“Strax” — Advancements in Topic Modeling

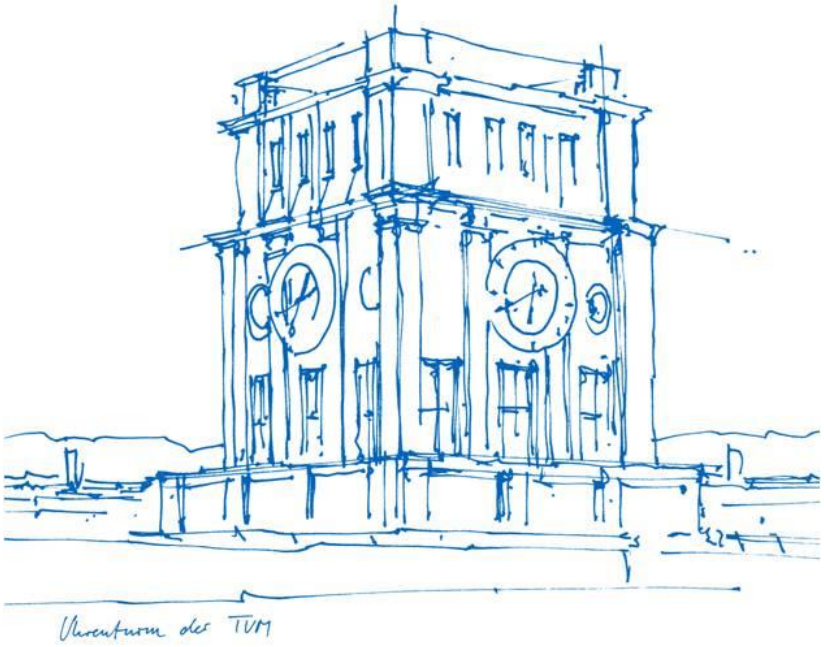
Technische Universität München

Fakultät für Informatik

NLP Lab Course, SS22

22.06.2022

Berk Sudan, Ferdinand Kapl, Yuyin Lang



Overview

- Datasets Exploration (Milestone #1)
- Preprocessing Steps (Milestone #2)
- Algorithms (Milestone #3 & #4) with Evaluation
  - LDA/NMF
  - Top2Vec & Visualization
  - BERTopic & Visualization
- Comparison (Milestone #6)
- Road Map

Datasets Exploration (Milestone # 1)

Short dataset

Resource Name	Is Suitable?	Type	Contains Tweet Text?	Event Count	Event Distribution
CRISIS NLP - Resource #01	Yes	Short Text Dataset	Yes	12	(2K - 2K - 2K - 2K - 2K - 1K - 2K - 1K - 2K - 2K - 2K - 3K)
CRISIS NLP - Resource #02	Yes	Short Text Dataset	Yes	7	(1K - 1K - 2K - 1K - 1K - 13K - 5K)
CRISIS NLP - Resource #03	Yes	Short Text Dataset	Yes	6	(2K - 1K - 9K - 1K - 2K - 2K)
CRISIS NLP - Resource #04	No	A Tool for LSTM RNNs	-	-	-
CRISIS NLP - Resource #05	Yes	Short Text Dataset	Yes	7	(1K - 4K - 4K - 4K - 0.5K - 1K - 1K)
⋮					
CRISIS NLP - Resource #12	Yes	Short Text Dataset for Eye Witness	Yes	4	(2K - 2K - 2K - 2K)
CRISIS NLP - Resource #13	No	Image Dataset	-	-	-
⋮					

Datasets Exploration (Milestone # 1)

index	Tweets
0	RT @diplo: Twerkbook pro #plurnt #earthquake http://t.co/5x5ya6wxF6
1	In @BBCUrdu #Balochistan EarthQuake - program NDMA's Military Official Continue to Refuse accepting outside help http://t.co/Xs8wK4gIP7 …
2	RT @ErumManzoor: People who wanna help #earthquake affectees in Baluchistan can contact @AsimBajwaISPR
3	MT @ARYNEWSOFFICIAL: #RedCrescent delivers relief in Awaran http://t.co/3KM9VFUGIO #Pakistan #earthquake #redcross
4	Another two islands emerge off coast http://t.co/iAI0mbVna2 via @zite #Earthquake

- They are tweets
- Often less than 30 words
- Contains lots of websites and urls

Datasets Exploration (Milestone # 1)

Long dataset

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

- 20 Newsgroups
- Roughly grouped into 6 big topics
- May overlap between small topics

Preprocessing steps (Milestone # 2)

For crisis dataset

```
def to_lowercase(text):
    def standardize_accented_chars(text):
        def remove_url(text):
            def expand_contractions(text):
                def remove_mentions_and_tags(text):
                    def remove_stop_words(text):
                        def lemmatize(text):
```

Specially for OCTIS initialization

```
def df_to_vocab(df):
    def list_to_vocab(word_list):
        tweets_df.to_csv(f'{save_dir}/corpus_with_header_reduced.tsv',\
                         sep = '\t', index=False)
```

Algorithms (Milestone # 3 & 4)

- LDA / NMF
- Top2Vec
- BERTopic

LDA / NMF

LDA/NMF on crisis dataset 12

Doc-Topic Output

LDA:

