Team #3

Trend Analysis and Event Tracking

Soyoung Yoon, Junseop Ji, Jihee Park

Contents

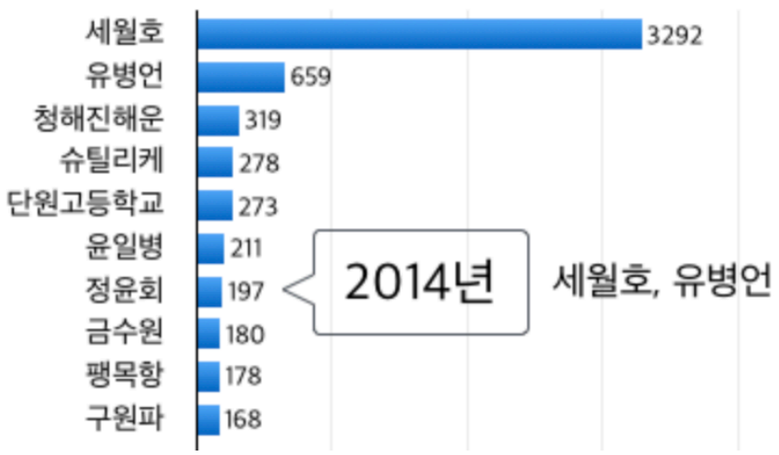
Introduction

Enormous Amount of Data

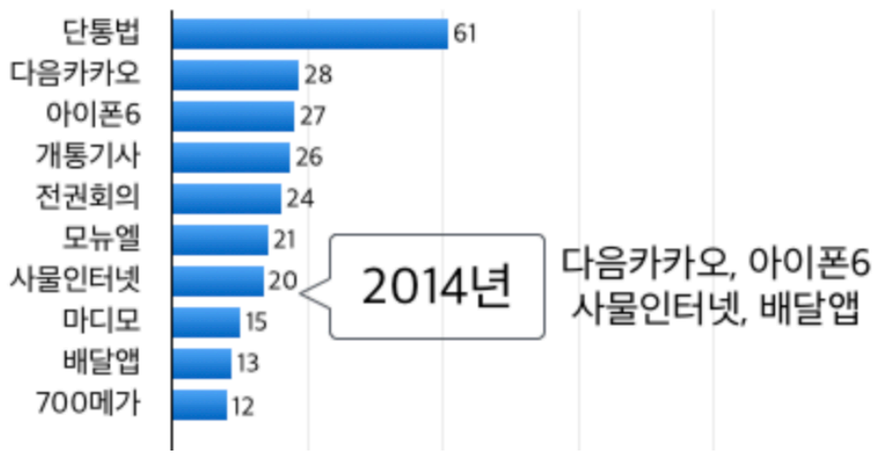
Need to Extract Trends/Issues

Using our text mining technique!

5년간 새로운 뉴스 주제어 트렌드
- 빈도순위 10위



5년간 새로운 뉴스 주제어 트렌드
- IT



Trend Analysis

Trend Analysis = Topic Modeling

ACL 2018 paper
Modify the input document-term matrix of standard LDA.

ACL Anthology FAQ Corrections Submissions Search...

Improving Topic Quality by Promoting Named Entities in Topic Modeling

Katsiaryna Krasnashchok, Salim Jouili

Abstract

News related content has been extensively studied in both topic modeling research and named entity recognition. However, expressive power of named entities and their potential for improving the quality of discovered topics has not received much attention. In this paper we use named entities as domain-specific terms for news-centric content and present a new weighting model for Latent Dirichlet Allocation. Our experimental results indicate that involving more named entities in topic descriptors positively influences the overall quality of topics, improving their interpretability, specificity and diversity.

PDF
BibTeX
Search
Poster

Trend Analysis = Topic Modeling

ACL 2018 paper
Modify the input document-term matrix of standard LDA.

1. Independent Named Entity Promoting.

$$tf_w = \begin{cases} \alpha * tf_w & \text{if } w \text{ is NE} \\ tf_w & \text{otherwise} \end{cases} \quad (1)$$

For example:

d\w	good	time	ne_nhl	play	ne_espn
D_1	4	2	1*α	6	0*α
D_2	5	3	2*α	2	1*α
D_3	8	4	0*α	4	2*α

By varying the value of α, we can control the importance of named entities in the corpus.

2. Document Dependent Named Entity Promoting.

$$tf_{dw} = \begin{cases} tf_{dw} + \max_w tf_{dw} & \text{if } w \text{ is NE} \\ tf_{dw} & \text{otherwise} \end{cases} \quad (2)$$

where $\max_w tf_{dw}$ is the most frequent term in the document. For example:

d\w	good	time	ne_nhl	play	ne_espn
D_1	4	2	1+**6**	**6**	0
D_2	**5**	3	2+**5**	2	1+**5**
D_3	**8**	4	0	4	2+**8**

Preferred method, since it does not introduce any new parameters into LDA.

Trend Analysis = Topic Modeling

ACL 2018 paper

Modify the input document-term matrix of standard LDA.

1. Independent Named Entity Promoting.

$$tf_w = \begin{cases} \alpha * tf_w & \text{if } w \text{ is NE} \\ tf_w & \text{otherwise} \end{cases} \quad (1)$$

For example:

d\w	good	time	ne_nhl	play	ne_espn
D_1	4	2	1*α	6	0*α
D_2	5	3	2*α	2	1*α
D_3	8	4	0*α	4	2*α

By varying the value of α, we can control the importance of named entities in the corpus.

2. Document Dependent Named Entity Promoting.

$$tf_{dw} = \begin{cases} tf_{dw} + \max_w tf_{dw} & \text{if } w \text{ is NE} \\ tf_{dw} & \text{otherwise} \end{cases} \quad (2)$$

where $\max_w tf_{dw}$ is the most frequent term in the document. For example:

d\w	good	time	ne_nhl	play	ne_espn
D_1	4	2	1+**6**	**6**	0
D_2	**5**	3	2+**5**	2	1+**5**
D_3	**8**	4	0	4	2+**8**

Preferred method, since it does not introduce any new parameters into LDA.

t-d # 1	play	paper	tree	Sally	apple	Japan	river
before	20	12	3	2	15	5	30
after (1)							
after (2)							

Trend Analysis = Topic Modeling

ACL 2018 paper

Modify the input document-term matrix of standard LDA.

1. Independent Named Entity Promoting.

$$tf_w = \begin{cases} \alpha * tf_w & \text{if } w \text{ is NE} \\ tf_w & \text{otherwise} \end{cases} \quad (1)$$

For example:

d\w	good	tim
D_1	4	2
D_2	5	3
D_3	8	4

By varying the value
tance of named entit

2. Document Dependent Named Entity Promoting.

$$tf_{dw} = \begin{cases} tf_{dw} + \max\limits_{w} tf_{dw} & \text{if } w \text{ is NE} \\ tf_{dw} & \text{otherwise} \end{cases} \quad (2)$$

where $\max tf_{dw}$ is the most frequent term in the doc-

α : 10, $\max\limits_{w} tf_{dw}$: 30,

play	ne_espn
6	0
2	1+**5**
4	2+**8**

not introduce any

t-d # 1	play	paper	tree	Sally	apple	Japan	river
before	20	12	3	2	15	5	30
after (1)							
after (2)							

Trend Analysis = Topic Modeling

ACL 2018 paper

Modify the input document-term matrix of standard LDA.

1. Independent Named Entity Promoting.

$$tf_w = \begin{cases} \alpha * tf_w & \text{if } w \text{ is NE} \\ tf_w & \text{otherwise} \end{cases} \quad (1)$$

For example:

d\w	good	tim
D_1	4	2
D_2	5	3
D_3	8	4

By varying the value
tance of named entit

2. Document Dependent Named Entity Promoting.

$$tf_{dw} = \begin{cases} tf_{dw} + \max_w tf_{dw} & \text{if } w \text{ is NE} \\ tf_{dw} & \text{otherwise} \end{cases} \quad (2)$$

where $\max tf_{dw}$ is the most frequent term in the doc-

α : 10, $\max_w tf_{dw}$: 30,

Named Entities: Sally, Japan

play	ne_espn
6	0
2	1+**5**
4	2+**8**

not introduce any

t-d # 1	play	paper	tree	Sally	apple	Japan	river
before	20	12	3	2	15	5	30
after (1)							
after (2)							

Trend Analysis = Topic Modeling

ACL 2018 paper

Modify the input document-term matrix of standard LDA.

