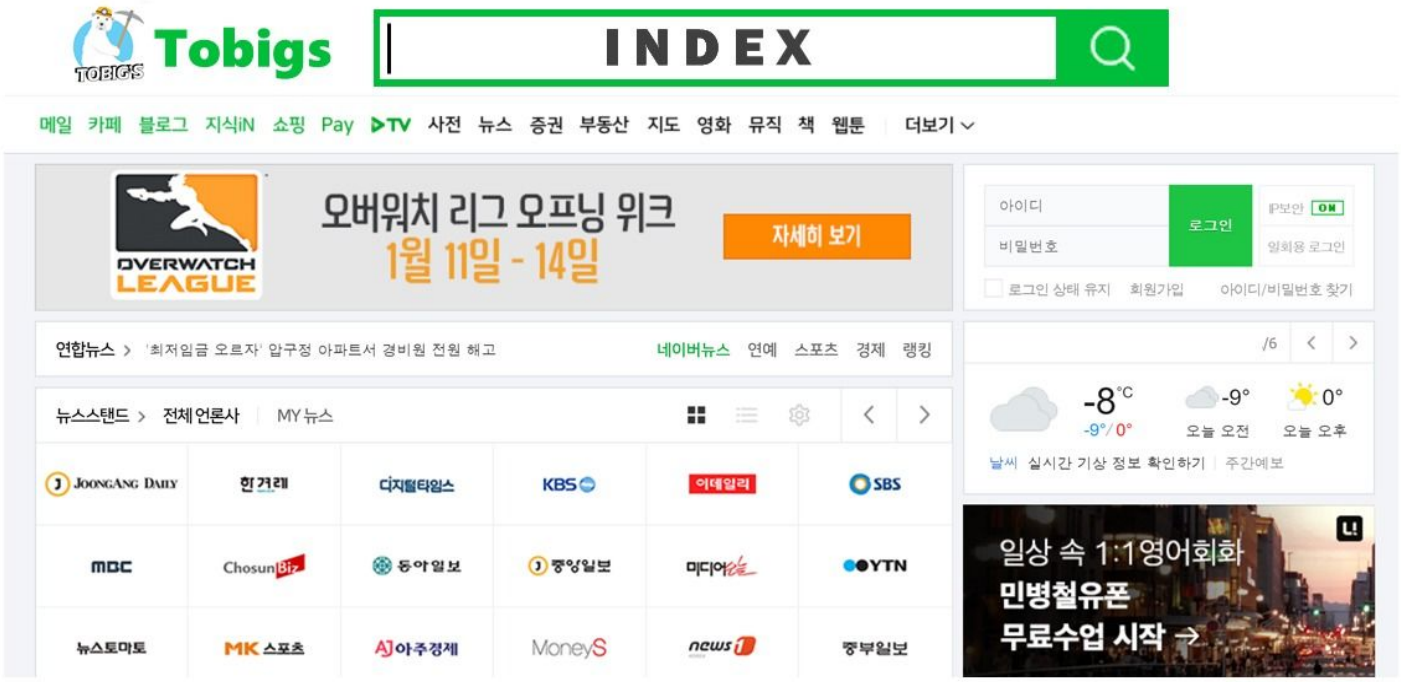

Topic network
NAVER RANK
Tobigs 5th conference
RE;PLACE
김수현 김은서 김지수 이제형 장정주


Tobigs
INDEX
메일 카페 블로그 지식iN 쇼핑 Pay TV 사전 뉴스 증권 부동산 지도 영화 뮤직 책 웹툰 더보기
오버워치 리그 오프닝 위크
1월 11일 - 14일
자세히 보기
OVERWATCH LEAGUE
아이디
비밀번호
로그인
IP보안 ON
일회용 로그인
로그인 상태 유지 회원가입 아이디/비밀번호 찾기
연합뉴스 '최저임금 오르자' 압구정 아파트서 경비원 전원 해고
네이버뉴스 연예 스포츠 경제 랭킹
/6
-8°C
-9°/0°
-9°
오늘 오전
0°
오늘 오후
날씨 실시간 기상 정보 확인하기 주간예보
뉴스스탠드 전체언론사 MY뉴스
JOONGANG DAILY
한겨레
디지털타임스
KBS
이데일리
SBS
MBC
ChosunBiz
동아일보
중앙일보
미디어오늘
YTN
뉴스토마토
MK 스포츠
AJ아주경제
MoneyS
news1
중부일보
일상 속 1:1영어회화
민병철유폰
무료수업 시작 →