	Document	Real Label	Assigned Topic Num	Assignment Score
0	unprecedented wildfire break northern southern...	forestfires	forestfires	0.700985
1	heatwaves flood warning going place thought to...	floods	floods	0.737232
2	wildfire devastate northern southern	forestfires	forestfires	0.848227
3	uploaded month patreon little earlier case cyc...	hurricanes	floods	0.662079
4	morrison still extreme wildfire behaviour	forestfires	forestfires	0.723109

NMF:

	Document	Real Label	Assigned Topic Num	Assignment Score
0	unprecedented wildfire break northern southern...	forestfires	forestfires	1.000000
1	heatwaves flood warning going place thought to...	floods	floods	0.883809
2	wildfire devastate northern southern	forestfires	forestfires	1.000000
3	uploaded month patreon little earlier case cyc...	hurricanes	hurricanes	0.992597
4	morrison still extreme wildfire behaviour	forestfires	forestfires	0.999595

Gives higher scores than LDA

LDA/NMF on crisis dataset 12

Topic-Word Output

LDA:

	num_given_topics	topic_num	topic_words
0	4	0	[wildfire, california, flood, fort, mcmurray, ...
1	4	1	[flood, earthquake, flash, like, warning, goin...
2	4	2	[earthquake, flood, wildfire, felt, california...
3	4	3	[cyclone, debbie, ex, flood, queensland, schoo...

NMF:

	num_given_topics	topic_num	topic_words
0	4	0	[cyclone, debbie, ex, queensland, school, trop...
1	4	1	[earthquake, flood, like, felt, feel, home, wa...
2	4	2	[flood, rain, flash, queensland, amp, time, fl...
3	4	3	[wildfire, california, fort, mcmurray, fire, c...

LDA/NMF on crisis dataset 12

Documents in topics

LDA:

```
Topic 0:
hurricanes     1253
floods          449
forestfires     347
earthquake       44
Name: Real Label, dtype: int64
--------------------------------
Topic 1:
forestfires    1136
floods          498
hurricanes      265
earthquake      116
Name: Real Label, dtype: int64
--------------------------------
Topic 2:
floods         555
hurricanes     233
forestfires    230
earthquake     140
Name: Real Label, dtype: int64
--------------------------------
Topic 3:
earthquake     1195
floods          474
forestfires     191
hurricanes      180
Name: Real Label, dtype: int64
--------------------------------
```

NMF:

```
Topic 0:
floods         1910
hurricanes      153
forestfires      30
earthquake        4
Name: Real Label, dtype: int64
--------------------------------
Topic 1:
earthquake     1481
forestfires     183
hurricanes      174
floods           29
Name: Real Label, dtype: int64
--------------------------------
Topic 2:
forestfires    1682
hurricanes      116
floods           29
earthquake        9
Name: Real Label, dtype: int64
--------------------------------
Topic 3:
hurricanes     1488
forestfires       9
floods            8
earthquake        1
Name: Real Label, dtype: int64
--------------------------------
```

LDA/NMF on 20 News Dataset

Doc-Topic Output

LDA:

	Document	Real Label	Assigned Topic Num	Assignment Score
0	fax modem card sell mail	misc.forsale	10	0.728531
1	run server server install run add	comp.windows.x	12	0.668365
2	live part lead wait important remember judge j...	soc.religion.christian	0	0.499256
3	doesn pain deserve die lie rape	talk.religion.misc	0	0.538701
4	sale mile good condition good condition player...	rec.autos	13	0.676404

NMF:

	Document	Real Label	Assigned Topic Num	Assignment Score
0	fax modem card sell mail	misc.forsale	3	0.479223
1	run server server install run add	comp.windows.x	0	0.656878
2	live part lead wait important remember judge j...	soc.religion.christian	15	0.634671
3	doesn pain deserve die lie rape	talk.religion.misc	17	0.373880
4	sale mile good condition good condition player...	rec.autos	9	0.218945

LDA/NMF on 20 News Dataset

Topic-Word Output

Not very informative :(

LDA:

	num_given_topics	topic_num	topic_words
0	20	0	[people, time, make, child, give, religion, ma...
1	20	1	[people, government, gun, state, law, weapon, ...
2	20	2	[people, make, time, blue, work, ticket, hand,...
3	20	3	[man, homosexual, make, light, people, good, t...
4	20	4	[key, chip, encryption, make, clipper, post, s...
5	20	5	[armenian, space, turkish, launch, year, russi...
6	20	6	[color, driver, bit, mode, card, run, work, di...

NMF:

	num_given_topics	topic_num	topic_words
0	20	0	[server, include, base, support, send, widget,...
1	20	1	[system, user, list, information, internet, ap...
2	20	2	[system, key, question, atheist, post, argumen...
3	20	3	[graphic, image, mail, send, package, format, ...
4	20	4	[government, turkish, russian, administration,...
5	20	5	[window, widget, call, application, subject, s...
6	20	6	[privacy, key, internet, encryption, computer,...

