# Topic Model Advancements

What has been done and should be done

# Tweets preprocessing

Use Crisis_12 as an example:

Original:

| | Tweets | partition | Topics |
|---|---|---|---|
| **0** | One in eight new apartment buildings in New Yo... | test | floods |
| **1** | jhopesgalaxy mumbai gon flood soon | train | floods |
| **2** | Space! Storms! Cyclone Debbie seen from spac... | train | hurricanes |
| **3** | Htc wildfire mutable segmental phoneme inanima... | train | forestfires |
| **4** | Lismore, Northern Rivers hit in #TCDebbie afte... | train | hurricanes |
| **...** | ... | ... | ... |
| **7995** | #RT #Follow Fort McMurray Wildfire in Alberta ... | train | forestfires |
| **7996** | Massive cyclone makes landfall in northeastern... | train | hurricanes |
| **7997** | who felt that??? #earthquake #sanfrancisco #sf... | train | earthquakes |
| **7998** | KTVU we just had a earthquake in oakland ca at... | dev | earthquakes |
| **7999** | I felt that... #earthquake | train | earthquakes |

8000 rows × 3 columns

# Tweets preprocessing

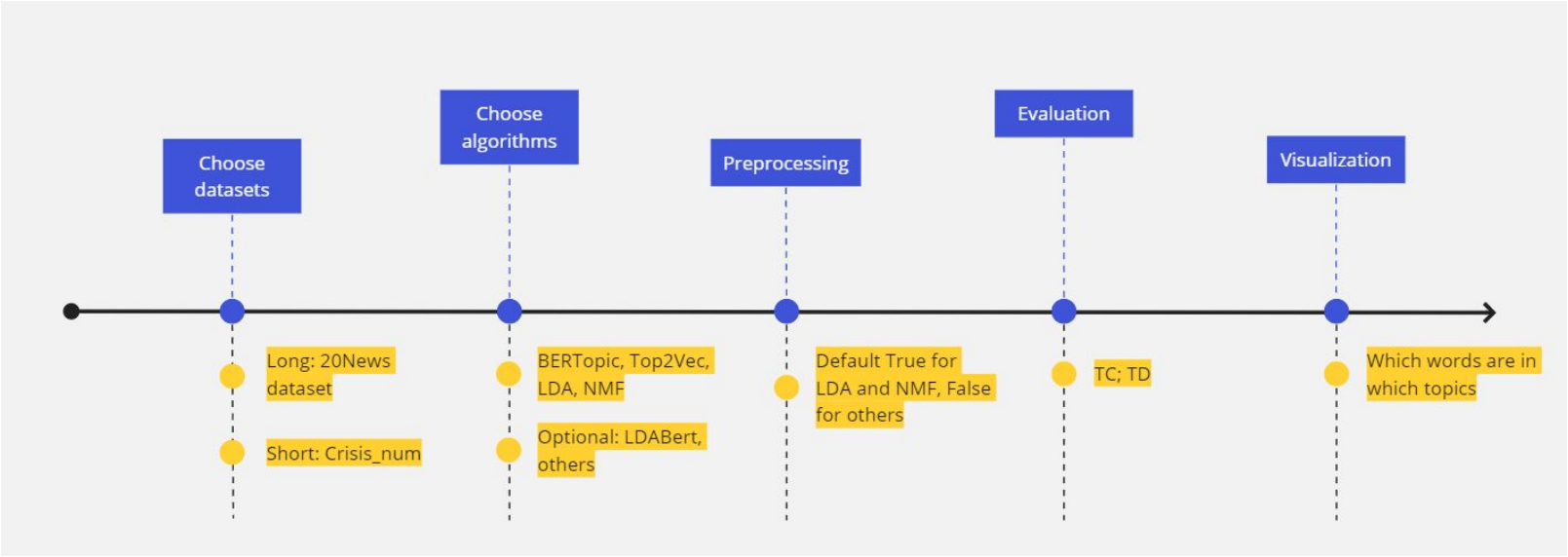
Preprocessed:

(Including lower case, remove urls, stop words, lemmatization, remove short sentences etc.)