01 주제 소개

02 데이터 수집
및 전처리

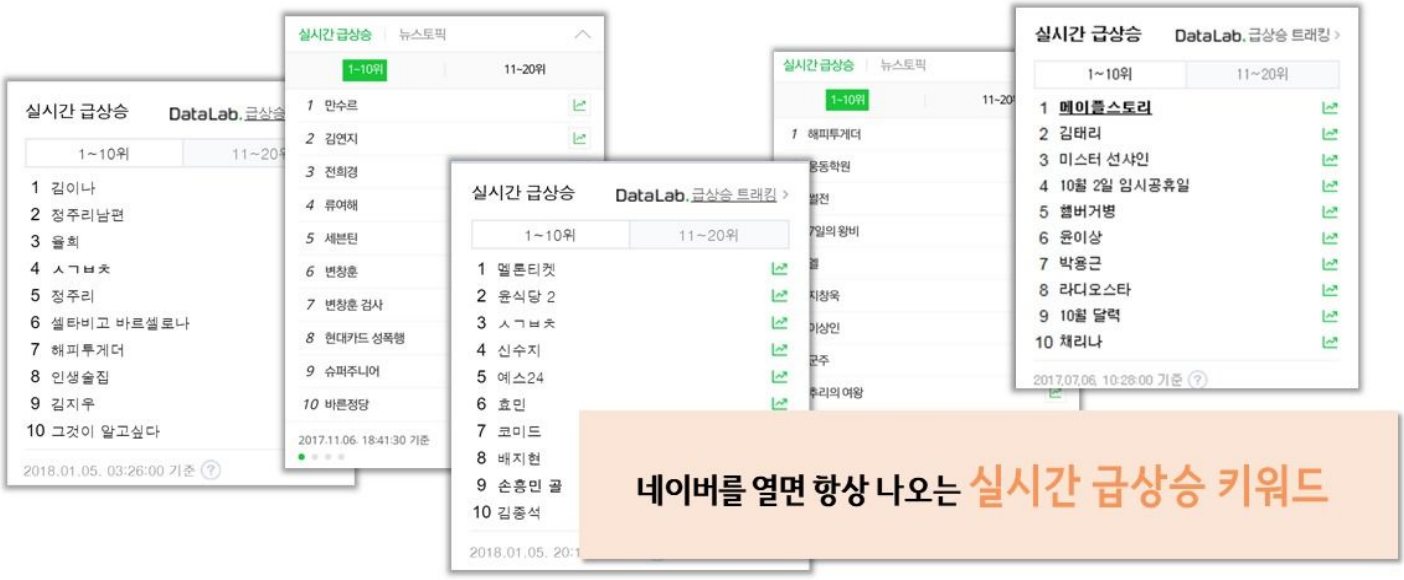
03 모델링

04 결과

05 한계점

01

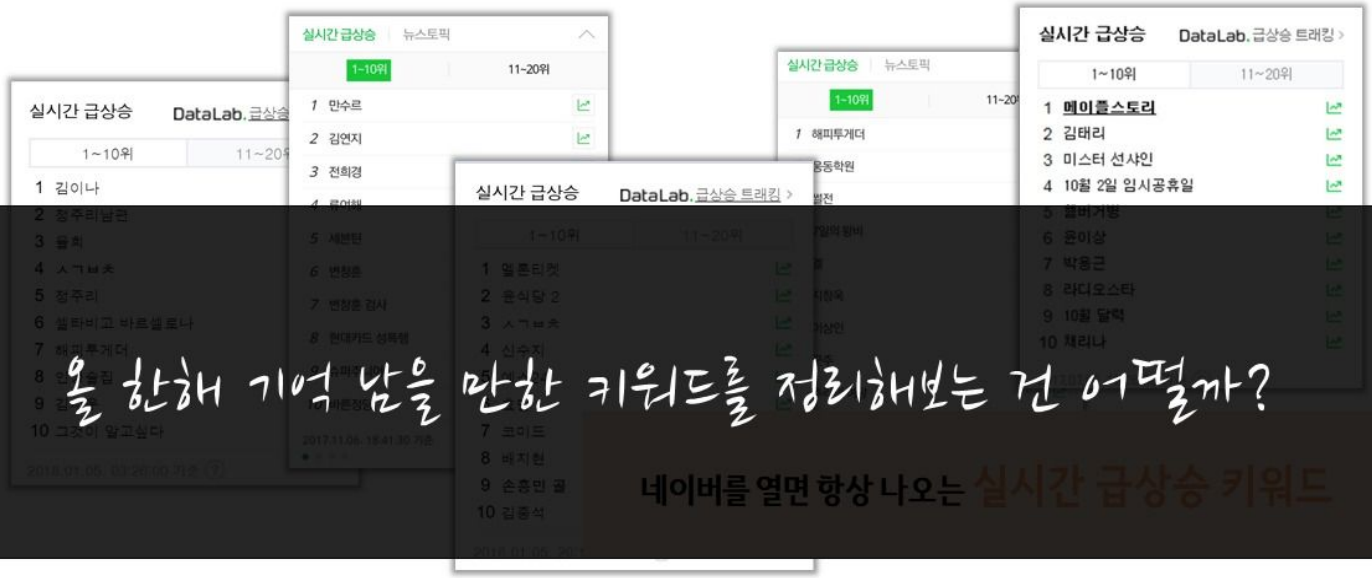
주제 선정 배경



네이버를 열면 항상 나오는 실시간 급상승 키워드

01

주제 선정 배경



올 한해 기억 남을 만한 키워드를 정리해보는 건 어떨까?

네이버를 열면 항상 나오는 실시간 급상승 키워드

“올 한해 Hot Topic 의 시간 흐름 에 따른 정리 및 요약”

01 주제 소개

02 데이터 수집 및 전처리

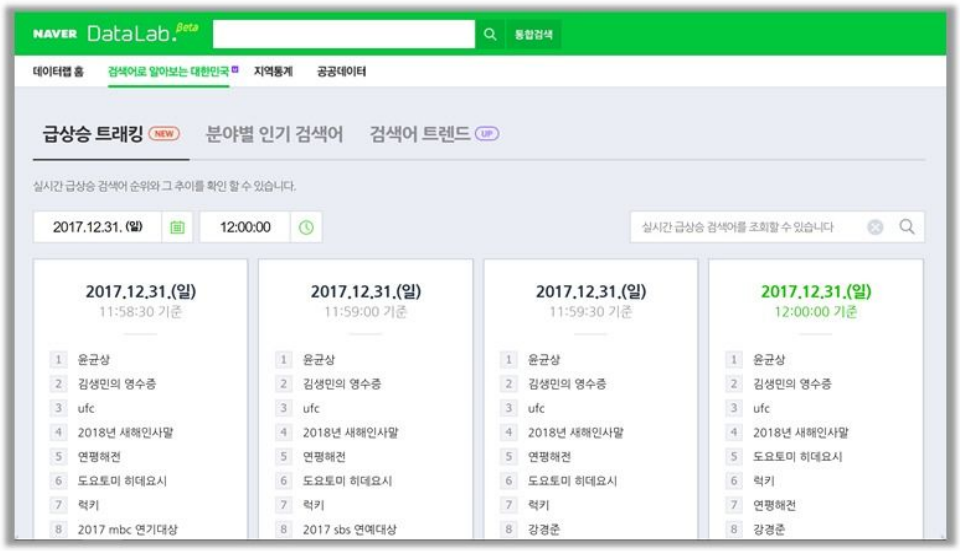
03 모델링

04 결과

05 한계점

02

데이터 수집 및 전처리



네이버에서 제공하는 데이터

2017.03.29 - 2017.12.31

기간 데이터 크롤링

시카고 타자기	111179
맨투맨	102755
프로듀스 101 시즌2	101412
역적	101370
장문복	99710
추리의 여왕	99380
로또	99293
손흥민	98713
썰전	95489
북한	90883
무한도전	90297
그것이 알고싶다	88935
ufc	85479
귓속말	83393

