

# Invest Our Technology 4.0

뉴스기사 텍스트 마이닝을 활용한 4차 산업혁명 투자영역 발견

2018 spring, Data analytics

이준송조\_ 이준송 박영준 배수영 이창수 조성욱 전명욱

# Contents



## 001 Team Introduction

## 002 Motivation & Objective

- Motivation
- Objective

## 003 Project process

- Data Acquisition
- Data Pre-Processing
- Data Modeling

## 004 Conclusion

- Finding
- Contribution

## 005 WBS & Gantt chart

# Team Introduction



이준송

Resource investigator  
Coordinator

- 기회를 발굴 및 탐색
- 목표를 명확히 하고  
팀원들에게 업무를 위임



박영준

Plant

- 창조적 아이디어 제안



배수영

Specialist  
Completer

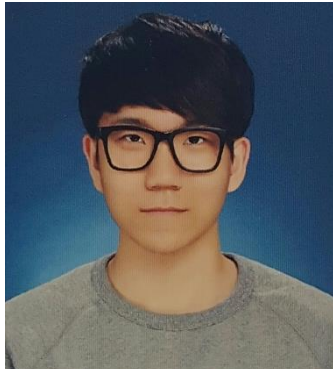
- 전문분야의 지식 제공
- 실수나 빠진 것을  
찾아내는 업무 수행



이창수

Shaper

- 장애에 봉착할 때 이를  
극복, 추진



조성욱

Implementer  
Monitor evaluator

- 아이디어를 실행에 옮김
- 모든 안을 살피고 판단



전명욱

Teamworker

- 경청하고 마찰을 피하며  
조직을 평온하게 함

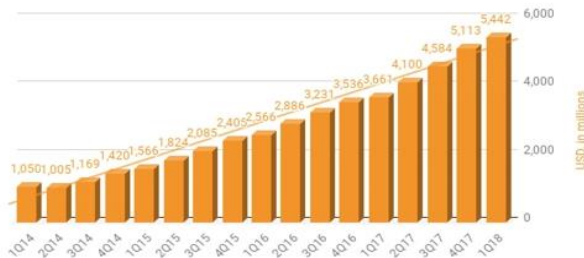
# Motivation & Objectives



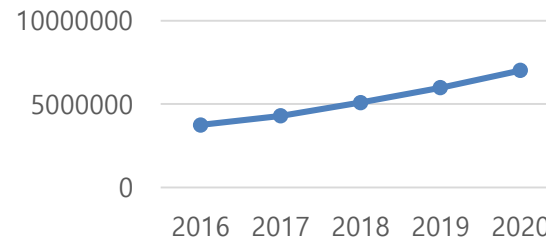
기술 초기단계인 2006년, Cloud 컴퓨팅 사업 런칭 시작

現, 전체 영업이익의 **73%**, 동일 업계 **독보적 1위**

매년 15%이상의 가파른 성장세를 보이는 Public cloud시장에서 점유율 32%를 자랑



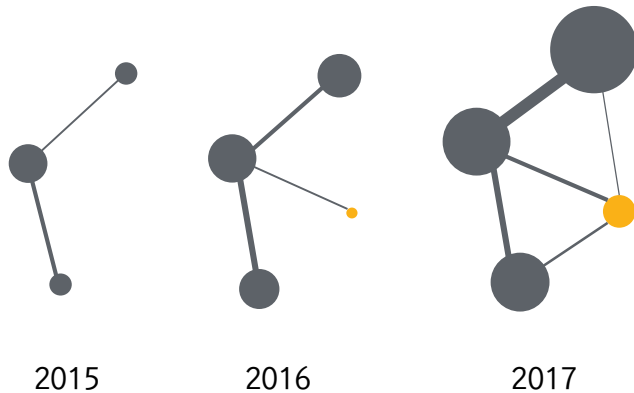
〈아마존의 전체 영업이익 중 클라우드 수익〉



〈전세계 Public cloud 사용자 지출 전망〉

이러한 AWS의 성공동력은 초기 유망기술에 대한  
**부상성 평가**에 기반한 **선지적인 투자**

# Motivation & Objectives



4차 산업혁명 관련

시간의 흐름에 따른 **item, keyword** 도출

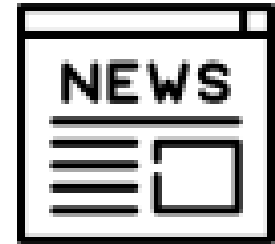
동향 파악, 잠재 성장 키워드 파악

# Data Acquisition

## What?

누적성, 다양성, 공신력 등과 같은 특징과 더불어, 제일 큰 특징으로 **특정 주제가 어떠한 성과나 가능성을 드러냈을 때 출현 및 빈도수가 증가**한다는 특징을 갖고 있다.

※ **뉴스를 데이터 획득 매체로 선정**



## Where?

1990년 이후 약 **3,000만 건 이상의 기사**를 축적하고 있는

**[Bigkinds]** 사이트를 이용하여 뉴스 데이터 획득하기로 한다.



## How?

**2016 ~ 2018년 5월**, [4차 산업혁명] 키워드를 포함하는 기사 약 3000여 건 수집, CSV파일로 확보  
(단, 추출과정의 공정성을 위해 Random 추출 실시)

기간	검색 건수	샘플링
2016. 1 ~ 2016. 12	3,489건	약 <b>500</b> 건
2017. 1 ~ 2017. 12	17,252건	약 <b>1,500</b> 건
2018. 1 ~ 2018. 5	6,578건	약 <b>1,000</b> 건

상세 검색

4차 산업혁명

정확히 일치하는 단어/문장 (" ")

4차 산업혁명

제외하는 단어 단어 (-)



검색 기간

2016-01-01



~

2016-12-31



전체

1일

1주

1개월

3개월

6개월

1년

언론사 ( 46 / 46 )

지역선택



☒ 중앙지

경향신문

국민일보

내일신문

문화일보

서울신문

세계일보

한겨레

한국일보

☒ 경제지

매일경제

머니투데이

서울경제

파이낸셜뉴스

한국경제

헤럴드경제

☒ 지역종합지

강원도민일보

경기일보

경남도민일보

경남신문

경상일보

경인일보

광주일보

국제신문

대구일보

대전일보

매일신문

무등일보

부산일보

영남일보

울산매일

전남일보

전북도민일보

전북일보

세민일보

중도일보

중부매일

충부일보

충북일보

충청일보

충청투데이

한라일보

☒ 방송사

MBC

CBS

SBS

YTN

☒ 전문지

디지털타임스

전자신문

주제 분류 ( 7 / 93 )

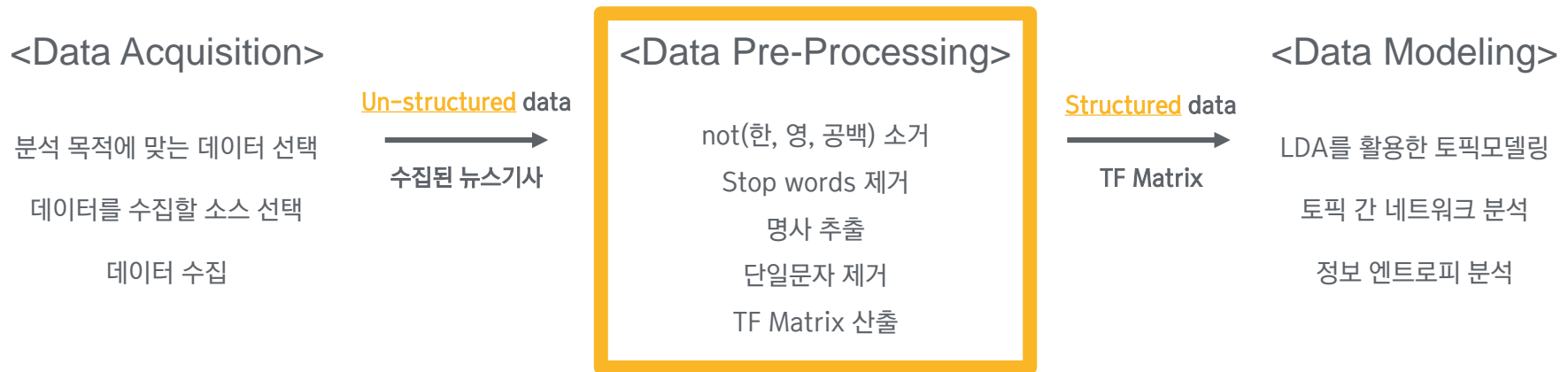


- ☐ 정치
- ☐ 경제
- ☐ 사회
- ☐ 문화
- ☐ 국제
- ☐ 지역
- ☐ 스포츠
- ☒ IT\_과학

# Data Pre-Processing

본 프로젝트에서는 Data acquisition 단계에서 수집된 비정형 데이터인 뉴스기사들을 활용하여, 해석적으로 의미가 있는 4차 산업관련 키워드와 그 빈도를 전처리 과정을 통해 추출한다.

