statement.
- **"Countries supporting Ukraine"**: We can see tweets like “America supports the diplomatic end of the war” and that “countries in the side lines support Ukraine”. There are also other tweets asking which country you support. In general, matching with what we expected.
- **"Humanitarian aid in Ukraine":** There are tweets talking about the delivery of humanitarian aid to Russia, Ukrainian volunteers helping with humanitarian aid, tweets requesting humanitarian aid for Ukraine and Russian attacks on humanitarian convoys in Ukraine. As we can see “humanitarian aid” is commonly used in tweets during this war.
- **"Citizens fleeing Ukraine"**: We discovered Mongolia is one of the countries Ukraine citizens are fleeing, most of the fleas are from Lyman, Zelensky tells Russian occupiers to flee for their lives, Moscow tries to draft fleeing Russian men at the borders.
- **"Putin and Zelensky peace talks"**: Putin says "We are ready for talks", Ukraine is ready to talk with Russia, Zelensky did not accept Putin's peace proposal, calls grow for China and India to talk sense into Putin and other tweets related with peace talks are found in this query as expected.

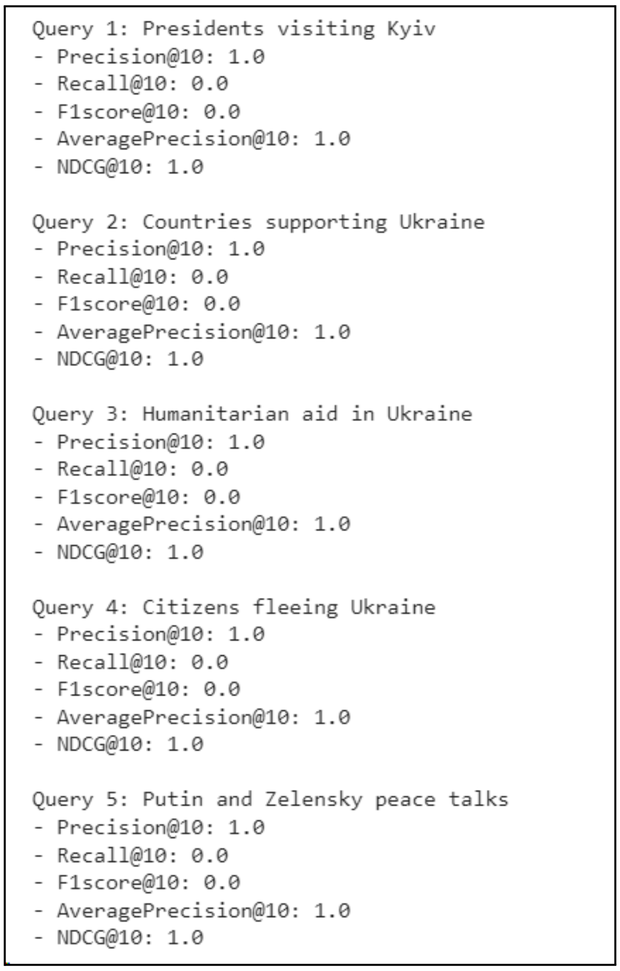
Evaluation

To evaluate each query we used the functions precision_at_k, avg_precision_at_k, map_at_k, rr_at_k, mrr_at_k, dcg_at_k, ndcg_at_k implemented in the second laboratory and we added the f1_score and the recall_k functions.

Evaluating our proposed queries:

The results of the search and rank of our queries show a very consistent and “perfect” performance, as we can observe below in the evaluation metrics for each one of the queries:

Gerard Morales - 242781
Patricia Garay - 229260
Maren Clapers - 243397
GROUP 102_6

