Mar de la Fuente - u199328 - 253535
Alba Yerga - u198634 - 252197
Alejandro Vílchez - u189522 - 242557

Report Part 2: Indexing and Evaluation

Github repository:
https://github.com/albayerga/G_102_6/releases/tag/IRWA-2024-u198634-u189522-u199328-part-2

1. Indexing

The next step to build our search engine is to construct the inverted index using the TF-IDF algorithm.

We first create an inverted index dictionary, which will map terms to their occurrences in documents. The structure will look like: [term -> {doc_id: [positions], doc_id: [positions], ...}]. This allows us to track where each term appears across different documents. Although we initially used a function provided in class, we found that it performed poorly with larger datasets. To address this, we redesigned the code to enhance performance.

Then to rank the documents, we represent the query as a weighted tf-idf vector and each document as a weighted tf idf vector. Then computing the cosine similarity score for the query vector and each document vector and ending ranking the given documents with respect to the query by score.

Finally, we implement the search function. Since we are dealing with conjunctive queries (AND) (each of the returned documents should contain all the words in the query) our search function will return the intersection of the lists of documents corresponding to each term in the query.

Propose test queries

These are the queries we propose to evaluate our search engine. We chose the following:

- Farmer protest
- Modi govt
- diesel price
- indian farmer
- Disha ravi

Mar de la Fuente - u199328 - 253535
Alba Yerga - u198634 - 252197
Alejandro Vílchez - u189522 - 242557

Ranked results for proposed test queries:

When you type for example 'farmers protest' you end up with the following result, we return for each document the Tweet | Date | Hashtags| Likes | Retweets | Url :

```
Processing query: 'Farmer protest'
Top 10 docs ids for query: 'Farmer protest'
doc_id= doc_32542
doc_id= doc_22865
doc_id= doc_7859
doc_id= doc_6060
doc_id= doc_5727
doc_id= doc_5183
doc_id= doc_46137
doc_id= doc_45189
doc_id= doc_43199
doc_id= doc_42188
Docs info for query 'Farmer protest':
```

	Tweet	Date	Hashtags	Likes	Retweets	Url
5183	[farmer, protest]	2021-02-22 22:32:21+00:00	[#FarmersProtest]	0	1	https://twitter.com/Rajnika78115125/status/136...
5727	[farmer, protest, farmer, protest]	2021-02-22 16:13:08+00:00	[#MSP_किसान_का_हक, #FarmersProtest]	0	0	https://twitter.com/JSekhupuria/status/1363884...
6060	[farmer, protest]	2021-02-22 12:50:47+00:00	[#FarmersProtest]	1	0	https://twitter.com/rwhrao_singh/status/136383...
7859	[farmer, protest, farmer, protest]	2021-02-22 02:00:20+00:00	[#FarmersProtests, #FarmersProtest]	0	0	https://twitter.com/Rajnika78115125/status/136...
22865	[largest, protest, kanganateam, farmer, protes...	2021-02-17 18:28:33+00:00	[#FarmersProtest, #FarmersProtest, #FarmersPro...	0	0	https://twitter.com/karm16200070/status/136210...
32542	[timesnow, farmer, protest, protest, protest]	2021-02-15 14:12:46+00:00	[#FarmersProtest]	0	0	https://twitter.com/45kHz/status/1361317533173...
42188	[farmer, protest]	2021-02-13 11:18:49+00:00	[#FarmersProtest]	0	0	https://twitter.com/khairaBet/status/136054898...
43199	[farmer, protest]	2021-02-13 05:07:42+00:00	[#FarmersProtest]	1	0	https://twitter.com/tash_kmb/status/1360455587...
45189	[farmer, protest]	2021-02-12 15:58:08+00:00	[#FarmersProtest]	3	1	https://twitter.com/Manpree52519810/status/136...
46137	[farmer, protest]	2021-02-12 10:08:02+00:00	[#FarmersProtest]	1	1	https://twitter.com/Parikau16692134/status/136...

```
Processing query: 'Modi govt'
Top 10 docs ids for query: 'Modi govt'
doc_id= doc_30979
doc_id= doc_32524
doc_id= doc_32236
doc_id= doc_4021
doc_id= doc_31725
doc_id= doc_32016
doc_id= doc_35049
doc_id= doc_32570
doc_id= doc_30292
doc_id= doc_28222
Docs info for query 'Modi govt':
```

