Team Miracle CS474 Term Project Report

Issue Trend Analysis & Event Tracking for on-issue, related-issue in Korea Herald news

Phatarapran Saraluck
School of Computing
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea
phatarapran9155@kaist.ac.kr

Wirittipol Supdateyarnnakorn
School of Computing
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea
kwsnarakz@kaist.ac.kr

HyeongSeok Seo
School of Computing
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea
wimpr@kaist.ac.kr

ABSTRACT

The project is to design a complete system to analyze a corpus of news from the Korea Herald over three years (2015~2017). The Result of this project includes finding the top ten most significant issues for each year and ranking them based on a specific criterion, tracking the two most suitable issues followed by sequencing events through time specifically tied to the issues, and representing events topically related to the issues.

KEYWORDS

Clustering, Topic Modeling, Dimensionality Reduction, NER

1 Issue Trend Analysis

The first task is Issue Trend Analysis which is to find the top 10 most significant issues for each year and rank them.

0. Visualization

We first made a visualization of the data:title-body to figure out the entire form, applying two dimensionality reduction technique, PCA and t-SNE for each year.

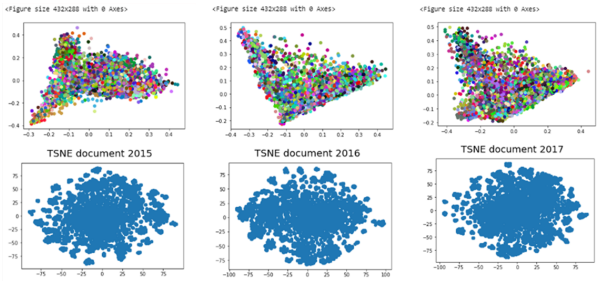


Figure 1. Visualization of Title-Body for each year, applying PCA & t-SNE

The figure shows the data has no pattern and clustered together in one bunch. We cannot directly identify the exact shape. Thus it needs to be splitted into various clusters,

1. Keyword extraction & Clustering

We apply DBSCAN as a clustering method because it doesn't need a number of clusters as a hyperparameter, and automatically finds it.

To find proper Epsilon value (hyperparameter for DBSCAN), We plotted relation between epsilon and the number of cluster for each year data

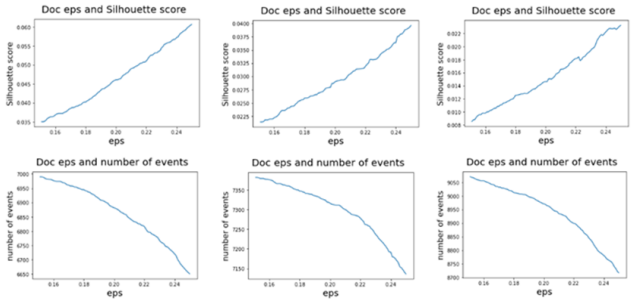


Figure 2–1. Relation between Epsilon & Silhouette score, Epsilon & number of clusters for each year

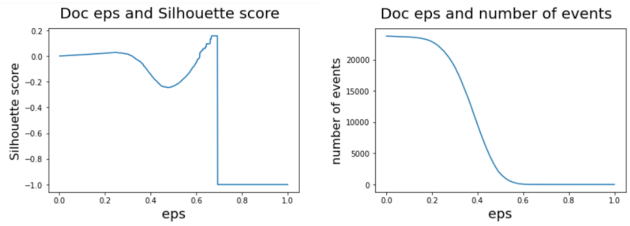


Figure 2-2. Relation between Epsilon & Silhouette score, Epsilon & Number of clusters for entire body data

Since Figure 2-1 shows a similar relation for each year, it doesn't connote any meanings but Figure 2-2, relation between Epsilon & Number of clusters have two maximum values of change of slope. First maximum point of the change of slope with the Epsilon value around 0.2 indicates starting point of number of documents are combined to distinct clusters together, while second maximum point of the change of slope with the Epsilon value around 0.5 indicates starting point of distinct clusters are become mixed together. Thus taking the Epsilon value around 0.20 will help to exclude outliers.