1. Independent Named Entity Promoting.

$$tf_w = \begin{cases} \alpha * tf_w & \text{if } w \text{ is NE} \\ tf_w & \text{otherwise} \end{cases} \quad (1)$$

For example:

d\w	good	tim
D_1	4	2
D_2	5	3
D_3	8	4

By varying the value
tance of named entit

2. Document Dependent Named Entity Promoting.

$$tf_{dw} = \begin{cases} tf_{dw} + \max_w tf_{dw} & \text{if } w \text{ is NE} \\ tf_{dw} & \text{otherwise} \end{cases} \quad (2)$$

where $\max tf_{dw}$ is the most frequent term in the doc-

α : 10, $\max_w tf_{dw}$: 30,

Named Entities: Sally, Japan

play	ne_espn
6	0
2	1+**5**
4	2+**8**

not introduce any

t-d # 1	play	paper	tree	Sally	apple	Japan	river
before	20	12	3	2	15	5	30
after (1)	20	12	3	2 * 10	15	5 * 10	30
after (2)							

Trend Analysis = Topic Modeling

ACL 2018 paper

Modify the input document-term matrix of standard LDA.

1. Independent Named Entity Promoting.

$$tf_w = \begin{cases} \alpha * tf_w & \text{if } w \text{ is NE} \\ tf_w & \text{otherwise} \end{cases} \quad (1)$$

2. Document Dependent Named Entity Promoting.

$$tf_{dw} = \begin{cases} tf_{dw} + \max_w tf_{dw} & \text{if } w \text{ is NE} \\ tf_{dw} & \text{otherwise} \end{cases} \quad (2)$$

For example:

d\w	good	tim
D_1	4	2
D_2	5	3
D_3	8	4

By varying the value
tance of named entit

where $\max tf_{dw}$ is the most frequent term in the doc-

α : 10, $\max_w tf_{dw}$: 30,

Named Entities: Sally, Japan

-> Named Entities get promoted!

play	ne_espn
6	0
2	1+**5**
4	2+**8**

not introduce any

t-d # 1	play	paper	tree	Sally	apple	Japan	river
before	20	12	3	2	15	5	30
after (1)	20	12	3	2 * 10	15	5 * 10	30
after (2)	20	12	3	2 + 30	15	5 + 30	30

Topic Modeling Process

Overview

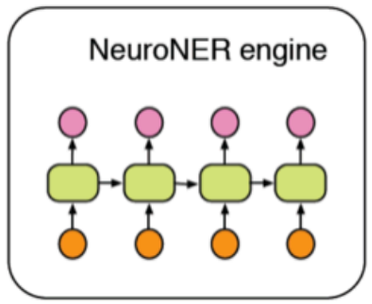
Tokenization

Named Entity Recognition

LDA + NER

Topic Modeling Process

Overview

Tokenization	NER	neuroNER	LDA with NER	Tune LDA
~~- lemmatization~~ - nltk.tokenize	- pos_tag - multi-word information	 - multi-word info	- NER token * 10	- Remove stopwords - Remove (,), ., \xec… - num_topics: 10 - num_topics: 10

Tokenization

1. Reduce Vocab Size

   - Lemmatization
     was -> be
     better -> good
     meeting -> meeting

   - Remove Stopwords [I, am, for, so, …]

Tokenization

1. Reduce Vocab Size

- Lemmatization

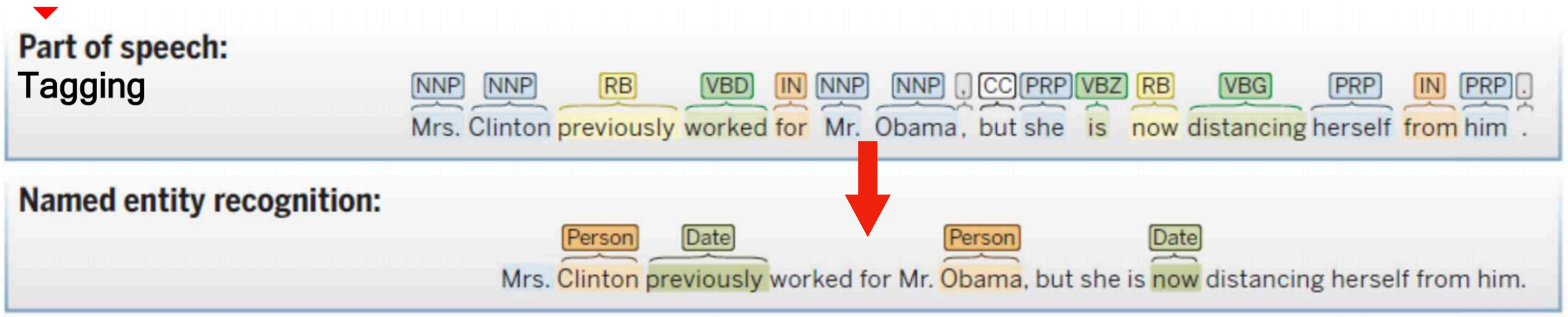
was -> be
better -> good
meeting -> meeting

CANNOT do POS tagging

- Remove Stopwords

[I, am, for, so, …]

POS tagging needs to be done before NER!



Tokenization

1. Reduce Vocab Size

   - Lemmatization — was -> be, better -> good, meeting -> meeting

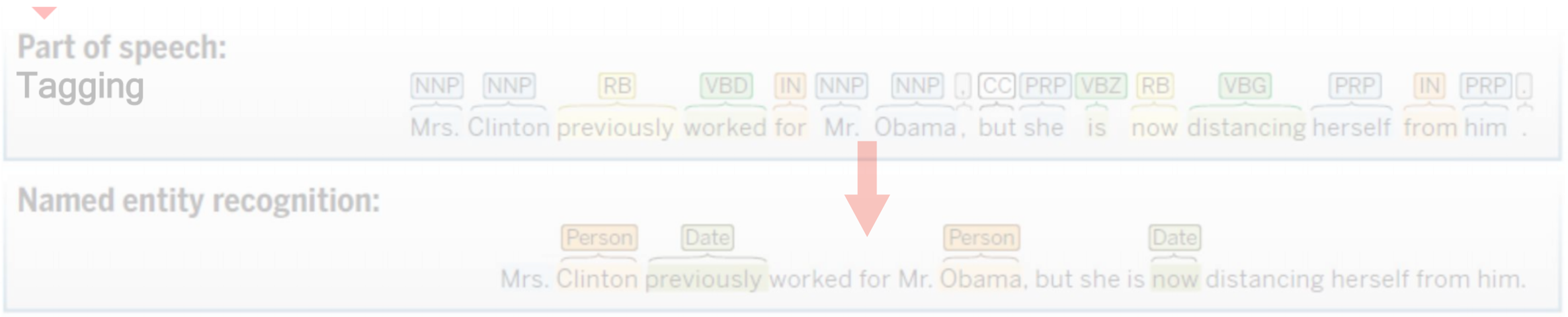
   - Remove Stopwords — [I, am, ...]

CANNOT do POS tagging

1. nltk.tokenize()

in order to not lose any information before POS tagging

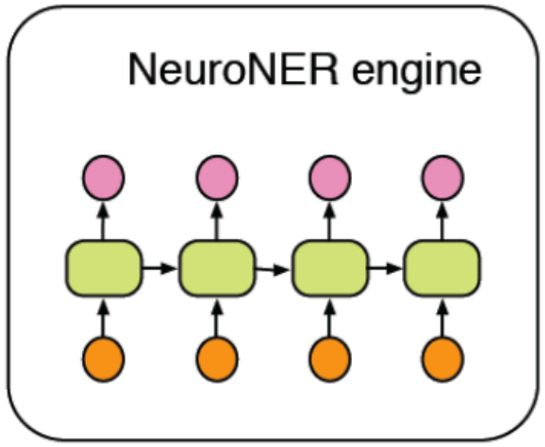
POS tagging needs to be done before NER!