| | Tweets | partition | Topics |
|---|---|---|---|
| **0** | one eight new apartment building new york buil... | test | floods |
| **1** | jhopesgalaxy mumbai gon flood soon | train | floods |
| **2** | space storm cyclone debbie seen space | train | hurricanes |
| **3** | htc wildfire mutable segmental phoneme inanima... | train | forestfires |
| **4** | lismore northern river hit aftermath nsw premi... | train | hurricanes |
| **...** | ... | ... | ... |
| **7982** | sydney weather wet wild condition ahead ex cyc... | train | hurricanes |
| **7983** | jeffflake pas law bum stop flood murderer rapi... | train | floods |
| **7984** | fort mcmurray wildfire alberta grows firefight... | train | forestfires |
| **7985** | massive cyclone make landfall northeastern aus... | train | hurricanes |
| **7987** | ktvu earthquake oakland ca pm | dev | earthquakes |

7306 rows × 3 columns

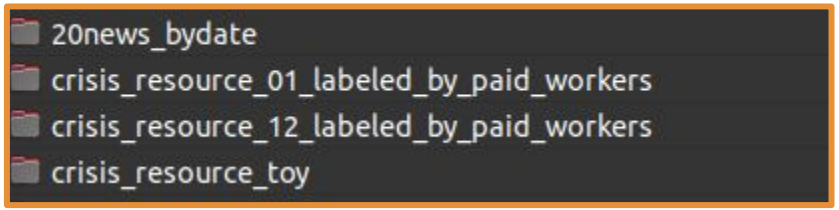
# Code integration plan



# Research as a Team

- LDA and NMF: Link
- BERTopic: Link
- Top2Vec: Link

# Datasets and Generic Dataset Loader




```python
def load_documents(dataset_dir: str, dataset_text_col: str) -> List[str]:
    if '20news_bydate' in dataset_dir:
        dataset_data_path = [path for path in Path(dataset_dir).iterdir() if path.suffix == '.pkz'][0]
        decompressed_pkl = zlib.decompress(open(dataset_data_path, 'rb').read())

        data = pkl.loads(decompressed_pkl)
        return data['train'].data

    dataset_data_paths = [path for path in Path(dataset_dir).iterdir() if path.suffix in {'.csv', '.tsv'}]
    dfs = []
    for data_path in dataset_data_paths:
        csv_delimiter = '\t' if data_path.suffix == '.tsv' else ','
        df = pd.read_csv(data_path, delimiter=csv_delimiter)
        # print(f'[INFO] Dataset from "{data_path}":', tabulate(df.head(5), headers="keys", tablefmt="psql"), sep='\n')
        dfs.append(df)

    merged_df = pd.concat(dfs, axis=0)
    documents = list(map(lambda doc: '' if pd.isna(doc) else doc, merged_df[dataset_text_col]))  # Replace nan with ''
    return documents
```

**Will be added:** Yahoo Dataset and other CRISIS short text datasets

# Model Output Design

**Doc-Topic Table:**

| Document ID | Document | Real Label | Assigned Topic ID | Assignment Score |
|---|---|---|---|---|
| 0 | omg an earthquake happened | earthquake | 2 | 0.9141560793 |
| 1 | Was that an earthquake lmao | earthquake | 2 | 0.8884242773 |
| 2 | Did we just have an earthquake? #concord #earthquake | earthquake | 2 | 0.8798669577 |
| 3 | forests are burning, help!? | wildfires | 3 | 0.8726007938 |

**Topic-Word Table:**

| method | method_specific_params | dataset_name | data_col | num_given_topics | reduced | topic_num | topic_size | topic_words | word_scores | num_detected_topics | num_final_topics | duration_secs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| top2vec | {'speed': 'fast-learn', 'embedding_model': 'doc2vec'} | crisis_resource_toy | text | 4 | FALSE | 0 | 204 | ['in' 'that' 'just' 'wildfire' 'earthquake' 'cyclone' 'flood' 'is' 'and' 'my' 'from' 'debbie' 'https' 'smoke' 'co' 'the' 'of' 'to' 'was' 'it' 'this'] | [ 0.0893354 0.08309343 0.07133193 0.06872502 0.05634531 0.05487543 0.05098582 0.03539272 0.03511589 0.02916614 0.01412737 0.00634522 0.00464096 0.00405226 | 2 | 2 | 9.47 |
| top2vec | {'speed': 'fast-learn', 'embedding_model': 'doc2vec'} | crisis_resource_toy | text | 4 | FALSE | 1 | 192 | ['that' 'https' 'in' 'flood' 'cyclone' 'this' 'was' 'the' 'smoke' 'to' 'is' 'it' 'earthquake' 'and' 'from' 'co' 'just' 'of' 'debbie' 'wildfire' 'my'] | [ 0.1190422 0.09579101 0.09119737 0.06858488 0.04900631 0.04673633 0.0354714 0.02776171 0.02330941 0.00390079 0.00081396 -0.00534165 -0.00726463 -0.01423862 -0.01439915 -0.01928411 -0.02175191 -0.02594034 -0.0530066 -0.07003934 -0.07053743] | 2 | 2 | 9.47 |

**Will be added:** More method specific parameters

# Top2Vec - Extracted Tunable Parameters Pt.1

**dataset_dir:** Dataset Directory

**min_count:** Set in the Top2Vec paper. Ignores all words with total frequency lower than this. For smaller corpora a smaller min_count is necessary. NOTE: This value largely depends on corpus size and its vocabulary.