그러나 ,

특정 요일마다 반복되는 TV프로그램이 대부분

주기성 제거 해주고자 결정

순위

	1	2	3		18	19	20
2017.11.06.(월)18:41:00 기준	만수르	김연지	전희경		한샘 카톡	공범	한샘
2017.11.06.(월)18:41:30 기준	만수르	김연지	전희경		한샘 카톡	공범	수능
2017.11.06.(월)18:42:00 기준	만수르	김연지	전희경		한샘 카톡	공범	한샘
2017.11.06.(월)18:42:30 기준	만수르	김연지	전희경	.....	손흥민	한샘 카톡	한샘
2017.11.06.(월)18:43:00 기준	만수르	김연지	전희경		한샘 카톡	손흥민	한샘
2017.11.06.(월)18:43:30 기준	만수르	김연지	전희경		공범	한샘 카톡	한샘
2017.11.06.(월)18:44:00 기준	만수르	김연지	전희경		수능	6시내고향	한샘 카톡

시간 (30초 간격)

검색어 랭킹화

차예련	24166
공각 기동대	21282
원라인	19740
최명길	19081
주상욱	18001
구룡마을	16823
스윙스	16382
박명수	14719
이상우	13617
김소연	13546

⋮

도가니	2
군복무기간 계산기	2
빠른길찾기	2
우체국	2
블랙박스 소비자원	1
티플	1
공무원연금관리공단	1

한번씩 등장했던 검색어 UNIQUE

↓

검색어마다 시간대별로 자신의 순위 기록
(20위에 들지 못한 시간대는 -1 기록)

↓

1위 -> 20점 ,,, 20위 -> 1점
을 부여하여 각 검색어 별 랭킹화

주기성 제거

키워드	키워드 등장횟수
그것이 알고싶다	35
무한도전	34
아는형님	34
나혼자산다	29
언니는 살아있다	25
오늘의운세	14
소사이어티 게임 2	12
기상청	10
품위있는 그녀	10

[토요일]

.....

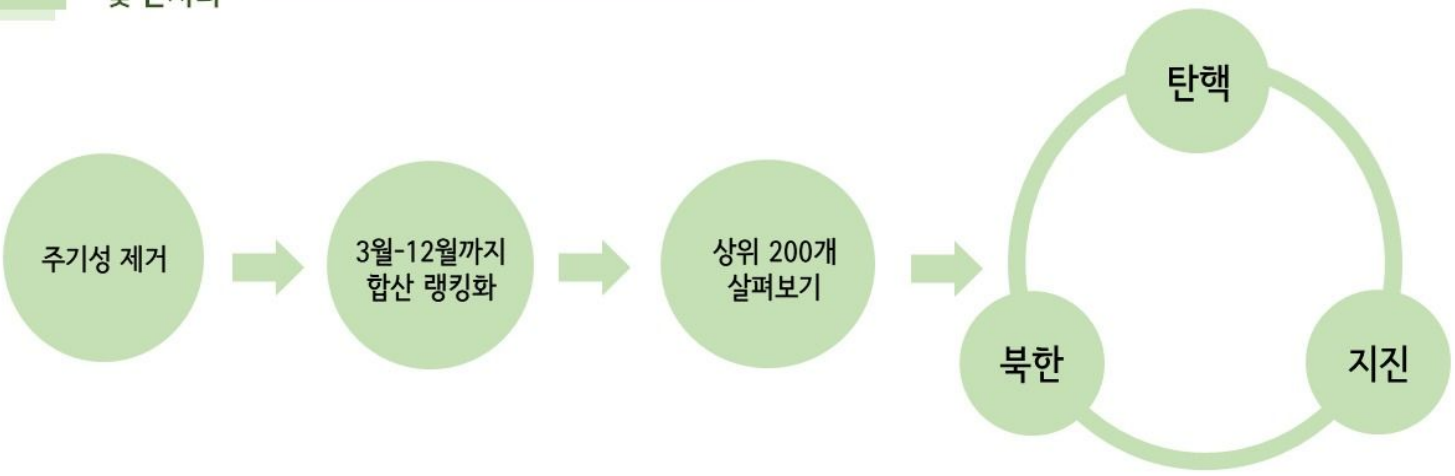
키워드	키워드 등장횟수
외모지상주의	35
썰전	31
뮤직뱅크	30
해피투게더	30
나혼자산다	27
인생술집	25
오늘의운세	18
어서와 한국은 처음이지?	15
아이돌학교	10

[금요일]

- 각 요일별 10회 이상 반복 단어를 **주기성**으로 간주하여 제거

- 단,
  토요일에 무한도전이 **34번**나오고
  금요일에 무한도전이 **5번** 나왔다면
  ~~토요일에 해당하는 무한도전만 제거~~

주기성 제거 → 3월-12월까지 합산 랭킹화 → 상위 200개 살펴보기



탄핵, 북한, 지진 등의 키워드가 자주 등장
데이터 양을 줄이기 위해 위 3개 주제에 대해서 분석

데이터 수집 및 전처리

SUBKEYWORD 추출

탄핵
박근혜
박지만
박근혜 영장심사
영장심사
자유한국당
강부영판사
박근혜 구속
안철수
조윤선
우병우
기소
안철수 동생
대선 tv토론
대선토론
대선 토론
심상정
유승민
북한
고영태
홍준표

지진
경주 지진
포항지진
일본지진
황사
구례 지진
북한 지진
일본 지진
쓰촨성
중국 지진
기상청
멕시코 지진
멕시코
부산 폭우
미세먼지
태풍
수능 연기
포항
포항지진피해
여진
포항여진

북한
문재인
문재인 주적
유시민
주적
박지원
랜섬웨어
불상발사체
북한 미사일 발사
북한 미사일
지대함 미사일
오토 웜비어
북한 중대발표
김정은
지진
북한 지진
미사일
북한 핵실험
귀순 북한 병사
jsa
귀순

