Project Presentation

Team LDA

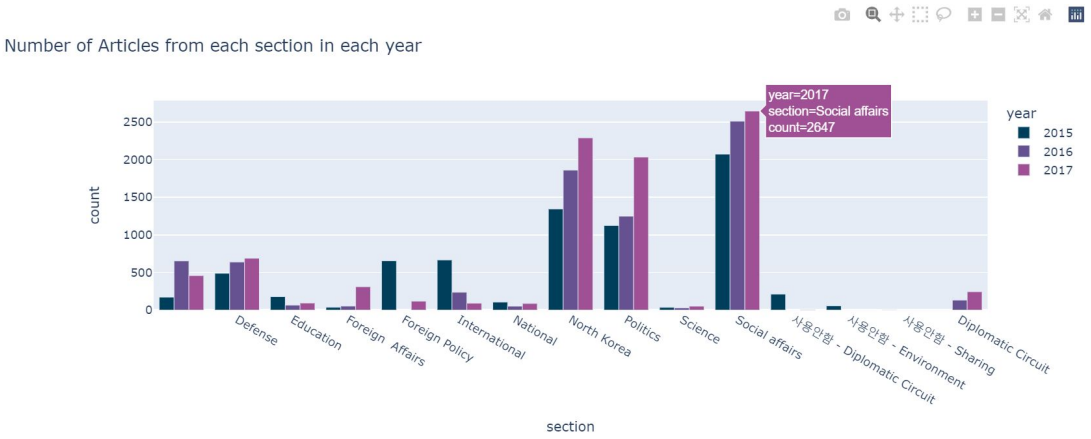
Ohjun Kwon (20160051)
Melka Dawit (20180754)
Webi Dabuse (20180883)

Dec 10, 2021

Contents

Problem Definition

- Extract the most trending issues for a given time interval and track some events by directly or indirectly related to them by applying leveraging TM techniques.
- Dataset
  - Korean newspaper articles from the year 2015, 2016, and 2017.



Article sections:

- Social Affairs - 7,233 articles
- North Korea - 5,497 articles
- Politics - 4,409 articles
- Others - 6,628 articles

Date Cleaning

- We cleaned the title and body of the articles
- Remove punctuations and symbols
- Change nk to North Korea
- Change sk to South Korea
- Change contractions to full forms
- Remove stop words
- Remove texts like ‘yonhap’
- Avoided stemming and lemmatization

```
def data_cleaning(self, text):
  text = text.lower()
  text = re.sub(r'http[s]?://(?:[a-z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-f][0-9a-f]))+', '', text) # clean url
  text = re.sub(r'#(\w+)', '', text)   # clean hashtags
  text = re.sub(r'@(\w+)', '', text)   # clean @s
  text = re.sub(r'<[^>]+>', '', text)  # clean tags
  text = re.sub(r'\w*[0-9]\w*', '', text)   # clean any number word combination
  text = re.sub(r'[’‘]', '\'', text)   # replace ’‘ with '
  text = re.sub(r's\'', '', text)      # clean s'
  text = re.sub(r'[£₹$€₩]', ' ', text) # clean currency symbols
  text = re.sub(r'[δ∫βωδσ∊∆≡απθ+*-=°^×√÷]', '', text) # clean math symbols
  text = re.sub(r'[/(),!@"“”?%_&#:;><{}~\[\]|…]', '', text)   # clean punctuation
  text = text.split()

  temp = []
  for word in text:
    if word=="u.s.":
      word="us"
    if word=="graphic" or word=="news":
      continue
    temp.append(word)
  text = " ".join(temp)
  text = re.sub(r'[.]', '', text)
  text = [self.contractions[word] if word in self.contractions else word for word in text.split()]  # change contractions to full forms

  temp = []
  for word in text:
    if word=="n":
      word="north"
    elif word=="s":
      word="south"
    elif word=="k":
      word="korea"
    elif word=="nk":
      word="north korea"
    if word not in self.stop_words:
      temp.append(word)
  text = temp
  text = " ".join(text)
  text = re.sub(r'\'s', '', text)      # clean 's
  text = re.sub(r'\'', '', text)       # clean '
  text = re.sub(r'yonhap','', text)    # clean yonhap
  return text
```

Related Works

- Document Level Relation Extraction on a paragraph trained on DocRED dataset
- DocRED: A Large-Scale Document-Level Relation Extraction Dataset
  - https://aclanthology.org/P19-1074/
- A Novel Document-Level Relation Extraction Method Based on BERT and Entity Information
  - https://ieeexplore.ieee.org/document/9098945

Approaches

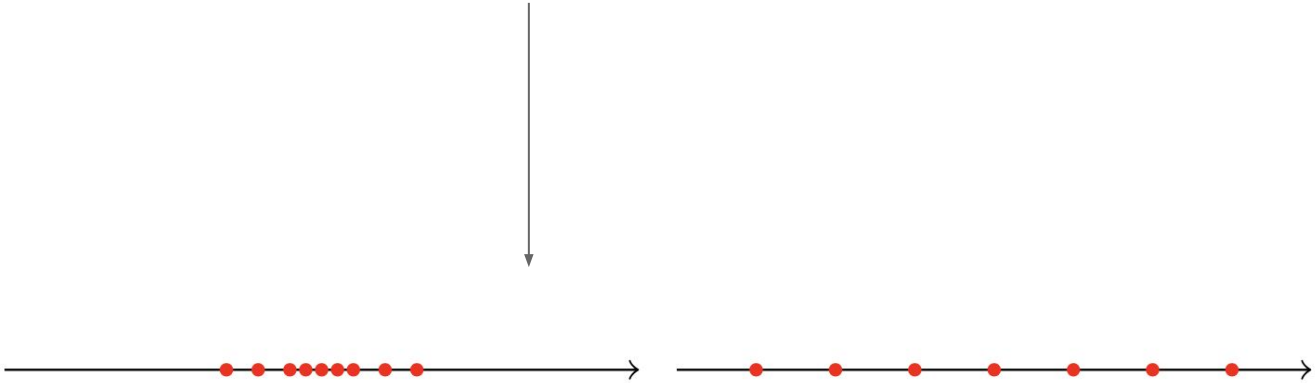
- Clustering
  - K-means, DBSCAN, HDBSCAN
- Dimensionality reduction
  - PCA, UMAP
- Topic Modelling and pretrained models
  - BERTopic, BERT, RoBERTa
- Relation Extraction and Entity Recognition
  - Spacy, Stanza
- Document similarity

Task-1

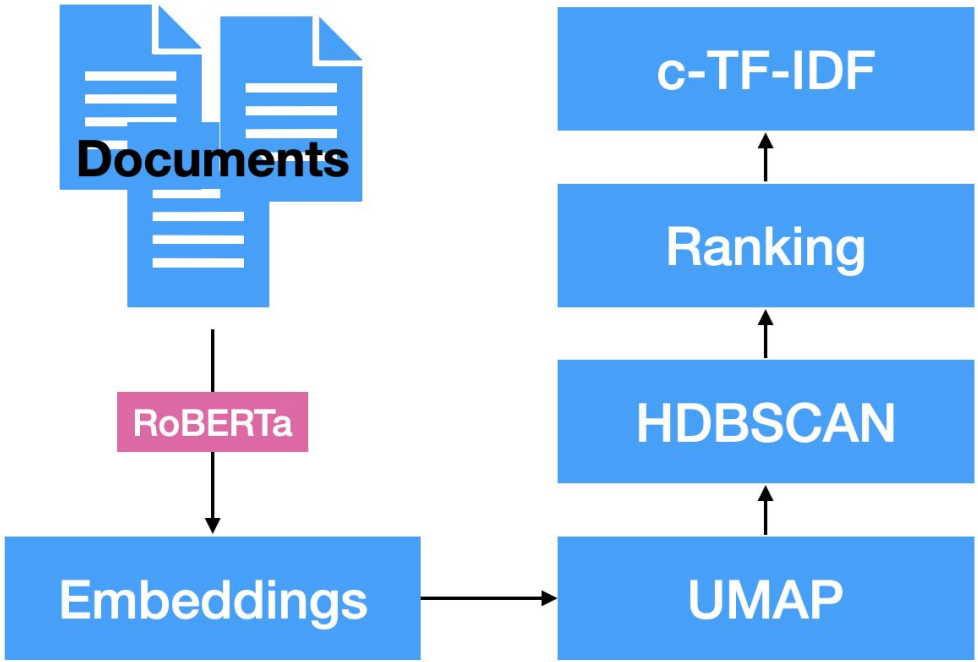
Issue Trend Analysis

Problem Definition

Find the top ten most **significant** issues for each year and rank them



Overall Pipeline



Data Loading

```
data = pd.concat(map(pd.read_json,
↪  glob('data/*.json'))).reset_index(drop=True)
data = data.drop(columns=[' author', ' description', '
↪  section'])
data['year'] = data[' time'].apply(lambda x:
↪  int(x.split('-')[0]))
data['days'] = data.apply(lambda x:
↪  (dateutil.parser.isoparse(x[' time']).date() -
↪  datetime.date(x.year, 1, 1)).days, axis=1)
```