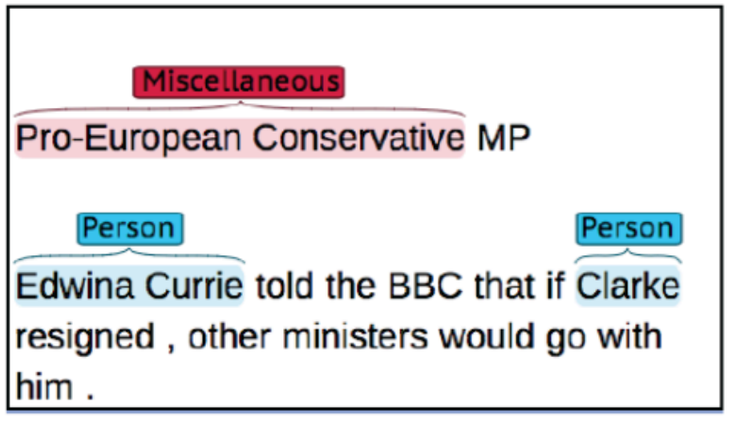


Named Entity Recognition

By NeuroNER

Pro-European Conservative MP Edwina Currie told the BBC that if Clarke resigned , other ministers would go with him .





http://neuroner.com/

- Neural network trained on CONLL 2003 dataset

- 4 types of NE: person, location, organization and miscellaneous

LDA + NER

Topics	Test	20 Newsgroups			Reuters-2013		
		C_v	Lift	Excl.	C_v	Lift	Excl.
20	Baseline Unigram	0,534	3,390	0,788	0,539	3,891	0,610
	Baseline NE	0,503	3,273	0,767	0,559	4,059	0,598
	NE Independent (x1,3)	0,494	3,394	0,755	0,551	4,209	0,563
	NE Independent (x1,5)	0,527	3,464	0,770	0,552	4,308	0,618
	NE Independent (x2)	0,525	3,756	0,797	0,548	4,449	0,640
	NE Independent (x2,5)	0,539	3,779	0,765	0,550	4,661	0,635
	NE Independent (x5)	**0,543**	5,071	0,898	0,517	5,701	0,708
	NE Independent (x10)	0,486	**6,416**	**0,950**	0,511	**6,560**	**0,773**
	NE Doc. Dependent	**0,543**	4,600	0,780	**0,566**	5,749	0,625

Independent Named Entity Promoting works best when α : 10!

-> Modify document-term matrix
with named entities promoted by the factor of 10.

Overall Example

Donald Trump and Interstate Properties said they were holding preliminary discussions regarding possible joint acquisition of Alexanders Inc at 47 dlrs per share. The possible acquisition is subject to any applicable real estate gains and transfer taxes, the joint statement said. … be offered, the statement said.

Tokenization: [Donald, Trump, and, Interstate, Properties, said, ….. statement, said]

Overall Example

Donald Trump and Interstate Properties said they were holding preliminary discussions regarding possible joint acquisition of Alexanders Inc at 47 dlrs per share. The possible acquisition is subject to any applicable real estate gains and transfer taxes, the joint statement said. … be offered, the statement said.

Tokenization: [Donald, Trump, and, Interstate, Properties, said, ….. statement, said]

Named Entities: ALEXANDERS, Donald Trump and Interstate Properties, Alexanders Inc, Trump and Interstate, Alexander, …

Overall Example

Donald Trump and Interstate Properties said they were holding preliminary discussions regarding possible joint acquisition of Alexanders Inc at 47 dlrs per share. The possible acquisition is subject to any applicable real estate gains and transfer taxes, the joint statement said. … be offered, the statement said.

Tokenization: [Donald, Trump, and, Interstate, Properties, said, ….. statement, said]

Named Entities: ALEXANDERS, Donald Trump and Interstate Properties, Alexanders Inc, Trump and Interstate, Alexander, …

Make doc2bow for each document

Donald Trump	and	said	ALEXANDERS	statement	…
1	10	5	2	3	…

Overall Example

Donald Trump and Interstate Properties said they were holding preliminary discussions regarding possible joint acquisition of Alexanders Inc at 47 dlrs per share. The possible acquisition is subject to any applicable real estate gains and transfer taxes, the joint statement said. … be offered, the statement said.

Tokenization: [Donald, Trump, and, Interstate, Properties, said, ….. statement, said]

Named Entities: ALEXANDERS, Donald Trump and Interstate Properties, Alexanders Inc, Trump and Interstate, Alexander, …

Remove Stopwords

Donald Trump	~~and~~	said	ALEXANDERS	statement	…
1	~~1~~	5	2	3	…

Overall Example

Donald Trump and Interstate Properties said they were holding preliminary discussions regarding possible joint acquisition of Alexanders Inc at 47 dlrs per share. The possible acquisition is subject to any applicable real estate gains and transfer taxes, the joint statement said. … be offered, the statement said.

Tokenization: [Donald, Trump, and, Interstate, Properties, said, ….. statement, said]

Named Entities: ALEXANDERS, Donald Trump and Interstate Properties, Alexanders Inc, Trump and Interstate, Alexander, …

LDA Promoting

Donald Trump	[illegible]	said	ALEXANDERS	statement	…
1 * 10	[illegible]	5	2 * 10	3	…

Overall Example

Donald Trump and Interstate Properties said they were holding preliminary discussions regarding possible joint acquisition of Alexanders Inc at 47 dlrs per share. The possible acquisition is subject to any applicable real estate gains and transfer taxes, the joint statement said. … be offered, the statement said.

Tokenization: [Donald, Trump, and, Interstate, Properties, said, ….. statement, said]

Named Entities: ALEXANDERS, Donald Trump and Interstate Properties, Alexanders Inc, Trump and Interstate, Alexander, …

Feed it each year

Donald Trump	and	said	ALEXANDERS	statement	…
20	10	5	20	3	…
2	9	7	12	5	…
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	…
5	[illegible]	89	42	1	…

Overall Example

LDA model

Donald Trump	and	said	ALEXANDERS	statement	…
20	10	5	20	3	…
2	9	7	12	5	…
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	…
5	[illegible]	89	42	1	…

Overall Example

LDA model

GET Top 10 trend results each year

2015	2016	2017
Roh Moo-hyun's Oppression by NIS	Political scandal about President Park	PyeongChang Olympics
Thaad placement	Zika Virus	Choi sun-sil gate and presidential impeachment
MERS	Seoul and Gyeonggi Province issue	North Korea relationship
...	...	...