NMF on 20 News Dataset

Documents in topics

Topic 1:	
comp.os.ms-windows.misc	17
comp.sys.ibm.pc.hardware	14
comp.windows.x	14
comp.sys.mac.hardware	14
sci.space	12
rec.motorcycles	9
misc.forsale	9
sci.electronics	7
comp.graphics	6

Topic 11:	
comp.graphics	27
comp.os.ms-windows.misc	26
comp.windows.x	21
comp.sys.ibm.pc.hardware	8
sci.crypt	6
sci.electronics	5
comp.sys.mac.hardware	4

Topic 6:	
sci.crypt	22
sci.electronics	4
misc.forsale	2
comp.os.ms-windows.misc	1
comp.sys.ibm.pc.hardware	1
sci.med	1

Topic 8:	
sci.space	39
misc.forsale	6
rec.autos	4
rec.motorcycles	4
rec.sport.baseball	3
talk.politics.misc	2

Topic 12:	
talk.politics.guns	41
rec.autos	8
talk.politics.misc	7
sci.crypt	6
rec.motorcycles	6
sci.electronics	4

NMF on 20 News Dataset

Documents in topics

```
Topic 16:
comp.os.ms-windows.misc     4
comp.sys.ibm.pc.hardware    3
sci.crypt                   1
comp.windows.x              1
sci.electronics             1
misc.forsale                1
```

```
Topic 19:
sci.space                   2
soc.religion.christian      2
talk.politics.guns          2
rec.sport.baseball          1
rec.autos                   1
sci.electronics             1
sci.med                     1
Name: Real Label, dtype: int64
```

```
Topic 18:
comp.graphics               6
comp.sys.mac.hardware       3
misc.forsale                2
comp.sys.ibm.pc.hardware    2
comp.windows.x              2
comp.os.ms-windows.misc     1
```

Top2Vec

Top2Vec - Available Parameters

- **Dataset**
- **Minimum Topic Words:** Depends on corpus size and its vocabulary.
- **Embedding Model:** 8 models tested
- **umap_args:**
  - **# of Neighbours**
  - **# of Components**
  - **Distance Metric**
- **hdbscan_args:**
  - **Minimum Cluster Size**
  - **Distance Metric**
  - **Cluster Selection Method**
- **Number of Topics:** To force Topic Reduction

Top2Vec - Topic Assignments

	Document ID	Document	Real Label	Assigned Topic Num	Assignment Score
0	3802	This literally just made me SO ANGRY I was NOT heading into any flood-prone areas soooooooo https://t.co /R1KkbkBggH	floods	0	0.799604
1	3846	i hope the flood gets here soon https://t.co/m5xSmesu8e	floods	0	0.794808
2	2097	SimonMaloy JeremyMcLellan There s a flood here that s pretty bad. It hasn t flooded like this since hurricane Beu https://t.co/FOmjwUkBoy	floods	0	0.792194
3	3818	Flash Flood WARNING continues until 10am. Use caution getting around this morning. Watch for flooded roads. Get https://t.co/sGDfuH7mvj	floods	0	0.776224
4	3094	martycormack SeanBakerMN NWS Hi all, it doesn t feel quite right to hit like on these posts about the flood. I https://t.co/mARQBvjKrC	floods	0	0.772866

	Document ID	Document	Real Label	Assigned Topic Num	Assignment Score
2899	329	Wow just had a earthquake	earthquakes	1	0.914956
2900	821	Ohhh shit earthquake	earthquakes	1	0.914789
2901	615	ummm earthquake anyone??	earthquakes	1	0.913467
2902	794	Holy shit earthquake	earthquakes	1	0.913255
2903	1121	Um earthquake anyone?	earthquakes	1	0.912174

Parameters: Embedding=”universal-sentence-encoder-large”, Data=”CRISIS-12”

Assignment Score: The cosine similarity of the document and topic vector.

Top2Vec - Evaluation

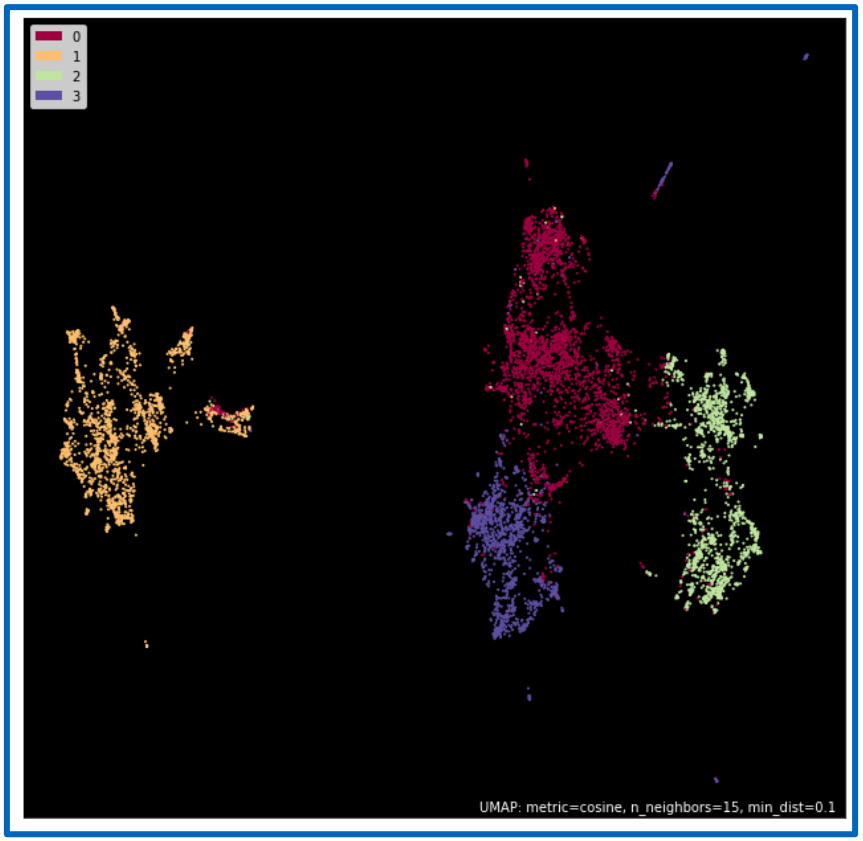
```
Topic 0:
floods_eyewitness_crowdflower_2000        1955
hurricanes_eyewitness_crowdflower_2000     526
forestfires_eyewitness_crowdflower_2000    358
earthquakes_eyewitness_crowdflower_2000     60
Name: Real Label, dtype: int64
--------------------------------
Topic 1:
earthquakes_eyewitness_crowdflower_2000   1926
floods_eyewitness_crowdflower_2000          11
forestfires_eyewitness_crowdflower_2000      1
Name: Real Label, dtype: int64
--------------------------------
Topic 2:
forestfires_eyewitness_crowdflower_2000   1640
floods_eyewitness_crowdflower_2000          13
earthquakes_eyewitness_crowdflower_2000      9
hurricanes_eyewitness_crowdflower_2000       6
Name: Real Label, dtype: int64
--------------------------------
Topic 3:
hurricanes_eyewitness_crowdflower_2000    1468
floods_eyewitness_crowdflower_2000          21
earthquakes_eyewitness_crowdflower_2000      5
forestfires_eyewitness_crowdflower_2000      1
Name: Real Label, dtype: int64
```