Data Loading

	title	time	body	year	days
0	A snapshot of multiculturalism in South Korea	2018-01-01 17:07:00	With birthrates persistently low and the senio...	2018	0
1	[Weekender] Korea's dynamic 2017	2018-01-01 13:22:00	From North Korea's nuclear weapons program nea...	2018	0
2	People's Party members support Ahn's push for ...	2017-12-31 16:18:00	The leader of the center-left People's Party g...	2017	364
3	[Newsmaker] Panamanian vessel probed over susp...	2017-12-31 14:55:00	PYEONGTAEK -- South Korea has seized and insp...	2017	364
4	Hong Kong ship crew questioned in S. Korea for...	2017-12-30 15:44:00	The crew of a Hong Kong-registered ship have b...	2017	363
...	...	...	...	...	...
23764	N. Korean leader's speech arouses cautious opt...	2015-01-01 13:36:00	North Korean leader Kim Jong-un's New Year's D...	2015	0
23765	N. Korean leader open to inter-Korean summit t...	2015-01-01 10:05:00	North Korean leader Kim Jong-un said Thursday ...	2015	0
23766	Ex-U.S. envoy calls for clearer communication ...	2015-01-01 09:27:00	The United States should make its thoughts on ...	2015	0
23767	U.S. imposes sanctions on N. Korean firm	2015-01-01 09:25:00	The United States has imposed sanctions on a N...	2015	0
23768	Park calls for military readiness amid tension...	2015-01-01 09:24:00	President Park Geun-hye called on the military...	2015	0

Document Embedding

- Pre-trained RoBERTa-large model
  - Using sentence_transformers framework
  - *all-roberta-large-v1*

```
from sentence_transformers import SentenceTransformer

model = SentenceTransformer('all-roberta-large-v1')
embeddings = model.encode(data[' body'].tolist())
```

Document Embedding

- Pre-trained RoBERTa-large model
  - Using *sentence_transformers* framework
  - *all-roberta-large-v1*

```
[[ 0.01112086 -0.02989836  0.00035584 ...  0.0024787   0.00300203
   0.07555952]
 [-0.00669455  0.02810164 -0.00719905 ...  0.01459044 -0.02864629
   0.0169838 ]
 [ 0.00992763 -0.01956357  0.03801578 ... -0.01523898  0.00299372
   0.06408203]
 ...
 [-0.02489829 -0.01752306  0.00754775 ... -0.00106896  0.0083714
   0.01421531]
 [-0.06365523  0.0124728   0.04411678 ...  0.02108982 -0.00537032
  -0.01242058]
 [-0.01416792 -0.02924789  0.00203461 ...  0.02693999  0.00627369
   0.04826877]]
(23769, 1024)
```

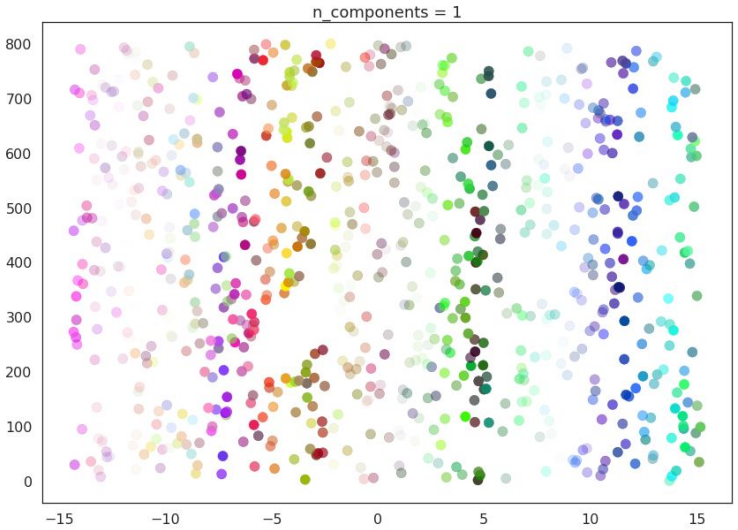
Document Clustering

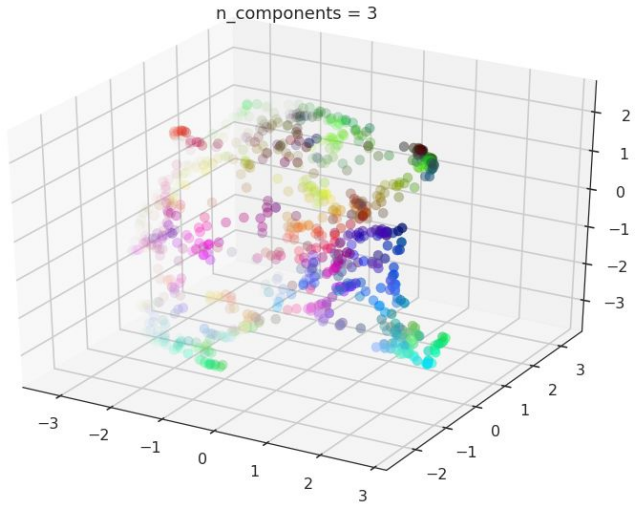
- Dimensionality Reduction
  - Clustering algorithms are prone to have the curse of dimensionality
  - UMAP (Uniform Manifold Approximation and Projection)

```
dim_reduction = UMAP(n_components=5, min_dist=0,
↪  metric='cosine').fit_transform(embeddings)
```

Document Clustering

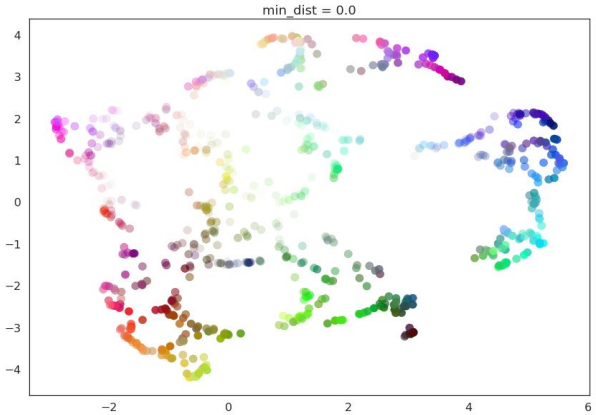
```
dim_reduction = UMAP(n_components=5, min_dist=0,
↪  metric='cosine').fit_transform(embeddings)
```

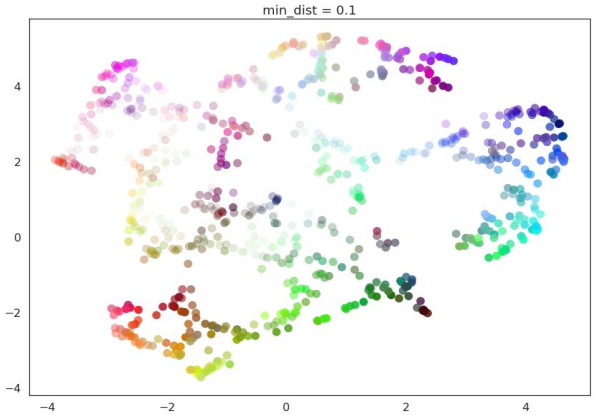


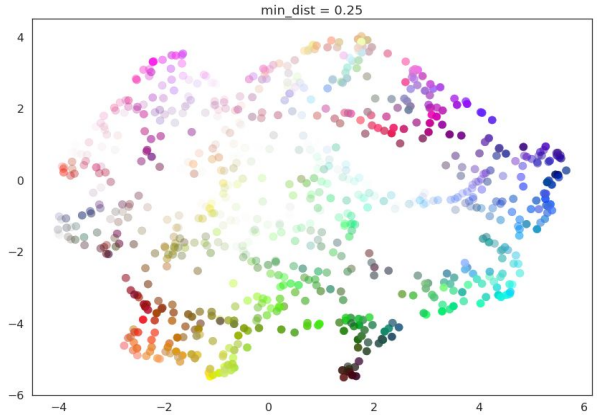


Document Clustering

```
dim_reduction = UMAP(n_components=5, min_dist=0,
↪  metric='cosine').fit_transform(embeddings)
```