On-issue Event Tracking

Process

Monthly Division

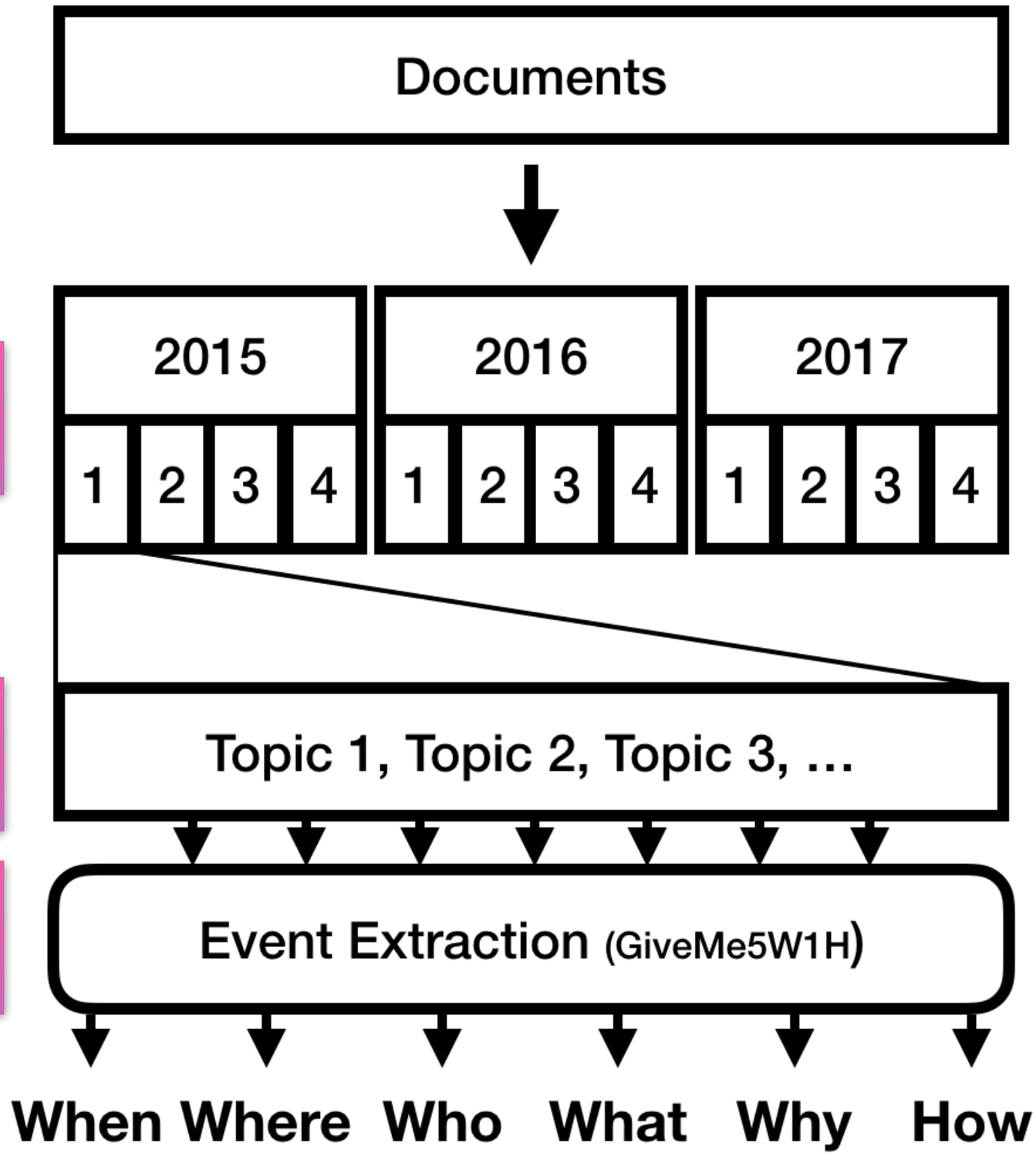
Articles in the Group Classification

Event Extraction

Monthly Division

Articles in the Group Classification

Event Extraction



Monthly Division

Divide all news articles monthly.
2015 Jan, 2015 Feb, ..., 2017 Dec, ~~2018 Jan~~.

```
def split_by_month(df):
    df['timestamp'] = pandas.to_datetime(df[' time'])
    df = df.sort_values(by=['timestamp'])
    month = df['timestamp'].dt.to_period('M')
    return df.groupby(month)
```

Extract 1 event from 1 monthly group.

Why "Monthly"?

Monthly Division

For example:

the issue MERS: May 20th. 2015 ~ Jan 26th. 2016
(Approximately 9 months)

Yearly: 1 Event
Quarterly: 3 Events
Monthly: 9 Events —> Just Fit
Daily: Too Much

Articles in the Group Classification

With yearly LDA model, trained at Trend Analysis process, classify the documents in the monthly groups

```
body = dictionary.doc2bow(row['tokenized_body'])
vector = model[body]
cat = max(vector, key=lambda x: x[1])[0]
```

10 classified groups for each month
Total 3 x 12 x 10 = 360 groups

Event Extraction

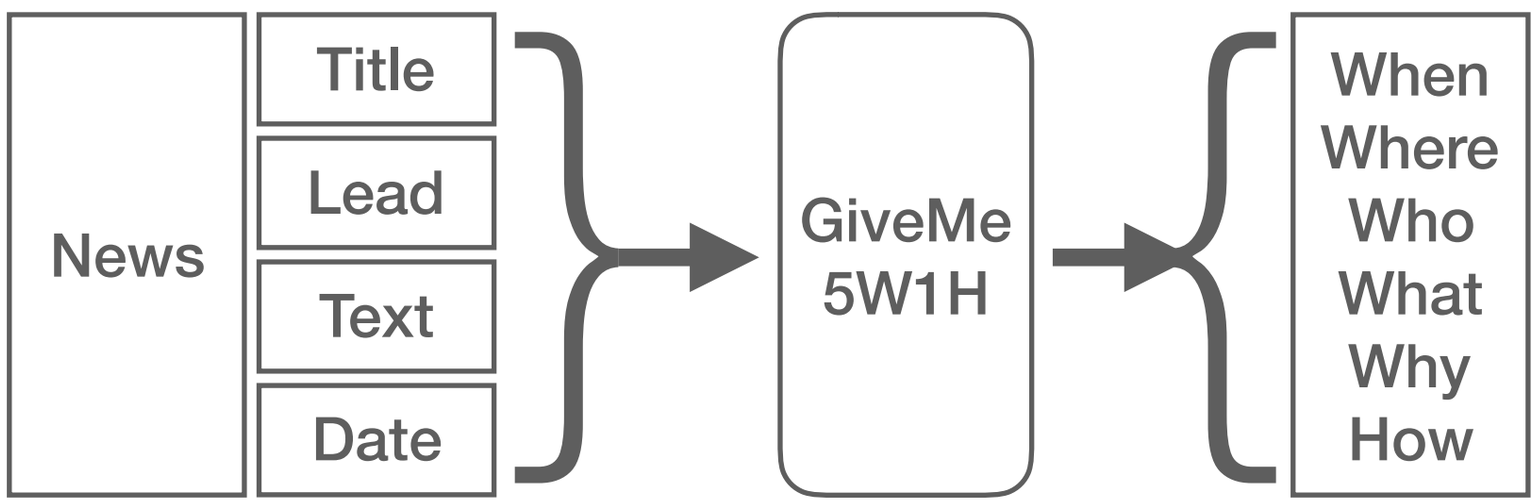
For each group, extract the events with a library "GiveMe5W1H"

Taliban attacks German consulate in northern Afghan city of Mazar-i-Sharif with truck bomb

The death toll from a powerful Taliban truck bombing at the German consulate in Afghanistan's Mazar-i-Sharif city rose to at least six Friday, with more than 100 others wounded in a major militant assault.

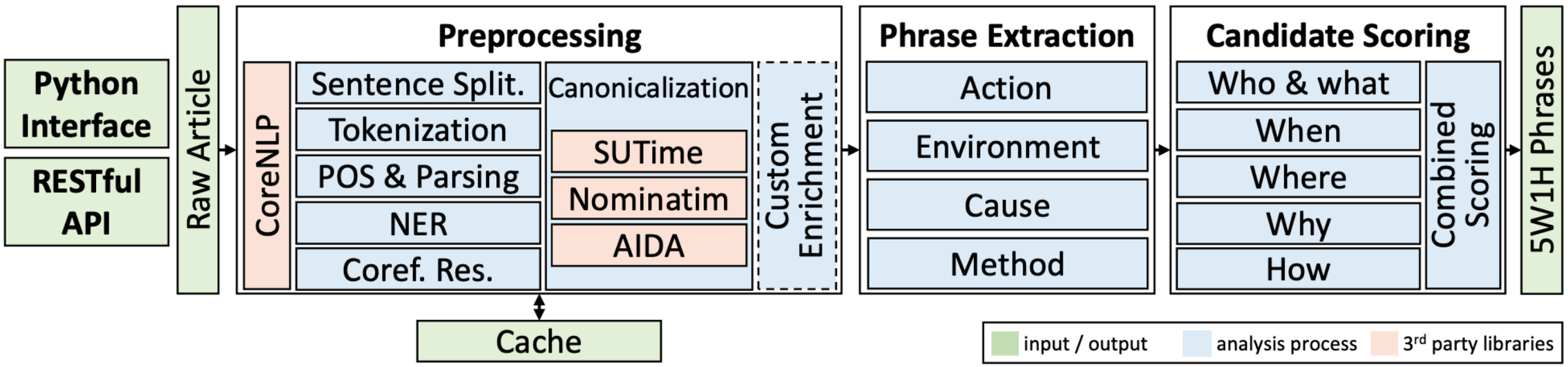
The Taliban said the bombing late Thursday, which tore a massive crater in the road and overturned cars, was a "**revenge attack**" for US air strikes this month in the volatile province of Kunduz that left 32 civilians dead. […]

The state-of-the-art(2019) tool for extracting when/where/who/what/why/how features from the document by giving title/lead/text/published date.