With the exact approximation of Epsilon, the value is 0.190

Year	Number of clusters	Total number of documents
2015	6919	7156
2016	7335	7485
2017	9000	9128

Table 1. Number of clusters out of total number of documents for each year after applying Epsilon value 0.190

After taking the Epsilon value to 0.195, Rank by number of documents clustered together. Followings are Top 20 clusters ranked by number of documents for each year. Left quantity is the cluster index, the right quantity is the number of documents clustered together.

	counter_2015_event	counter_2016_event	counter_2017_event
	[(2954, 6), (132, 5), (5034, 4), (5452, 4), (6013, 4), (1223, 3), (1837, 3), (2166, 3), (2864, 3), (3613, 3), (5363, 3), (5849, 3), (5859, 3), (6482, 3), (18, 2), (34, 2), (42, 2), (43, 2), (45, 2), (145, 2)]	[(16, 5), (5095, 4), (431, 3), (518, 3), (706, 3), (5782, 3), (6234, 3), (7166, 3), (11, 2), (90, 2), (178, 2), (310, 2), (477, 2), (494, 2), (526, 2), (558, 2), (592, 2), (714, 2), (768, 2), (837, 2)]	[(67, 11), (617, 7), (3706, 5), (4733, 5), (5, 4), (6652, 4), (813, 3), (911, 3), (4110, 3), (5033, 3), (6414, 3), (44, 2), (113, 2), (114, 2), (190, 2), (313, 2), (334, 2), (591, 2), (606, 2), (849, 2)]

Figure 3. Top 20 clusters ranked by number of documents for each year

To form an issue, it's clear that there is a need to extract keywords for each cluster. There are two possible approaches. One is using keyword extraction tools, another is summarizing the cluster into a fixed number of words using a summarizer.

We applied Yake for keyword extraction tools. It has 5 main steps that are (1) text pre-processing, (2) feature extraction, (3) computing term score, (4) n-gram generation and computing candidate keyword score, (5) data deduplication and ranking. Following Figure is the result after applying Yake that extraction of 4 keywords from top 10 clusters for each year.

Figure 4. Four keywords extracted from top 10 clusters for each year

2. Summarize & Clustering

Another approach is using a summarizer, using a pre-trained BART model which is a seq2seq structured denoising auto-encoder made for applying various fields.

Unlike previous processing, we first summarized each body of articles. After that, embed them using TF-IDF, cluster the embedded vectors using DBSCAN (eps=0.190) to get a label to each document, which group it belongs to. Rank top 10 most frequent labels

Then concatenate the summarized body of each document to form a cluster, expecting to form a similar meaning, closing the distance between documents. And summarize each cluster.

	2015	label2015	freq2015	2016	label2016	freq2016	2017	label2017	freq2017
0	South Korea reported no additional cases of MI...	2957	6	Radio Pyongyang, the North's state-run radio s...	16	5	Temperatures across the country plummeted abou...	67	11
1	In October, some 36,900 babies were born, up 1...	132	5	Some 200 chickens were found dead on Monday mo...	518	4	A South Korean research team says it has uncov...	617	7
2	Cho Hyun-ah, former vice president of Korean A...	6018	4	Voter turnout in South Korea's parliamentary e...	5095	4	South Korean scientists have developed an adhe...	4735	5
3	Seoul-Tokyo ties have plunged to lowest levels...	5457	4	Walkway along Deoksu Palace in downtown Seoul ...	705	3	About 30,300 babies were born in May, down 11....	3708	5
4	South Korea's top financial regulator said Fri...	5039	4	Pigs at two swine farms in Nonsan in the centr...	5782	3	The search for the missing South Korean ship, ...	6654	4
5	Activists from the Humane Society Internationa...	3617	3	Suh Ye-won is the director of the state-run Na...	7166	3	The H5N6-strain bird flu was detected on a far...	5	4
6	The worst winter seasonal yellow dust in five ...	5864	3	North Korea renewed its calls for peace treaty...	6234	3	100 Seoul residents, art connoisseurs and city...	4112	3
7	101 Audi owners filed the suit with a Seoul d...	1224	3	Some 1.7 million people gathered in central Se...	431	3	Moon Jae-in will ask a parliamentary committee...	911	3
8	The body of a 47-year-old man was found 7 kilo...	1840	3	South Korea on Thursday released a set of meas...	4376	2	South Korean mixed martial arts fighter Bang T...	813	3
9	The 8.5 trillion won ($8.3 billion) project ca...	5854	3	The ejection test of a KN-11 missile from a su...	7255	2	A 33-year-old Korean woman was confirmed to ha...	5035	3

Figure 5. Result after concatenating each cluster after applying DBSCAN.