	Tweet	Date	Hashtags	Likes	Retweets	Url
4021	[think, modi, govt, wors, british, govt, pagdi...	2021-02-23 04:33:50+00:00	[#Pagdi_Sambhal_Jatta, #FarmersProtest]	0	0	https://twitter.com/777sattiSingh/status/13640...
28222	[farmer, protest, india, domest, affairsupport...	2021-02-16 15:13:34+00:00	[#FarmersProtest, #ToolkitCase]	0	0	https://twitter.com/umakantsingh_IN/status/136...
30292	[shame, india, govt, modi, fuel, scam, iam, mo...	2021-02-16 04:12:17+00:00	[#ModiFuelScam, #IamAgainstModiGovt, #PetrolDi...	4	7	https://twitter.com/nishamirok/status/13615288...
30979	[govt, dictatorship, iam, modi, govt, farmer, ...	2021-02-16 01:03:08+00:00	[#IamAgainstModiGovt, #FarmersProtest]	0	0	https://twitter.com/Sahibpreet1111/status/1361...
31725	[govt, go, dilut, polic, total, salwa, judum, ...	2021-02-15 18:39:34+00:00	[#farmers!, #IamAgainstModiGovt, #FarmersProtest]	0	1	https://twitter.com/umakantsingh_IN/status/136...
32016	[strictli, condemn, unlaw, action, modi, govt,...	2021-02-15 16:45:03+00:00	[#FarmersProtest, #IamAgainstModiGovt]	2	0	https://twitter.com/AmrinderS_13/status/136135...
32236	[shame, india, govt, iam, modi, govt, farmer, ...	2021-02-15 15:42:31+00:00	[#IndiaGovt., #IamAgainstModiGovt, #FarmersPro...	0	0	https://twitter.com/legendjatt007/status/13613...
32524	[im, modi, govt, india, threat, undeclar, emer...	2021-02-15 14:18:51+00:00	[#IamAgainstModiGovt, #farmersprotest]	8	13	https://twitter.com/PrinceR98409311/status/136...
32570	[world, modi, govt, hate, passion, sing, song,...	2021-02-15 14:01:51+00:00	[#FarmersProtest, #IamAgainstModiGovt]	66	56	https://twitter.com/iMani_KaurRai/status/13613...
35049	[shame, act, modi, govt, india, silenc, farmer...	2021-02-15 02:46:08+00:00	[#IndiaBeingSilenced, #FarmersProtest, #ShameO...	0	1	https://twitter.com/JSMaan18/status/1361144734...

```
Processing query: 'diesel price'
Top 10 docs ids for query: 'diesel price'
doc_id= doc_12933
doc_id= doc_28458
doc_id= doc_20756
doc_id= doc_27520
doc_id= doc_27290
doc_id= doc_12511
doc_id= doc_12825
doc_id= doc_29721
doc_id= doc_28119
doc_id= doc_41265
Docs info for query 'diesel price':
```

	Tweet	Date	Hashtags	Likes	Retweets	Url
12511	[myogiadityanath, takefarmlawsback, farmerspro...	2021-02-20 12:05:50+00:00	[#takefarmlawsback, #farmersprotest, #standwit...	0	0	https://twitter.com/chahalcs28/status/13630975...
12825	[thursday, went, past, mark, madhya, pradesh, ...	2021-02-20 10:05:21+00:00	[#PetrolDieselPriceHike, #PetrolPrice, #Petrol...	1	1	https://twitter.com/abuzargaffarisr/status/136...
12933	[bjp, rule, state, better, control, price, pet...	2021-02-20 09:12:34+00:00	[#BJP, #Petrol, #Diesel, #ModiHaiToMehngaiHai,...	2	0	https://twitter.com/suneet7954/status/13630539...
20756	[month, farmer, protest, peopl, petrol100, pro...	2021-02-18 07:10:45+00:00	[#FarmersProtest, #petrol100, #BJP, #Days, #di...	2	0	https://twitter.com/RishabRath/status/13622984...
27290	[ashey, din, petrol, diesel, price, hike, farm...	2021-02-16 20:01:20+00:00	[#PetrolDieselPriceHike, #FarmersProtest]	1	0	https://twitter.com/Majhakisansang1/status/136...
27520	[farmer, protest, petrol, diesel, price, hike]	2021-02-16 18:46:37+00:00	[#FarmersProtest, #PetrolDieselPriceHike]	0	0	https://twitter.com/sangharusski/status/136174...
28119	[godi, media, explain, petrol, diesel, price, ...	2021-02-16 15:49:47+00:00	[#Godi, #Explained:, #Petrol,, #Diesel, #Modi,...	0	0	https://twitter.com/Shahidlived/status/1361704...
28458	[petrol, diesel, price, hike, diesel, also, sc...	2021-02-16 13:53:04+00:00	[#PetrolDieselPriceHike, ##FarmersProtest, #Pe...	2	0	https://twitter.com/ranjit1442/status/13616749...
29721	[andhbhakt, like, petrol, price, hike, petrol,...	2021-02-16 06:16:40+00:00	[#Andhbhakts, #PetrolPriceHike, #PetrolDieselP...	0	0	https://twitter.com/Gurjot20523956/status/1361...
41265	[give, best, price, petrol, diesel, go, back, ...	2021-02-13 16:58:09+00:00	[#GoBackModi, #FarmersProtest]	3	2	https://twitter.com/dinumeena73/status/1360634...

```
Processing query: 'indian farmer'
Top 10 docs ids for query: 'indian farmer'
doc_id= doc_30112
doc_id= doc_5374
doc_id= doc_9022
doc_id= doc_34729
doc_id= doc_30122
doc_id= doc_31839
doc_id= doc_12469
doc_id= doc_17156
doc_id= doc_40010
doc_id= doc_44653
Docs info for query 'indian farmer':
```