Event Extraction

Pipeline structure of the library:



Count the term(5W1H) frequency, and treat it as a score, extract the article's title which has highest score.
—> Representative article of the group

Off-issue Event Tracking

Process

Articles in the Group Classification

BoW Feature Extraction

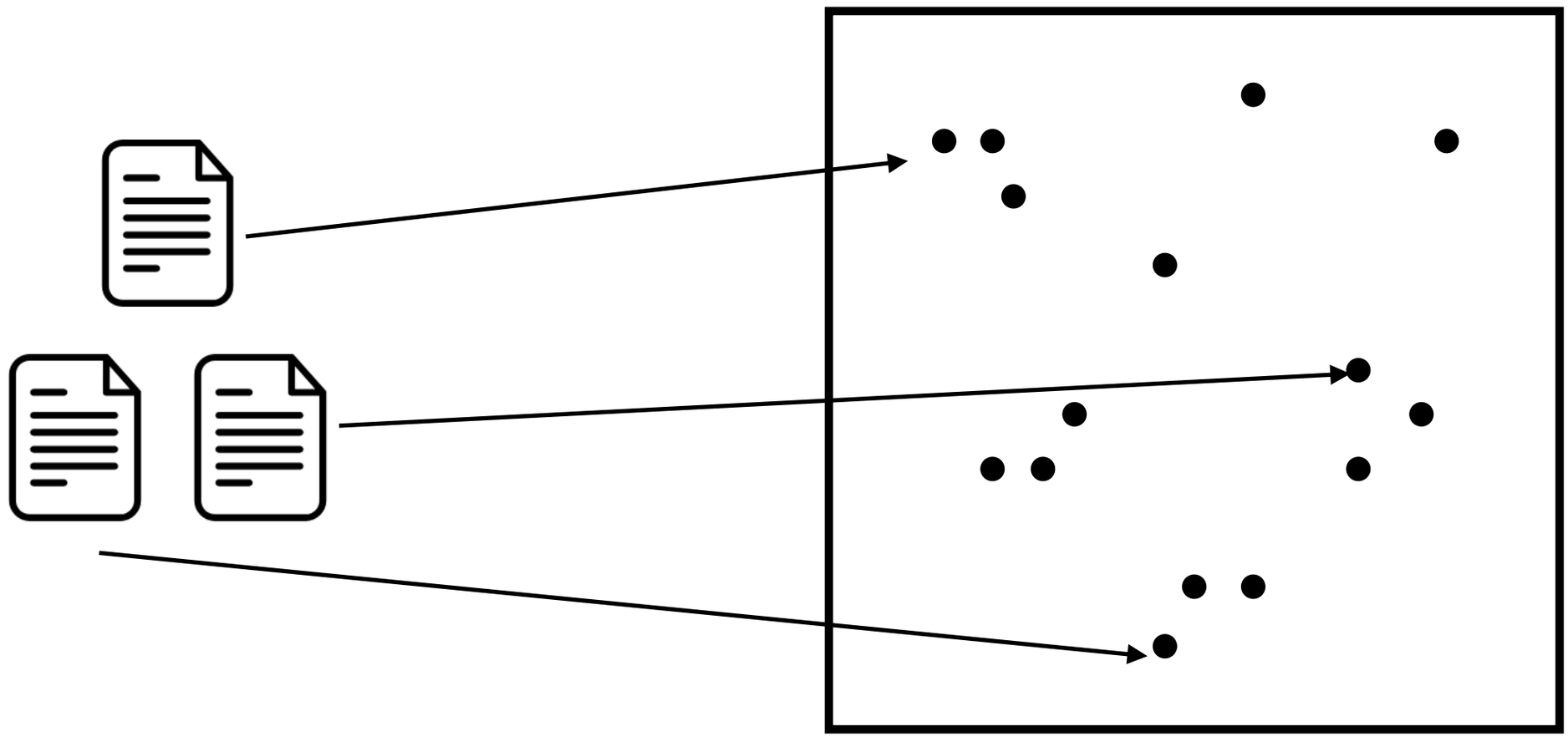
Document Clustering

Get Representative Document

Event Extraction

BoW Feature Extraction

With documents of specific issue in a year, we map the documents to vector space to calculate its similarity.



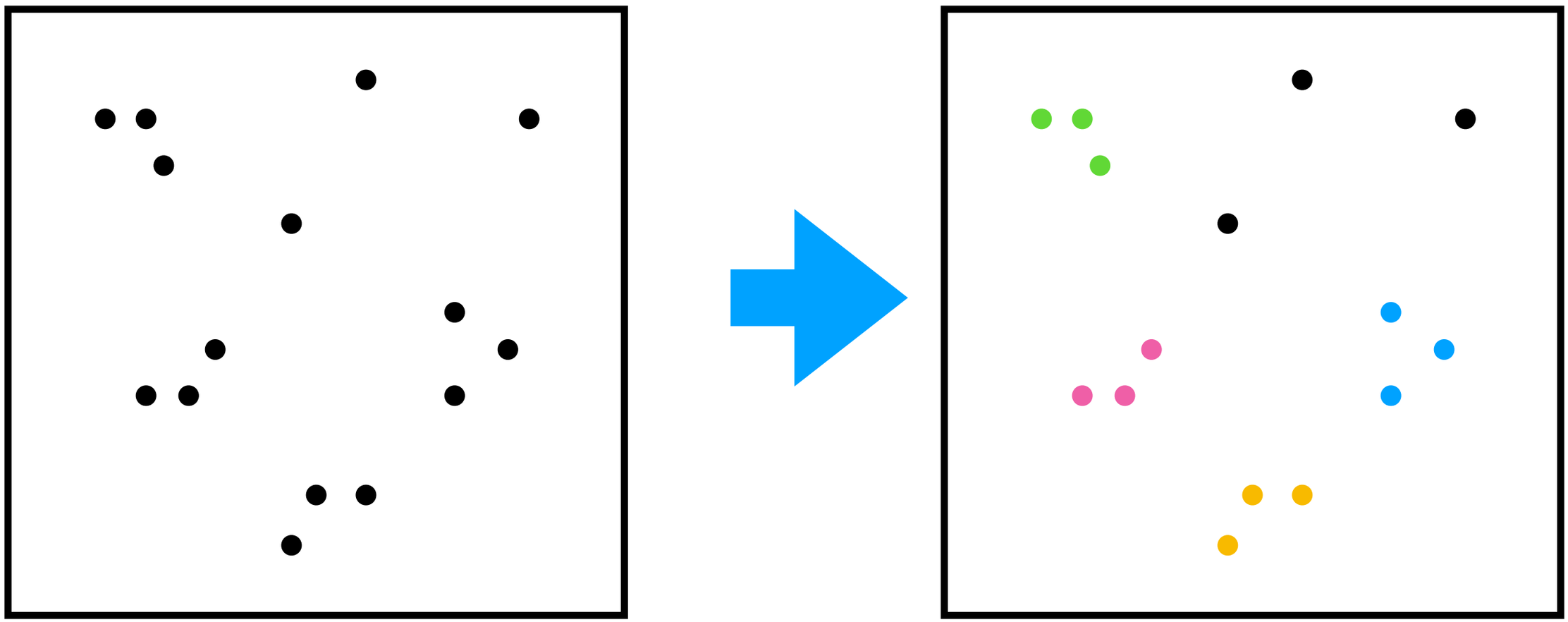
BoW Feature Extraction

We use BoW feature extraction algorithm to convert document to vector space of normalized word frequency.

```
sp = sparse.dok_matrix((tlen, clen + 1))
for i in range(tlen):
    bow = dictionary.doc2bow(texts.iloc[i])
    wordSum = sum(map(lambda x: x[1], bow))
        for wordIdx, wordCount in bow:
        sp[i, wordIdx]  = wordCount * WVec / wordSum
return sp
```

Document Clustering

Given vector space with transformed document, we need to cluster similar documents which may refer same event.