2015
['South Korea reported no additional cases of Middle East Respiratory Syndrome for the 14th straight day on Sunday. The number of people diagnosed with MERS in the country remained unchanged at 186 with the death toll also staying flat at 36. The disease has claimed over 530 lives globally, posting a fatality rate of over 36 percent.',

'In October, some 36,900 babies were born, up 1.1 percent from the same month last year. The rebound follows newborn numbers falling 3.6 percent and 3.7 percent in August and September. South Korea has been trying to push up its birthrate to prevent a decline in the national workforce.',

'Cho Hyun-ah, former vice president of Korean Air, sentenced to one year in prison. She caused a public uproar by forcing a cabin crew chief to disembark from a flight. The de facto heiress of the flag carrier was found to have ordered the taxiing plane to return to the gate.',

'Seoul-Tokyo ties have plunged to lowest levels in recent years mainly due to the sex slavery issue. Historians estimate the number of sex slaves at about 200,000 with only 53 South Korean victims alive today. Japan angered Seoul and Beijing by saying that its 1993 apology was the outcome of a political compromise.',

"South Korea's top financial regulator said Friday that the government-backed loan scheme has contributed to helping household borrowers. Local lenders had extended 20 trillion won in loans to help borrowers convert their short-term floating-rate mortgages into longer, fixed-rate ones. The size and soundness of household borrowing has come under serious question as the amount has ballooned.",

'Activists from the Humane Society International and the Change for Animals Foundation freed the dogs from a dog meat farm. Debate persists over whether recognizing dog as food would protect the eaters or the dogs. Unconfirmed tallies by dog meat lovers in 2012 put annual consumption at 2 million dogs.',

Figure 6. Summarized body of each cluster.

3. Combining the result and form issue

With two approaches applied distinctly, we processed issues by following criteria. 1. Extracting the same keyword, 2. Sentence in summarized results having the same keyword with 1. The Following are the results of top 10 issues for each year.

2015 : MERS, Birthrate decline, Seoul-Tokyo sex slavery issue, Vice President of Korean Air forcing cabin crew to disembark, Loan limit extension, Activist freed dogs, Yellow-dust, History textbook, Audi Volkswagen, dead man body at Jeju island.

2016 : Pyeongyang broadcasting messages, Chickens death in Chungcheong, South Korea's parliamentary election, Deoksu Palace restoration, contagious disease in pigs, National Research Center for Gifted and Talented Education, peace treaty negotiation, Park Geun-hye step down, citizen's sugar consumption, North Korea missile test

2017 : Cold wave, dinosaur footprint, adhesive patch, lowest birthrate, the Stellar Daisy ship missing, bird flu outbreak, art connoisseurs and city officials gathered at Seoul City Hall, Venture minister nominee Hong Jong-haak in Moon Jae-in committee, Bang Tae-hyun taking a bribe, mosquito-borne virus,

2 Issue Tracking

For Issue tracking, there are two tasks. First step is to choose two issues most suitable from the first task. Then automatically extract events related to each issue from news articles, with the number of at least five for each issue. And extract people, organizations, and places.
To perform the task, choosing the best suitable issues should be prioritized. In the issues extracted from Task 1, a suitable one should have an influential effect. To meet this standard, we set a criterion that picking the issue related to different years of issues, and having a high number of different issues related together. This process is done manually. First issue is North Korea, that appears highly related to 2 issues in 2016, Pyongyang broadcasting messages and North Korea missile test, President Park which is highly related to 1 issues in 2016, Park Geun-Hye step down, and related to 6 issues through 2015 and 2017, MERS, Seoul-Tokyo sex slavery issue, History textbook, South Korea's parliamentary election, peace treaty negotiation, Moon Jae-in committee,