	Tweet	Date	Hashtags	Likes	Retweets	Url
5374	[vp, dear, madam, indian, farmer, need, justic...	2021-02-22 20:12:48+00:00	[#FarmersProtest]	0	0	https://twitter.com/Amandeepjohal11/status/136...
9022	[modirojgardo, indian, youth, farmer, protest,...	2021-02-21 16:14:23+00:00	[#modi_rojgar_do, #FarmersProtest-, #Petrol100...	2	0	https://twitter.com/Roshan575002/status/136352...
12469	[indian, cricket, son, got, msp, mumbai, india...	2021-02-20 12:21:15+00:00	[#MSP, #MumbaiIndians, #IPLAuctions2021, #Tool...	0	0	https://twitter.com/dapinder_brar/status/13631...
17156	[indian, daughter, support, farmer, protest, c...	2021-02-19 07:56:29+00:00	[#FarmersProtest]	0	1	https://twitter.com/karim_mewati/status/136267...
30112	[themanikgoyalb, indian, govt, indian, system,...	2021-02-16 04:49:51+00:00	[#FarmersProtest]	1	0	https://twitter.com/iRavinderSG/status/1361538...
30122	[indian, govt, indian, system, farmer, protest...	2021-02-16 04:48:26+00:00	[#FarmersProtest, #ReleaseDetainedFarmersAndAc...	3	0	https://twitter.com/iRavinderSG/status/1361537...
31839	[disha, ravi, jail, indian, activist, link, gr...	2021-02-15 17:49:00+00:00	[#FarmersProtest, #IndianInjustice]	41	14	https://twitter.com/UK51NGH/status/13613719494...
34729	[indian, farmer, protest, matter, british, ind...	2021-02-15 04:00:26+00:00	[#FarmersProtest]	2	1	https://twitter.com/manjitghuman58/status/1361...
40010	[nandini, actorsiddharth, buy, decid, priceamb...	2021-02-13 22:45:35+00:00	[#FarmersProtest]	0	0	https://twitter.com/Kamalpr70500608/status/136...
44653	[think, indian, farmer, today, indian, farmer,...	2021-02-12 18:19:26+00:00	[#FarmersProtest]	1	0	https://twitter.com/bishbishN/status/136029244...

```
Processing query: 'Disha ravi'
Top 10 docs ids for query: 'Disha ravi'
doc_id= doc_24045
doc_id= doc_28075
doc_id= doc_35524
doc_id= doc_35520
doc_id= doc_32720
doc_id= doc_36598
doc_id= doc_34540
doc_id= doc_24250
doc_id= doc_8824
doc_id= doc_36243
Docs info for query 'Disha ravi':
```