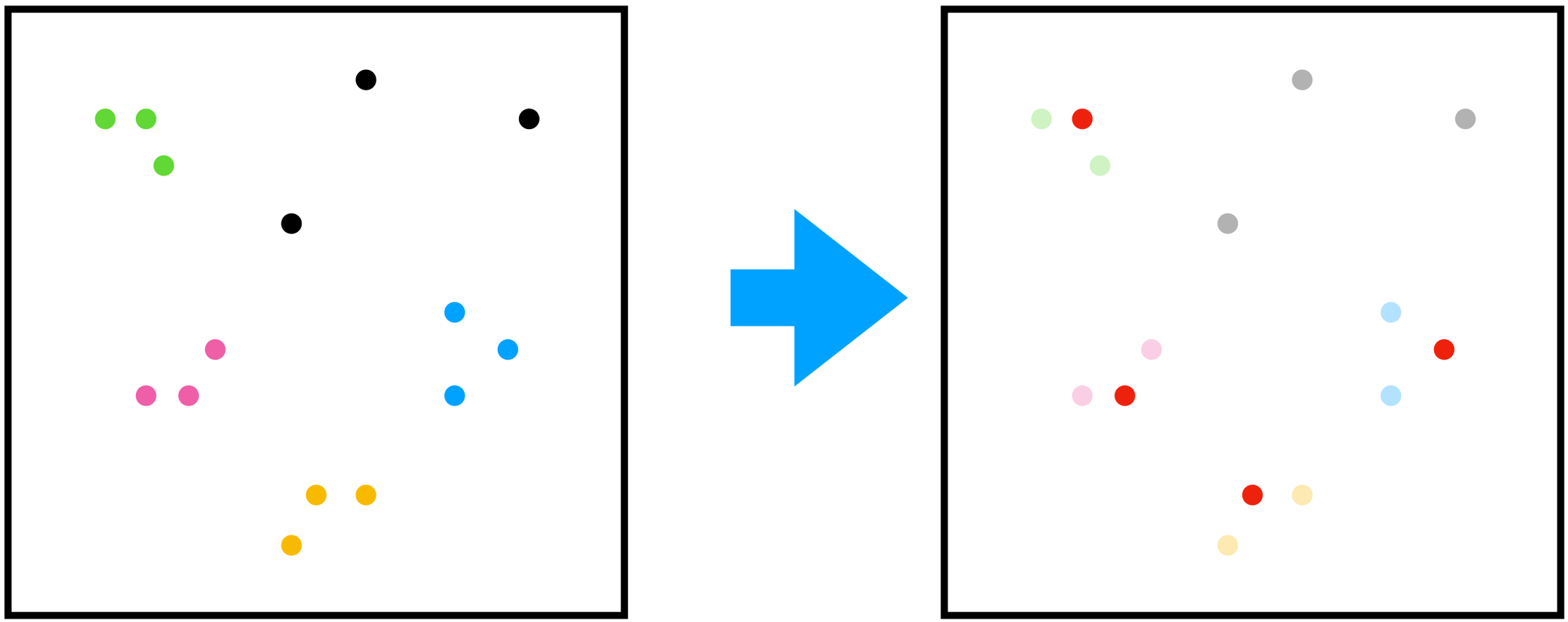
Document Clustering

We use DBSCAN algorithm for density-based clustering.

```
tc = bow2vec(dictionary, topic0['neuroner_tokenized'], clen)
for i in range(tlen):
    tc[i, clen] = topic0['timestamp'].iloc[i] * WTime
clustering = DBSCAN(eps=EPS, min_samples=MSAMPLE).fit_predict(tc)
return clustering
```

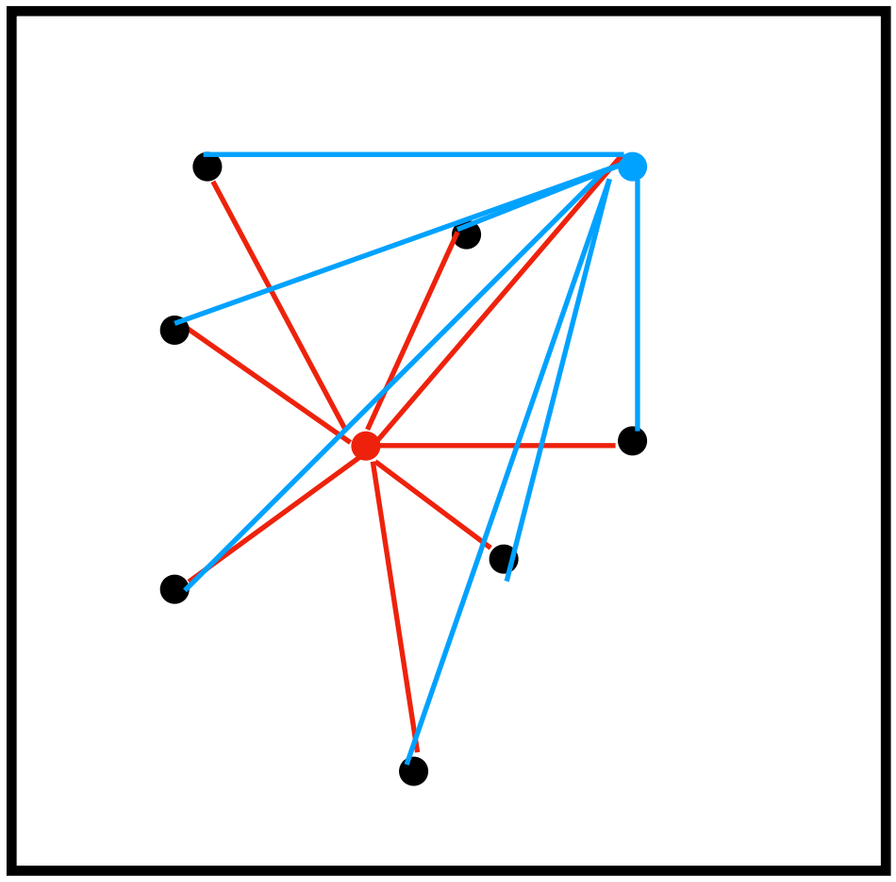
Get Representative Document

Now we have groups of issue. To extract event description, we use one representative document to extract event.



Get Representative Document

In a group, we define representative document as a document which has smallest difference with each document.