And then, we applied LDA, one powerful method of topic modeling, to make issues become narrower, for the documents which contain keywords "North Korea" and "President Park" each.

```
In [11]: print_top_words(lda, nk_vectorizer, n_top_words=25)
Topic #1: korean north say north korea missile south border complex official city government source factory kaesong ship se
oul system range south korea report zone nuclear industrial artillery gps

Topic #2: india modi railway sinbo membership osjd trans university narendra student organization hyong choson jik unanimou
s pro bolster join yeo course member siberian attend shipyard interoperable

Topic #3: north korean complex south industrial say wage government firm worker kaesong park factory ministry company north
korea seoul decision gaeseong joint pay inter official south korea month

Topic #4: hwang ahn prime kyo minister act poland kuwait national polish cabinet security al provocation president posture
ensure court effort stress session grow european carry thornberry

Topic #5: visit world hiroshima south korea space say mexico peace city nobel korea plan official mexican program sri memor
ial year u.s obama prize camp lanka north korea state

Topic #6: intelligence nis agency spy lawmaker national lee service committee parliamentary say rep session file surveillan
ce byung official hacking team assembly chief closed program woo cheol

Topic #7: defense missile thaad system south korea deployment say u.s china seoul threat deploy altitude terminal security
high area south issue washington decision north korea battery korean official

In [12]: '''
         LDA 50 topics
         Topic 1: nulear/ missile
         Topic 7: Security
         Topic 9: nuclear
         Topic 10: People
         Topic 23: Chinese
         Topic 25: Olympic
         Topic 28: reactor
         Topic 32: nuclear
         Topic 35: military
         Topic 37: security
         Topic 39: nuclear
         Topic 40: bomb
         Topic 41: Trump
         Topic 44: economic
         Topic 45: north and south relation
         Topic 46: launching missile
         Topic 47: launching rocket
         Topic 48: attack/ military
         Topic 49: ?
         '''
```

Figure 7-1. Top 50 topic keyword after applying LDA for documents with "North Korea"

```
In [11]: print_top_words(lda, pp_vectorizer, n_top_words=25)
Topic #10: park economic government say people year economy country korea president public national policy state new growth
geun hye effort official meeting call nation minister south korea

Topic #11: house home park residence samseong dong president hye chung late stay geun semester fire southern country friday
year student office hee gumi yonhap sunday father

Topic #12: ferry sewol say president park kim family disaster people death sinking government victim ship memorial sink yea
r ceremony accident life rescue leave april kill geun

Topic #13: tillerson rex li leg faithful secretary acting ironclad shoulder final enrique inaugural biennial pena march tou
r importance trip scrutinize summit.the nieto jinjun threat mexico added.hwang

Topic #14: defense military say ministry official plan system force thaad air han resident seongju south korea government s
ite country launcher project base radar technology area year operation

Topic #15: rally park police seoul say president protest hold people saturday geun group hye street korea central protester
public gwanghwamun square citizen country take government old

Topic #16: document presidential wa dae cheong president park official leak say record cho office law aide officer chung ar

In [12]: '''
         LDA 50 topics
         Topic 0: Iran
         Topic 1: development
         Topic 2: ?
         Topic 3: ?
         Topic 4: city?
         Topic 5: disaster
         Topic 6: -
         Topic 7: Between country
         Topic 8:
         Topic 9: education
         Topic 10: economic
         Topic 11: residence
         Topic 12: sewol
         Topic 16: president
         Topic 19: MERS
         Topic 21: president park
         Topic 26: Party
         Topic 27: nuclear
         Topic 30: Electronics
         Topic 34: unification
         Topic 36: mongolia
         Topic 48: North Korea
         '''
```

Figure 7-2. Top 50 topic keyword after applying LDA for documents with "President Park"

The topic keyword on Figure 7-1 and Figure 7-2 shows ordered by most frequent keyword. We chose “missile” on the result of North Korea, so that it turns into “North Korea Missile” because “missile” is top 1 result of Figure 7-1. While maintaining President Park as an issue because there were no such proper keywords that make the issue narrower.

2.1 On-issue Event Tracking

For on-issue event tracking specifically, the events should be related to the issue in chronological order.

First we restricted the data with the section in the data table, to “North Korea” in terms of dimensionality reduction cause it directly contains keywords of selected issue, “North Korea Missile”, thus have high possibility that meaningful results are contained in it, also less meaningful results are not contained in it. Also for Park Geun-Hye, restrict the section “politics” for the perspective of dimensionality reduction, cause restricting the section can effectively reduce the dimension via excluding articles with another section.