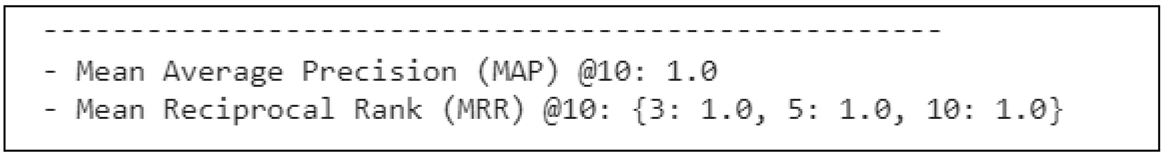


```
Query 1: Presidents visiting Kyiv
- Precision@10: 1.0
- Recall@10: 0.0
- F1score@10: 0.0
- AveragePrecision@10: 1.0
- NDCG@10: 1.0

Query 2: Countries supporting Ukraine
- Precision@10: 1.0
- Recall@10: 0.0
- F1score@10: 0.0
- AveragePrecision@10: 1.0
- NDCG@10: 1.0

Query 3: Humanitarian aid in Ukraine
- Precision@10: 1.0
- Recall@10: 0.0
- F1score@10: 0.0
- AveragePrecision@10: 1.0
- NDCG@10: 1.0

Query 4: Citizens fleeing Ukraine
- Precision@10: 1.0
- Recall@10: 0.0
- F1score@10: 0.0
- AveragePrecision@10: 1.0
- NDCG@10: 1.0

Query 5: Putin and Zelensky peace talks
- Precision@10: 1.0
- Recall@10: 0.0
- F1score@10: 0.0
- AveragePrecision@10: 1.0
- NDCG@10: 1.0
```

For all queries, the precision, average precision and NDCG values are 1. So, the queries achieve the maximum score possible in these metrics meaning that every tweet that our rank and search returns is relevant. Also, that there are no false positives.

However, for all queries, the recall value is 0. This means that none of the top 10 tweets for each query are able to deliver all the relevant information contained in the tweet collection about the war. So, there are no true positives either.

In conclusion, while our search excels in terms of precision, there is a lot of room for improvement in terms of recall and fine-tuning the ranking of the most relevant tweets for our queries.



```
--------------------------------------------------
- Mean Average Precision (MAP) @10: 1.0
- Mean Reciprocal Rank (MRR) @10: {3: 1.0, 5: 1.0, 10: 1.0}
```

As we can observe above, we also got a MAP score of 1 and a MRR score of 1 which means that, on average for each query, the system is returning a ranked list where all relevant tweets appear at the very top of the list. In other words, for every query, the search and rank is achieving a perfect precision at every position.

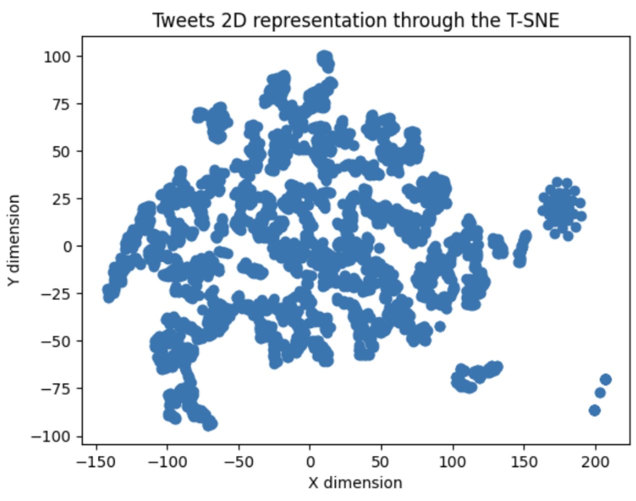
This is very rare, so we supposed that there might be some issues with our model because we are getting perfect precision but very bad recall/sensitivity in every query.

Tweets representation in a two-dimensional scatter plot

We have chosen the word2vec algorithm to represent words and tweets in vector form. To do this, we have created a function called *vectorize* which for each term in the tweet, retrieves its vector representation of Word2Vec words and adds it to a list of values.

After processing all terms in the tweet, it calculates the average vector by adding the word vectors and dividing them by the number of terms, returning the list of average vectors for all tweets.

Additionally, we trained the word2vec model with the data from the tweets column of our dataset, we called the *vectorize* function, converting the text data into numerical vectors, we reduced the dimensionality of the vectors for visualization purposes using t-SNE and we created a 2D scatter. The resulting scatter plot:



GitHub link

https://github.com/PatriciaGaray/IRWA-2023-u242781-u192945-u190355

TAG: IRWA Project Part 2