- 1차 탐색

뉴스 기사에서 같이 들어가 있는 지의 여부에 따른 subkeyword 탐색

- 2차 탐색

실시간 검색어에 등장한 횟수가 **5회 이상인** 키워드를 추출

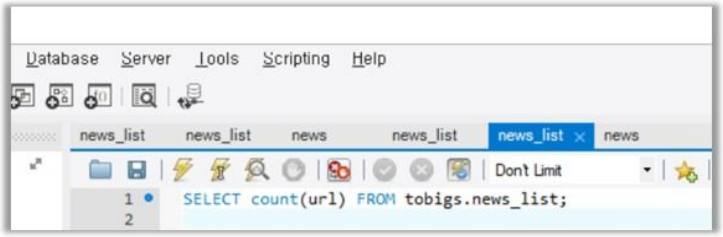
01 주제 소개

02 데이터 수집
및 전처리

03 모델링

04 결과

05 한계점



- 각 주제마다
  실시간 급상승 검색어에 등장했던 날짜를
  기준으로 뉴스 기사 크롤링

TOT (Topic modeling)

LexRank/TextRank

Word2Vec

TOT

LexRank/TextRank

- 여러 기사 중 **대표기사** 추출
- 여러 문장 중 **대표문장** 추출

Word2Vec

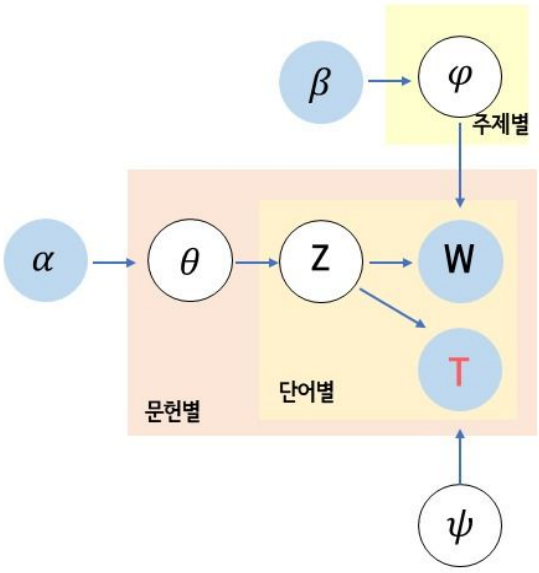
TOT = LDA + Time

왜?
시간을 연도별로 이산화 한다면,
범주들이 서로 독립이어서
시간 간격 고려 불가

2017년 1월 1일

2017년 12월 31일

2018년 1월 1일



TOT (Topic modeling)

- 핵심 키워드에 대한 **흐름별 Topic** 확인

LexRank/TextRank

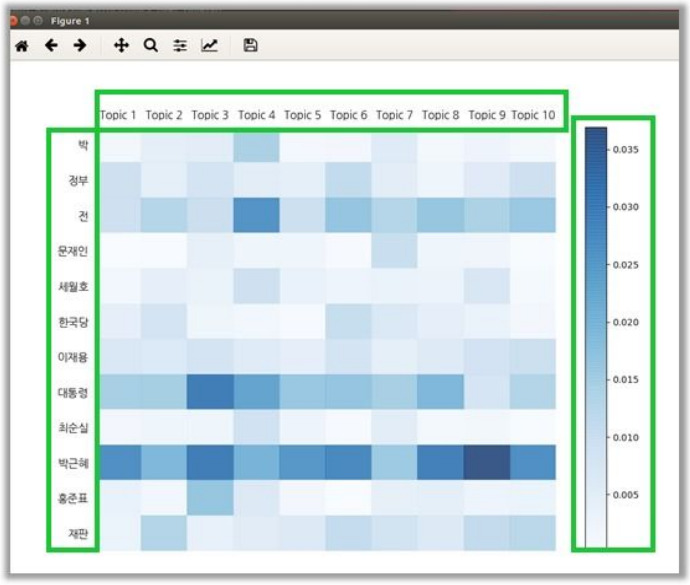
Word2Vec

TOT

LexRank/TextRank

Word2Vec

- Keyword간의 **거리 계산**



- Column - Topic 1,2 … 10은 기간을 의미
  -단, 기간은 순서대로 나열 되어있지 않다.
- Row - 기사 속 비중 있는 subkeyword
- 색 - 색 변화 별 value크기

TextRank 는 **문장** summarization

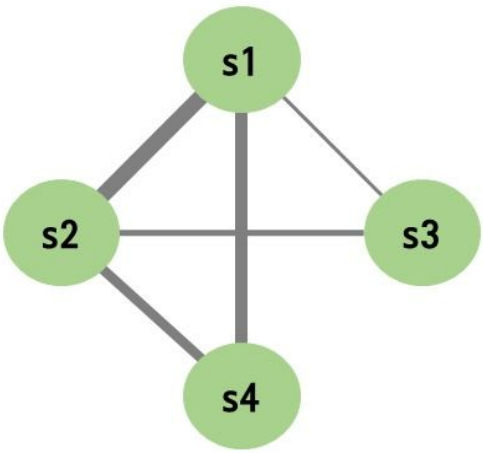
- 문장간 유사도를 계산하여 유사한 문장의 centroid에 해당하는 문장만 선택

LexRank 는 **다문서** summarization

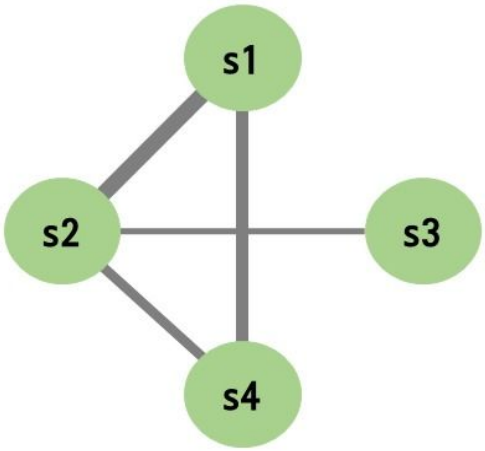
- TextRank와 동일한 방법 + 노드간 클러스터링이 들어간다.

문장	1	2	3	4
1	1	0.5	0.1	0.4
2	0.5	1	0.2	0.3
3	0.1	0.2	1	0
4	0.4	0.3	0	1

1. tf-idf 기반으로
수정된 cosine 유사도 계산



2. 유사도간 관계를 그려주고
-엣지weight 0.2이상인 선만 남긴다.