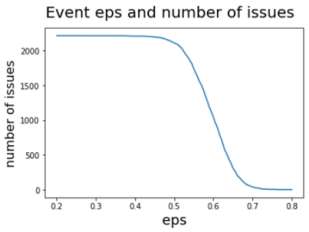
Although there were 1285 Nan-section data, even can contains the keyword “North Korea”, most of them has each section, so it could be ignored.

```
df['section'].isna().value_counts()
```

```
False    22484
True      1285
Name: section, dtype: int64
```

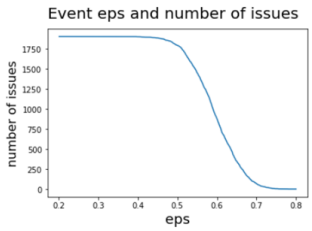
Figure 8. Number of articles with Nan-Section

Then after applying DBSCAN and taking the first maximum point value of change of slope, rank by order of number of documents clustered together in a cluster, similarly with the process of Task 1 already performed above.



	summary	label	freq
0	North Korea appears to be advancing its submar...	100	13
1	Group of 13 North Korean defectors who worked ...	1265	13
2	Around 200 propaganda leaflets from the North ...	547	10
3	254 people of 90 families from the South met w...	2216	9
4	84 percent of Americans think President Donald...	203	8
5	UN nuclear watchdog should revise its view on ...	778	7
6	U.N. Security Council adopted Resolution 2270 ...	1403	7
7	IOC President Thomas Bach is seeking to visit ...	13	7
8	North Korea has featured a flight crew of its ...	751	6
9	Shallow 3.5-magnitude earthquake hits North Ko...	342	6

Figure 9-1. Relation between Epsilon value & Number of documents clustered together, Top 10 cluster ranked by number of documents in a cluster after applying DBSCAN on Section NK and “Missile”

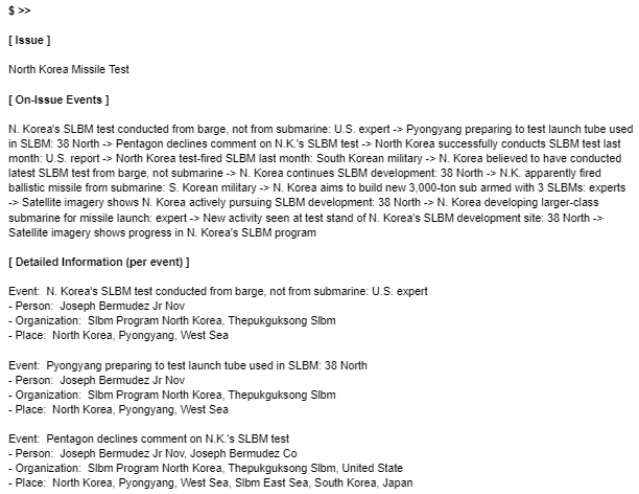


	summary	label	freq
0	Moon Jae-in of the liberal Democratic Party ga...	360	11
1	Former UN Secretary-General Ban Ki-moon's retu...	642	9
2	President Park Geun-hye on Friday granted spec...	1212	9
3	President Park Geun-hye vetoed the controversi...	1329	8
4	67.8 percent of respondents said they approved...	37	8
5	South Korean President Park Geun-hye returned ...	1939	8
6	Eight former renegade lawmakers returned to th...	28	7
7	Seoul court on Tuesday held the first preparat...	525	7
8	President Park Geun-hye arrived in Seoul on We...	1373	7
9	The Federation of Korean Trade Unions said Tue...	1504	7

Figure 9-2. Relation between Epsilon value & Number of documents clustered together, Top 10 cluster ranked by number of documents in a cluster after applying DBSCAN on Section Politics and “Missile”

Then, we restricted the data region to just focus on exactly one cluster, which is highly clustered together but actually constructed with different documents. On the first issue, we chose the first cluster due to its highest number of documents, and the sixth cluster on the second issue for its content is the Middle East tour so it will highly contain chronological data.