	Tweet	Date	Hashtags	Likes	Retweets	Url
8824	[proud, disha, ravi, free, disha, ravi, farmer...	2021-02-21 17:11:55+00:00	[#DishaRavi, #FreeDishaRavi, #FarmersProtest]	0	0	https://twitter.com/Me13015931/status/13635369...
24045	[releasedisha, releas, disha, ravi, disha, rav...	2021-02-17 10:25:41+00:00	[#releasedisha, #ReleaseDishaRavi, #DishaRavi,...	2	1	https://twitter.com/nikysaji/status/1361985162...
24250	[rais, voic, disha, ravi, disha, ravi, arrest,...	2021-02-17 08:54:23+00:00	[#DishaRaviArrest, #DishaRaviArrested, #Disha_...	3	1	https://twitter.com/ImAliQureshi/status/136196...
28075	[farmer, protest, disha, ravi, disha, ravi, ar...	2021-02-16 16:01:37+00:00	[#FarmersProtest, #DishaRavi, #DishaRaviArrest...	0	1	https://twitter.com/actuallyshivom/status/1361...
32720	[dictatorship, farmersprotest, releas, disha, ...	2021-02-15 13:01:03+00:00	[#Farmersprotest, #ReleaseDishaRavi, #DishaRavi]	4	0	https://twitter.com/aakash_du/status/136129948...
34540	[disha, ravi, arrest, support, farmer, protest...	2021-02-15 04:34:54+00:00	[#DishaRaviArrested, #FarmersProtest, #Release...	0	0	https://twitter.com/bijoshpv/status/1361172109...
35520	[releas, disha, ravi, disha, ravi, farmer, pro...	2021-02-15 00:08:20+00:00	[#शहीद_जवान_शहीद_किसान, #DishaRavi, #FarmersPr...	1	1	https://twitter.com/Baldev52633391/status/1361...
35524	[stand, disha, ravi, disha, ravi, farmer, prot...	2021-02-15 00:04:36+00:00	[#DishaRavi, #FarmersProtest]	1	0	https://twitter.com/Baldev52633391/status/1361...
36243	[releas, disha, ravi, justic, disha, ravi, far...	2021-02-14 18:12:01+00:00	[#ReleaseDishaRavi, #FarmersProtest]	1	0	https://twitter.com/MaanDee08215437/status/136...
36598	[arrest, disha, ravi, cowardli, releas, disha,...	2021-02-14 16:44:42+00:00	[#DishaRavi, #ReleaseDishaRavi, #FreeDishaRavi...	0	0	https://twitter.com/MannnatKaur/status/1360993...

2. Evaluation

We evaluated the 2 queries suggested in the latest email:

- query 1: "people's rights"
- query 2: "Indian Government"

For query 1, P = 0.428 and R = 0.933.

	Relevant	Nonrelevant
Retrieved	14	313
Not retrieved	1	-

For query 2, P = 0.026 and R = 0.933.

	Relevant	Nonrelevant
Retrieved	14	516
Not retrieved	1	-

Now, we compute again the ranked-based measures with the new queries and k=150.

- **Precision@K (P@K)**
    - P@150 for query 1 = 0.0466
    - P@150 for query 2 = 0.02
- **Recall@K (R@K)**
    - R@150 for query 1 = 0.4666
    - R@150 for query 2 = 0.2
- **Average Precision@K (P@K)**
    - Average Precision@150 for query 1: 0.0668
    - Average Precision@150 for query 2: 0.0407
- **F1-Score@K**
    - F1 Score@150 for query 1: 0.0848
    - F1 Score@150 for query 2: 0.0363
- **Mean Average Precision (MAP)**
    - MAP@150 for the queries: 0.0537
- **Mean Reciprocal Rank (MRR)**
    - MRR for the test queries: 0.0645
- **Normalized Discounted Cumulative Gain (NDCG)**
    - NDCG for query 1: 0.3679
    - NDCG for query 2: 0.3218

With the new queries, we obtain the similar results as before for query 1, but we obtain better results for query 2. In this case, for query 2, we get a high recall and low precision (not zero) because a lot of non relevant documents are being retrieved (>500) but 14 out of the 15 relevant ones are retrieved.

3. T-SNE

We started by choosing the Word2Vec embedding, which creates a vector representation for each word in our vocabulary. For example, words like "farmer," "protest," and "modi" that were tokenized from the tweets, and during the training phase, the Word2Vec model learned to generate embeddings based on the context of these words.

