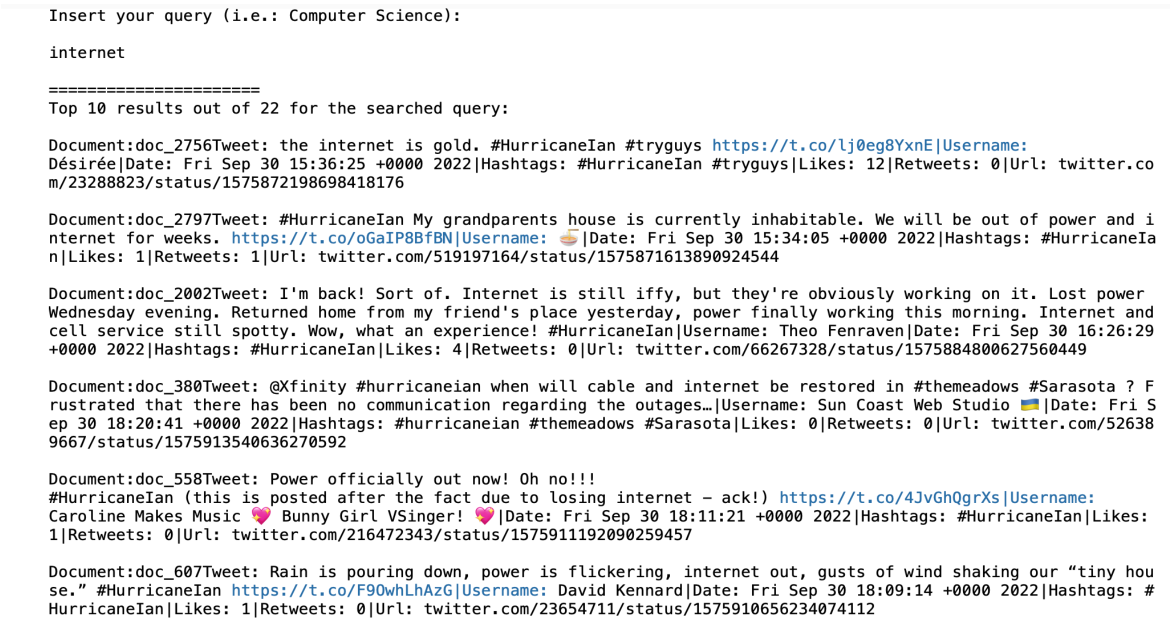
IR Project: Part 2

We are continuing working with the part 1 of the project. The final of Part 1 was a ranked search (example of word: internet):