초기 뉴스기사에서 기술적으로 관련이 없는 의미 단위들을 걷어낸 후, 유의미한 의미집합을 키워드 벡터로 만든다.



< Framework of IOT project >



# Data Pre-Processing

## 정규식처리1

정규식처리2

정규식처리3

정규식처리4

TF Matrix

### 한글, 영어 소문자 및 대문자, 공백문자 남기고 제거

특수기호와 숫자 등이 제거되고 한글, 영어,  
공백문자만 남게 된다.

```
def cut_trash(corpse):  
    for i in range(0,len(corpse)):  
        corpse[i]=re.sub("[^₩sa-zA-Z가-힣]", "",corpse[i])
```

#### Before

정밀의료 시장을 향한 격전은 이미 시작됐다.  
관련 기술 최고 보유국인 미국은 정밀의료  
발전계획 발표와 함께 열성적으로 추진 중이  
며, 영국은 10만 게놈프로젝트를 필두로 국  
가 주도로 의료데이터 수집 · 분석을 시작했  
다.



#### After

정밀의료 시장을 향한 격전은 이미 시작됐다  
관련 기술 최고 보유국인 미국은 정밀의료  
발전계획 발표와 함께 열성적으로 추진 중이  
며 영국은 10만 게놈프로젝트를 필두로 국가  
주도로 의료데이터 수집 · 분석을 시작했  
다

# Data Pre-Processing

정규식처리1

정규식처리2

정규식처리3

정규식처리4

TF Matrix

## Stop words 제거 ( IOT\_stoplist.txt (Apendix 첨부) )

반복적으로 수행되는 전처리 단계로,  
불용어 사전에 등록된 문자를 제거한다.

```
for i in range(0,len(corpse)):
    corpse[i]=re.sub("있|것|되|않|아니|같|우리|때|년|그|월|때|문|그|것|두|말|그|러|내|받|못|그|린|또|많|그|리|고|줄|따|르|가|지|시|키|지|금|생|각|그|레|모|르|어|텐|경|우|생|각|
    . . .
```

```
으로서|그|만|이|대|할 따름|이|대|쿵|탕탕|광|광|둥|둥|봐|봐|라|아|아|야|아|아|니|와|아|응|년|이|천|육|이|천|칠|이|천|팔|이|천|구|하|나|둘|셋|넷|다|섯|여|섯|일|곱|여|덟|아|홉|형|하|라|","|corpse[i])
```

### Before

정밀의료 시장을 향한 격전은 이미 시작됐다  
관련 기술 최고 보유국인 미국은 정밀의료  
발전계획 발표와 함께 열성적으로 추진 중이  
며 영국은 만 게놈프로젝트를 필두로 국가  
주도로 의료데이터 수집 분석을 시작했다



### After

정밀의료 시장을 향한 은  
관련 기술 보유국인 은 정밀의료  
발전계획 적으로 중이  
며 은 만 게놈프로젝트를 로  
로 의료데이터 수집 분석을

# Data Pre-Processing

정규식처리1

정규식처리2

정규식처리3

정규식처리4

TF Matrix

## 명사 추출 (KonlPy.Kkma 모듈 활용)

KonlPy의 Kkma 모듈을 활용하여 명사만을 추출한다.  
추출과정에서 남아있던 조사 등이 제거된다.

```
for i in range(0,len(corpse)):
    corpse[i]=engine.nouns(corpse[i])
```

### Before

정밀의료 시장을 향한 은  
관련 기술 보유국인 은 정밀의료  
발전계획 적으로 중이  
며 은 만 게놈프로젝트를 로  
로 의료데이터 수집 분석을



### After

‘정밀’, ‘의료’, ‘정밀의료’, ‘시장’, ‘은’,  
‘기술’, ‘보유국’, ‘인’, ‘은’, ‘정밀’,  
‘의료’, ‘정밀의료’, ‘발전’, ‘계획’, ‘발전  
계획’, ‘중’, ‘은’, ‘만’, ‘게놈’, ‘프로젝트’,  
‘게놈프로젝트’, ‘의료’, ‘데이터’, ‘의료데  
이터’, ‘수집’, ‘분석’

# Data Pre-Processing

정규식처리1

정규식처리2

정규식처리3

**정규식처리4**

TF Matrix

## 분석 시 다루기 어려운 단일문자 제거

단일문자는 분석 시 방해가 되므로 제거한다.

```
for i in range(0,len(corpse)):
    for j in range(0,len(corpse[i])):
        if(len(corpse[i][j])==1):
            corpse[i][j]=re.sub(".", " ",corpse[i][j])
```

### Before

‘정밀’, ‘의료’, ‘정밀의료’, ‘시장’, ‘은’,  
‘기술’, ‘보유국’, ‘인’, ‘은’, ‘정밀’,  
‘의료’, ‘정밀의료’, ‘발전’, ‘계획’, ‘발전  
계획’, ‘중’, ‘은’, ‘만’, ‘게놈’, ‘프로젝트’,  
‘게놈프로젝트’, ‘의료’, ‘데이터’, ‘의료데  
이터’, ‘수집’, ‘분석’



### After

‘정밀’, ‘의료’, ‘정밀의료’, ‘시장’,  
‘기술’, ‘보유국’, ‘정밀’, ‘의료’,  
‘정밀의료’, ‘발전’, ‘계획’, ‘발전계획’,  
‘게놈’, ‘프로젝트’, ‘게놈프로젝트’, ‘의  
료’, ‘데이터’, ‘의료데이터’, ‘수집’, ‘분석’

# Data Pre-Processing

정규식처리1

정규식처리2

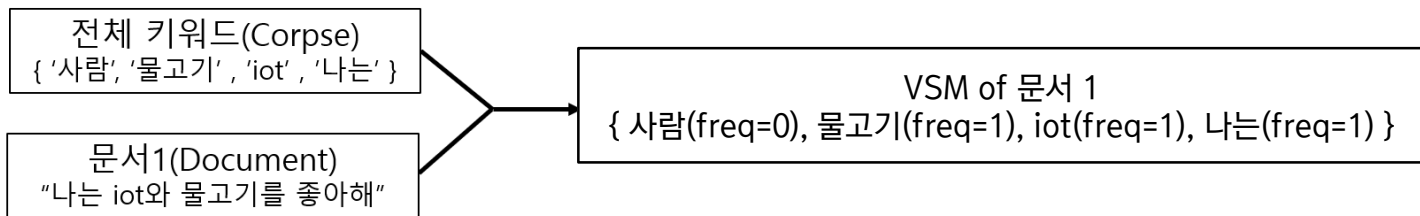
정규식처리3

정규식처리4

TF Matrix

## Vector space model(Term-Frequency matrix)

전체 문서에서 출현한 키워드들에 대해 한 문서를 하나의 벡터로 만드는 모델.  
본 프로젝트에서는 각 기사들마다 대응되는 키워드 벡터를 생성한 후, 모델링에서 활용한다.