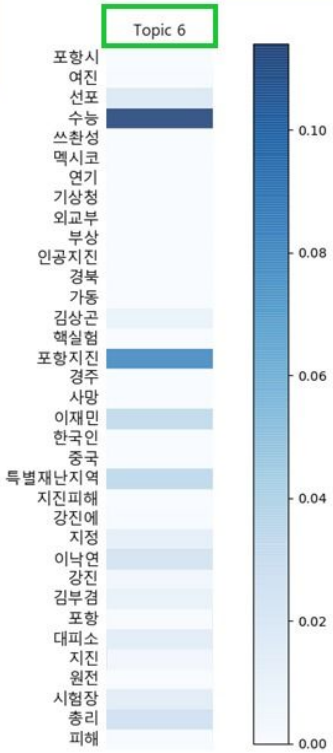


3. 각 노드별 value를 인접한 노드에 나눠주는 계산 반복
-반복하면 수렴값 발생 , 큰 수렴값을 갖는 것이 중요문장

03

모델링

기간추출



키워드	value
수능	0.12
포항지진	0.07
특별재난지역	0.04

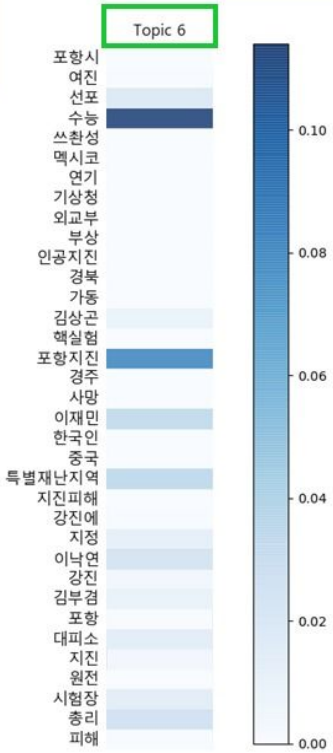
⋮

1 Value의 크기별로 키워드를 나열

- 앞의 sentence 1개 -> 기사 1개 로 생각
- LexRank의 주요문장을 뽑아내듯
  주요 기사를 뽑아낼 것이다.

03

모델링 기간추출

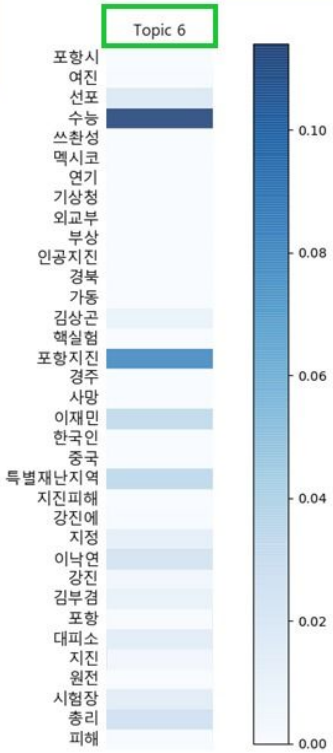


[지진] TOT결과 중 Topic6을 이용해
핵심 키워드와 대표기사를 추출 하려한다

03

모델링

기간추출



키워드	value
수능	0.12
포항지진	0.07
특별재난지역	0.04

⋮

2017.11.15 ~ 2017.11.16

1 Value의 크기별로 키워드를 나열

2 기간별 기사의 제목/본문 에서
키워드를 counting하여 유사도가 높은 기간추출

LexRank결과 총 10개의 기사

- '지진으로 연기된 수능'
- 지진피해모습
- 지진 피해 차량
- 지진 피해 다세대주택
- 지진으로 대피한 이재민들
- 지진 피해복구 대민 지원하는 해병대
- 지진으로 붕괴위험에 처한 아파트
- 지진 피해로 집 나서는 주민
- 지진피해현장 살피는 우원식-유승민
- 지진 대피소에 라면 지원하는 관계자들