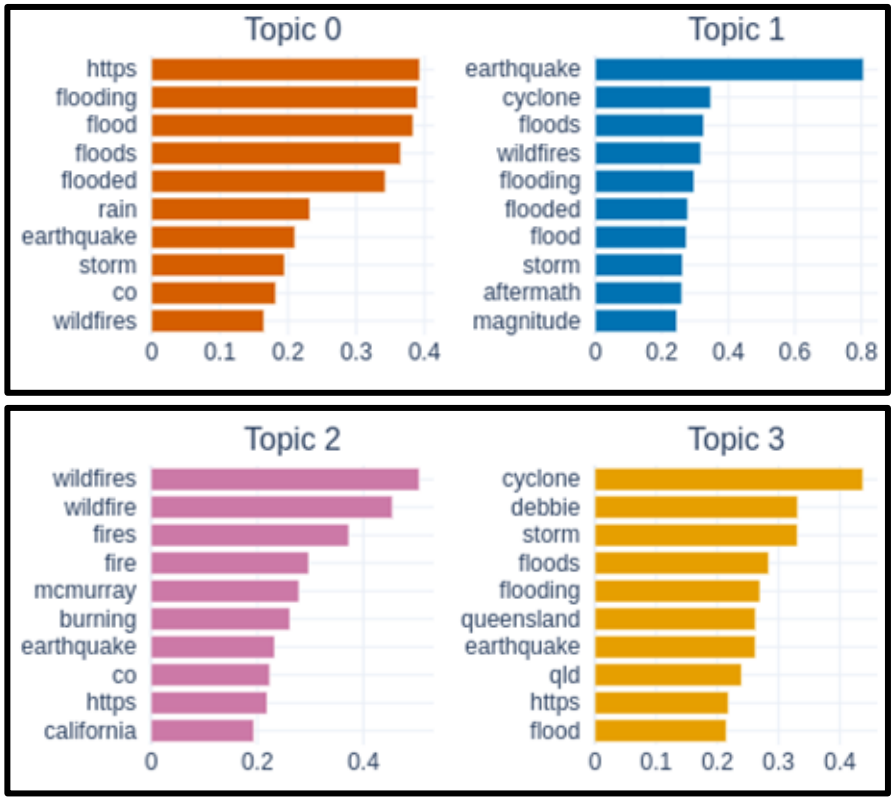
Diversity Score : 0.5500
Coherence Score: 0.4860

Parameters: Embedding=”universal-sentence-encoder-large”, Data=”CRISIS-12”

Assignment Score: The cosine similarity of the document and topic vector.

Top2Vec - Visualization





Parameters: Embedding=”universal-sentence-encoder-large”, Data=”CRISIS-12”

Word Score: The cosine similarity of the document and topic word.

Top2Vec - Evaluation - 20News

```
Topic 0:
rec.sport.hockey          569
rec.sport.baseball        533
talk.politics.misc          3
comp.sys.mac.hardware       2
sci.med                     2
sci.electronics             1
soc.religion.christian      1
sci.crypt                   1
misc.forsale                1
rec.motorcycles             1
Name: Real Label, dtype: int64
```

