Part 1: Text Processing

You are provided with a document corpus which is a set of tweets related to Hurricane Ian (tw_hurricane_data.json). You can see an example document in the appendix.

As a first step, you must pre-process the documents by

- Removing stop words
- Tokenization
- Removing punctuation marks
- Stemming
- and... anything else you think it's needed (bonus point)

HINTS:

1. Take into account that for future queries, the final output must return (when present) the following information for each of the selected documents: **Tweet | Username | Date | Hashtags | Likes | Retweets | Url** (here the "Url" means the tweet link).

2. Think about how to handle the hashtags from your pre-processing steps (e.g., removing the "#" from the word), since it may be useful to involve them as separated terms inside the inverted index.

3. Suggested library that may help you in stemming and stop words: **nltk**

Make sure you map the tweet's Ids with the document ids as the document Ids will be considered for the evaluation stage of the project (tweet_document_ids_map.csv).

Appendix

Example document extracted from Twitter:


```
{
  "created_at": "Fri Sep 30 18:39:08 +0000 2022",
  "id": 1575918182698979328,
  "id_str": "1575918182698979328",
  "full_text": "So this will keep spinning over us until 7 pm…go
away already. #HurricaneIan https://t.co/VROTxNS9rz",
  "truncated": false,
  "display_text_range": [
    0,
    76
  ],
  "entities": {
    "hashtags": [
      {
        "text": "HurricaneIan",
        "indices": [
          63,
          76
        ]
      }
    ],
    "symbols": [

    ],
    "user_mentions": [

    ],
    "urls": [

    ],
    "media": [
      {
        "id": 1575918178261254162,
        "id_str": "1575918178261254162",
        "indices": [
          77,
          100
        ],
        "media_url":
"http://pbs.twimg.com/media/Fd7JO8pXwBI9HPw.jpg",
```

```
        "media_url_https":
"https://pbs.twimg.com/media/Fd7JO8pXwBI9HPw.jpg",
        "url": "https://t.co/VROTxNS9rz",
        "display_url": "pic.twitter.com/VROTxNS9rz",
        "expanded_url":
"https://twitter.com/suzjdean/status/1575918182698979328/photo/1",
        "type": "photo",
        "sizes": {
          "small": {
            "w": 521,
            "h": 680,
            "resize": "fit"
          },
          "thumb": {
            "w": 150,
            "h": 150,
            "resize": "crop"
          },
          "medium": {
            "w": 919,
            "h": 1200,
            "resize": "fit"
          },
          "large": {
            "w": 1284,
            "h": 1677,
            "resize": "fit"
          }
        }
      }
    ]
  },
  "extended_entities": {
    "media": [
      {
        "id": 1575918178261254162,
        "id_str": "1575918178261254162",
        "indices": [
          77,
          100
        ],
        "media_url":
"http://pbs.twimg.com/media/Fd7JO8pXwBI9HPw.jpg",
        "media_url_https":
"https://pbs.twimg.com/media/Fd7JO8pXwBI9HPw.jpg",
        "url": "https://t.co/VROTxNS9rz",
        "display_url": "pic.twitter.com/VROTxNS9rz",
        "expanded_url":
"https://twitter.com/suzjdean/status/1575918182698979328/photo/1",
        "type": "photo",
        "sizes": {
          "small": {
            "w": 521,
            "h": 680,
            "resize": "fit"
```

```
          },
          "thumb": {
            "w": 150,
            "h": 150,
            "resize": "crop"
          },
          "medium": {
            "w": 919,
            "h": 1200,
            "resize": "fit"
          },
          "large": {
            "w": 1284,
            "h": 1677,
            "resize": "fit"
          }
        }
      }
    ]
  },
  "metadata": {
    "iso_language_code": "en",
    "result_type": "recent"
  },
  "source": "<a href=\"http://twitter.com/download/iphone\"
rel=\"nofollow\">Twitter for iPhone</a>",
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 28709505,
    "id_str": "28709505",
    "name": "Suz□",
    "screen_name": "suzjdean",
    "location": "Charleston, SC & DC",
    "description": "MY #NATS #Caps #gamecocks family! I stand with
#Ukraine □□ IG: sjdean74",
    "url": null,
    "entities": {
      "description": {
        "urls": [

        ]
      }
    },
    "protected": false,
    "followers_count": 3811,
    "friends_count": 2868,
    "listed_count": 74,
    "created_at": "Sat Apr 04 01:35:19 +0000 2009",
    "favourites_count": 320543,
    "utc_offset": null,
    "time_zone": null,
```

```
    "geo_enabled": true,
    "verified": false,
    "statuses_count": 165706,
    "lang": null,
    "contributors_enabled": false,
    "is_translator": false,
    "is_translation_enabled": false,
    "profile_background_color": "0099B9",
    "profile_background_image_url":
"http://abs.twimg.com/images/themes/theme4/bg.gif",
    "profile_background_image_url_https":
"https://abs.twimg.com/images/themes/theme4/bg.gif",
    "profile_background_tile": false,
    "profile_image_url":
"http://pbs.twimg.com/profile_images/1513709638323257351/NuehKDmA_no
rmal.jpg",
    "profile_image_url_https":
"https://pbs.twimg.com/profile_images/1513709638323257351/NuehKDmA_n
ormal.jpg",
    "profile_banner_url":
"https://pbs.twimg.com/profile_banners/28709505/1649038002",
    "profile_link_color": "0099B9",
    "profile_sidebar_border_color": "FFFFFF",
    "profile_sidebar_fill_color": "95E8EC",
    "profile_text_color": "3C3940",
    "profile_use_background_image": true,
    "has_extended_profile": true,
    "default_profile": false,
    "default_profile_image": false,
    "following": false,
    "follow_request_sent": false,
    "notifications": false,
    "translator_type": "none",
    "withheld_in_countries": [

    ]
  },
  "geo": null,
  "coordinates": null,
  "place": {
    "id": "6057f1e35bcc6c20",
    "url":
"https://api.twitter.com/1.1/geo/id/6057f1e35bcc6c20.json",
    "place_type": "admin",
    "name": "South Carolina",
    "full_name": "South Carolina, USA",
    "country_code": "US",
    "country": "United States",
    "contained_within": [

    ],
    "bounding_box": {
      "type": "Polygon",
      "coordinates": [
        [
```

```
            [
              -83.353955,
              32.04683
            ],
            [
              -78.499301,
              32.04683
            ],
            [
              -78.499301,
              35.215449
            ],
            [
              -83.353955,
              35.215449
            ]
          ]
        ]
      },
      "attributes": {

      }
    },
    "contributors": null,
    "is_quote_status": false,
    "retweet_count": 0,
    "favorite_count": 0,
    "favorited": false,
    "retweeted": false,
    "possibly_sensitive": false,
    "lang": "en"
}
```