Document Clustering

```
dim_reduction = UMAP(n_components=5, min_dist=0,
↪  metric='cosine').fit_transform(embeddings)
```

`metric`

The final UMAP parameter we will be considering in this notebook is the `metric` parameter. This controls how distance is computed in the ambient space of the input data. By default UMAP supports a wide variety of metrics, including:

Document Clustering

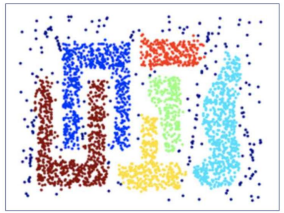
- Clustering
  - HDBSCAN
    - Hierarchical clustering with DBSCAN

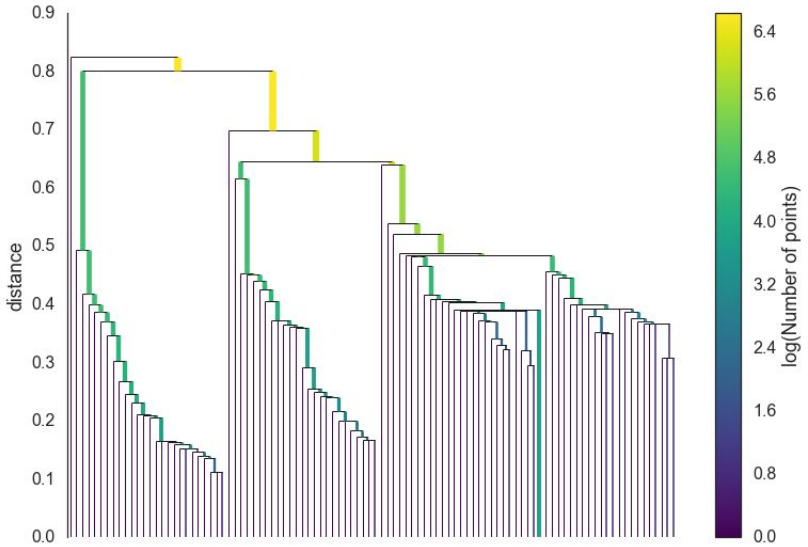
DBSCAN Clustering Algorithm

- Density-based Spatial Clustering of Applications with Noise
- Density-based clustering locates regions of high density and separates outliers

c.f. Partition-based clustering
Outliers affect the centroids

Density:
the number of points in a specified radius.





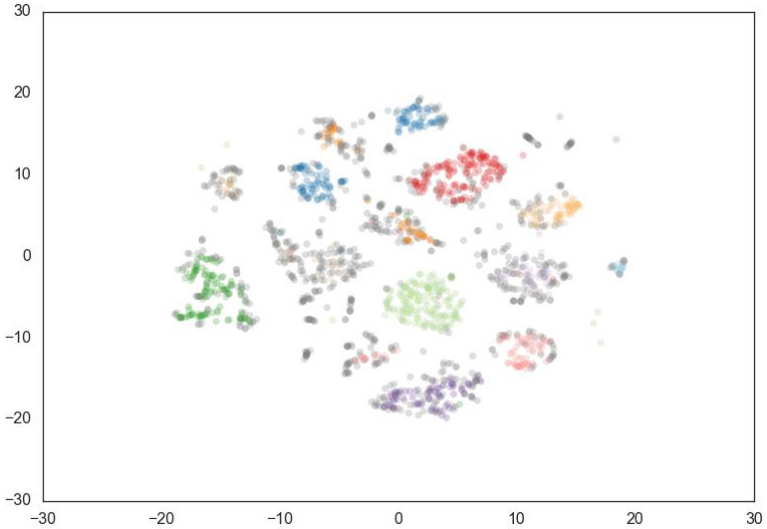
Document Clustering

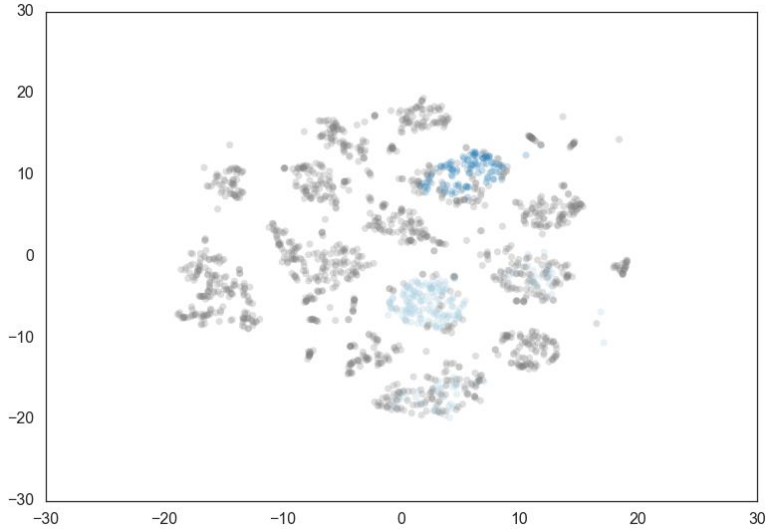
- Clustering
  - HDBSCAN

```
clusterer = HDBSCAN(min_cluster_size=40)
clusterer.fit(dim_reduction)
data['topic'] = clusterer.labels_
data['prob'] = clusterer.probabilities_
```

Document Clustering

```
clusterer = HDBSCAN(min_cluster_size=40)
clusterer.fit(dim_reduction)
data['topic'] = clusterer.labels_
data['prob'] = clusterer.probabilities_
```





Document Clustering

	title	time	body	year	days	topic	prob
0	A snapshot of multiculturalism in South Korea	2018-01-01 17:07:00	With birthrates persistently low and the senio...	2018	0	84	1.000000
1	[Weekender] Korea's dynamic 2017	2018-01-01 13:22:00	From North Korea's nuclear weapons program nea...	2018	0	-1	0.000000
2	People's Party members support Ahn's push for ...	2017-12-31 16:18:00	The leader of the center-left People's Party g...	2017	364	70	1.000000
3	[Newsmaker] Panamanian vessel probed over susp...	2017-12-31 14:55:00	PYEONGTAEK -- South Korea has seized and insp...	2017	364	64	0.915748
4	Hong Kong ship crew questioned in S. Korea for...	2017-12-30 15:44:00	The crew of a Hong Kong-registered ship have b...	2017	363	64	0.930521
...	...	...	...	...	...	...	...
23764	N. Korean leader's speech arouses cautious opt...	2015-01-01 13:36:00	North Korean leader Kim Jong-un's New Year's D...	2015	0	87	0.753503
23765	N. Korean leader open to inter-Korean summit t...	2015-01-01 10:05:00	North Korean leader Kim Jong-un said Thursday ...	2015	0	87	0.758473
23766	Ex-U.S. envoy calls for clearer communication ...	2015-01-01 09:27:00	The United States should make its thoughts on ...	2015	0	-1	0.000000
23767	U.S. imposes sanctions on N. Korean firm	2015-01-01 09:25:00	The United States has imposed sanctions on a N...	2015	0	64	1.000000
23768	Park calls for military readiness amid tension...	2015-01-01 09:24:00	President Park Geun-hye called on the military...	2015	0	-1	0.000000

Topic Extraction

- Before ranking topics, we need to label each group with keywords (topics) in order to actually examine the results.