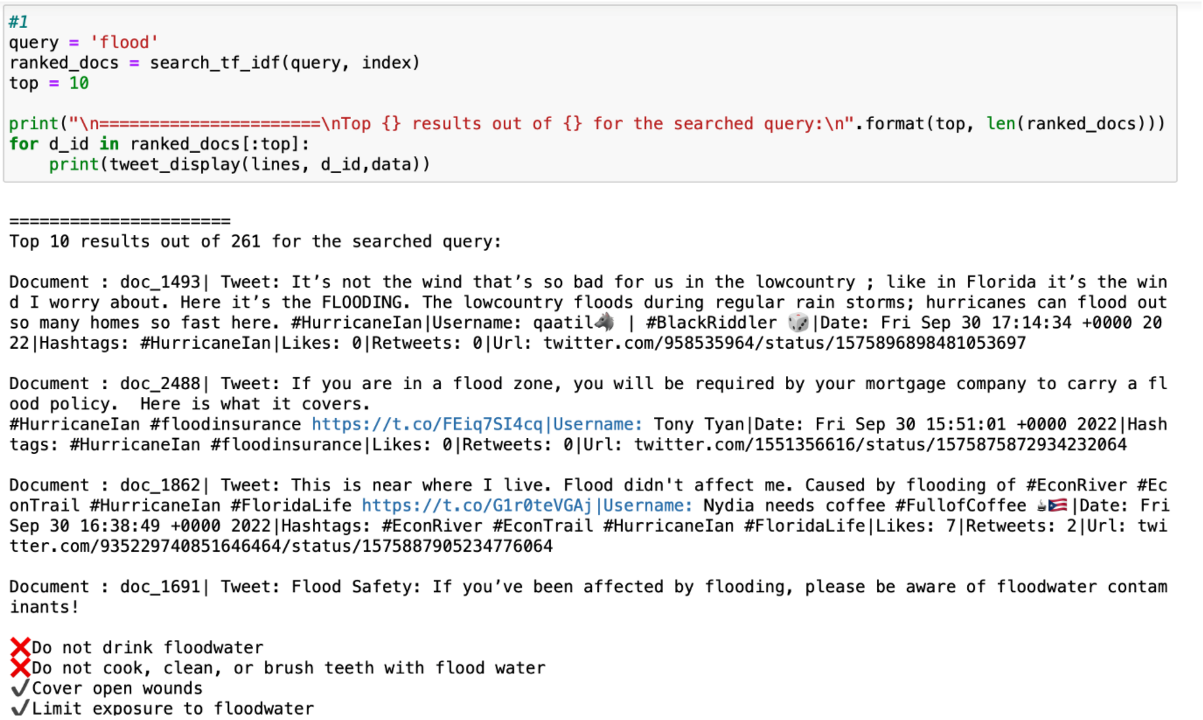


```
Insert your query (i.e.: Computer Science):

internet

======================
Top 10 results out of 22 for the searched query:

Document:doc_2756Tweet: the internet is gold. #HurricaneIan #tryguys https://t.co/lj0eg8YxnE|Username:
Désirée|Date: Fri Sep 30 15:36:25 +0000 2022|Hashtags: #HurricaneIan #tryguys|Likes: 12|Retweets: 0|Url: twitter.co
m/23288823/status/1575872198698418176

Document:doc_2797Tweet: #HurricaneIan My grandparents house is currently inhabitable. We will be out of power and i
nternet for weeks. https://t.co/oGaIP8BfBN|Username: 🍜|Date: Fri Sep 30 15:34:05 +0000 2022|Hashtags: #HurricaneIa
n|Likes: 1|Retweets: 1|Url: twitter.com/519197164/status/1575871613890924544

Document:doc_2002Tweet: I'm back! Sort of. Internet is still iffy, but they're obviously working on it. Lost power
Wednesday evening. Returned home from my friend's place yesterday, power finally working this morning. Internet and
cell service still spotty. Wow, what an experience! #HurricaneIan|Username: Theo Fenraven|Date: Fri Sep 30 16:26:29
+0000 2022|Hashtags: #HurricaneIan|Likes: 4|Retweets: 0|Url: twitter.com/66267328/status/1575884800627560449

Document:doc_380Tweet: @Xfinity #hurricaneian when will cable and internet be restored in #themeadows #Sarasota ? F
rustrated that there has been no communication regarding the outages…|Username: Sun Coast Web Studio 🇺🇦|Date: Fri S
ep 30 18:20:41 +0000 2022|Hashtags: #hurricaneian #themeadows #Sarasota|Likes: 0|Retweets: 0|Url: twitter.com/52638
9667/status/1575913540636270592

Document:doc_558Tweet: Power officially out now! Oh no!!!
#HurricaneIan (this is posted after the fact due to losing internet – ack!) https://t.co/4JvGhQgrXs|Username:
Caroline Makes Music 💖 Bunny Girl VSinger! 💖|Date: Fri Sep 30 18:11:21 +0000 2022|Hashtags: #HurricaneIan|Likes:
1|Retweets: 0|Url: twitter.com/216472343/status/1575911192090259457

Document:doc_607Tweet: Rain is pouring down, power is flickering, internet out, gusts of wind shaking our “tiny hou
se.” #HurricaneIan https://t.co/F9OwhLhAzG|Username: David Kennard|Date: Fri Sep 30 18:09:14 +0000 2022|Hashtags: #
HurricaneIan|Likes: 1|Retweets: 0|Url: twitter.com/23654711/status/1575910656234074112
```

As we can see we print the tweets like this:

Doc_Name | Tweet | Username | Date | Hashtags | Likes | Retweets | Url

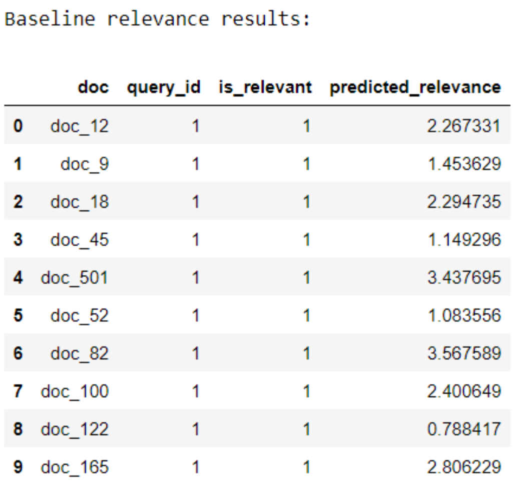
After that we make the 5 queries. Example of our first query (flood):



```
#1
query = 'flood'
ranked_docs = search_tf_idf(query, index)
top = 10

print("\n======================\nTop {} results out of {} for the searched query:\n".format(top, len(ranked_docs)))
for d_id in ranked_docs[:top]:
    print(tweet_display(lines, d_id,data))
```