03

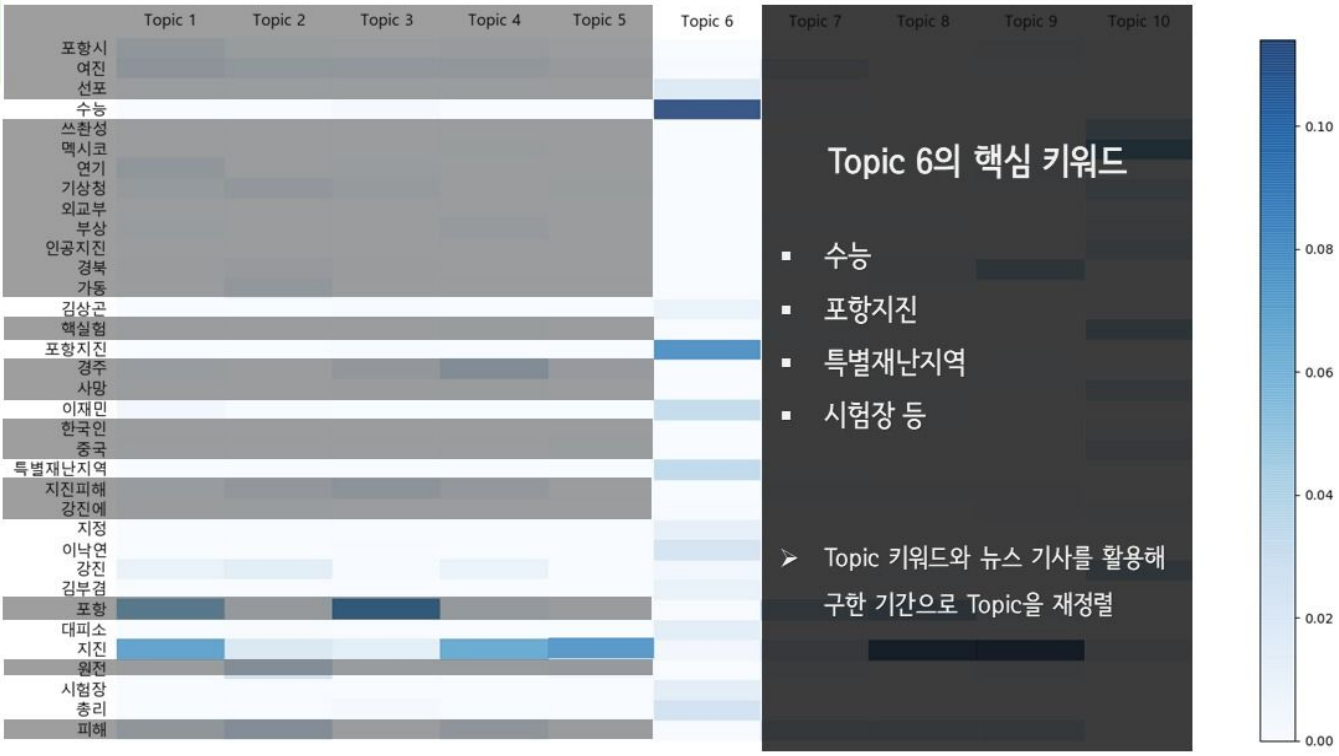
모델링

TOT + LexRank

03

모델링

TOT + LexRank



Topic 6의 핵심 키워드

- 수능
- 포항지진
- 특별재난지역
- 시험장 등

➢ Topic 키워드와 뉴스 기사를 활용해 구한 기간으로 Topic을 재정렬

03

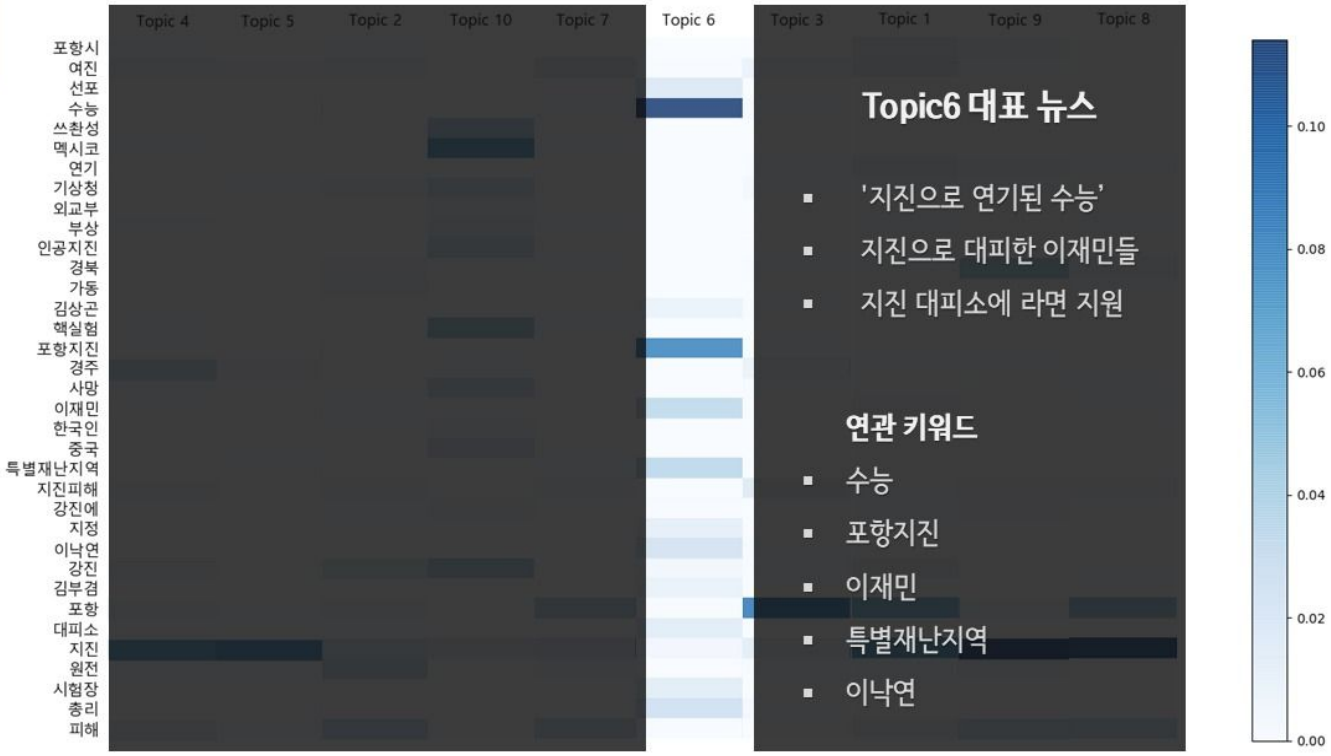
모델링

TOT + LexRank

03

모델링

TOT + LexRank



Topic6 대표 뉴스

- '지진으로 연기된 수능'
- 지진으로 대피한 이재민들
- 지진 대피소에 라면 지원

연관 키워드

- 수능
- 포항지진
- 이재민
- 특별재난지역
- 이낙연

Word2Vec 은

문장의 순서에 따라서 학습을 하여 단어들을 벡터화한다
이를 통해 단어들의 유사도를 계산할 수 있다

- DB에 등록되어 있는 기사들 학습
- 각 주제의 keyword들이 잘 인식되게 하려고 cKONLPY 사용
- 품사는 전체 다 사용

```
In [69]: textrank = TextRank(rows[0]['url'])
         for row in textrank.summarize(3):
             print(row)
```

■ 방송 : YTN 뉴스와이드■ 진행 : 유석현 앵커■ 출연 : 채문석 YTN 선임기자, 서양호 두문정치전략연구소장- 최순실 딸 정유라 구속영장 기각- 법원 "정유라, 구속 필요성 없다"- 정유라 "심려 끼쳐 정말 죄송"▶앵커> 정유라 씨 구속영장, 기각이 되었습니다
심지어 검찰 수사과정에서 독일에서 덴마크로 도피한 행각에 대해서도 도피가 아니라 엄마가 시켜서 갔다라고 했을 정도로 일관되게 엄마 말을 잘 듣는 학생의 모습을 취함으로써 검찰의 구속영장을 잘 피해나갔다는 생각이 드는데요
그런데 이번에 기각이 되다 보니까 다른 보강수사를 해야겠지만 특히 최근에도 문재인 대통령께서 기본적으로 특검 수사 기한이 연장되지 못해서 검찰에서 이걸 좀 국정농단 사건을 좀더 자세히 들여다 볼 필요가 있지 않느냐, 이런 언급을 했고 그 이후에 특검 수사팀의 수수팀장인 윤석열 검사가 서울중앙지검장으로 갔습니다

TEXTRANK로 키워드에 대한 기사요약 + 유사도 계산 > 이후 네트워크 학습

```
model.wv.distance('박근혜/Noun','구속/Noun')
```