Topic Extraction

```
group = data.groupby('topic', as_index=False)
topics = group.agg({' body': ' '.join})
```

topic	body
-1	From North Korea's nuclear weapons program nea...
0	The education ministry under the former Park G...
1	Globally beloved Nutella has been around since...
2	The number of patients infected with malaria i...
3	A woman who suffered from a rare disease calle...
...	...
97	South Korea's Foreign Minister Kang Kyung-wha ...
98	US President Donald Trump has promised to help...
99	A group of South Korea's social and political ...
100	WASHINGTON -- The United States would still be...
101	Despite sharing the common goal of achieving N...

Topic Extraction

```
count_vectorizer = CountVectorizer(ngram_range=(1, 3),
↪  preprocessor=partial(preprocess, stem_lemmatize=False))
count = count_vectorizer.fit_transform(topics[' body'])
words = count_vectorizer.get_feature_names_out()
```

Topic Extraction

```
['aa',
 'aa aa',
 'aa aa december',
 'aa aa highest',
 'aa aa last',
 'aa aa october',
 'aa citing',
 'aa citing country',
 'aa concluded',
 'aa concluded july',
 'aa december',
 'aa december last',
 'aa equivalent',
 'aa equivalent higher',
 'aa existed',
 'aa existed years',
 'aa guided',
 'aa guided weapon',
```

```
r =
r=p
ect
ect
```

```
'abandon biggest',
'abandon biggest stumbling',
'abandon bill',
'abandon bill turning',
'abandon boy',
'abandon boy body',
'abandon boy received',
'abandon called',
'abandon called seoul',
'abandon calls',
'abandon calls hostile',
'abandon candidacy',
'abandon candidacy general',
'abandon candidacy month',
'abandon citizenship',
'abandon citizenship rather',
'abandon commitment',
'abandon commitment allies',
'abandon complete',
'abandon complete verifiable',
'abandon confrontational',
'abandon confrontational policy',
'abandon construction',
'abandon construction shingori',
'abandon continue',
'abandon continue shin',
'abandon continued',
'abandon continued development',
'abandon controversial',
'abandon controversial counterterrorism',
'abandon core',
'abandon core policies',
'abandon country',
'abandon country weapons',
```

```
am_
tem
(tc
ame
```

```
'abandonment body seek',
'abandonment came',
'abandonment came france',
'abandonment chief',
'abandonment chief state',
'abandonment cooperative',
'abandonment cooperative politics',
'abandonment corpse',
'abandonment corpse suspected',
'abandonment denuclearization',
'abandonment denuclearization principles',
'abandonment duties',
'abandonment duties immediately',
'abandonment early',
'abandonment early date',
'abandonment gap',
'abandonment gap larger',
'abandonment government',
'abandonment government enacted',
'abandonment jasonyeo',
'abandonment jasonyeo com',
'abandonment korean',
'abandonment korean council',
...]
```

Topic Extraction (RoBERTa)

Aggregated Documents

	title	time	body	year	days
2913	Cheong Wa Dae demands apology from lawmaker ov...	2017-08-29 17:11:00	The presidential office Cheong Wa Dae on Tuesd...	2017	240
6163	National Assembly to convene first plenary ses...	2016-12-29 10:46:00	South Korea's National Assembly on Thursday is...	2016	363
6436	Lawmaker submits to court signatures from citi...	2016-12-12 11:51:00	An opposition lawmaker submitted the signature...	2016	346
6456	[News Analysis] Impeachment heralds fall of Pa...	2016-12-09 20:03:00	Despite the lingering last minute uncertaintie...	2016	343
6472	Impeachment vote as it happened	2016-12-09 09:51:00	To impeach the embattled head of state, at lea...	2016	343
...	...	...	...	...	...
18573	Watchdog says ministers did not violate electi...	2015-09-15 09:37:00	The election watchdog has decided that separat...	2015	257
18582	NPAD pushes to impeach home affairs minister	2015-09-14 19:18:00	The main opposition New Politics Alliance for ...	2015	256
18594	Opposition party tables impeachment motion aga...	2015-09-14 13:40:00	The main opposition party initiated an impeach...	2015	256
18884	NPAD seeks to impeach home affairs minister	2015-08-28 18:14:00	The main opposition New Politics Alliance for ...	2015	239
18889	Opposition party seeks to impeach home affairs...	2015-08-28 13:17:00	The home affairs minister offered an apology F...	2015	239

RoBERTa

```
[ 0.00992763 -0.01956357  0.03801578 ... -0.01523898  0.00299372
  0.06408203]
```

Words (Topics)

‘impeach park’

RoBERTa

```
[ 0.01112086 -0.02989836  0.00035584 ...  0.0024787   0.00300203
  0.07555952]
```

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Topic Extraction (RoBERTa)

```
[('dispute china korean', 0.6162),
 ('thaad ratification korea', 0.6165),
 ('thaad targeting korea', 0.6276),
 ('thaad tension korea', 0.6285),
 ('korea thaad deployment', 0.6289),
 ('china urges korea', 0.6314),
 ('korea urges china', 0.6315),
 ('china retaliation korea', 0.6358),
 ('korean threats china', 0.6507),
 ('korean missiles china', 0.6561)],
```

```
[('impeach park', 0.5932),
 ('support park impeachment', 0.5938),
 ('park impeachment vote', 0.5945),
 ('impeachment least saenuri', 0.5986),
 ('park impeachment', 0.599),
 ('vote park impeachment', 0.606),
 ('park shun impeachment', 0.6064),
 ('park impeachment opposition', 0.6187),
 ('impeachment park saenuri', 0.648),
 ('park prepares impeachment', 0.648)],
```

```
[('toxicity report humidifier', 0.6848),
 ('disinfectant toxic humidifier', 0.6939),
 ('humidifier disinfectant victims', 0.6982),
 ('toxic humidifier sterilizer', 0.7023),
 ('toxic humidifier sterilizers', 0.7047),
 ('toxic humidifier disinfectant', 0.7052),
 ('exposed toxic humidifier', 0.7066),
 ('humidifier sterilizers victims', 0.7156),
 ('deadly humidifier sterilizers', 0.7159),
 ('victims toxic humidifier', 0.7416)],
```

Topic Extraction (RoBERTa)

```
[[('hospital korea', 0.5152),
  ('mers outbreak', 0.5188),
  ('korea mers', 0.5256),
  ('mers infection', 0.5309),
  ('hospital mers', 0.5405),
  ('virus mers', 0.5415),
  ('virus koreans', 0.5418),
  ('patients korea', 0.544),
  ('mers patients', 0.5539),
  ('infection mers', 0.5697)],
```

ngram_range=(1,2)

Topic Extraction (RoBERTa)

- However, this method is too slow.
  - Need to embed all documents and topic words separately.
- We just want to examine all groups temporarily.
- We can use this method for only top 10 topics after ranking them.

Topic Extraction (TF-IDF)

- Simple method to extract keywords for a document.
- However, it only considers and extracts keywords for one document.

Topic Extraction (c-TF-IDF)

- Considering each document as a concatenation of all documents in a group.
- TF-IDF can be used to extract keywords for each concatenated document.
- TF-IDF should take the number of classes instead of the number of documents since we merged documents.

$$c - TF - IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_j^n t_j}$$

Topic Extraction (c-TF-IDF)

	topic	body	count	keyword
72	71	Prosecutors asked an appellate court on Wednes...	1040	samsung scandal prosecution prosecutors former...
47	46	JEONJU, South Korea -- Police on Friday found ...	892	abuse sex woman child year old victim police
71	70	The leader of the center-left People's Party g...	871	percent candidate saenuri conservative preside...
14	13	Five Chinese military planes on Monday entered...	701	south korea missile defense defense system chi...
32	31	The leader of the ruling Democratic Party on F...	629	slavery comfort women issue tokyo victims abe ...
...	...	...	...	...
80	79	President Moon Jae-in's senior politics secret...	42	bin saudi arabia uae oil middle east al kuwait
44	43	A suspect in the case of a couple who went mis...	42	denmark mother daughter chung yoo ra danish choi
67	66	President Moon Jae-in accepted the resignation...	41	hwang kyo ahn prime minister acting president
39	38	The number of refugees entering South Korea ha...	40	country justice ministry immigration refugee s...
26	25	North Korea has appointed a senior communist p...	40	il leader kim north korean kim jong un