## The number of keywords?

TF Matrix 생성과정에서 빈도 값의 하한의  
설정값에 따라서 키워드 숫자는 상이함

1→2 : 35000개 감소, 2→3 : 10000개 감소

Min : 3~5개 구간에서 의미 없는  
키워드들이 많이 소거되었음



# Data Pre-Processing

정규식처리1

정규식처리2

정규식처리3

정규식처리4

TF Matrix

## Term-Frequency matrix

TF Matrix 최종 도출

A	B	C	D	E	F	G	H	I	J
	가갈	가격	가게	가공	가공기술	가교	가구	가능성	가동
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0
19	0	0	0	1	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0
22	0	0	0	0	1	0	0	0	0
23	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0

TF Matrix(2981\*10291)

Edge list 생성

기사	키워드	빈도
	0 개설	1
	0 경제	1
	0 경제산업기	1
	0 경제취업창	1
	0 계약	1
	0 공유	1
	0 공학	1
	0 과목	1
	0 과정	1
	0 과학	1

모델링에서 분석을 위해 도출된  
TF매트릭스를 0 이상의 값을  
가지는 Edge-list로 표현

# Data Modeling

모델링에서는 이전 단계에서 생성한 대량의 기사데이터에 대한 TF Matrix를 바탕으로 **의미 있는 통찰**을 얻어내기 위해 Topic modeling과 다양한 방법론을 이용한다.

## 〈TF Matrix〉

	B	C	D	E	F	G	H	I	J
가	가	가	가	가	가	가	가	가	가
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0
19	0	0	1	0	0	0	0	1	0
20	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0
22	0	0	0	1	0	0	0	1	0
23	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	1	0
25	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0

Topic modeling &  
other method



## 〈Meaningful intelligence〉

4차 산업혁명 관련

시간의 흐름에 따른 **item, keyword** 도출

**동향** 파악, **잠재 성장 키워드** 파악

## 5 Steps of Modeling

Topic decision

토픽 모델링에서 쓰일 최적 토픽 수를 결정

Topic modeling

LDA를 이용한 Topic modeling

Topic Labeling & Coherence

LDA를 통해 도출된 Topic에 대한 이름 짓기

Topic network

T-D Matrix를 활용, 기간 별 토픽 간 상관관계 모니터링

Information entropy

T-T Matrix를 활용, 토픽간 영향관계와 특징 분석

# Data Modeling

## Topic Decision Topic Modeling Topic Labeling Topic Coherence Topic Network Information Entropy

적정 토픽 수를 결정할 때, 1) Perplexity 또는  
2) avg. Cosine-similarity elbow point method를 이용하여 결정한다.

### 1) Perplexity 반등 시점

$$Perp = 2^{-\frac{\sum_w LL(w)}{N}}$$

특정 확률 모델이 실제로 관측되는 값을 얼마나 잘 예측하는지를 평가할 때 사용된다.

Perplexity 값의 추세를 보았을 때, 해당 값이 갑작스레 반등하는 시점 부근의 Topic 수가 최적 Topic 수라 할 수 있다.

Topic의 수가 2~100일 때의 Perplexity 값의 추세를 살펴본 결과,  
**눈에 띄는 반등하는 시점을 찾아볼 수 없었다.**

따라서, 2) avg. Cosine-similarity elbow point method를 활용하기로 한다.

B
perplexity
2649.03453
2516.48207
2463.84451
2434.1759
2424.8959
2377.11748
2358.98689
2322.66565
2311.94397
2310.38647
2289.89811
2280.36816
2263.27801
2258.2461
2236.2558
2279.95487
2269.14543
2189.75595
2219.94584
2161.36133

〈Topic 수에 따른 Perplexity값의 추세〉



# Data Modeling

## Topic Decision

Topic Modeling

Topic Labeling

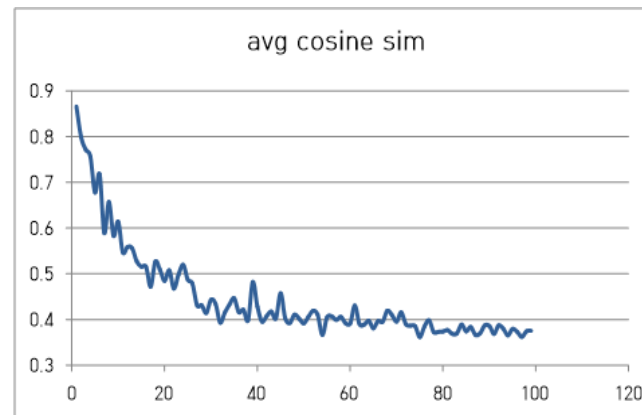
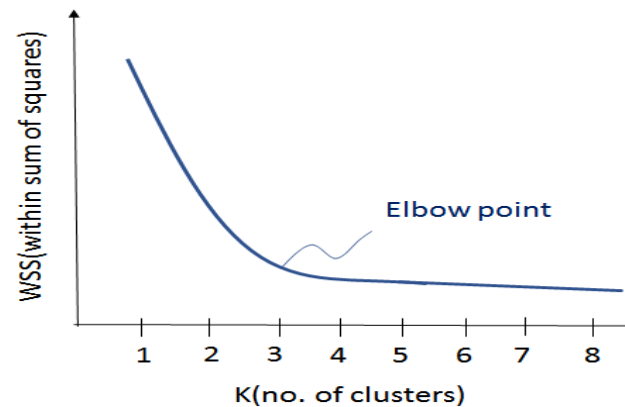
Topic Coherence

Topic Network

Information Entropy

## 2) Elbow point method

Avg.cosine 값이  
최하점에 도달 후, 다시 반등하는 지점  
혹은  
안정화되기 시작하는 포인트  
를 최적 토픽으로 결정하는 기법



본 프로젝트에선 눈에 띄는 반등 지점은 없으나,  
토픽수가 32개 부근 이후로 개형이 안정화  
되는 경향을 발견,  
본 프로젝트에서는 33개를 토픽의 개수로 선정

	A
1	avg cosine sim
2	0.865856954
3	0.799689067
4	0.771585759
5	0.756501165
6	0.678233305
7	0.717475452
8	0.589333274
9	0.657163155
10	0.583168158
11	0.614132337
12	0.547104591
13	0.55865006
14	0.556273917
15	0.527469771
16	0.515476115
17	0.515596379
18	0.471867038
19	0.52674717
20	0.509232684
21	0.484529483
22	0.507677728
23	0.46788259
24	0.498159744
25	0.519880673
26	0.487295532
27	0.478507387
28	0.43105093
29	0.431188201
30	0.414188201
31	0.443328437
32	0.454106099
33	0.383244454
34	0.415754631
35	0.433049052
36	0.446814654
37	0.41592875
38	0.421466475
39	0.398183477
40	0.482189439
41	0.428637784
42	0.395533018
43	0.407726882
44	0.417518735
45	0.402238537
46	0.45770943

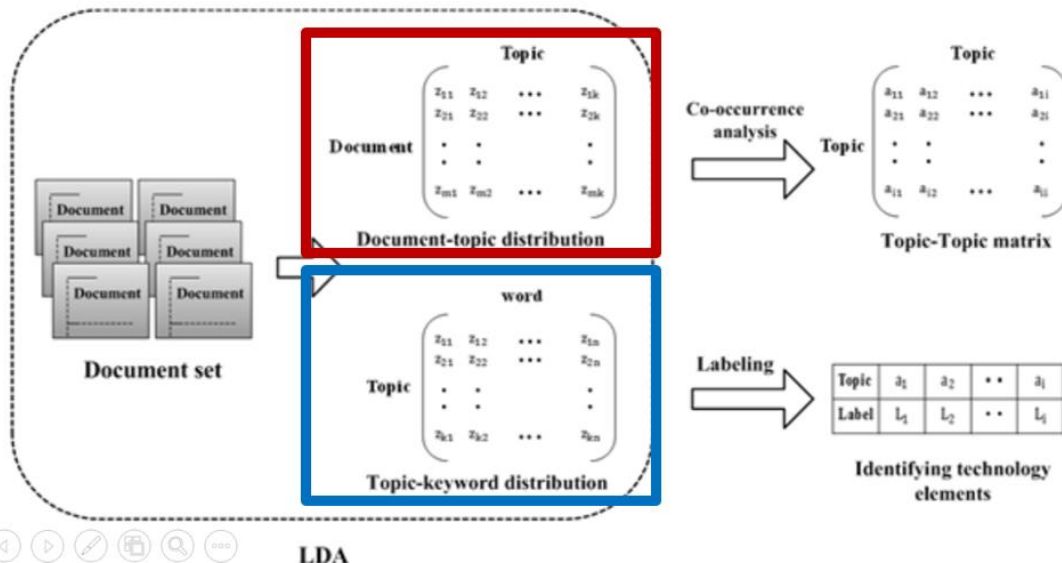
# Data Modeling

Topic Decision **Topic Modeling** Topic Labeling Topic Coherence Topic Network Information Entropy