Sum of red line < Sum of blue line

Experiment

Reuters Dataset

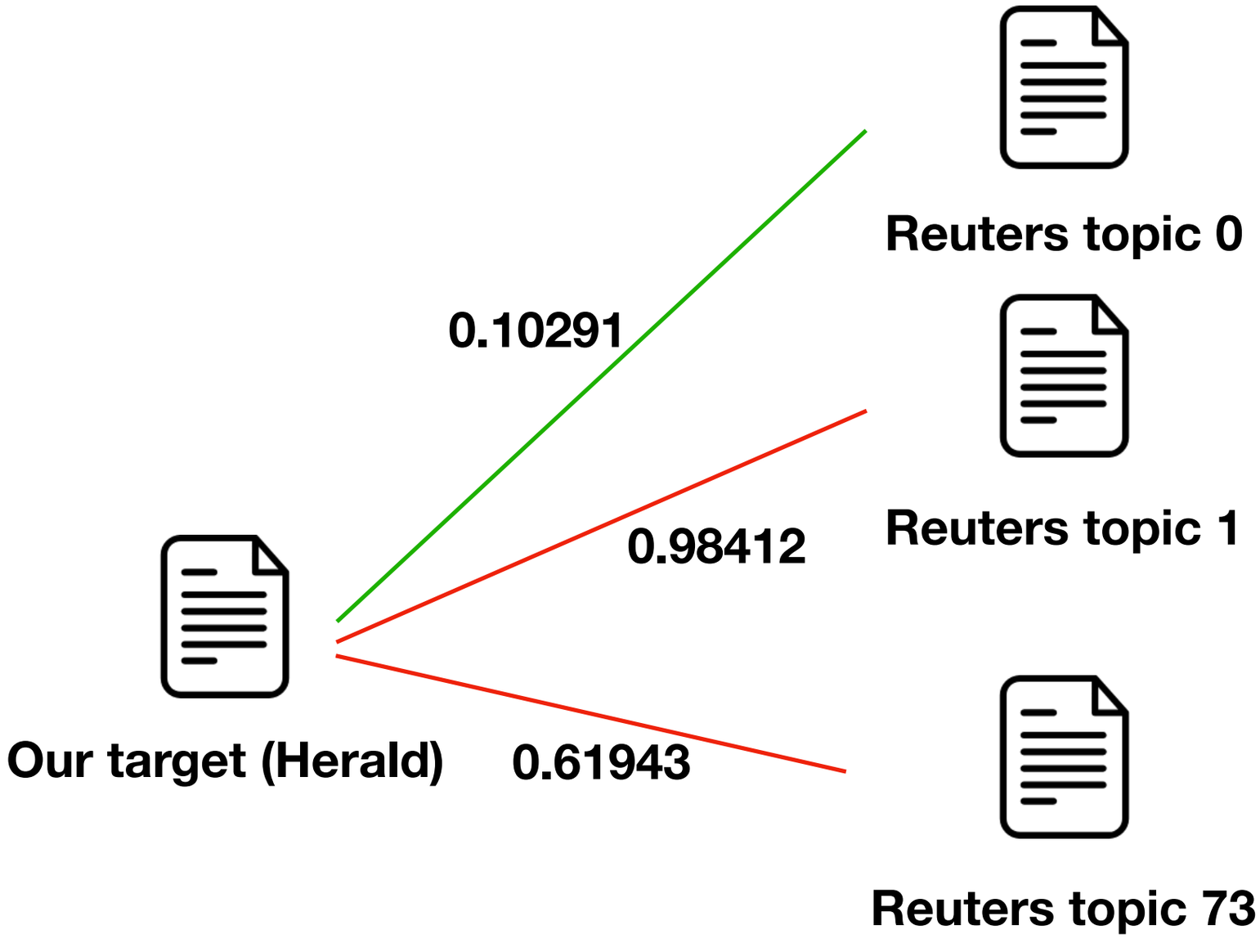
Reuters-21578 benchmark

benchmark dataset for document classification based on NEWS stories

Consists of several kinds of topics, and about 9,000 documents which is labeled by topics

Selecting Test Data

To ensure similarity between our target and test dataset, we compute cosine similarity for each topics in Reuters dataset, and choose 10 topics which has lowest distance.



Selecting Test Data

We select 10 topics in Reuters dataset, and try to recover classified topics using LDA result.

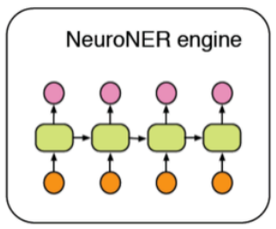
Tokenization

- ~~lemmatization~~
- nltk.tokenize

NER

- pos_tag
- multi-word information

neuroNER



- multi-word info

LDA with NER

- NER token * 10

Tune LDA

- Remove stopwords
- Remove (,), ., \xec…
- num_topics: 10
- num_topics: 10

Topic information

Test Result

In result, we success to recover 7 topics out of 10 topics.

Labeled topic	Reconstructed
Yen	O
Lumber	O
Veg-oil	O
Strategic-metal	O
Carcass	X
Meal-feed	X
Gold	O
Ship	X
Cocoa	O
Oilseed	O

Result

Trends Result

2015	2016	2017
Roh Moo-hyun's Oppression by NIS	Political scandal about President Park	PyeongChang Olympics
Thaad placement	Zika Virus	Choi sun-sil gate and presidential impeachment
MERS	Seoul and Gyeonggi Province issue	North Korea relationship
Saenuri Party and Park Geun-hye	Relationship with Ban Kimoon and UN	Sewol ho
Korea compared with OECD	Thaad and missile	Donald Trump
Statement of ICC	National Police Agency,	The next presidential candidate
Japan and sex slaves	South Korea-ASEAN relationship	Korea-China relationship and DAPA
Mount Geumgangsan-relationship with North Korea	Saenuri Party and general elections	Elected candidate Moon Jae-in
Indonesian Air Force	US elections	European Union
South China Issue	KATUSA	ASEAN and USFK

Trends Result

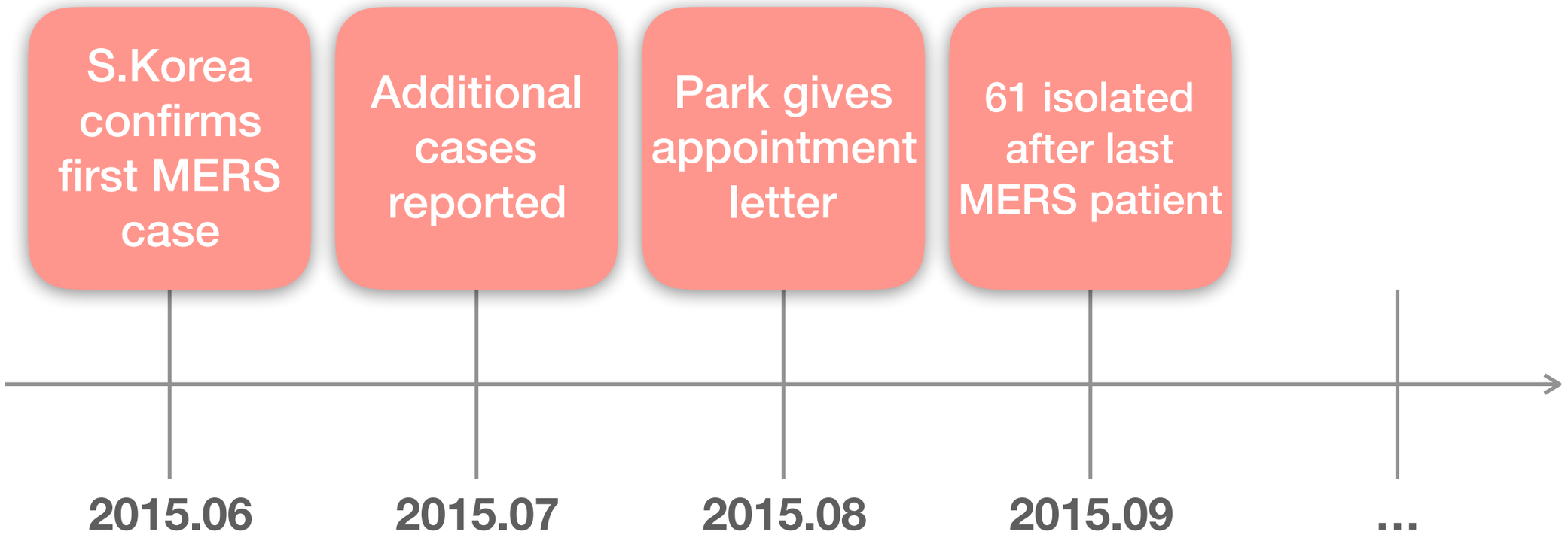
Choose 2 Issues for Event tracking

2015	2016	2017
Roh Moo-hyun's Oppression by NIS	Political scandal about President Park	PyeongChang Olympics
Thaad placement	Zika Virus	Choi sun-sil gate and presidential impeachment
MERS	Seoul and Gyeonggi Province issue	North Korea relationship
Saenuri Party and Park Geun-hye	Relationship with Ban Kimoon and UN	Sewol ho
Korea compared with OECD	Thaad and missile	Donald Trump
Statement of ICC	National Police Agency,	The next presidential candidate
Japan and sex slaves	South Korea-ASEAN relationship	Korea-China relationship and DAPA
Mount Geumgangsan-relationship with North Korea	Saenuri Party and general elections	Elected candidate Moon Jae-in
Indonesian Air Force	US elections	European Union
South China Issue	KATUSA	ASEAN and USFK

On-issue Tracking: MERS

For specific series of month (because we have total 12 summaries, one for each month):

Event: S.Korea confirms first MERS case -> Additional cases reported -> Park gives appointment letter to new health minister -> 61 isolated....