Topic Extraction + Clustered Articles

MERS

	title	time	body	year	days	topic	prob
692	Woman dies after getting tooth extraction in G...	2017-11-30 14:26:00	A woman who suffered from a rare disease calle...	2017	333	3	1.0
19927	No MERS cases reported for four straight days	2015-07-01 18:21:00	The Middle East respiratory syndrome outbreak ...	2015	181	3	1.0
20288	17 percent of MERS patients medical staff: gov't	2015-06-15 11:02:00	So far, 17 percent of the Middle East respirat...	2015	165	3	1.0
20465	S. Korea confirms 1st student MERS patient	2015-06-08 11:29:00	South Korea on Monday confirmed a male high sc...	2015	158	3	1.0
20392	Police book eight for spreading false rumors	2015-06-10 19:48:00	Police said Wednesday eight people were booked...	2015	160	3	1.0
20393	Government urges public cooperation to end MERS	2015-06-10 19:48:00	Acting Prime Minister Choi Kyung-hwan on Wedne...	2015	160	3	1.0
20282	17 percent of MERS patients medical staff: gov't	2015-06-15 14:03:00	So far, 17 percent of the Middle East respirat...	2015	165	3	1.0
20032	S. Korean MERS patient in China recovered, rel...	2015-06-26 15:26:00	A South Korean man, who had been treated at a ...	2015	176	3	1.0
20037	Two more die of MERS, including caregiver	2015-06-25 18:00:00	A caregiver who had not been informed about he...	2015	175	3	1.0
20057	S. Korea reports 2 more deaths from MERS, one ...	2015-06-25 09:32:00	South Korea reported an additional two deaths ...	2015	175	3	1.0

Topic Extraction + Clustered Articles

impeach

	title	time	body	year	days	topic	prob
6625	Saenuri dissenters propose Park resign by April	2016-11-30 09:55:00	A group of ruling Saenuri Party lawmakers supp...	2016	334	67	1.0
6659	Opposition not sure of impeachment motion passage	2016-11-28 17:25:00	Opposition parties Monday put the final touche...	2016	332	67	1.0
6593	People's Party seeks to pass impeachment motio...	2016-12-01 20:27:00	The third-biggest People's Party urged the Nat...	2016	335	67	1.0
10912	Aide dismisses rumors about Park's voluntary r...	2017-02-23 18:08:00	The presidential office on Thursday dismissed ...	2017	53	67	1.0
6600	April emerges as Park's departure time	2016-12-01 17:17:00	Following President Park Geun-hye's ambiguous ...	2016	335	67	1.0
6607	Presidential office urges parliament to initia...	2016-12-01 11:18:00	The presidential office Cheong Wa Dae on Thurs...	2016	335	67	1.0
6618	[Newsmaker] Anti-Park Saenuri MPs key to impea...	2016-11-30 17:04:00	President Park Geun-hye's allies and foes in t...	2016	334	67	1.0
6622	Opposition rejects Park's offer, sets next Fri...	2016-11-30 11:51:00	Heads of South Korea's three opposition partie...	2016	334	67	1.0
18884	NPAD seeks to impeach home affairs minister	2015-08-28 18:14:00	The main opposition New Politics Alliance for ...	2015	239	67	1.0
6163	National Assembly to convene first plenary ses...	2016-12-29 10:46:00	South Korea's National Assembly on Thursday is...	2016	363	67	1.0

Topic Extraction + Clustered Articles

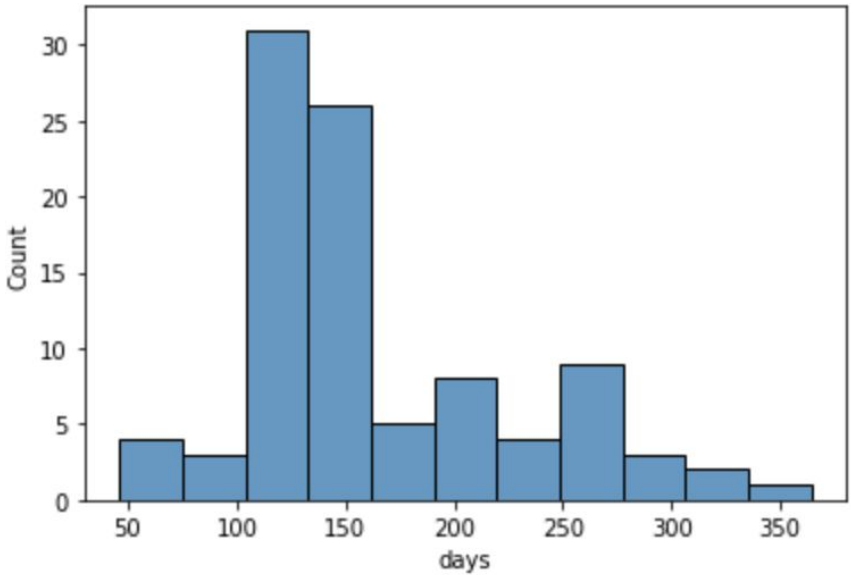
Sewol

	title	time	body	year	days	topic	prob
10039	Crucial preparations complete to relocate Sewo...	2017-03-26 16:11:00	Preparations have been completed for the retri...	2017	84	30	1.0
10133	Salvage operators raise Sewol ferry	2017-03-22 18:41:00	South Korea on Wednesday took a long-awaited s...	2017	80	30	1.0
19174	Chinese ships arrive for Sewol ferry recovery	2015-08-16 11:26:00	Two Chinese ships have arrived at the sinking ...	2015	227	30	1.0
19128	S. Korea to begin work to recover sunken ferry...	2015-08-18 10:54:00	South Korea will launch its multimillion dolla...	2015	229	30	1.0
9381	Search teams enter Sewol ferry to find victims...	2017-04-18 15:12:00	MOKPO -- Search parties entered the wreckage o...	2017	107	30	1.0
9324	Search underway to find remains of missing peo...	2017-04-19 16:45:00	MOKPO -- Salvage workers are examining the fou...	2017	108	30	1.0
5370	Sewol fact-finding to begin in earnest	2017-05-24 15:47:00	Hopes rise for a new investigation to clear li...	2017	143	30	1.0
10071	Sewol set to be loaded on transport vessel	2017-03-24 16:30:00	Salvage operators raised the sunken ferry Sewo...	2017	82	30	1.0
10090	Salvage effort to raise Sewol ferry enters cri...	2017-03-24 09:08:00	The government said Friday it is making all ou...	2017	82	30	1.0
9075	Missing ferry victim's school uniform possibly...	2017-04-27 20:09:00	MOKPO -- A team of workers on Thursday found a...	2017	116	30	1.0

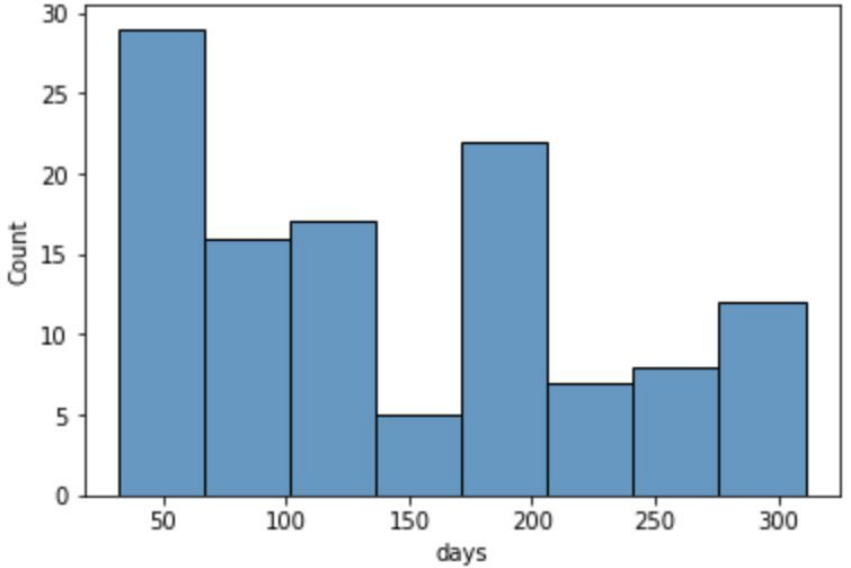
Topic Ranking

For topics occured 2016

Toxic humidifier disinfectant

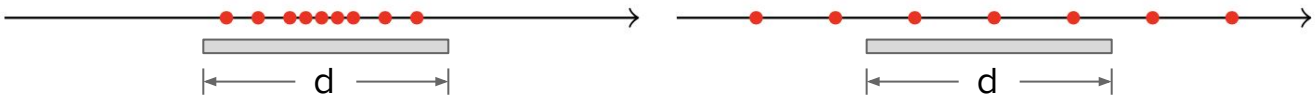


North Korea ballistic missile



Topic Ranking

$$\text{significance} = \max_{i \in [0,366-d)} \sum_{j=i}^{i+d} j$$



9 **3**

Topic Ranking

$$\text{significance} = \max_{i \in [0, 366-d)} \sum_{j=i}^{i+d} j$$

```
max([sum(days[i:i + length]) for i in range(len(days) -
↪  length)])
```