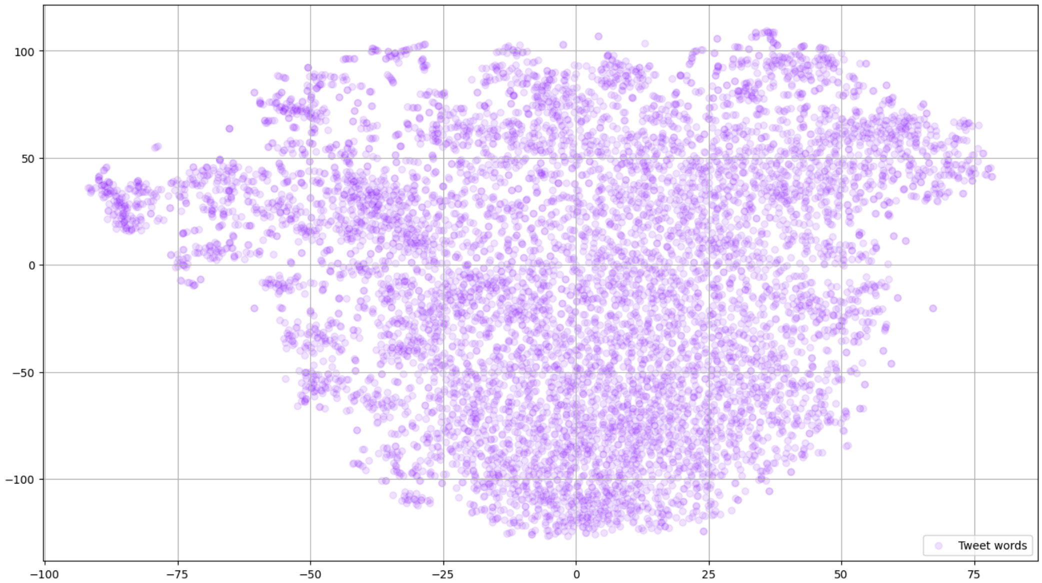
After training our Word2Vec model using the tokenized tweets, we were able to access the vocabulary and their respective embeddings. For instance, the vector representation of the word "farmer" might look something like this:

```
1 print(model.wv['farmer'])
```

```
[-0.8102957   0.2618334   0.8918676  -0.06767294  0.56045157 -0.9858518
  0.80591416  0.16014314  0.2548356   0.1930234   1.5862317   0.14496033
  0.2877727  -0.14814053 -0.2650403  -0.6690193   0.8960199   0.2575949
  0.3201867  -0.60504323  1.1974939  -0.0140938   0.5508737   0.50993866
  0.20442455  0.21301873 -0.21454617 -0.21113239 -0.17473501 -0.39202377
 -0.3636043  -0.32570627 -0.7168867   0.7079975   1.0433404   0.16586031
  1.1488253  -0.2932556   0.26584333  0.9688709  -0.65826374  0.26151526
 -1.3506787   0.46038145 -0.69139665  0.44286144 -0.51054925  0.22303668
  1.1928465   0.72005343  0.11505359  0.5513009   1.1184659  -0.15108535
 -0.00760236 -0.08723183  0.81308377  0.19185309  0.3666345   0.77422816
 -0.6846749   0.41146547  0.14136773 -1.7956593   0.58632994  0.45630875
 -0.58735853  0.2810279   0.26765344 -1.6035405   0.07173382 -0.47895584
  0.34642684  0.34920487 -0.655811   -0.17869727 -0.2333072   1.7451488
  0.28018543 -0.6502301   0.06000979  0.31267357  0.17617449  0.48980263
 -0.3535136   0.14745513 -0.07142348 -1.3253479   0.40022257  0.750446
  0.27373025  0.186043    0.54979616 -0.16656649  0.17673878  0.4622967
 -0.596005    0.38415787  0.9970127   0.52492785]
```

Fig 2 Example of the vector of the word farmer

To visualize the word embeddings, we applied the T-SNE algorithm, which is well-suited for reducing high-dimensional data to two while preserving the local structure of the data. We trained T-SNE using the embeddings generated by Word2Vec.



- **<u>Extra:</u>** evaluation with the queries given in the original assignment:

Before computing the evaluation functions, let’s evaluate the Precision and Recall table for both queries:

For query 1, P = 0.428 and R = 0.933.

	Relevant	Nonrelevant
Retrieved	14	313
Not retrieved	1	-

For query 2, P = 0 and R = 0.

	Relevant	Nonrelevant
Retrieved	0	3
Not retrieved	15	-

Then, we compute the following ranked-based measures. We used k=150.

- **Precision@K (P@K)**
    - P@150 for query 1 = 0.0466
    - P@150 for query 2 = 0

- **Recall@K (R@K)**
    - R@150 for query 1 = 0.4666
    - R@150 for query 2 = 0
    -

- **Average Precision@K (P@K)**
    - Average Precision@150 for query 1: 0.0488
    - Average Precision@150 for query 2: 0

- **F1-Score@K**
    - F1 Score@150 for query 1: 0.0727
    - F1 Score@150 for query 2: 0

- **Mean Average Precision (MAP)**
    - MAP@150 for the queries: 0.0244

- **Mean Reciprocal Rank (MRR)**
    - MRR for the test queries: 0.0454

- **Normalized Discounted Cumulative Gain (NDCG)**
    - NDCG for query 1: 0.3529
    - NDCG for query 2: 0

The evaluation shows that the model works pretty well for query 1, as it retrieves many relevant documents (high recall) but also includes a lot of irrelevant ones (low to moderate precision), especially at the top. For query 2, however, the model doesn't perform well, finding no relevant documents. This suggests that the model might need improvements to rank relevant results higher or detect the keywords better.