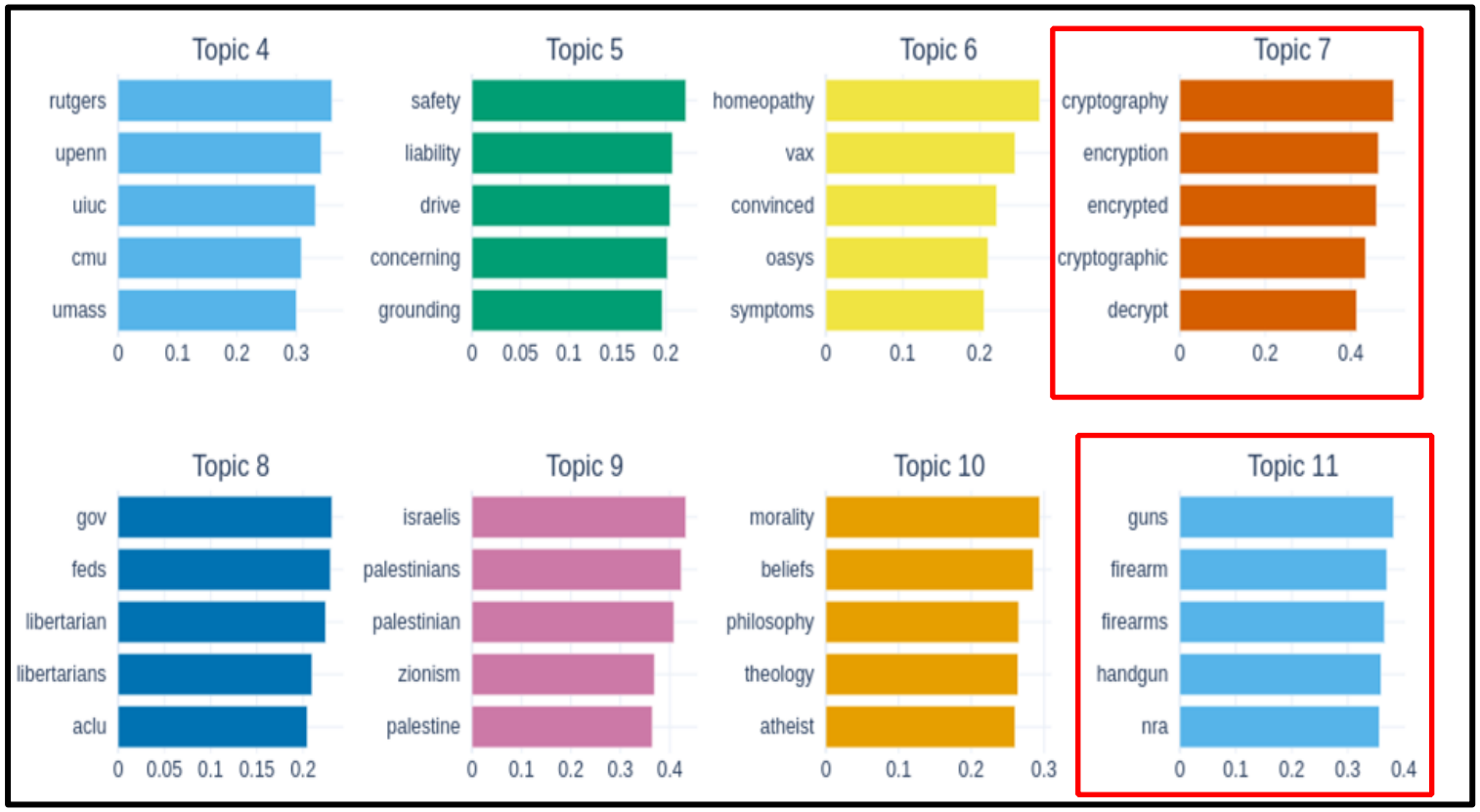
```
Topic 7:
alt.atheism               274
soc.religion.christian    140
talk.religion.misc        102
sci.med                    27
talk.politics.misc          9
talk.politics.mideast       6
sci.space                   6
talk.politics.guns          1
comp.graphics               1
sci.crypt                   1
Name: Real Label, dtype: int64
```

```
Topic 11:
comp.os.ms-windows.misc   217
comp.windows.x            137
comp.graphics              60
comp.sys.ibm.pc.hardware   30
sci.electronics            16
sci.crypt                  15
sci.space                  13
misc.forsale               12
comp.sys.mac.hardware      11
sci.med                     3
talk.politics.mideast       2
rec.sport.baseball          1
rec.autos                   1
alt.atheism                 1
talk.religion.misc          1
Name: Real Label, dtype: int64
```

Diversity Score : 0.8750
Coherence Score: 0.3797

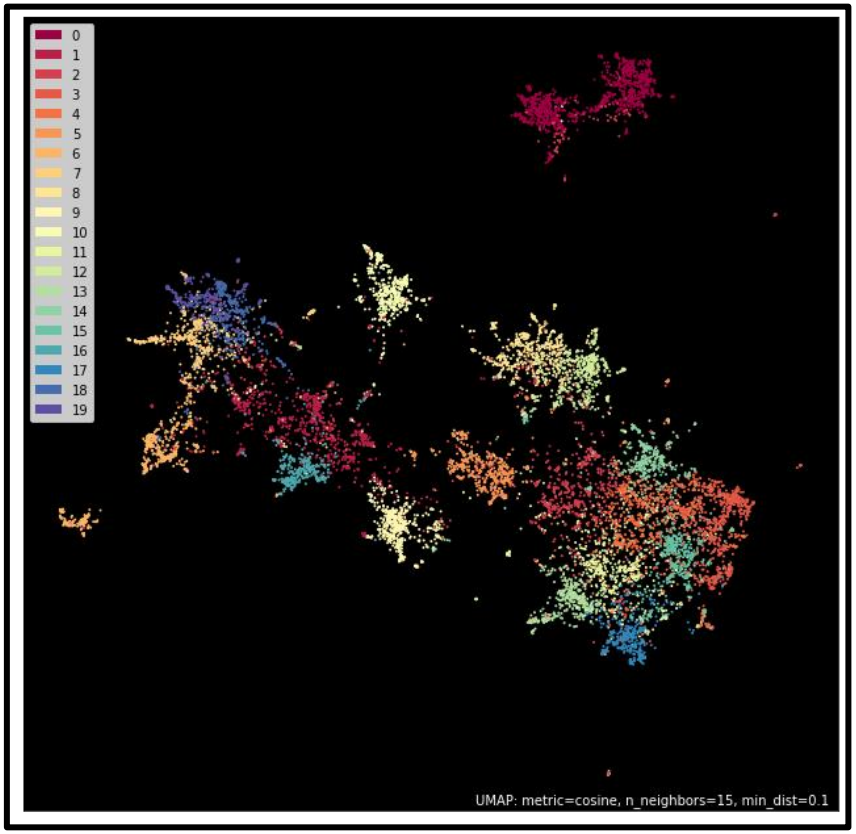
Observation: Similar labels in the same topic