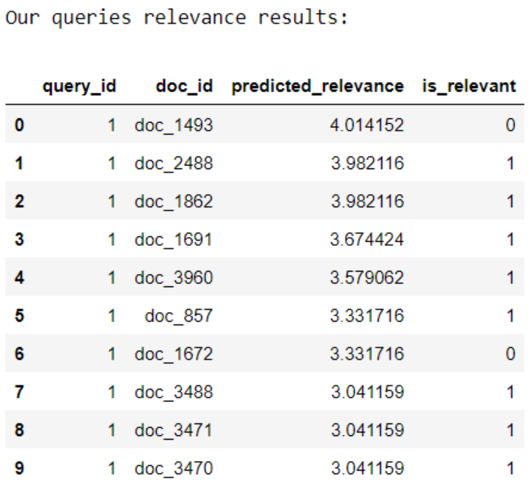
```
======================
Top 10 results out of 261 for the searched query:

Document : doc_1493| Tweet: It’s not the wind that’s so bad for us in the lowcountry ; like in Florida it’s the win
d I worry about. Here it’s the FLOODING. The lowcountry floods during regular rain storms; hurricanes can flood out
so many homes so fast here. #HurricaneIan|Username: qaatil🐺 | #BlackRiddler 🎲|Date: Fri Sep 30 17:14:34 +0000 20
22|Hashtags: #HurricaneIan|Likes: 0|Retweets: 0|Url: twitter.com/958535964/status/1575896898481053697

Document : doc_2488| Tweet: If you are in a flood zone, you will be required by your mortgage company to carry a fl
ood policy.  Here is what it covers.
#HurricaneIan #floodinsurance https://t.co/FEiq7SI4cq|Username: Tony Tyan|Date: Fri Sep 30 15:51:01 +0000 2022|Hash
tags: #HurricaneIan #floodinsurance|Likes: 0|Retweets: 0|Url: twitter.com/1551356616/status/1575875872934232064

Document : doc_1862| Tweet: This is near where I live. Flood didn't affect me. Caused by flooding of #EconRiver #Ec
onTrail #HurricaneIan #FloridaLife https://t.co/G1r0teVGAj|Username: Nydia needs coffee #FullofCoffee ☕🇵🇷|Date: Fri
Sep 30 16:38:49 +0000 2022|Hashtags: #EconRiver #EconTrail #HurricaneIan #FloridaLife|Likes: 7|Retweets: 2|Url: twi
tter.com/935229740851646464/status/1575887905234776064

Document : doc_1691| Tweet: Flood Safety: If you’ve been affected by flooding, please be aware of floodwater contam
inants!

❌Do not drink floodwater
❌Do not cook, clean, or brush teeth with flood water
✔Cover open wounds
✔Limit exposure to floodwater
```

After this we start this new part where we load one dataframe with the queries given and another one with the queries chosen by us and decide if is relevant or not (for our queries):



Baseline relevance results:

	doc	query_id	is_relevant	predicted_relevance
0	doc_12	1	1	2.267331
1	doc_9	1	1	1.453629
2	doc_18	1	1	2.294735
3	doc_45	1	1	1.149296
4	doc_501	1	1	3.437695
5	doc_52	1	1	1.083556
6	doc_82	1	1	3.567589
7	doc_100	1	1	2.400649
8	doc_122	1	1	0.788417
9	doc_165	1	1	2.806229



Our queries relevance results:

	query_id	doc_id	predicted_relevance	is_relevant
0	1	doc_1493	4.014152	0
1	1	doc_2488	3.982116	1
2	1	doc_1862	3.982116	1
3	1	doc_1691	3.674424	1
4	1	doc_3960	3.579062	1
5	1	doc_857	3.331716	1
6	1	doc_1672	3.331716	0
7	1	doc_3488	3.041159	1
8	1	doc_3471	3.041159	1
9	1	doc_3470	3.041159	1

Then we have implemented different functions to calculate:

- Precision@K
- Recall@K
- F1 score
- Average Precision@K
- NDGC
- MAP (Mean Average Precision)
- MRR (Mean Reciprocal Rank)