Topic Ranking

For topics occured 2016

Count

count	max_subset_sum	keyword
388	73	court abuse woman child year old victim police
313	96	allegations presidential prosecution court pro...
300	104	south korea missile defense defense system chi...
272	93	former opposition kim saenuri party minjoo par...
195	63	security council north korean pyongyang china ...
141	29	slavery foundation tokyo comfort women deal ja...
116	26	ballistic missile north korea missiles test mu...
115	25	thae defected restaurant defection north korea...
107	31	exercises joint south north korea navy drills ...
103	28	nuclear test sanctions pyongyang chinese north...

New significance measure

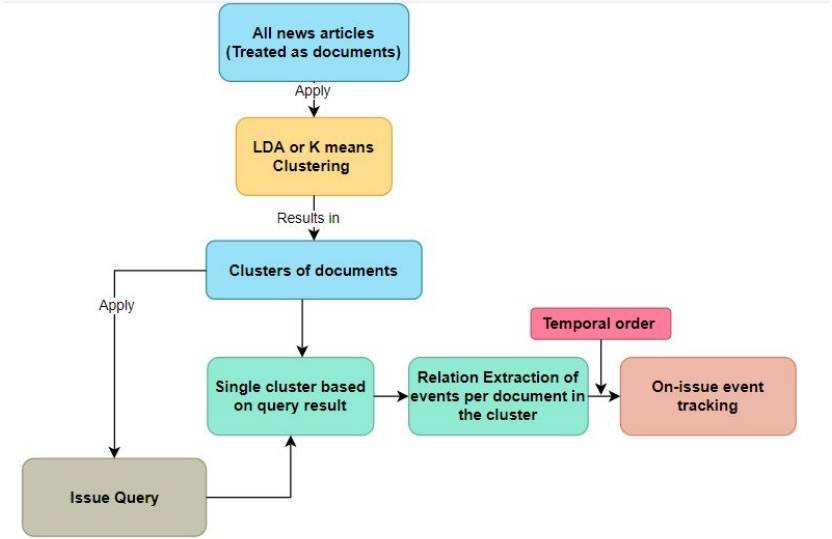
count	max_subset_sum	keyword
300	104	south korea missile defense defense system chi...
313	96	allegations presidential prosecution court pro...
272	93	former opposition kim saenuri party minjoo par...
388	73	court abuse woman child year old victim police
195	63	security council north korean pyongyang china ...
51	48	parties vote party opposition president saenur...
83	44	park geun protesters gwanghwamun president par...
96	42	reckitt benckiser humidifier disinfectant prod...
59	32	north south korean factory firms gaeseong indu...
107	31	exercises joint south north korea navy drills ...

Task-2

On-issue Event Tracking

Pipeline

- Clustering
  - K-means clustering
- **Issue Query**
  - Search the issue in each cluster to choose the relevant cluster
- Relevant Cluster Selection
- Event Selection
  - Using Euclidean distance from the center of relevant cluster.



Tools Used

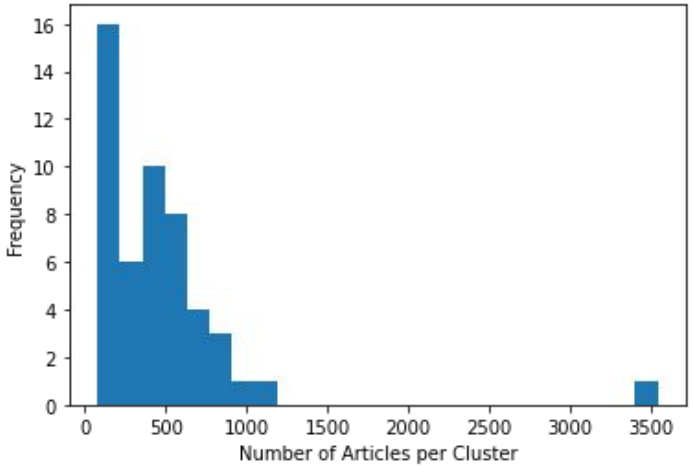
- K-means clustering form Scikit Learn.
  - After vectorizing the documents using term frequency inverse document frequency.
  - **K = 50**
- **Stanza CoreNLPClient** and Pipeline.
  - Name Entity Recognition

K-means Clustering

- Clustering shows that the dataset has a wide range of topics with different number of articles per cluster.
- Used different K to avoid anomalous clusters
- **For K= 50, one anomalous cluster has 3,538 articles.**

K-means Clustering

- After inspecting most of the articles' body, we found that several articles are related to North Korea and Crimes in Korea (Social Affairs).
- This irregularity has not improved with different k values.



K = 50

Results

- Event extraction phase the embedding of articles to measure the relation between issue and an event.
- We use Stanza to annotate the sentences in the selected articles.
- Stanza pipeline is used for the task of named entity recognition.

Results

- Sample output

```
[ Issue ]

hackers north korea cyberattacks internet hacking sony

[ On-Issue Events ]


N. Korea slams U.S. over new sanctions -> ‘The Interview’ i

[ Detailed Information (per event) ]

Event: N. Korea slams U.S. over new sanctions

        Person:  Kim Jong - un, Kim Jong - il, Kim
        place:  North Korea, United States, North, U.S.
        organization:  Sony, the Korean Central News Agenc

Event: ‘The Interview’ is no laughing matter for N. Korean

        Person:  Park, Kim Sung - min, Kim Jong - un.The,
        place:  North Korea, South, North, States
        organization:  Sony, AFP, anti-Kim, anti-Pyongyang
```

Task-3

Related Issue Event Tracking

Task

- Here, we will be identifying events that are related to the top issues from Task 1 but not directly tied
- We will be looking at the entire corpus and mainly rely on clustering
- Extract factors such as individuals' names, organizations, locations, and countries mentioned within a document - an article

Task

[ISSUE 1]

Deployment deploy radar

[Related-Issue Events]

Lee kang party

-> PERSON: Jun byunghun, Lee sangdon, Hyun kihwan, Kim moosung, Kim daejung, Chung uihwa, Chung, Cheong wa dae, Jeong yeonguk, Officelee, Bae hyunjung
-> ORGANIZATION: Saenuri party special committee, Federal reserve, Saenuri party
-> PLACE:

Members roh late

-> PERSON: Kim junghyun, Npad, Hwang qualification, Hwang, Kim npad, Chung uihwa, Cheong wa dae, Syndrome npad, Hwang kyoahn, Secondincommand, Soohyun yeo junsuk
-> ORGANIZATION: Npad attend, Saenuri party floor spokesperson, Saenuri, Saenuri party
-> PLACE:

Trump diplomatic xi

-> PERSON: Donald trump national security, Mcmaster, Kim sundong, Moon, Cheong wa dae, Thaad trump, Jung minkyung
-> ORGANIZATION: Democratic party, Leader liberty korea party, Democratic party korea, Park yongjin democratic party
-> PLACE:

[ISSUE 2]

Virus mers disease

[Related-Issue Events]

Carnegie kerry tank

-> PERSON: Hastings fl loretta sanchez, Peter roskam, Alan lowenthal, Denny, Shinzo abe, Linda sanchez, Christopher gibson ny blake, Wa donald beyer, Meng ny charles rangel ny gerald connolly, Bill pascrell, Tom reed ny rob woodall, Steve israel ny, Abe
-> ORGANIZATION: Nj adam schiff, Farenthold tx barbara comstock, Dca ed royce rca, Albio sires, Royce foreign affairs committee, Honda
-> PLACE: Asia