Top2Vec - Visualization - 20News

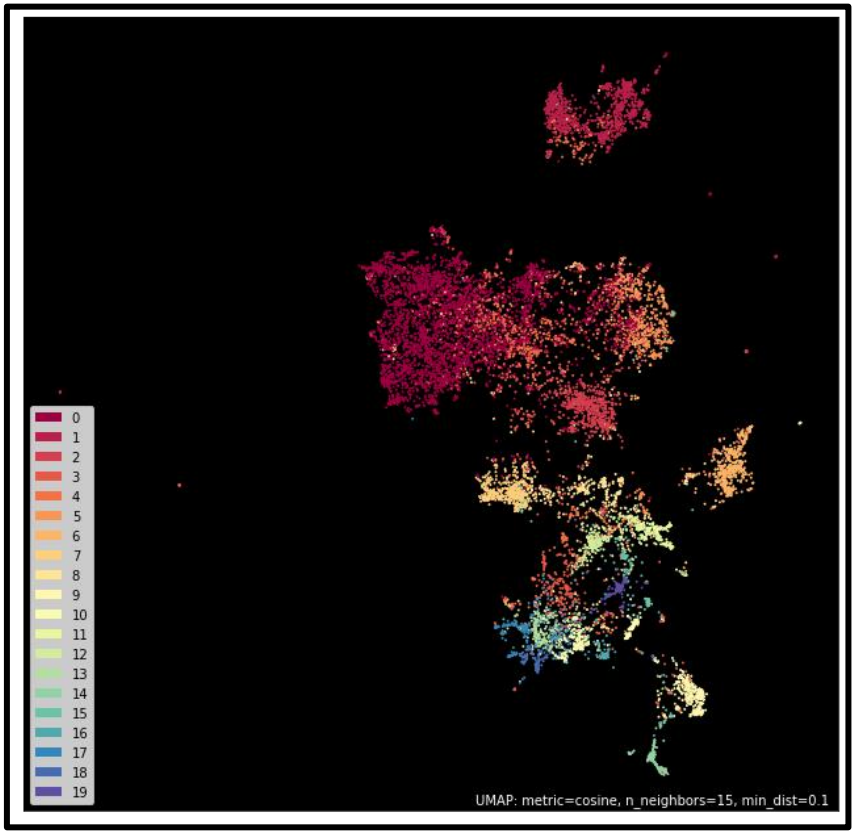


Observation: Similar words in the same topic

Top2Vec - Visualization - 20News



universal-sentence-encoder



all-MiniLM-L6-v2

BERTopic

BERTopic - Experiments

So far: four embedding models and the three datasets with & without preprocessed crisis datasets

	data	embedding_model	topic_diversity	topic_coherence
5	crisis_12	all-MiniLM-L12-v2	0.775	0.462321
15	crisis_12	all-mpnet-base-v2	0.775	0.423647
10	crisis_12	all-distilroberta-v1	0.750	0.478502
0	crisis_12	all-MiniLM-L6-v2	0.750	0.460125

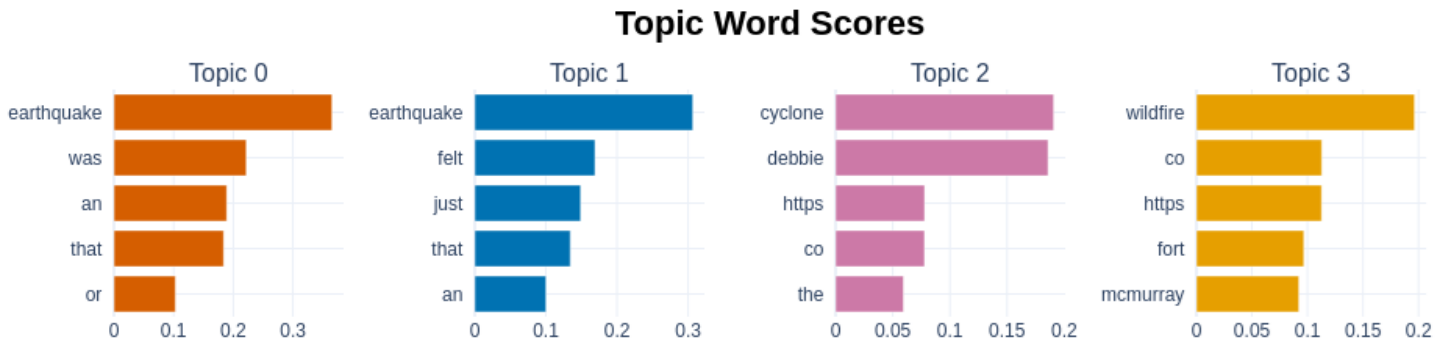
Often: No clear winner and performances are close (care: only one run)