We calculate this for each query of each Dataframe, we calculate precision, recall and F1 score separately. And the results of the evaluation of the system are the following:

```
Query : Landfall in South Carolina
==> Precision@5: 1.0
==> Recall@5: 0.5
==> Average Precision@5: 0.5
==> F1-score of first 5: 0.6666666666666666
==> NDCG@5: 1.0

Query : Help and recovery during the hurricane disaster
==> Precision@5: 0.8
==> Recall@5: 0.4
==> Average Precision@5: 0.38
==> F1-score of first 5: 0.5333333333333333
==> NDCG@5: 0.8539

Query : Floodings in South Carolina
==> Precision@5: 1.0
==> Recall@5: 0.5
==> Average Precision@5: 0.5
==> F1-score of first 5: 0.6666666666666666
==> NDCG@5: 1.0
----------------------------------------------------
==> Mean Average Precision (MAP) @5: 0.45999999999999996
==> Mean Reciprocal Rank (MRR) @5: 1.0
```

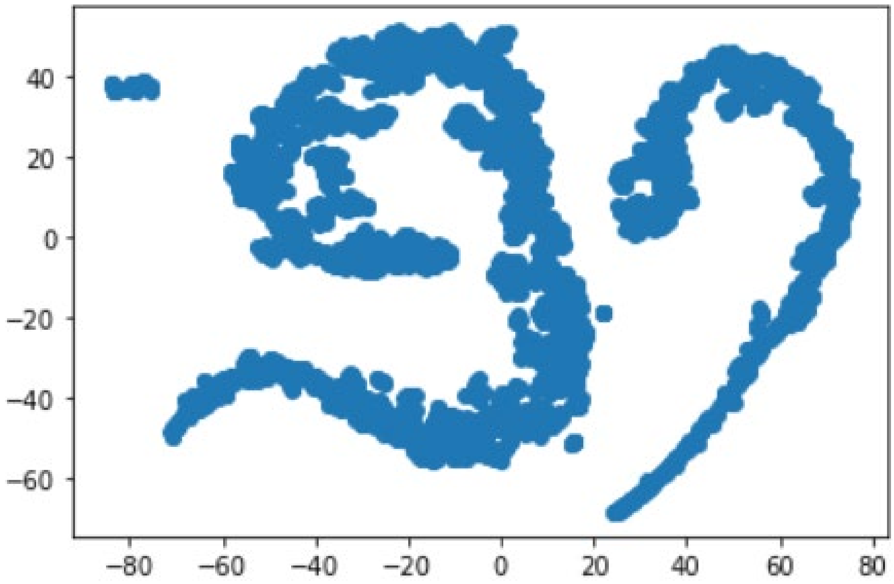
And here about our queries:

```
Query 1: flood
==> Precision@5: 0.8
==> Recall@5: 0.5
==> Average Precision@5: 0.33958333333333335
==> F1-score of first 5: 0.6153846153846154
==> NDCG@5: 0.6608

Query 2: emergency
==> Precision@5: 0.4
==> Recall@5: 0.2857142857142857
==> Average Precision@5: 0.19999999999999998
==> F1-score of first 5: 0.3333333333333333
==> NDCG@5: 0.4704

Query 3: hurricane
==> Precision@5: 0.4
==> Recall@5: 0.4
==> Average Precision@5: 0.18
==> F1-score of first 5: 0.4000000000000001
==> NDCG@5: 0.3452

Query 4: florida
==> Precision@5: 0.6
==> Recall@5: 0.6
==> Average Precision@5: 0.2866666666666666
==> F1-score of first 5: 0.6
==> NDCG@5: 0.4469

Query 5: landfall
==> Precision@5: 0.8
==> Recall@5: 0.5714285714285714
==> Average Precision@5: 0.5714285714285714
==> F1-score of first 5: 0.6666666666666666
==> NDCG@5: 0.8688
----------------------------------------------------
==> Mean Average Precision (MAP) @5: 0.31553571428571425
==> Mean Reciprocal Rank (MRR) @5: 0.6666666666666667
```

Here we can say that query 1 and 5 are more precise than the others with a k= 5, but also query 4 has a good precision associated with the best recall. That will be interesting to do a P/R graph.
However the better recall of query 4 is not enough as we could see by looking at the f1-score that is the highest for query 5 that has also the highest average prediction.
If we care more about finding the best result in the first place, query 5 is the best because has the highest NDCG, while query 3 is the worst, in fact we can find two non-relevant documents in first places.
Also mean average precision is lower than the given queries.
Mean Reciprocal rank is also lower because not all the queries found a relevant doc in first place.

We can conclude that the query Landfall in South Carolina is the one with the better results and from our selected queries, the query one (landfall) is the one with better results.

After this we are making a plot of vectors-tweets, this is the result:



THIS IS THE REPOSITORY: https://github.com/JordiBadia01