## LDA

사람이 문서를 작성한다는 가정 하에 문서의 잠재적인 주제를 찾아내는 확률기반 생성모형.  
벡터들의 조합은 문서의 주제를 나타낸다. 토픽모델링은 구조화되지 않은 방대한 문서 집단에서  
잠정적 주제를 찾아내고 대량 문서를 분류하는 데 사용된다

## LDA in our project



각 문서의 토픽 분포를 나타내는  
**기사-토픽 매트릭스**  
네트워크 분석을 진행하기 위하여  
토픽들의 동시 출현 매트릭스를  
구축하는데 활용

토픽 별 구성 단어의 분포를 나타내는  
**토픽-키워드**  
토픽의 높은 구성 확률을 가지는  
상위 키워드를 확인하여 해당  
토픽에 대한 Labeling에 활용될 것이다.

# Data Modeling

Topic Decision **Topic Modeling** Topic Labeling Topic Coherence Topic Network Information Entropy

토픽

1	Topic1	Topic2	Topic3	Topic4	Topic5
2	0.208210488	0.000229568	0.000229568	0.000229568	0.000229568
3	0.000234907	0.000234907	0.000234907	0.000234907	0.000234907
4	5.97E-05	5.97E-05	0.338575945	5.97E-05	5.97E-05
5	5.97E-05	5.97E-05	0.338575945	5.97E-05	5.97E-05
6	0.000150015	0.000150015	0.000150015	0.000150015	0.000150015
7	0.000126263	0.000126263	0.000126263	0.059014407	0.000126263
8	0.000261233	0.000261233	0.000261233	0.000261233	0.505531117
9	0.000150015	0.000150015	0.000150015	0.000150015	0.164056692
10	0.000175162	0.000175162	0.000175162	0.10494539	0.156982981
11	0.000148544	0.000148544	0.000148544	0.039161333	0.000148544
12	0.000234907	0.000234907	0.000234907	0.000234907	0.000234907
13	0.513283644	0.000208986	0.000208986	0.000208986	0.19735122
14	0.000226142	0.000226142	0.000226142	0.000226142	0.000226142

기사-토픽 매트릭스

키워드

1	글로벌융합	글자	금리	금속	금액
2	5.22E-06	5.22E-06	0.000225	5.22E-06	5.22E-06
3	5.02E-06	5.02E-06	5.02E-06	5.02E-06	5.02E-06
4	1.53E-06	1.53E-06	1.53E-06	0.001107	1.53E-06
5	1.47E-06	9.82E-05	1.47E-06	1.47E-06	1.47E-06
6	1.02E-06	1.02E-06	1.02E-06	1.02E-06	0.000437
7	9.45E-06	9.45E-06	9.45E-06	0.002745	9.45E-06
8	6.47E-06	6.47E-06	6.47E-06	6.47E-06	6.47E-06
9	5.80E-06	5.80E-06	5.80E-06	0.000318	5.80E-06
10	9.10E-07	9.10E-07	1.17E-06	9.10E-07	9.10E-07
11	5.17E-07	5.17E-07	5.17E-07	5.17E-07	5.17E-07
12	8.73E-06	8.73E-06	8.73E-06	8.73E-06	8.73E-06
13	5.85E-06	5.85E-06	5.85E-06	5.85E-06	5.85E-06
14	1.11E-05	1.11E-05	1.11E-05	1.11E-05	1.11E-05

토픽-키워드 매트릭스

	A	B	C	D	E	F	G	H	I
1	1stWord	1stProb	2ndWord	2ndProb	3rdWord	3rdProb	4thWord	4thProb	5thWord
2	과정	0.005016	전문가	0.004303	경제	0.003702	분야	0.003696	정책
3	기술	0.009026	지능	0.0083	인공지능	0.008154	인공	0.008096	시대
4	스마트	0.007989	기술	0.006739	제품	0.006008	사물	0.005513	생산
5	정보	0.005639	기술	0.005368	시스템	0.005287	스마트	0.005005	구축
6	개발	0.005606	지능	0.005166	기술	0.005157	인공	0.005101	인공지능
7	항공	0.009493	우주	0.007557	전기	0.006461	선정	0.005923	화학
8	주제	0.01058	주최	0.008695	개최	0.008443	행사	0.008389	코엑스
9	체인	0.009453	블록	0.009129	블록체인	0.008911	기술	0.007862	정보
10	추진	0.007229	계획	0.00658	개발	0.006117	기술	0.006114	분야
11	사람	0.005278	때문	0.005226	지능	0.005188	인공	0.005186	인공지능
12	기술	0.008491	주제	0.007847	강연	0.007159	기업	0.006262	세계
13	체결	0.008164	개발	0.007699	기술	0.007436	추진	0.006621	연구
14	팀장	0.007663	취재	0.007474	임성	0.007366	임성현	0.006998	특별
15	정보	0.005555	관련	0.005395	제도	0.005332	활용	0.00524	개인
16	기술	0.005933	블록체인	0.005845	화폐	0.005712	금융	0.005707	기업
17	중소	0.006422	기술	0.00619	사업	0.006129	개최	0.005867	기업
18	주제	0.007537	인공	0.006952	인공지능	0.006949	기술	0.00674	지능
19	기술	0.007207	인터넷	0.005876	올해	0.005791	분야	0.005493	예정

각 토픽에 대한 키워드



Topic 별 키워드 출현확률  
내림차순 정리

# Data Modeling

Topic Decision   Topic Modeling   **Topic Labeling**   Topic Coherence   Topic Network   Information Entropy

## Good to label

1stWord	2ndWord	3rdWord	4thWord	5thWord	6thWord	7thWord	8thWord
스마트	기술	제품	사물인터넷	생산	제조	인터넷	시스템
개발	지능	기술	인공	인공지능	주행	자율	세계
항공	우주	정밀	선정	화학	연구	공학부	공학

→ Topic 1 : 제조 관련 사물인터넷  
Topic 2 : 인공지능 관련 자율주행  
Topic 3 : 항공 및 우주 정밀기술

## Bad to label

1stWord	2ndWord	3rdWord	4thWord	5thWord	6thWord	7thWord	8thWord
주제	주최	개최	행사	코엑스	기술	엑스	시대
추진	계획	개발	기술	분야	지원	정부	과학
주제	인공	개발	기술	지능	공유	행사	기업

→ Topic coherence를 이용

# Data Modeling

Topic Decision

Topic Modeling

Topic Labeling

**Topic Coherence**

Topic Network

Information Entropy

## Topic Coherence

토픽 모델링의 처리된 결과가 사람이 바라는 결과와 얼마나 일치하는지를 평가하는 방법

뉴스데이터의 경우,  
Google-titles-match가  
가장 높은 효율을 보임

Google	TITLES	<b>0.80</b>	
	LOGHITS	0.46	
Gold-standard	IAA	0.79	0.73

Table 2: Spearman rank correlation  $\rho$  values for the different scoring methods over the NEWS dataset (best-performing method for each resource underlined; best-performing method overall in boldface)

출처 : Automatic Evaluation of Topic Coherence(2010)-David Newman et al.