North korea said

-> PERSON: Yun byungse, Yu myunghwan, Sangho, Yoshiro mori, Cheong wa dae, Takeo kawamura, Shinzo abe
-> ORGANIZATION: Park geunhye hold, Seoul hotelsaying
-> PLACE:

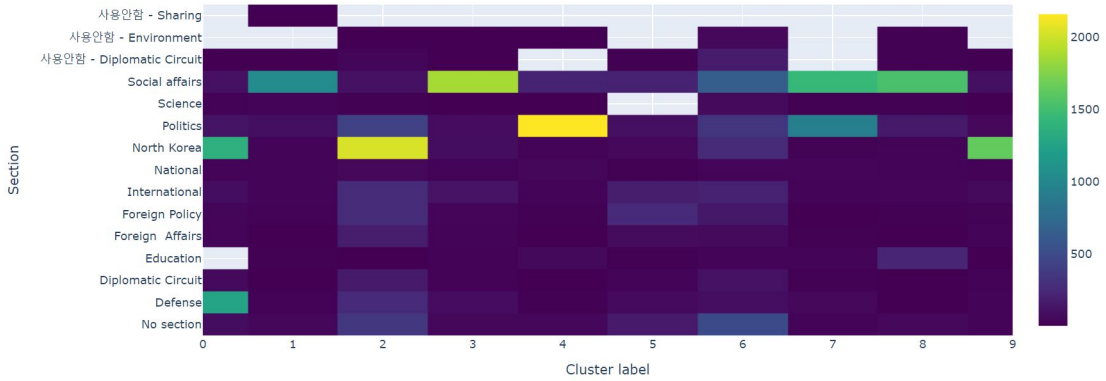
Israel palestine mission

-> PERSON:
-> ORGANIZATION: Assembly interparliamentary union
-> PLACE:

Approaches

- **Step 1**
  - Dimensionality reduction **(to 5)**
  - K-means clustering (**number of clusters = 10**)
  - Evaluated by the Elbow method

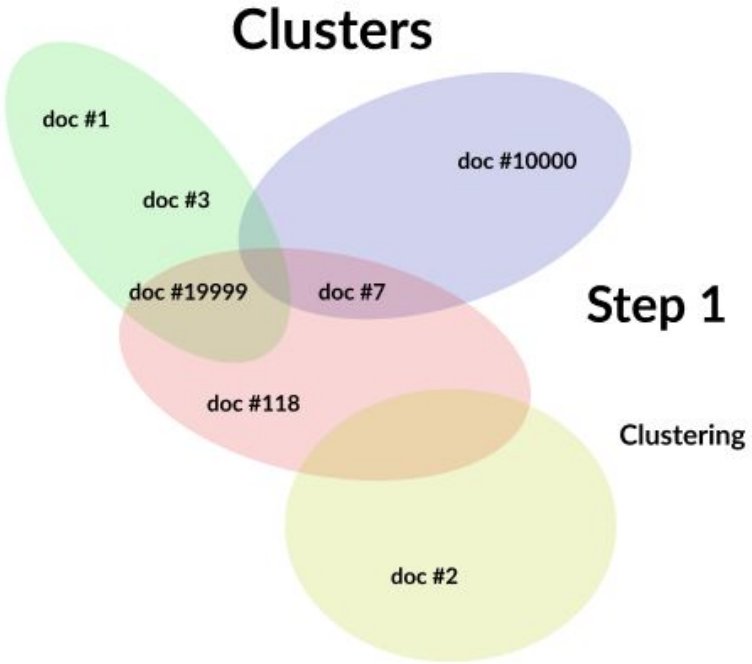
```
args = edict()
args.reduceTo = 5
args.clusters = 10
args.max_iter = 500
args.docs = 10
args.how_many_docs = 20
args.how_many_related = 3
args.cluster_freq = 0
args.neighbouring_cluster = 1
```



Document distribution after clustering

Approaches

- **Step 1**

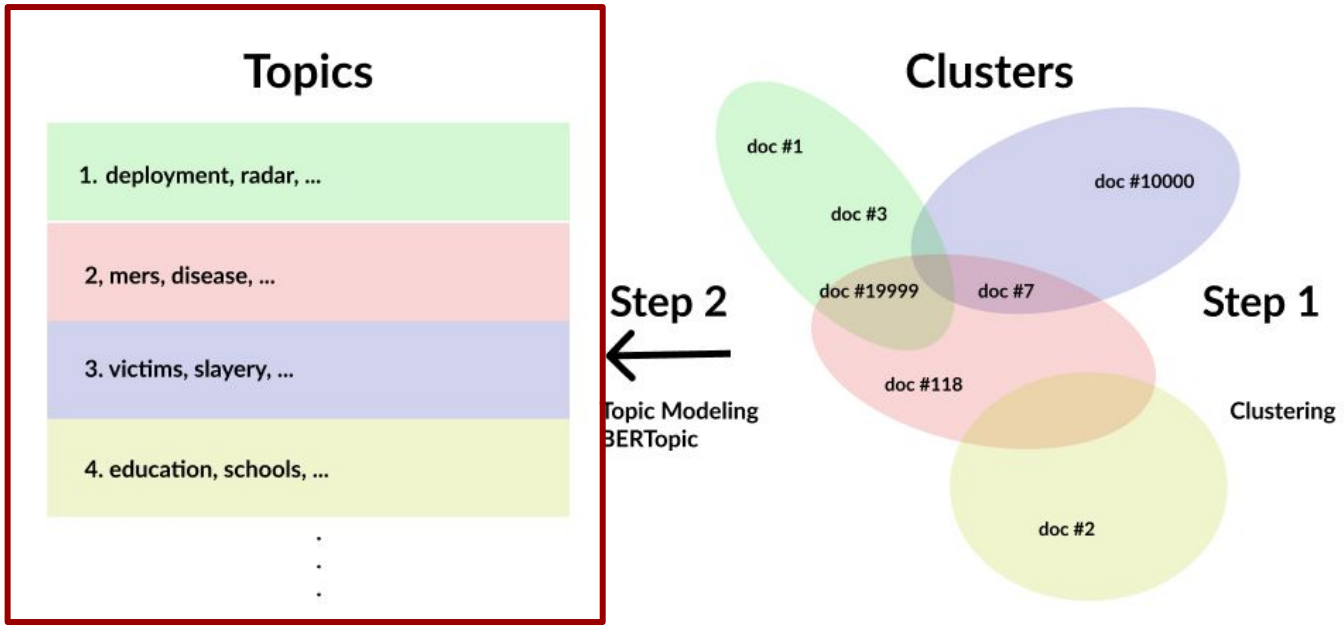


Approaches

- **Step 2**
  - Topic Modeling with **BERTopic**
  - Pretrained models such as ***all-mpnet-base-v2***, all-roberta-large-v1, *distilbert-base-nli-mean-tokens*, *all-MiniLM-L6-v2* are used for embedding the documents