BERTopic - Experiments: Best Results

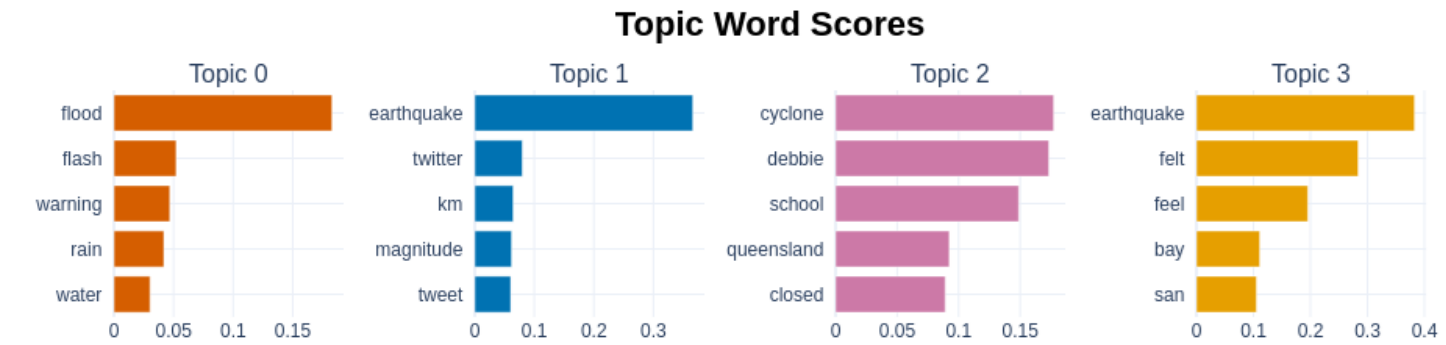
	data	embedding_model	topic_diversity	topic_coherence
0	20news	all-distilroberta-v1	0.460000	0.454753
1	crisis_1	all-mpnet-base-v2	0.633333	0.408661
4	crisis_1_preprocessed	all-distilroberta-v1	0.850000	0.699861
2	crisis_12	all-MiniLM-L12-v2	0.775000	0.462321
3	crisis_12_preprocessed	all-distilroberta-v1	0.975000	0.611101

Again: No clear winner ⇨ Different embeddings achieve best results

BERTopic - Compare Best Results on Crisis 12



Unprocessed Crisis 12



Preprocessed Crisis 12

Impression: Topic words for the preprocessed version are more descriptive

BERTopic - Results on Crisis 12: Doc Distribution

	Topic	Count	Name
0	-1	5919	-1_co_https_the_flood
1	0	579	0_earthquake_was_an_that
2	1	578	1_earthquake_felt_just_that
3	2	485	2_cyclone_debbie_https_co
4	3	439	3_wildfire_co_https_fort

Unprocessed Crisis 12

	Topic	Count	Name
0	-1	3483	-1_wildfire_debbie_cyclone_california
1	0	2055	0_flood_flash_warning_rain
2	1	810	1_earthquake_twitter_km_magnitude
3	2	554	2_cyclone_debbie_school_queensland
4	3	404	3_earthquake_felt_feel_bay

Preprocessed Crisis 12

Problem: Many docs classified as “noise” (especially for unprocessed dataset)

BERTopic - Results on Crisis 12: Documents

```
['Let\'s play "Earthquake or someone dropped a really big keg." My money\'s on earthquake...',
 'UHH earthquake??',
 'earthquake :0',
 'Was that the earthquake lol',
 'Was that an earthquake ???']
```

```
['A small earthquake just happened',
 'Small earthquake just now',
 'That was a nice little earthquake. Got my attention.',
 'who just felt that earthquake',
 'who the fuck just felt that earthquake']
```

Unprocessed Crisis 12

```
['Debbie the injured cockatoo is making a full recovery after getting caught in cyclone Debbie. #7News https://t.co/sbskOa38h
V',
 'Debbie, the cockatoo who lost her feathers during the cyclone that lashed Far North Queensland, has died. #9News https://t.c
o/EFkM7KYb02',
 'Battered Cyclone Debbie cockatoo loses fight https://t.co/dyCG0eQ7py',
 "Cyclone Debbie is a crazy betch but she got campus closed so I can't complain ",
 "dom followed me and dm'd me back, thx cyclone debbie for cancelling school"]
```

```
["Akshay T.U: Syrians who fled war now flee Canada's wildfire: Amid the  https://t.co/4x5LB8KNol #ChennaiInsider https://t.co/W
PciSDUCBw",
 'Syrian refugees aid Canadians caught in massive wildfire - Middle East Eye https://t.co/CJn3DuU2s0',
 'Syrian refugees in Canada step up to help Fort McMurray wildfire evacuees - CNN https://t.co/g7WPnWttbG',
 'The Massive Wildfire Burning in Alberta https://t.co/LvWsSJpT0W',
 "I was in Alberta three weeks ago. Wildfire smoke caused issues on one part of the highway, but I can't imagine what it's like
now. Scary!"]
```

BERTopic - Results on Crisis 12: Documents

```
[['guess explains last apartment consistently flooded',
  'aspinwallboro declared state emergency area still flood watch stay safe neighbor',
  'think harvey fuckin local would know drive fucking flood zone',
  'acosta showed total lack compassion thought like spike heel view flood area',
  'time thought filthy vile disgusting democrat want flood country ple'],
 ['dat earthquake though',
  'ooo another earthquake',
  'omg tiny earthquake',
  'either house shook reason fucking huge earthquake',
  'building shook idea earthquake'],
 ['school closed brisbane brace nightmare cyclone debbie wake',
  'school closed brisbane brace nightmare cyclone debbie wake',
  'school closed brisbane brace nightmare cyclone debbie wake via',
  'southeast queensland school closed due bad weather caused cyclone debbie gt gt',
  'work amp school vinaka cyclone debbie'],
 ['yoo anyone feel earthquake',
  'oh snap felt earthquake',
  'jus felt earthquake',
  'first earthquake ever felt',
  'pretty sure experienced first earthquake']]
```