## Google-titles-match

토픽 모델링으로 도출된 문서의 핵심 키워드를 Google에 검색하여 상위 100개 제목 속 키워드와 일치하는 횟수를 측정

## Topic Labeling에 활용

23번 Topic의 핵심 키워드 5개(센서, 개발, 시스템, 소프트, 서비스)를 3개씩 선택, 조합하여 5C<sub>3</sub>가지 경우에 대해 Google-titles-match 계산

$$\text{Google-titles-match}(\mathbf{w}) = \mathbf{1} [w_i = v_j]$$

( W: 검색 제목의 키워드, V: 모델링 결과의 키워드 )

결과

1. 센서, 개발, 시스템 -> 177
2. 소프트, 개발, 시스템 -> 150
3. 소프트, 개발, 서비스 -> 138
4. 개발, 서비스, 시스템 -> 138
5. 센서, 소프트, 개발 -> 124
6. 센서, 개발, 서비스 -> 124
7. 소프트, 서비스, 시스템 -> 124
8. 센서, 서비스, 시스템 -> 116
9. 센서, 소프트, 시스템 -> 104
10. 센서, 소프트, 서비스 -> 95

< Google-titles-match(센서,개발,시스템,소프트,서비스) >

1~10번 조합을 Searching한 결과, **바이오, 생체** 라는 키워드와 빈출



Topic Labeling 과정에 반영



23번 Topic  
**바이오 센서**

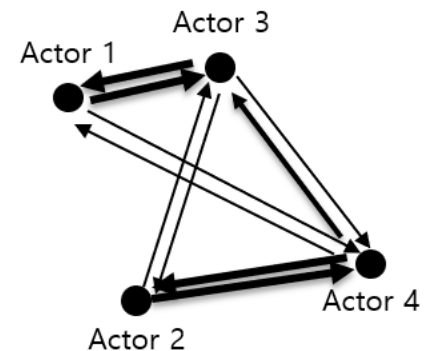
# Data Modeling

Topic Decision   Topic Modeling   Topic Labeling   Topic Coherence   **Topic Network**   Information Entropy

## Actor network theory(ANT)

특성을 가지는 행위자(Actor)와 노드들 간의 관계인 링크(link)로 구성요소 간 전체 및 부분적 패턴과 연결 관계를 설명하는 네트워크 분석방법

	Actor 1	Actor 2	Actor 3	Actor 4
Actor 1	1	0	3	1
Actor 2	0	1	1	3
Actor 3	3	1	1	2
Actor 4	1	3	2	1



## ANT of Our topic – topic matrix

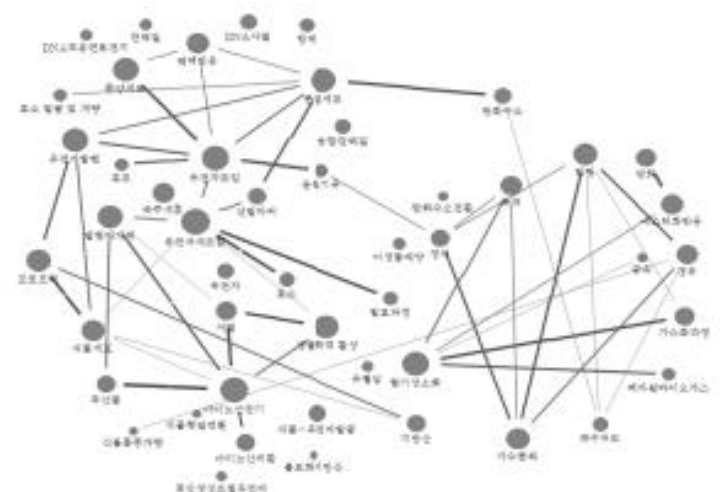
A =

	Topic1	Topic2	Topic3	Topic4	Topic5
1	0.208210488	0.000229568	0.000229568	0.000229568	0.000229568
2	0.000234907	0.000234907	0.000234907	0.000234907	0.000234907
3	5.97E-05	5.97E-05	0.338575945	5.97E-05	5.97E-05
4	5.97E-05	5.97E-05	0.338575945	5.97E-05	5.97E-05
5	0.000150015	0.000150015	0.000150015	0.000150015	0.000150015
6	0.000126263	0.000126263	0.000126263	0.059014407	0.000126263
7	0.000261233	0.000261233	0.000261233	0.000261233	0.505931117
8	0.000150015	0.000150015	0.000150015	0.000150015	0.164056692
9	0.000175162	0.000175162	0.000175162	0.10494539	0.156982981
10	0.000148544	0.000148544	0.000148544	0.039161333	0.000148544
11	0.000234907	0.000234907	0.000234907	0.000234907	0.000234907
12	0.513283644	0.000208986	0.000208986	0.000208986	0.19735122
13	0.000226142	0.000226142	0.000226142	0.000226142	0.000226142
14					

Topic-Document Matrix

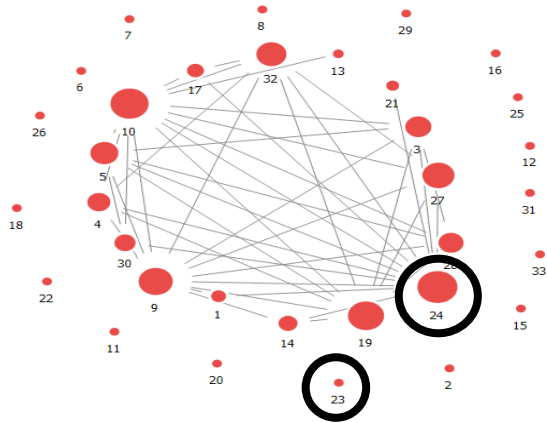
$$N = \text{trans}(A) * A$$

토픽들 간의  
영향관계 모니터링

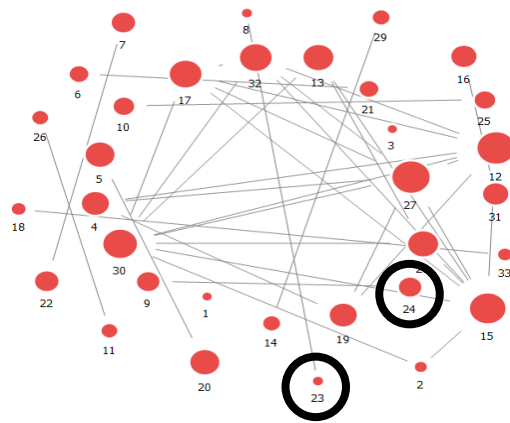


# Data Modeling

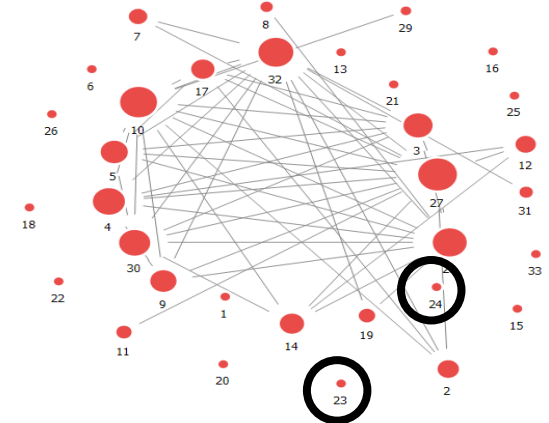
Topic Decision   Topic Modeling   Topic Labeling   Topic Coherence   **Topic Network**   Information Entropy