**embedding_model:** Embedding model for the part where semantic relationships of the data are being learned. Options: [ `doc2vec` , `universal-sentence-encoder` , `universal-sentence-encoder-large` , `universal-sentence-encoder-multilingual` , `universal-sentence-encoder-multilingual-large` , `distiluse-base-multilingual-cased` , `all-MiniLM-L6-v2` , `paraphrase-multilingual-MiniLM-L12-v2` ]

**umap_args:**

```
n_neighbors:    Set in the Top2Vec paper. The size of local neighborhood (in terms of number of neighboring
                sample points) used for manifold approximation. Larger values result in more global views of the
                manifold, while smaller values result in more local data being preserved. In general values
                should be in the range 2 to 100.

n_components:   Set in the Top2Vec paper. the dimension of the space to embed into. This defaults to 2 to
                provide easy visualization, but can reasonably be set to any integer value in the range 2 to
                100.

metric:     Set in the Top2Vec paper. Options: [`euclidean`, `manhattan`, `chebyshev`, `minkowski`,
            `canberra`, `braycurtis`, `mahalanobis`, `wminkowski`, `seuclidean`, `cosine`, `correlation`,
            `haversine`, `hamming`, `jaccard`, `dice`, `russelrao`, `kulsinski`, `ll_dirichlet`,
            `hellinger`, `rogerstanimoto`, `sokalmichener`, `sokalsneath`, `yule`].
```

# Top2Vec - Extracted Tunable Parameters Pt.2

**hdbscan_args:**

```
min_cluster_size:   Set in the Top2Vec paper. The minimum size of clusters; single linkage splits that contain
                    fewer points than this will be considered points "falling out" of a cluster rather than a
                    cluster splitting into two new clusters.

metric:     Set in the Top2Vec paper. The metric to use when calculating distance between instances in a feature
            array. If metric is a string or callable, it must be one of the options allowed by
            metrics.pairwise.pairwise_distances for its metric parameter. If metric is "precomputed", X is
            assumed to be a distance matrix and must be square. Options: [`cosine`, `euclidean`, `haversine`,
            `l2`, `l1`, `manhattan`, `precomputed`, `nan_euclidean`].

cluster_selection_method:   Set in the Top2Vec paper. The method used to select clusters from the condensed
                            tree. The standard approach for HDBSCAN* is to use an Excess of Mass algorithm
                            to find the most persistent clusters. Alternatively you can instead
                            select the clusters at the leaves of the tree -- this provides the
                            most fine-grained and homogeneous clusters. Options: [`eom`,`leaf`].
```

**doc2vec_speed:** This parameter is only used when using doc2vec as embedding_model. Options: [ `fast-learn` , `learn` , `deep-learn` ]

**num_topics:** Given number of topics. If model can reduce the number of topics, it can reduce to num_topics.

**data_col:** Data column of the given datasets. For 20newsgroup dataset, it is redundant.

# Top2Vec - Extracted Tunable Parameters  Application

```python
if __name__ == '__main__':
    args = {
        'dataset_dir': './data/crisis_resource_toy',
        'data_col': 'text',
        'num_topics': 4,

        # ####### Top2Vec Specific Arguments #########
        'embedding_model': 'doc2vec',
        'doc2vec_speed': 'fast-learn',
        'min_count': 50,
        'umap_args': {
            'n_neighbors': 15,
            'n_components': 5,
            'metric': 'cosine'
        },
        'hdbscan_args': {
            'min_cluster_size': 15,
            'metric': 'euclidean',
            'cluster_selection_method': 'eom'
        },
    }
    run(**args)
```

# Top2Vec - What Comes Next - Evaluation

**Evaluation Part (Draft)**

In [45]:

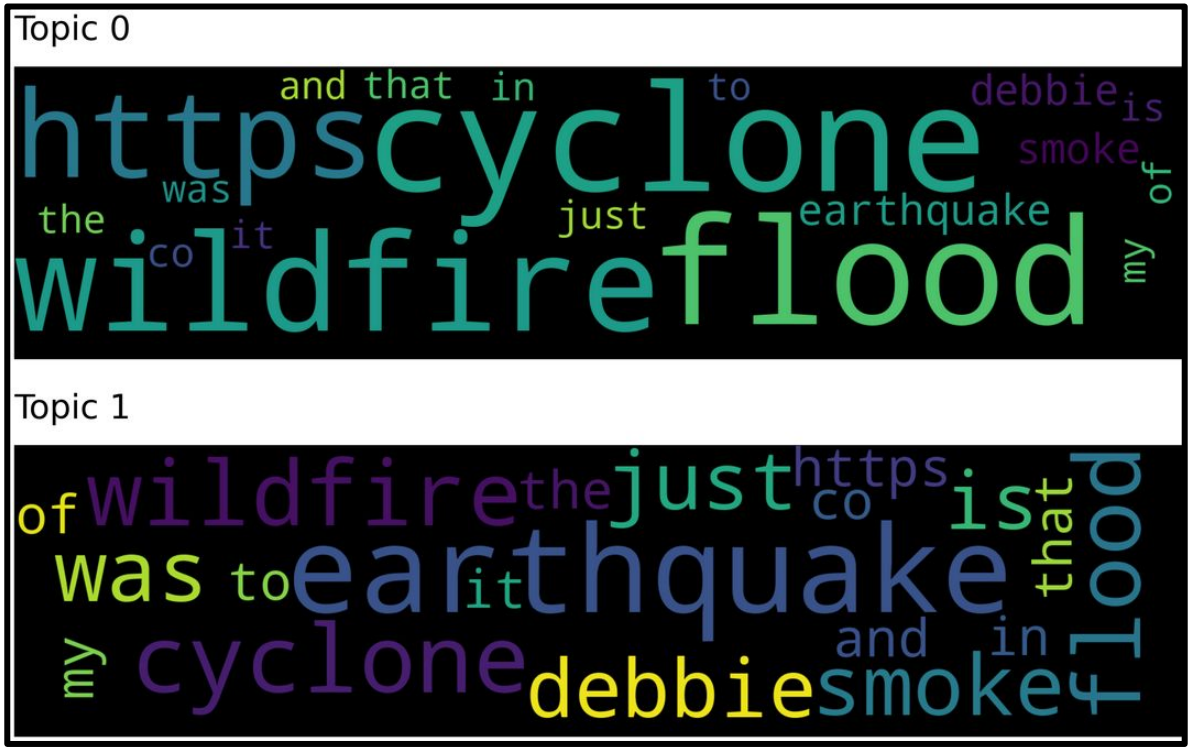
```
model_output = {"topics":[topic_stat['topic_words'] for topic_stat in topic_stats]}

from octis.evaluation_metrics.diversity_metrics import TopicDiversity
from octis.evaluation_metrics.coherence_metrics import Coherence
for topk_val in range(1,30):
    metric_diversity = TopicDiversity(topk=topk_val)
    metric_coherence = Coherence(texts = [d.split(" ") for d in documents], topk = topk_val, measure = "c_v")
    score_diversity = metric_diversity.score(model_output)
    score_coherence = metric_coherence.score(model_output)
    print(f'> topk={topk_val},score_diversity={"%.2f" % score_diversity},score_coherence={"%.2f" %score_coherence
```

```
> topk=1,score_diversity=1.00,score_coherence=1.00
> topk=2,score_diversity=0.75,score_coherence=0.33
> topk=3,score_diversity=0.67,score_coherence=0.26
> topk=4,score_diversity=0.62,score_coherence=0.40
> topk=5,score_diversity=0.60,score_coherence=0.40
> topk=6,score_diversity=0.58,score_coherence=0.40
> topk=7,score_diversity=0.57,score_coherence=0.40
> topk=8,score_diversity=0.62,score_coherence=0.40
> topk=9,score_diversity=0.61,score_coherence=0.40
> topk=10,score_diversity=0.55,score_coherence=0.40
> topk=11,score_diversity=0.55,score_coherence=0.40
> topk=12,score_diversity=0.54,score_coherence=0.40
> topk=13,score_diversity=0.54,score_coherence=0.40
> topk=14,score_diversity=0.54,score_coherence=0.40
> topk=15,score_diversity=0.53,score_coherence=0.40
> topk=16,score_diversity=0.56,score_coherence=0.40
> topk=17,score_diversity=0.56,score_coherence=0.40
> topk=18,score_diversity=0.53,score_coherence=0.40
> topk=19,score_diversity=0.50,score_coherence=0.40
```

