




Information retrieval project part 1

In order to execute the code you need to run the notebook cells in order. The project is prepared to work both in jupyter notebooks and google colab (if stored at the root of MyDrive), provided its folder structure is not modified in any way.


For this lab we used pandas dataframes to store the tweets and operate on them. So we decided to create a dataframe with only the information that would be useful in our case. After that we decided to make some modifications to extract correctly the url, hashtags and user names. In case there are no url for a concrete tweet, we have decided to leave it blank.

We also needed to process the text inside the tweet, this involves trimming, tokenizing and removing the stopwords using the nltk library, and more concretely using the function provided in class. We have also decided to remove links, users and hashtags since we have considered that they wouldn't be useful.

The last think we did was use tweets ids to convert to doc ids using the csv map file provided to us. This results in having the following dataframe:

	id	full_text	user	created_at	favorite_count	retweet_count	url	hashtags
0	doc_1	keep spin us 7 pm...go away already.	suz 	2022-09-30 18:39:08+00:00	0	0	https://twitter.com/suzjdean/status/1575918182...	[hurricaneian]
1	doc_2	heart go affect wish everyon road current brav...	lytx	2022-09-30 18:39:01+00:00	0	0		[hurricaneian]
2	doc_3	kissimme neighborhood michigan ave.	christopher heath	2022-09-30 18:38:58+00:00	0	0	https://twitter.com/CHeathWFTV/status/15759181...	[hurricaneian]
3	doc_4	one tree backyard scare poltergeist tree it' s...	alex 	2022-09-30 18:38:57+00:00	0	0		[scwx, hurricaneian]
4	doc_5	pray everyon affect associ winknews. sympathi ...	tess 	2022-09-30 18:38:53+00:00	0	0		[hurricaneian]

The final result as a dataframe is saved in a csv file called processed_tweets, with the following content:

	id	full_text	user	created_at	favorite_count	retweet_count	url	hashtags
0	doc_1	keep spin us 7 pm...go away already.	suz 	2022-09-30 18:39:08	0	0	https://twitter.com/suzjdean/status/1575918182...	[hurricaneian]
1	doc_2	heart go affect wish everyon road current brav...	lytx	2022-09-30 18:39:01	0	0		[hurricaneian]
2	doc_3	kissimme neighborhood michigan ave.	christopher heath	2022-09-30 18:38:58	0	0	https://twitter.com/CHeathWFTV/status/15759181...	[hurricaneian]