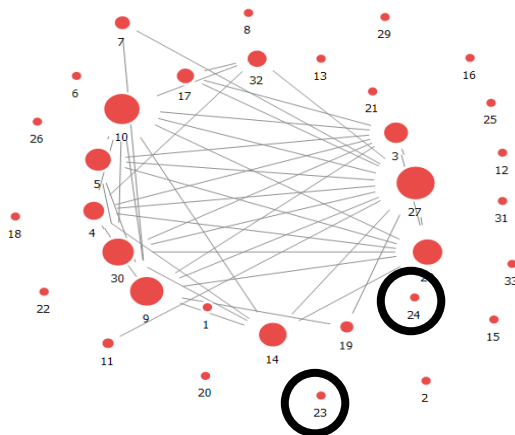
〈 2016 〉



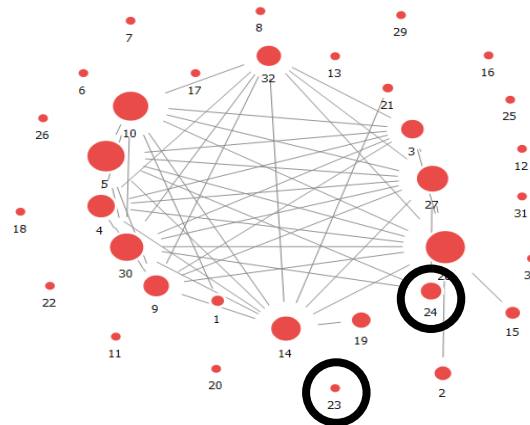
〈 2017/ 1~4 〉



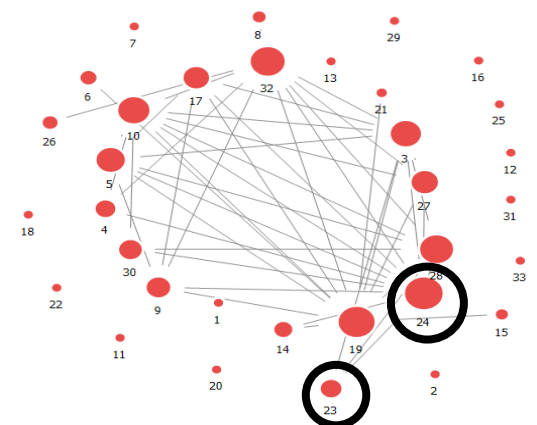
〈 2017/ 5~8 〉



〈 2017/ 9~12 〉



〈 2018/ 1~3 〉



〈 2018/ 4~5 〉

# Data Modeling

Topic Decision

Topic Modeling

Topic Labeling

Topic Coherence

Topic Network

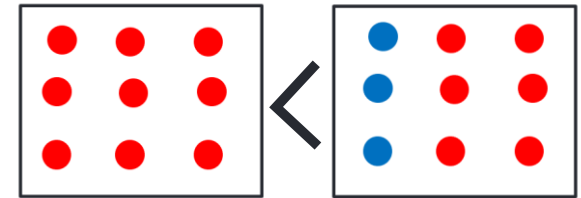
Information Entropy

## Information Entropy

무질서도를 의미하며 불확실성, 다양성을 말함. 정보 엔트로피(H)는 여러 종류의 요소로 구성되어 있는 집단에서 해당 요소의 군집내 비율(P)로 계산한다.

정보 엔트로피가 크다는 것은  
자료들이 균등분포를 하므로  
불확실성이 높다는 것을 의미  
한다.

$$H = - \sum_{i=1}^n P_i \log P_i$$



< 2개 군집에 대한 정보 엔트로피 비교 >

## Information Entropy of our project?

토픽-토픽 네트워크에서 연구구간에 따른 엔트로피 계산 후, 인접 토픽간 다양성과 영향성을 파악하고  
토픽들과의 연계적 측면에서 향후 발전 가능성을 측정

Topic 1  
직업교육



Topic 6  
우주항공



Topic 23  
바이오센서



# Data Modeling

Topic Decision Topic Modeling Topic Labeling Topic Coherence Topic Network

Information Entropy

		16	17상	17중	17하	18상	18중	전체	평균	증가율
Topic1	직업교육	0.10448	0.10732	0.11296	0.10727	0.11468	0.10539	0.1206	0.10868	0.00172
Topic2	빅데이터융합	0.11599	0.10957	0.114	0.12602	0.11191	0.11026	0.1242	0.11463	-0.0101
Topic3	스마트 팩토리	0.11835	0.10912	0.12039	0.11854	0.12347	0.11964	0.12693	0.11825	0.00217
Topic4	IOT	0.11057	0.09713	0.11606	0.11349	0.11943	0.118	0.12091	0.11245	0.01309
Topic5	자율주행	0.11251	0.05408	0.12125	0.12031	0.11796	0.10999	0.12399	0.10602	-0.0045
Topic6	우주항공	0.11055	0.08491	0.11631	0.11672	0.11039	0.09174	0.12046	0.1051	-0.0366
Topic7	코엑스 컨퍼런스	0.1114	0.0269	0.12686	0.12868	0.12465	0.11468	0.13061	0.10553	0.00583
Topic8	금융기업의 폐쇄형 블록체인	0.11106	0.0937	0.11754	0.1222	0.11626	0.0978	0.1269	0.10976	-0.0251
Topic9	데이터센터	0.11794	0.08688	0.11978	0.12853	0.12716	0.12099	0.13096	0.11688	0.00512
Topic10	인공지능에 의한 미래계획	0.12294	0.0497	0.12315	0.123	0.12429	0.12092	0.12832	0.11067	-0.0033
Topic11	기술관련 강연	0.10572	0.08654	0.11061	0.11166	0.11512	0.09653	0.12629	0.10436	-0.018
Topic12	연구협약	0.11534	0.06678	0.11725	0.12185	0.11954	0.10052	0.12619	0.10688	-0.0271
Topic13	인터넷뉴스	0.10778	0.09535	0.10714	0.07822	0.09884	0.10484	0.11757	0.0987	-0.0055
Topic14	의료서비스	0.11581	0.09361	0.1182	0.11432	0.12595	0.11399	0.12676	0.11365	-0.0032
Topic15	가상화폐	0.10459	0.09149	0.10241	0.11202	0.11087	0.09958	0.12077	0.10349	-0.0098
Topic16	중소기업 기술이전	0.10837	0.06182	0.11482	0.12666	0.11587	0.11115	0.12603	0.10645	0.00508
Topic17	인공지능기술 대회	0.1129	0.09301	0.1271	0.12591	0.11523	0.11784	0.12997	0.11533	0.0086
Topic18	공유기술 협약	0.10393	0.08817	0.11283	0.12402	0.11216	0.107	0.12602	0.10802	0.00583
Topic19	SNS	0.12667	0.09713	0.12435	0.1172	0.11645	0.13097	0.13213	0.1188	0.0067
Topic20	인공지능과 일자리의 연관성	0.10287	0.05408	0.09298	0.12162	0.11373	0.10953	0.12413	0.09914	0.01263
Topic21	특허재산권	0.09798	0.08491	0.12128	0.11286	0.11603	0.0954	0.12402	0.10474	-0.0053
Topic22	도시개발	0.09098	0.0269	0.09454	0.09722	0.09906	0.07144	0.11636	0.08002	-0.0472
Topic23	바이오센서	0.09111	0.0937	0.11571	0.11906	0.11284	0.10563	0.12173	0.10634	0.03
Topic24	SNS보안	0.12852	0.08688	0.11374	0.12082	0.10721	0.12542	0.12875	0.11376	-0.0049
Topic25	에너지산업	0.11343	0.0497	0.09364	0.1132	0.10894	0.11153	0.12415	0.09841	-0.0034