**Note:** It is planned in the Generic Evaluation Module Design

# Top2Vec - What Comes Next - Visualization Module



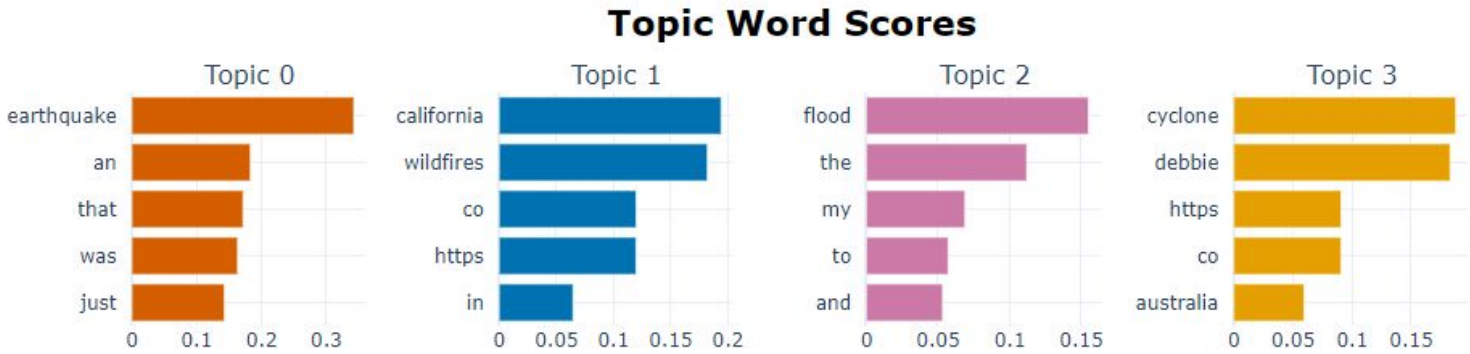

**Note:** It is planned in the Generic Visualization Module Design

# BERTopic - A short glimpse

Comparison for BERTopic on Crisis #12:

| | embedding_model | duration | n_topics | min_docs_per_topic | n_gram_range | topic_diversity | topic_coherence | duration_readable |
|---|---|---|---|---|---|---|---|---|
| 0 | all-MiniLM-L6-v2 | 140.808688 | 20 | 10 | (1, 1) | 0.635 | 0.481375 | 2min 20sec |
| 1 | all-MiniLM-L12-v2 | 268.787493 | 20 | 10 | (1, 1) | 0.625 | 0.439895 | 4min 28sec |
| 2 | all-distilroberta-v1 | 458.462295 | 20 | 10 | (1, 1) | 0.650 | 0.468073 | 7min 38sec |
| 3 | all-mpnet-base-v2 | 884.003001 | 20 | 10 | (1, 1) | 0.625 | 0.464636 | 14min 44sec |

Example Topics for Crisis #12:



# BERTopic - A short glimpse 2

Problem: with only a few topics a lot of the tweets are classified as noise

| | Topic | Count | Name |
|---|---|---|---|
| 0 | -1 | 6155 | -1_co_https_the_to |
| 1 | 0 | 715 | 0_earthquake_an_that_was |
| 2 | 1 | 479 | 1_california_wildfires_co_https |
| 3 | 2 | 328 | 2_flood_the_my_to |
| 4 | 3 | 323 | 3_cyclone_debbie_https_co |

Example docs for topic 0:

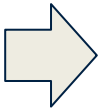
```
['#earthquake in #sanfrancisco just felt it in #southbeach. Shook the 7th floor pretty good...',
 'Anyone feel an #earthquake in #SanFrancisco just now?',
 'Umm... was that an earthquake I just felt  ? #SanFrancisco',
 'earthquake ?',
 'Uhhh earthquake?',
 'UHH earthquake??',
 'I think we just had a little earthquake.',
 'so we just had an earthquake Um',
 'We just dead ass had a earthquake',
 'Who felt that earthquake?']
```

# Topic Model Evaluation Research

- Topic coherence measures: “best” measure in the sense of correlation with human judgement
  -> C_v: big sliding window, word with whole window, indirect NPMI and mean
  (see: “Exploring the Space of Topic Coherence Measures”)

- Top2vec: compares different algs with Topic Information Gain and fixed #topics in stepsizes

- BERTopic: compares algs with Topic coherence (NPMI) and Topic diversity (% of unique words)



NEED not only metric evaluation but human judgment (with visualization) as well

# Our Questions on Clickup