Preprocessed Crisis 12

Comparison

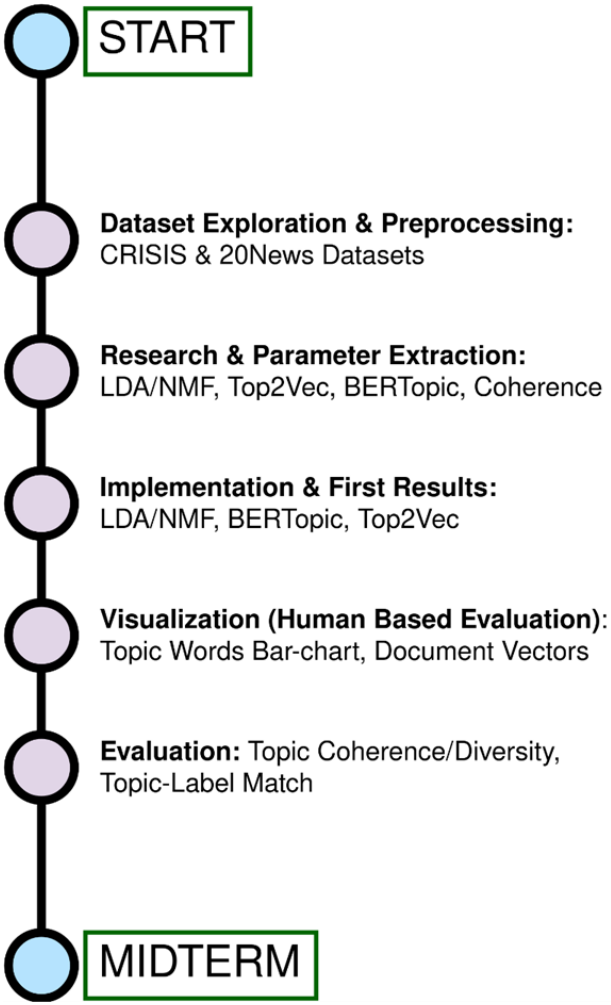
* Uses preprocessed data; other algs unprocessed (BERTopic preprocessed in parenthesis)

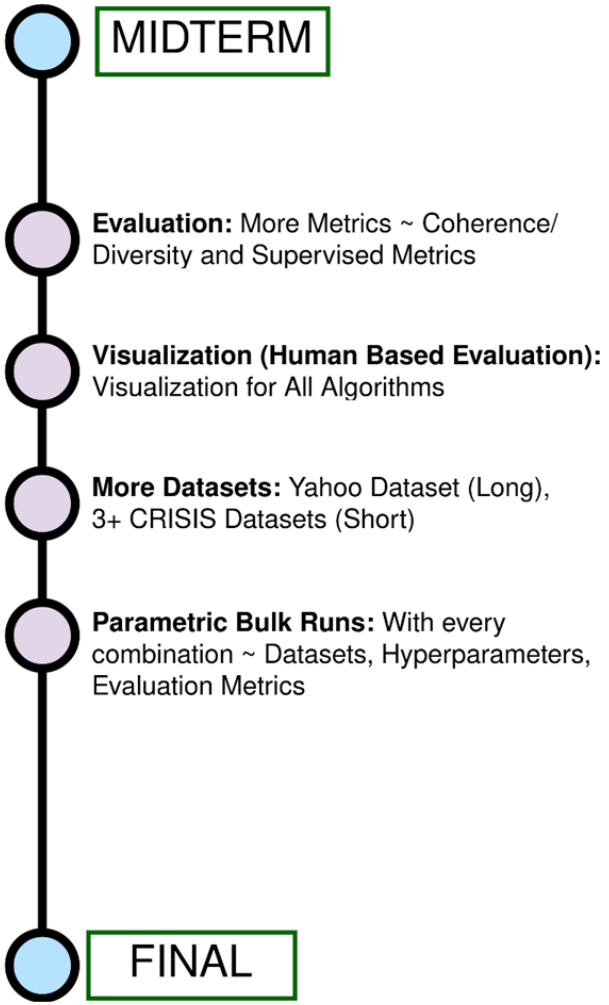
Quick Comparison

Algorithm	**20News**		**Crisis 12**	
	TC	**TD**	**TC**	**TD**
NMF*	**0.598**	0.695	0.456	0.682
LDA*	0.507	0.695	0.406	0.791
BERTopic [all-distilroberta-v1]	0.456	0.460	0.479 (0.611)	0.750 (0.975)
BERTopic [all-MiniLM-L12-v2]	0.502	0.360	0.462 (0.556)	0.775 (0.925)
BERTopic [all-mpnet-base-v2]	0.456	0.430	0.424 (0.561)	**0.775 (0.975)**
Top2Vec [all-MiniLM-L6-v2]	0.380	**0.875**	0.378	0.700
Top2Vec [universal-sentence-encoder]	0.376	0.865	-	-
Top2Vec [universal-sentence-encoder-large]	-	-	**0.486**	0.550

Road Map

Road Map





Optional: 1-2 New Algorithms ~ CTM and/or LDA-Bert

Q & A