Thus the final form of issue is “North Korea Missile test” and “President Park Geun-hye’s Middle-East tour” Divide each document and sort them in temporal lines, apply spaCy for NER tagging to extract person, organization, location, and finally form an event with the title and content of each article. The event was formed by simply using the title of each document with the assumption that title indicates core information of content of the body. The final outcome



$ >>

[Issue]

North Korea Missile Test

[On-Issue Events]

N. Korea's SLBM test conducted from barge, not from submarine: U.S. expert -> Pyongyang preparing to test launch tube used in SLBM: 38 North -> Pentagon declines comment on N.K.'s SLBM test -> North Korea successfully conducts SLBM test last month: U.S. report -> North Korea test-fired SLBM last month: South Korean military -> N. Korea believed to have conducted latest SLBM test from barge, not submarine -> N. Korea continues SLBM development: 38 North -> N.K. apparently fired ballistic missile from submarine: S. Korean military -> N. Korea aims to build new 3,000-ton sub armed with 3 SLBMs: experts -> Satellite imagery shows N. Korea actively pursuing SLBM development: 38 North -> N. Korea developing larger-class submarine for missile launch: expert -> New activity seen at test stand of N. Korea's SLBM development site: 38 North -> Satellite imagery shows progress in N. Korea's SLBM program

[Detailed Information (per event)]

Event: N. Korea's SLBM test conducted from barge, not from submarine: U.S. expert
- Person: Joseph Bermudez Jr Nov
- Organization: Slbm Program North Korea, Thepukguksong Slbm
- Place: North Korea, Pyongyang, West Sea

Event: Pyongyang preparing to test launch tube used in SLBM: 38 North
- Person: Joseph Bermudez Jr Nov
- Organization: Slbm Program North Korea, Thepukguksong Slbm
- Place: North Korea, Pyongyang, West Sea

Event: Pentagon declines comment on N.K.'s SLBM test
- Person: Joseph Bermudez Jr Nov, Joseph Bermudez Co
- Organization: Slbm Program North Korea, Thepukguksong Slbm, United State
- Place: North Korea, Pyongyang, West Sea, Slbm East Sea, South Korea, Japan

Figure 10. Result of On-issue event tracking

2.2 Related-issue Event Tracking

For related-issue event tracking specifically, the related events are not directly tied to particular issues. In order to perform it, Topics extracted from topic modeling (LDA) for each issue. Then we randomly selected 5 documents, excluding inner words that formed the issue. For the first issue, “Missile test” was excluded in the North Korea section articles, and for the second issue, we excluded the “Middle-East”, in the politics section, and finally applied spaCy to extract person, organization, location. The related-event name is simply used by the title of each document.


[Issue]

North Korea Missile Test

[Related-Issue Events]

Gyeonggi Province rebrands English Village to Change Up Campus, Korea to introduce 'spiciness' labels for instant noodle, Trump to deliver speech at National Assembly, China's special envoy to N. Korea returns to Beijing, [Photo News] Cold snap freezes Han River, Korean veteran confesses killing civilians at Vietnam War

[Detailed Information (per event)]

Event: Korean veteran confesses killing civilians at Vietnam War
- Person: Yoon Mi Hyang, Viet Cong, Eun Byel- Organization:
- Place: Vietnam, Korea, Japan
Event: [Photo News] Cold snap freezes Han River
- Person: Yoon Mi Hyang, Viet Cong, Eun Byel, Yanggu Inje Cheorwon Issue, Hyun Koo Korea Herald, Hyun Koo Korea, Park Hyun Koo
- Organization: - Place: Vietnam, Korea, JapanEvent: China's special envoy to N. Korea returns to Beijing
- Person: Yoon Mi Hyang, Viet Cong, Eun Byel, Yanggu Inje Cheorwon Issue, Hyun Koo Korea Herald, Hyun Koo Korea, Park Hyun Koo, Kim Jong Un, Xi Jinping, China Ji Jae Ryong, Choe Ryong Hae, Ri Su Yong, China Xi, Kim Il, Kim Jong Il, China Accord, Xi, Kim, Baik Tae Hyun- Organization: International Liaison Department, Communist Party, Congress, Xinhua News Agency, Central Committee North Rule Workers Party, Kcna Report- Place: Vietnam, Korea, Japan, China, Pyongyang, Beijing, North Korea, South Korea

Figure 11. Result of Related-issue event tracking

REFERENCES

[1] https://www.kaggle.com/danielwolffram/topic-modeling-finding-related-articles

[2] https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da

[3] https://huggingface.co/docs/transformers/model_doc/bart

[4] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

[5] https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b

[6] https://towardsdatascience.com/keyword-extraction-python-tf-idf-textrank-topicrank-yake-bert-7405d51cd839

[7] https://towardsdatascience.com/machine-learning-clustering-dbscan-determine-the-optimal-value-for-epsilon-eps-python-example-3100091cfbc

[8] https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/