연도별 각 Topic의 information Entropy를 도출

다른 분야와의 활용성과  
향후 발전가능성의 분석을 위해

Information Entropy의  
평균값과 증가율을 측정

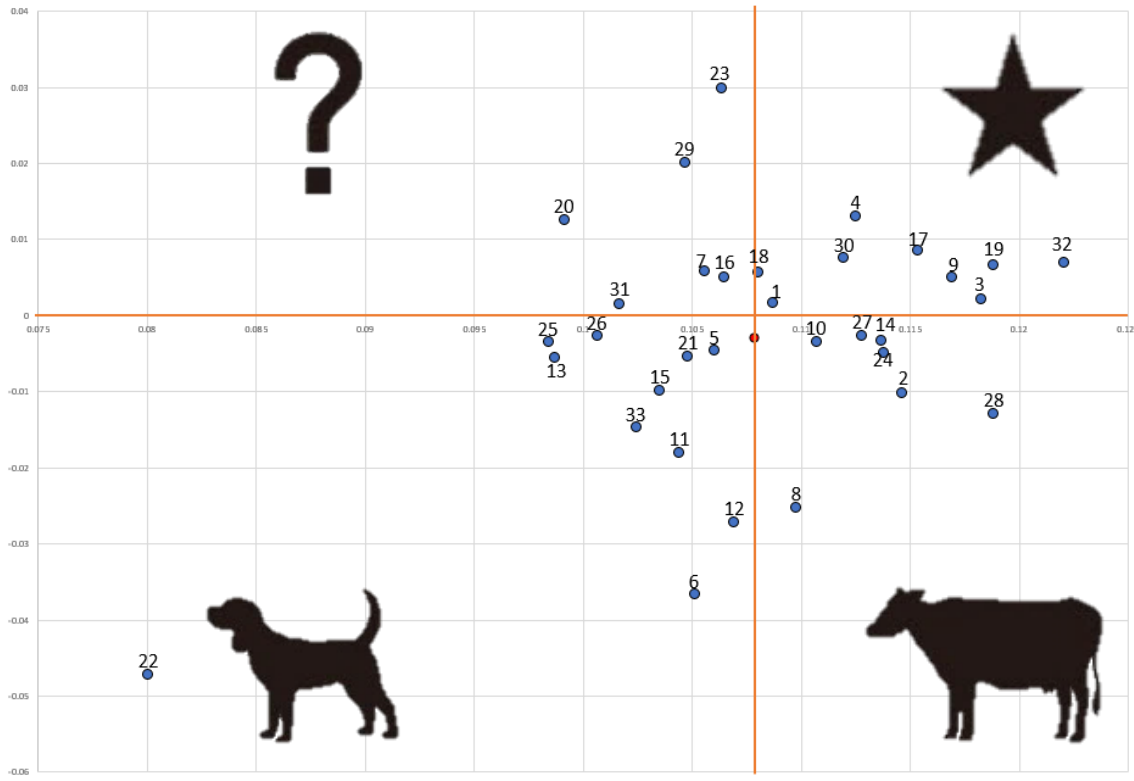
$$r = \left( \frac{a_n}{a_0} \right)^{1/n} - 1$$

$r$ : 증가율,  $a_0$ : 시작값,  $a_n$ : 끝값

# Data Modeling

Topic Decision Topic Modeling Topic Labeling Topic Coherence Topic Network

Information Entropy



( X : 평균값(활용도), Y : 증가율(성장가능성) )

## 1사분면

높은 평균값 + 높은 증가율  
활용도가 높고 성장가능성도 높아  
이미 투자가 많이 이루어지고 있는 영역

## 2사분면

낮은 평균값 + 높은 증가율  
활용도는 낮으나 성장가능성이 높아  
유망성이 높은 잠재투자영역

## 3사분면

낮은 평균값 + 낮은 증가율  
활용도와 성장가능성이 낮아  
투자 시 회피 해야 할 영역

## 4사분면

높은 평균값 + 낮은 증가율  
단기적 수익률을 높일 수 있는 투자영역

# Conclusion

## Finding

Information Entropy 도표의 1사분면에 위치한 토픽들은 바이오센서는 현재 활용도와 성장률 모두 긍정적인 수치를 보여주고 있다. 그러므로 안정적인 투자대상을 모색하고 있다면 위 토픽을 우선 고려해볼 수 있다.

Information Entropy 도표의 2사분면에 위치한 Topic23 바이오센서는 활용도는 낮지만 평균에 근접해있고, 성장률은 Topic중 가장 높으므로 장기적인 성장가능성을 우선한다면 Topic 23에 대한 투자를 고려해 볼 수 있을 것이다

Information Entropy 도표의 4사분면에 위치한 Topic24 SNS보안은 2017년 9~12월 이후 Entropy값과 degree값이 증가하는 추세를 보이고 있다. 이는 올해 초에 발생한 Facebook 개인정보 유출 사태를 시발점으로 관심이 높아지고 연구가 활발히 진행되었기 때문이라고 생각해볼 수 있다. 따라서 단기적 투자대상으로 선택 가능하다.

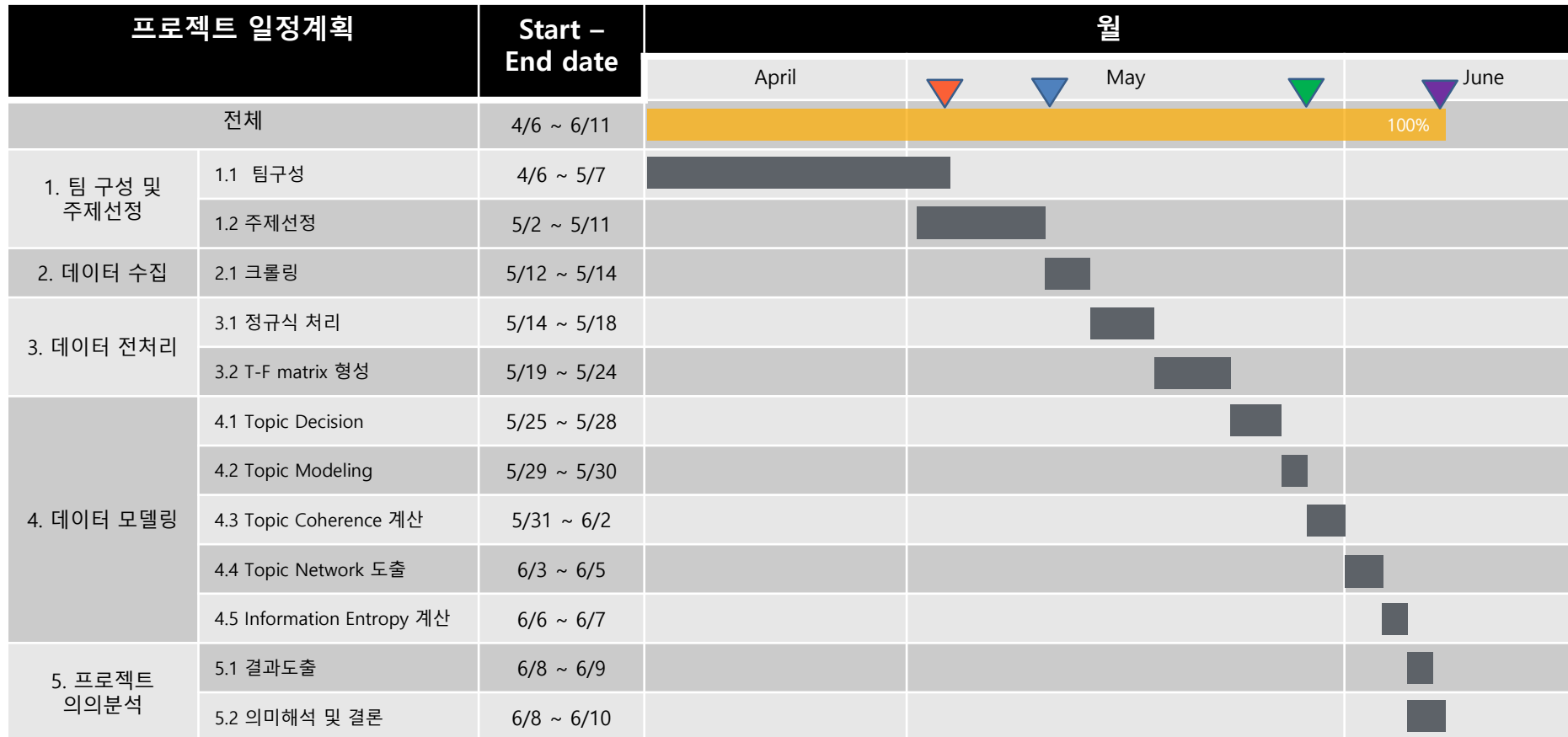
## Contribution





본 프로젝트에서 분석한 결과는 **4차 산업혁명 기술들의 세부 기술요소를 파악**할 수 있으며, **전체적인 동향 분석 결과를 제공**한다.