On-issue Tracking: MERS

Details

“Park gives appointment letter to new health minister"
When: Monday
Where: South Korea
What: gives appointment letter
Who: Park
How: to new health minister

"61 isolated after last MERS patient rediagnosed"
When: Sunday
Where: South Korea
What: rediagnosed
Who: last MERS patient
How: after last MERS patient rediagnosed

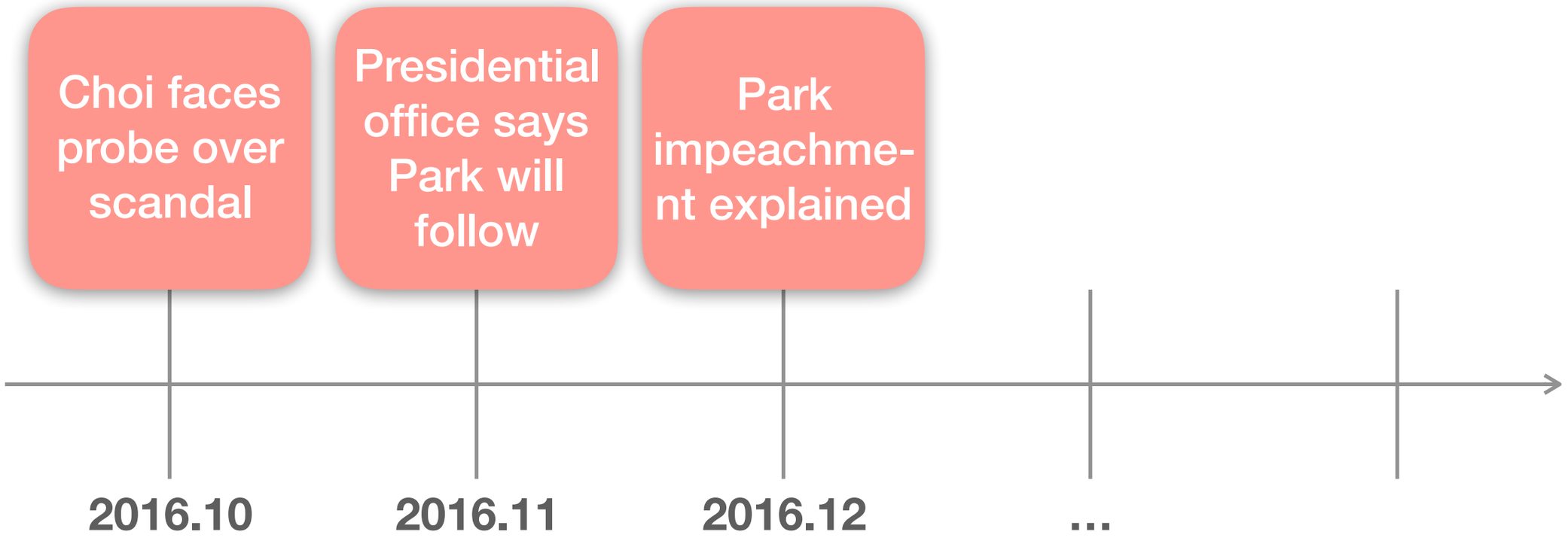
Off-issue Tracking: MERS

who	Korean Air	President Park	S. Korea	Minimum wage	the chief of the Korea Confederation of Trade Unions
what	heiress gets 1 year	welcomes labor reform deal	reports no new MERS cases	declared despite resistance	walked out of the temple
when	Thursday	Tuesday	the day before	Wednesday	11:20 a.m.
where	Seoul	unknown	S. Korea	South Korea	Seoul
why	Korean Air	President Park Geun-hye on Tuesday	S. Korea	Minimum wage	police
how	former vice president of Korean Air , to one	“ tough ” decision to compromise on reform measures that	no new MERS cases for 14th day .	The South Korean government Wednesday announced next year	Unions voluntarily walked out of the temple in central Seoul

On-issue Tracking:
Political scandal

For specific series of month (because we have total 12 summaries, one for each month):

Event: Choi faces probe over influence-peddling scandal -> Presidential office says Park will follow whatever decision parliament makes on her fate -> Park Geun-hye impeachment explained...



On-issue Tracking:
Political scandal

Details

"Choi faces probe over influence-peddling scandal"
When: Sunday
Where: Seoul
What: faces probe over influence-peddling scandal
Who: Choi
How: Presidential office dismisses claim over Park 's retirement residence

"Presidential office says Park will follow ... on her fate"
When: Tuesday
Where: Seoul
What: will follow whatever decision parliament makes on her fate
Who: President Park Geun-hye
How: Presidential secretary offer to resign

Off-issue Tracking: Political scandal

who	she	Independent counsel Park Young-soo	President Park Geun-hye	A formal arrest warrant	Office	a special inspector
what	repeatedly rejected to appear at a parliamentary	faces the daunting task	is impeached	has been issued Thursday	give the person	can face a jail term
when	Monday	Dec. 11	Saturday	Thursday	7:30 a.m.	Monday
where	Seoul	unknown	Gwanghwamun Square	Grand Korea	Seoul	Seoul
why	Lawmakers	They	because of the people around me . ”	Choi	Choi Soon-sil , the mysterious woman accused of interfering in state	a special inspector
how	she repeatedly rejected to appear at a parliamentary hearing	Can independent counsel untangle Choi scandal ?	only halfway through ‘ .	s longtime confidante , accused of collaborating with	Choi Soon-sil returns ; Blue House ‘ raid ’ by	Special inspector Lee Seok-su , who had been tasked with

Conclusion

Conclusion

- **We success to summarize 270K news articles and give trend, on/off issue on that trend.**

- **It is shown to be effective to minimize the manual efforts**

- **Can be used in not just news, but marketing, research, or the other fields as well.**

Summary

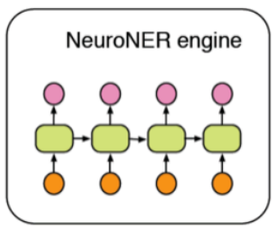
Tokenization

- ~~lemmatization~~
- nltk.tokenize

NER

- pos_tag
- multi-word information

neuroNER



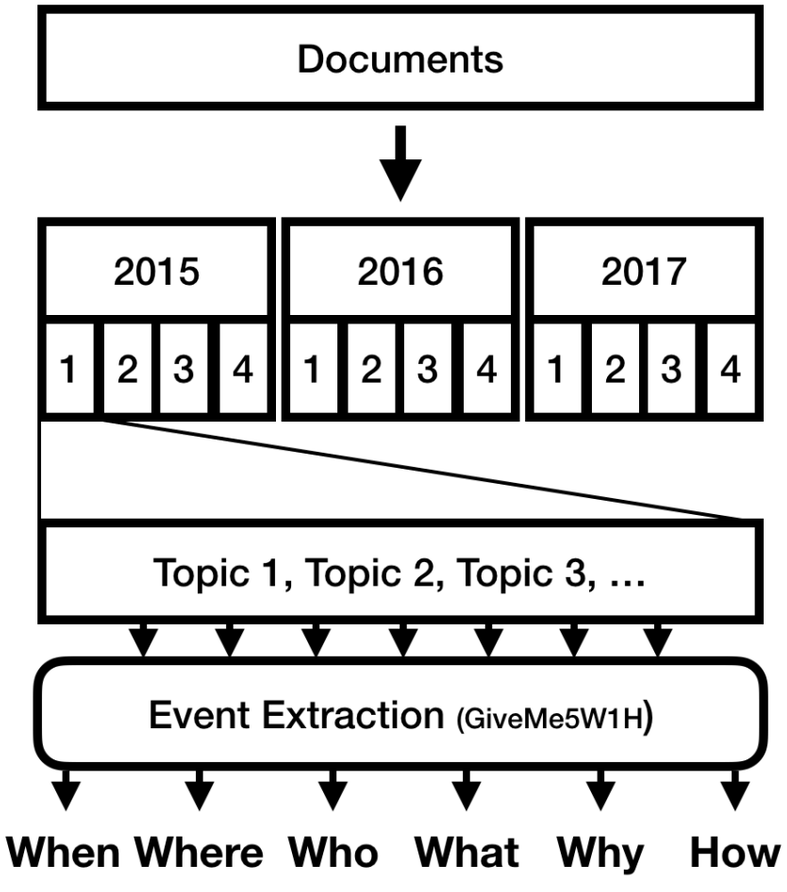
- multi-word info

LDA with NER

- NER token * 10

Tune LDA

- Remove stopwords
- Remove (,), ., \xec…
- num_topics: 10
- num_topics: 10



BoW Feature Extraction

Document Clustering

Representative Document

Event Extraction

Trend Analysis Evaluation by Reuters Dataset

REUTERS®

Thank You!