Approaches

- **Step 2**



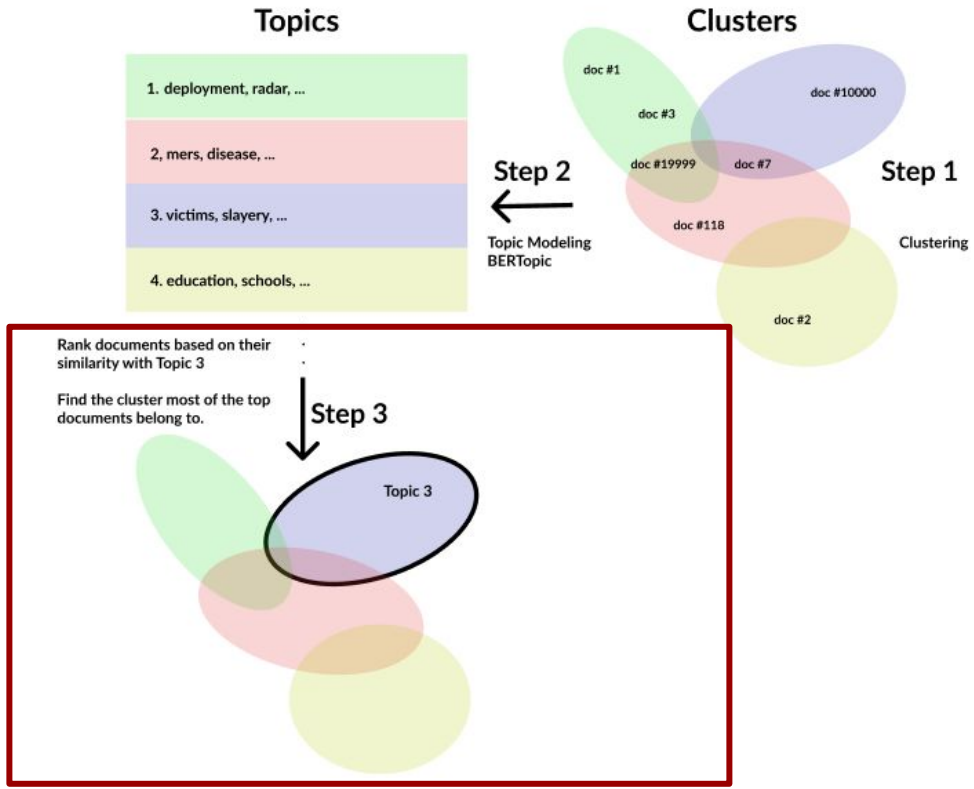
Approaches

- **Step 3**
  - Rank documents based on their cosine similarity to the top issues
  - Look at the top few documents (**20 documents**)
  - Move into the **top-first** cluster most of these documents belong to.

```
args = edict()
args.reduceTo = 5
args.clusters = 10
args.max_iter = 500
args.docs = 10
args.how_many_docs = 20
args.how_many_related = 3
args.cluster_freq = 0
args.neighbouring_cluster = 1
```

Approaches

- **Step 3**



Approaches

- **Step 4**
  - Move to the neighbouring cluster (**immediate first**)
  - Neighbouring clusters are ranked based on the euclidean distance of their centroids

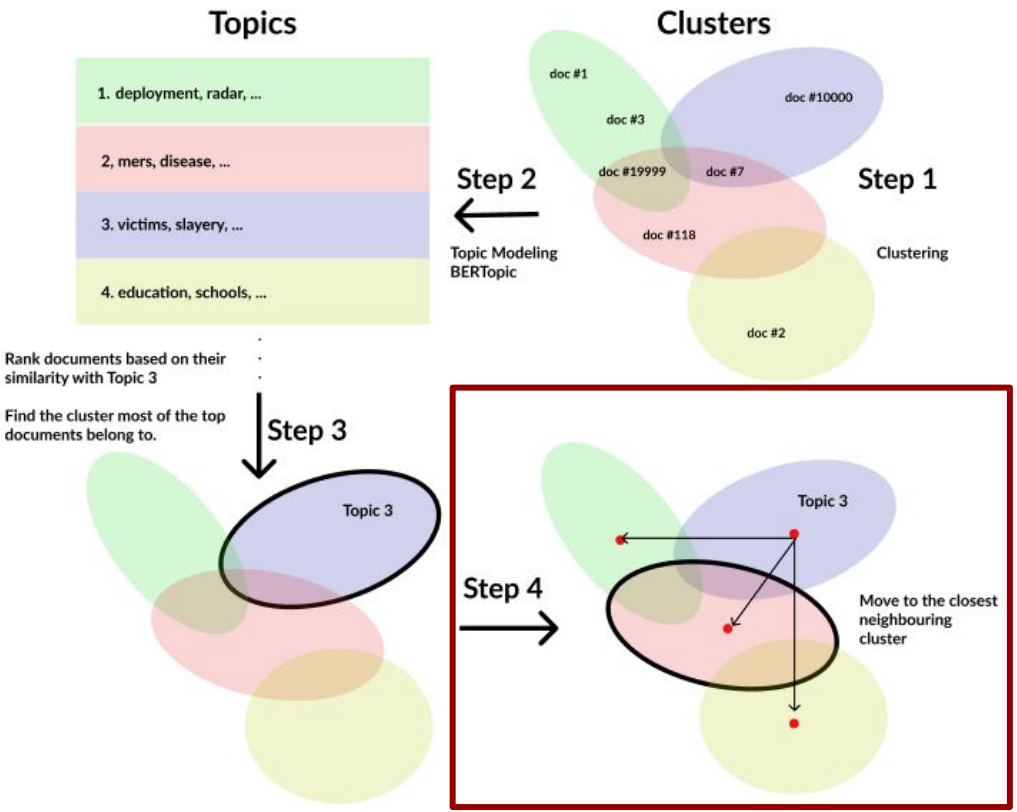
```
args = edict()
args.reduceTo = 5
args.clusters = 10
args.max_iter = 500
args.docs = 10
args.how_many_docs = 20
args.how_many_related = 3
args.cluster_freq = 0
args.neighbouring_cluster = 1
```

Assumption

- *Moving to other clusters neighboring the cluster the issue belongs to, we will likely encounter documents that are less related but indirectly related to it*

Approaches

- **Step 4**



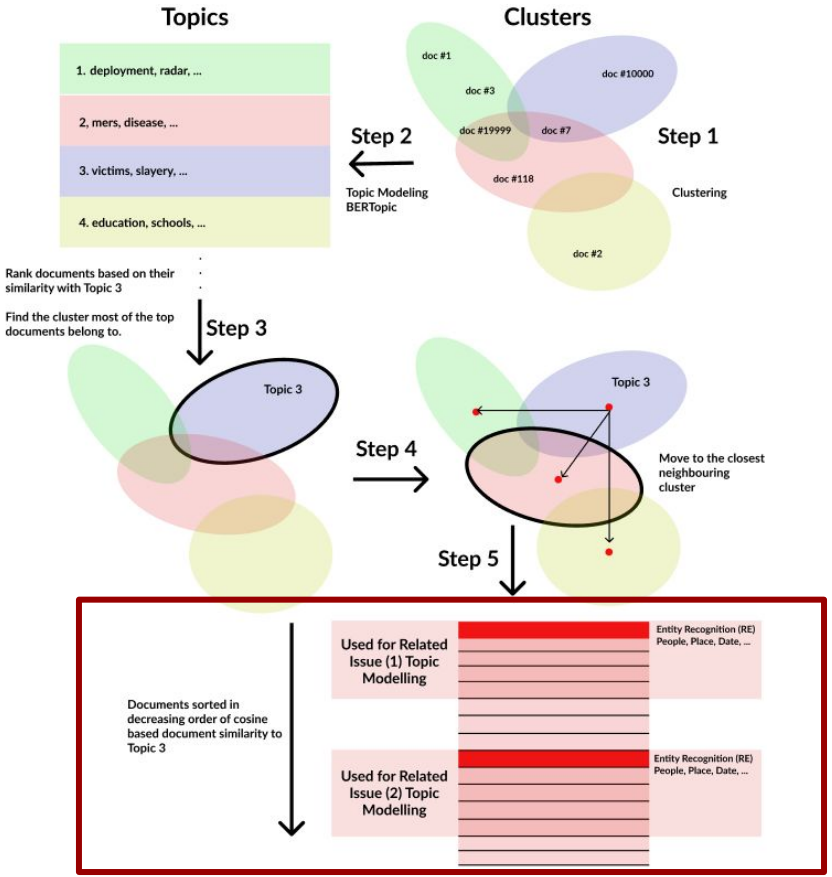
Approaches

- **Step 5**
  - Rank documents with in the neighbouring cluster based on their cosine similarity to the selected top issue
  - Pick every **n^{th} (10^{th})** document for relation extraction and entity recognition
  - Use the first few documents **(5)** that are ranked immediately under the selected document for topic extraction.
  - It is the extracted topic that is displayed as a related issue (**top-3**).

Approaches

- **Step 5**

```
args = edict()
args.reduceTo = 5
args.clusters = 10
args.max_iter = 500
args.docs = 10
args.how_many_docs = 20
args.how_many_related = 3
args.cluster_freq = 0
args.neighbouring_cluster = 1
```



Entity Recognition

- **spaCy** based

Entity recognition on cleaned titles

Thank you