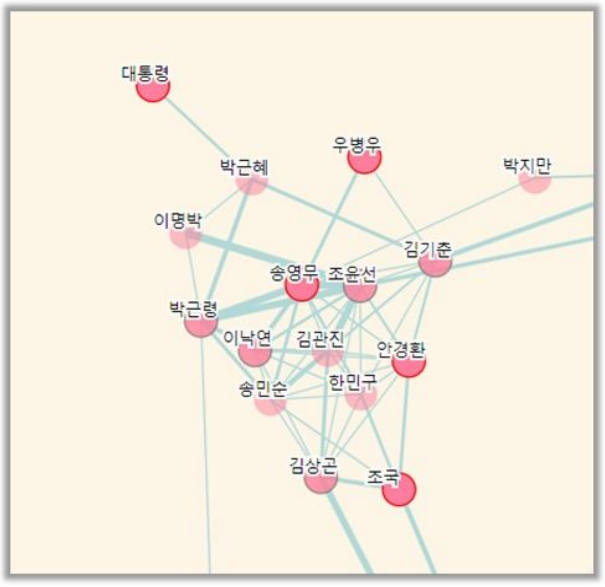
```
0.69645337245977323
```

X

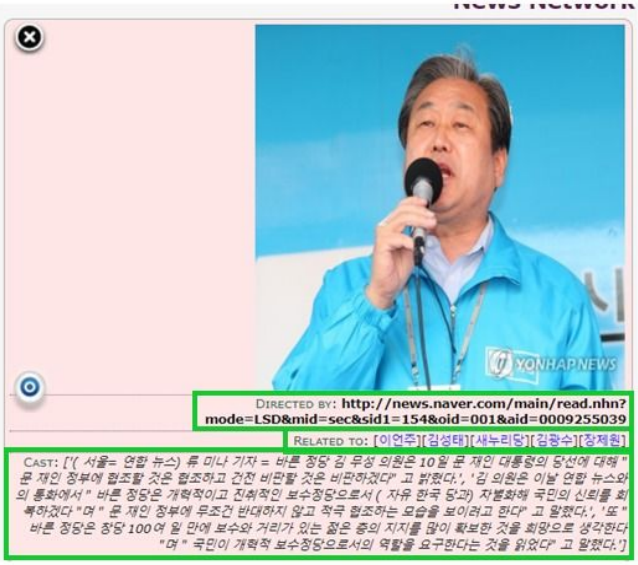
```
{'강부영판사': 0,
 '고영태': 534,
 '구속': 15761,
 '구속기소': 2557,
 '구형': 4862,
 '권순호': 435,
 '기소': 9575,
 '김관진': 1492,
 '김광수': 137,
 '김기춘': 2482,
 '김무성': 1670,
 '김상곤': 724,
 '김성태': 1069,
```

단어간 유사도

$= \frac{1}{\text{키워드 간 거리}}$ x 기사에서 동시에 나온 횟수 정규화



- Node는 subkeyword
- Edge는 단어간 유사도



- LexRank로 추출한 해당 키워드 대표기사
- 정의된 거리가 가까운 키워드
- TextRank로 요약된 기사

01 주제 소개

02 데이터 수집
및 전처리

03 모델링

04 결과

05 한계점

Topic 4
Topic 5
Topic 2
Topic10
Topic 7
Topic 6
Topic 3
Topic 1
Topic 9
Topic 8

8/9~9/3

中 쓰촨성 강진...13명 사망·170여 명 부상

이런 가운데 AFP통신은 중국 재난대응위원회를 인용해, 최대 백 명이 숨졌을 것으로 보이며 주택 13만 채가 피해를 봤다고 전했습니다

[종합]기상청 "北 인공지진, 역대 핵실험 최대…국내서도 감지"

이미선 지진화산센터장은 "지금까지 풍계리 인근 지역에서 총 6차 인공지진이 있었다"며 "6차에 대한 에너지 스펙트럼, 단층면 음파자료 분석 결과 인공지진이 확실하고 5차 대비 에너지 규모는 5~6배 정도 크다"라고 설명했다

규모 8.1 강진에 60여 차례 여진… 멕시코 국토 절반 '흔들'

7일 오후(현지시간) 멕시코 남부 치아파스주 피히히아판에서 남서쪽으로 87㎞ 떨어진 태평양 해상에서 규모 8.1의 강진이 발생했다고 미국 지질조사국(USGS)이 밝혔다

지진

Topic 4
Topic 5
Topic 2
Topic10
Topic 7
Topic 6 11/15~11/16
Topic 3
Topic 1
Topic 9
Topic 8

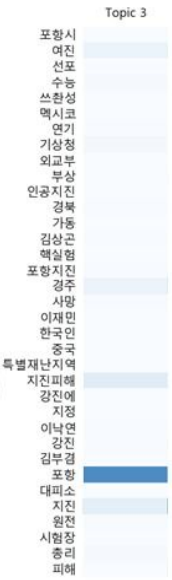
'지진으로 연기된 수능'

교육부가 포항에서 발생한 지진으로 인해 대학수학능력시험을 일주일 연기한 16일 오전 대전시교육청 제27지구 제28시험장인 동방고등학교에서 학교 관계자가 수험생 유의사항문을 정리하고 있다.

지진으로 대피한 이재민들

경북 포항시 북구 북쪽 9㎞ 지역에서 규모 5.4의 지진이 발생한 가운데 16일 흥해실내체육관 임시대피소에 이재민들이 대피해 있다.

Topic 4
Topic 5
Topic 2
Topic10
Topic 7
Topic 6
Topic 3 11/19~11/20
Topic 1
Topic 9
Topic 8



포항서 35시간여 만에 규모 2.0 여진 발생…여진 총 66회

지진의 발생 깊이는 8㎞다

탄핵, 대선

- Topic 3 4/17~4/19
- Topic 1
- Topic 8
- Topic 6
- Topic10
- Topic 7
- Topic 9
- Topic 2
- Topic 4
- Topic 5

[단독] 박근혜, 삼성 이재용에게 "손석희 갈아치우라" 외압

홍석현 대통령으로부터 두 번 외압 받았다 밝혀… 언론사주가 박근혜로부터 외압 받았다고 공개한 첫 사례

법원 넘어간 박근혜 '592억 뇌물'…무죄 아니면 실형뿐

아직 재판이 시작하지 않았지만, 박 전 대통령은 주요 혐의가 유죄로 인정될 경우 '특정범죄가중처벌에 관한 법률'에 따라 무기징역 또는 10년 이상의 형을 선고받게 된다