나아가, 본 팀은 **어떤 세부 기술들이 차후 경쟁력이 있을 기술인지에 대하여 통찰**하였다.

이는 4차 산업혁명의 파도 속에서 R&D를 연구하는 다양한 분야의 연구자들과 기업에게 기초 연구 자료로써 활용될 수 있으며 향후 연구 개발의 방향성을 설정하는 데 효과적일 것으로 기대된다.

# WBS & Gantt Chart



-  Team building (5/7)
-  Topic selection (5/15)
-  Interim report (5/29)
-  Final report (6/10)

1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53. 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65. 66. 67. 68. 69. 70. 71. 72. 73. 74. 75. 76. 77. 78. 79. 80. 81. 82. 83. 84. 85. 86. 87. 88. 89. 90. 91. 92. 93. 94. 95. 96. 97. 98. 99. 100. 101. 102. 103. 104. 105. 106. 107. 108. 109. 110. 111. 112. 113. 114. 115. 116. 117. 118. 119. 120. 121. 122. 123. 124. 125. 126. 127. 128. 129. 130. 131. 132. 133. 134. 135. 136. 137. 138. 139. 140. 141. 142. 143. 144. 145. 146. 147. 148. 149. 150. 151. 152. 153. 154. 155. 156. 157. 158. 159. 160. 161. 162. 163. 164. 165. 166. 167. 168. 169. 170. 171. 172. 173. 174. 175. 176. 177. 178. 179. 180. 181. 182. 183. 184. 185. 186. 187. 188. 189. 190. 191. 192. 193. 194. 195. 196. 197. 198. 199. 200. 201. 202. 203. 204. 205. 206. 207. 208. 209. 210. 211. 212. 213. 214. 215. 216. 217. 218. 219. 220. 221. 222. 223. 224. 225. 226. 227. 228. 229. 230. 231. 232. 233. 234. 235. 236. 237. 238. 239. 240. 241. 242. 243. 244. 245. 246. 247. 248. 249. 250. 251. 252. 253. 254. 255. 256. 257. 258. 259. 260. 261. 262. 263. 264. 265. 266. 267. 268. 269. 270. 271. 272. 273. 274. 275. 276. 277. 278. 279. 280. 281. 282. 283. 284. 285. 286. 287. 288. 289. 290. 291. 292. 293. 294. 295. 296. 297. 298. 299. 300. 301. 302. 303. 304. 305. 306. 307. 308. 309. 310. 311. 312. 313. 314. 315. 316. 317. 318. 319. 320. 321. 322. 323. 324. 325. 326. 327. 328. 329. 330. 331. 332. 333. 334. 335. 336. 337. 338. 339. 340. 341. 342. 343. 344. 345. 346. 347. 348. 349. 350. 351. 352. 353. 354. 355. 356. 357. 358. 359. 360. 361. 362. 363. 364. 365. 366. 367. 368. 369. 370. 371. 372. 373. 374. 375. 376. 377. 378. 379. 380. 381. 382. 383. 384. 385. 386. 387. 388. 389. 390. 391. 392. 393. 394. 395. 396. 397. 398. 399. 400. 401. 402. 403. 404. 405. 406. 407. 408. 409. 410. 411. 412. 413. 414. 415. 416. 417. 418. 419. 420. 421. 422. 423. 424. 425. 426. 427. 428. 429. 430. 431. 432. 433. 434. 435. 436. 437. 438. 439. 440. 441. 442. 443. 444. 445. 446. 447. 448. 449. 450. 451. 452. 453. 454. 455. 456. 457. 458. 459. 460. 461. 462. 463. 464. 465. 466. 467. 468. 469. 470. 471. 472. 473. 474. 475. 476. 477. 478. 479. 480. 481. 482. 483. 484. 485. 486. 487. 488. 489. 490. 491. 492. 493. 494. 495. 496. 497. 498. 499. 500. 501. 502. 503. 504. 505. 506. 507. 508. 509. 510. 511. 512. 513. 514. 515. 516. 517. 518. 519. 520. 521. 522. 523. 524. 525. 526. 527. 528. 529. 530. 531. 532. 533. 534. 535. 536. 537. 538. 539. 540. 541. 542. 543. 544. 545. 546. 547. 548. 549. 550. 551. 552. 553. 554. 555. 556. 557. 558. 559. 560. 561. 562. 563. 564. 565. 566. 567. 568. 569. 570. 571. 572. 573. 574. 575. 576. 577. 578. 579. 580. 581. 582. 583. 584. 585. 586. 587. 588. 589. 590. 591. 592. 593. 594. 595. 596. 597. 598. 599. 600. 601. 602. 603. 604. 605. 606. 607. 608. 609. 610. 611. 612. 613. 614. 615. 616. 617. 618. 619. 620. 621. 622. 623. 624. 625. 626. 627. 628. 629. 630. 631. 632. 633. 634. 635. 636. 637. 638. 639. 640. 641. 642. 643. 644. 645. 646. 647. 648. 649. 650. 651. 652. 653. 654. 655. 656. 657. 658. 659. 660. 661. 662. 663. 664. 665. 666. 667. 668. 669. 670. 671. 672. 673. 674. 675. 676. 677. 678. 679. 680. 681. 682. 683. 684. 685. 686. 687. 688. 689. 690. 691. 692. 693. 694. 695. 696. 697. 698. 699. 700. 701. 702. 703. 704. 705. 706. 707. 708. 709. 710. 711. 712. 713. 714. 715. 716. 717. 718. 719. 720. 721. 722. 723. 724. 725. 726. 727. 728. 729. 730. 731. 732. 733. 734. 735. 736. 737. 738. 739. 740. 741. 742. 743. 744. 745. 746. 747. 748. 749. 750. 751. 752. 753. 754. 755. 756. 757. 758. 759. 760. 761. 762. 763. 764. 765. 766. 767. 768. 769. 770. 771. 772. 773. 774. 775. 776. 777. 778. 779. 780. 781. 782. 783. 784. 785. 786. 787. 788. 789. 790. 791. 792. 793. 794. 795. 796. 797. 798. 799. 800. 801. 802. 803. 804. 805. 806. 807. 808. 809. 810. 811. 812. 813. 814. 815. 816. 817. 818. 819. 820. 821. 822. 823. 824. 825. 826. 827. 828. 829. 830. 831. 832. 833. 834. 835. 836. 837. 838. 839. 840.

# Appendix

## 2. Topic-Labeling

|         |                |
|---------|----------------|
| Topic1  | 직업교육           |
| Topic2  | 빅데이터융합         |
| Topic3  | 스마트 팩토리        |
| Topic4  | IOT            |
| Topic5  | 자율주행           |
| Topic6  | 우주항공           |
| Topic7  | 코엑스 컨퍼런스       |
| Topic8  | 금융기업의 폐쇄형 블록체인 |
| Topic9  | 데이터센터          |
| Topic10 | 인공지능에 의한 미래계획  |
| Topic11 | 기술관련 강연        |
| Topic12 | 연구협약           |
| Topic13 | 인터넷뉴스          |
| Topic14 | 의료서비스          |
| Topic15 | 가상화폐           |

|         |                |
|---------|----------------|
| Topic16 | 중소기업 기술이전      |
| Topic17 | 인공지능기술 대회      |
| Topic18 | 공유기술 협약        |
| Topic19 | SNS            |
| Topic20 | 인공지능과 일자리의 연관성 |
| Topic21 | 특허재산권          |
| Topic22 | 도시개발           |
| Topic23 | 바이오센서          |
| Topic24 | SNS보안          |
| Topic25 | 에너지산업          |
| Topic26 | 카이스트           |
| Topic27 | 융합기술관련 정책      |
| Topic28 | 금융데이터제공서비스     |
| Topic29 | 인공지능비서         |
| Topic30 | 기업분석           |
| Topic31 | 클라우드           |
| Topic32 | 교육             |
| Topic33 | 소셜커머스          |