Topic 3
Topic 1
Topic 8 5/5~5/6
Topic 6
Topic10
Topic 7
Topic 9
Topic 2
Topic 4
Topic 5

文, 강세지역 수도권서 마지막 주말유세…'이제는 굳히기'

특히 문 후보는 이날 자신이 공약했던 서울 홍대 거리에서 '프리허그' 행사를 진행, 지지층 결집에도 주력한다

'천만명 돌파' 사전투표 열기…"탄핵 거치며 참여의지 상승"

김만흠 한국정치아카데미 원장은 5일 뉴시스와 통화에서 "투표 참여 의지가 강하게 반영됐다"며 "국민들이 정치참여를 한 편으론 축제 참여라고 생각하고, 특히 탄핵 정부를 거치면서 역사적 현장에서 역할을 해보겠다는 시민들이 많아졌다"고 분석했다

Topic 3
Topic 1
Topic 8
Topic 6 5/20~5/20
Topic10
Topic 7
Topic 9
Topic 2
Topic 4
Topic 5

문재인 대통령, 취임 11일…'소통-협치' 평가

이른바 3철, 양정철 전 홍보기획비서관, 그리고 전해철 최고위원, 이호철 전 민정수석인데요

초유의 인사 태풍… 검찰, 칼날 위에 서다

'돈 봉투 만찬' 파문으로 감찰 대상에 오른 이영렬(사법연수원 18기) 서울중앙지검장 후임으로 다섯 기수나 후배인 윤석열(23기) 대전고검 검사가 파격적으로 발탁된 것이다

Topic 7 4/29~5/21

Topic 2
Topic 8
Topic 4
Topic 3
Topic 9
Topic 5
Topic 6
Topic 1
Topic10

북한, 평남 북창일대서 탄도미사일 1발 발사…공중폭발로 실패

이 미사일은 발사 직후 수초만에 공중에서 폭발한 것으로 알려졌다.

北, 文정부 출범 나흘만에 탄도미사일 1발 전격 발사(종합)

앞서 미국의소리(VOA) 방송은 지난달 27일 구성에 있는 방현비행장 북쪽에서
미사일 발사용 이동식발사대(TEL)가 인공위성 사진에 포착됐다고 보도한 바 있다

北, 문재인 정부 출범 12일만에 두번째 미사일 발사(종합2보)

북한이 21일 오후 평안남도 북창 일대서 '북극성 2형'으로 추정되는
탄도미사일 1발을 발사했으며 이 미사일은 500㎞를 비행했다"

Topic 7
Topic 2
Topic 8
Topic 4
Topic 3
Topic 9
Topic 5
Topic 6
Topic 1
Topic10

6/20~6/20

북한서 18개월 복역 후 혼수 상태로 풀려난 미 오토 웜비어 사망(종합)

미국 언론 보도에 따르면 오하이오 주 신시내티에 거주하는 웜비어의 가족들은
성명을 통해 병원에서 치료받던 웜비어가 이날 오후 3시20분 사망했다고 발표했다

Topic 3
Topic 1
Topic 8
Topic 6
Topic10
Topic 7
Topic 9
Topic 2
Topic 4
Topic 5

8/25~8/25

'박근혜 뇌물' 삼성 이재용, 1심 징역 5년…모든 혐의 유죄

이 부회장의 뇌물 공여 혐의가 유죄로 인정됨에 따라 뇌물수수자로 기소된 박 전 대통령도 이 부분에 대해 유죄 판단을 받을 가능성이 커졌다

북한

Topic 7
Topic 2
Topic 8
Topic 4
Topic 3
Topic 9
Topic 5
Topic 6
Topic 1
Topic10

9/3~9/5

文대통령 '대북기조' 바뀌었다…실효적 '제재·압박' 강화

문 대통령은 지난 4일 도널드 트럼프 미국 대통령과 전화통화를 통해
한국 미사일 탄두 중량 제한(500kg)을 해제하기로 합의했으며,
블라디미르 푸틴 대통령에게는 대북 원유 공급 중단과 북한 해외노동자 송출금지 등을
유엔 안보리에서 진지하게 검토해야 한다고 강조했다

Topic 7
Topic 2
Topic 8
Topic 4
Topic 3
Topic 9
Topic 5
Topic 6
Topic 1
Topic10 11/21~11/22

귀순 북한병사 자가호흡 · 의식명료…"주말쯤 일반병실로"

공동경비구역, JSA를 통해 귀순하다 총상을 입은 북한군 병사가 자가호흡을 하고 의식도 완전히 회복한 것으로 밝혀졌습니다'.

01 주제 소개

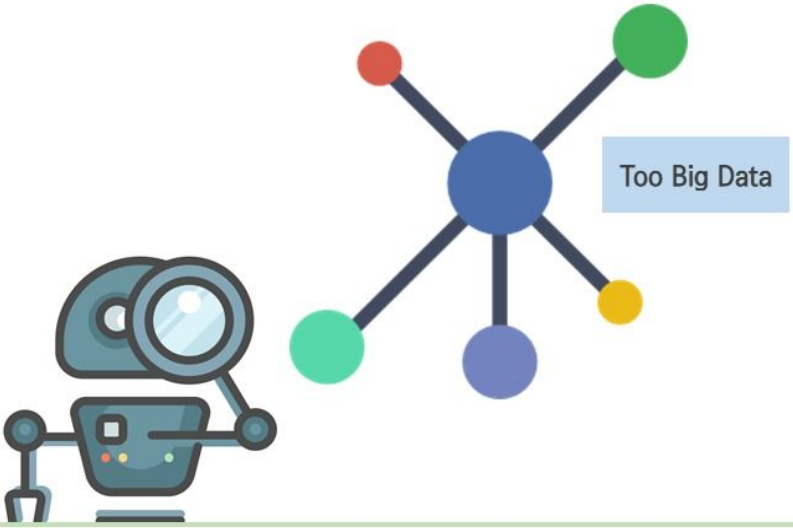
02 데이터 수집
및 전처리

03 모델링

04 결과

05 한계점

05 한계점



-데이터가 굉장히 많아
특정 Topic에 집중하여 프로젝트 진행

-TOT의 Topic 기간을 자의적으로 결정

Q & A

Thank you