# Midterm Report

**Department:** AI
**NAME:** JIN, CHANGLONG
**ID:** 2024314281

## 1. Basic Information and Objective

**Programming Language**: Python@3.11.12
**Main Libraries/Tools**: MeCab, Gensim (for LDA), and Tomotopy (for MDR)
**CPU**: M3 Chip
**CPU Architecture**: ARM64
**Dataset**: 먹거리_식품_안전(2017~2012).txt **(Medium-to-Long Text)**
**Project Repository**: https://github.com/kimchanglong0128/TextMining_Midterm.git

**Objective:**
**For LDA,**

- Which keywords frequently co-occur (**keywords extraction**)
- Which Agencies mainly focus on which topics
- Which documents strongly connected to a particular topic
- The general distribution pattern between documents and topics

**For DMR,**

- How the importance of topics changes across different agencies and time periods
- Certain topics are notably reinforced during periods or by specific institutions (Observe whether the data is sensitive to features such as temporal and agencies)

**Note:**
For this midterm task objective, I think it more close to clustering task(Topic Modeling), rather than a classification task. This means that detailed classification is not necessary; instead, we should focus on identifying the main clusters and analyzing their insights.

In my LDA-based analysis, I used **TF-IDF** instead of the original **BoW** representation to better reflect term importance. Therefore I used the gensim instead of tomotopy.

In my DMR-based analysis, I used the tomotopy to modeling. It still used **TF-IDF** to reflect term important.

Another concideration is whether it is necessary to assign names to the topic (Cluster Label). I think it is necessary.

Based on mathematical formulation of LDA:

Document topic distribution

$$\theta_d \sim Dirichlet(\alpha)$$

the d-th word distribution

$$z_{dn} = Dirichlet(\beta)$$

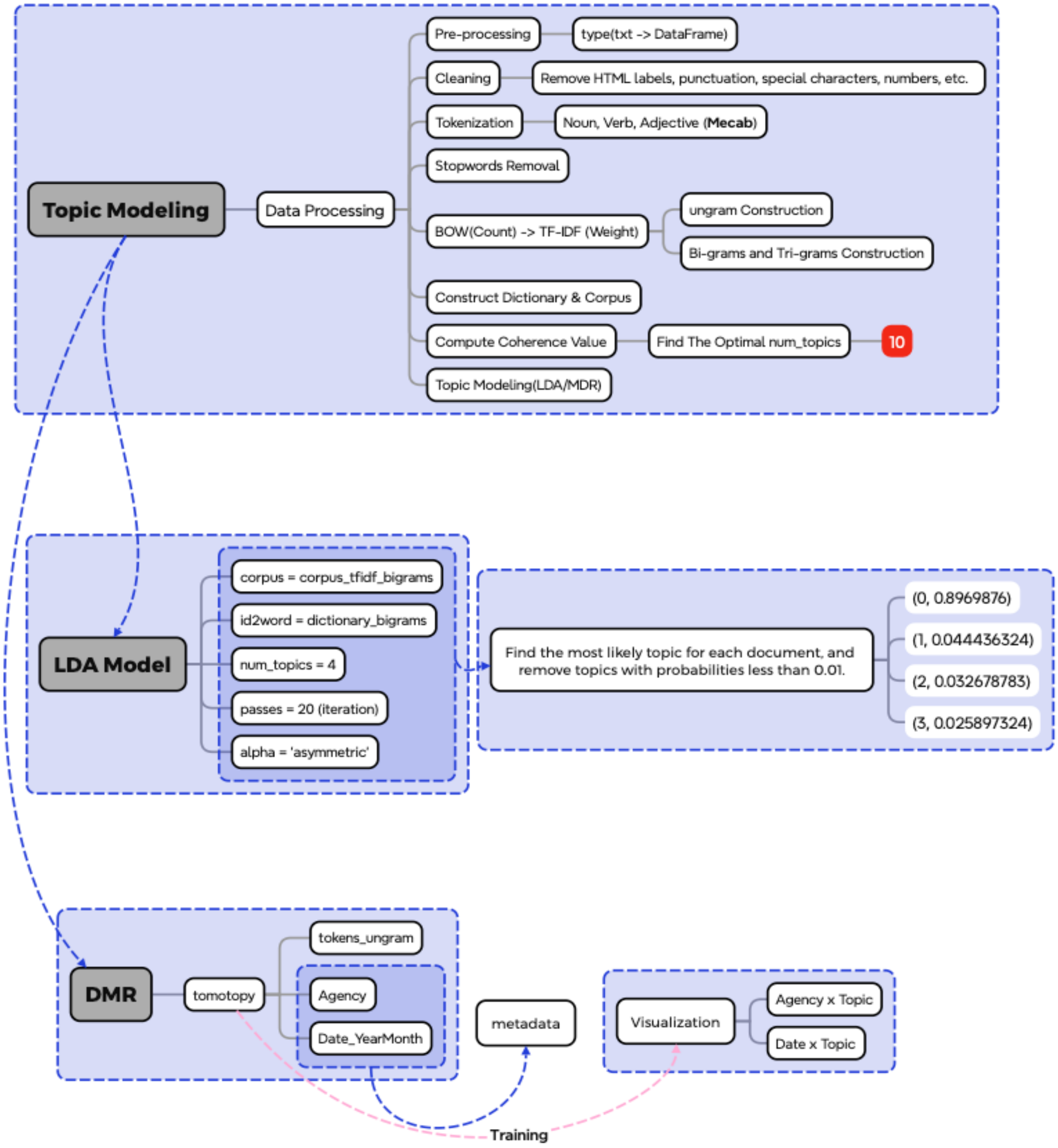
$$P(w, z, \theta | \alpha, \beta) = \prod_{d=1}^{D} P(\theta_d | \alpha) \prod_{n=1}^{N_d} P(z_{dn} | \theta_d) P(w_{dn} | z_{dn}, \beta)$$

w: set of words
z: topic assignment of all words
D: the number of documents

While we can predefine the number of topics, the actual semantic meaning (topic name) is unknown. Therefore, we need to infer the topic label based on the top keywords in each topic.

Finally, unigram, bigrams and trigrams model used in this midterm test. Since the performance differences were not much, the bigram model was selected for final analysis and presentation, Since the number of words for each document belongs to medium-to-long text (500-700 words).

## 2. Preprocessing Workflow



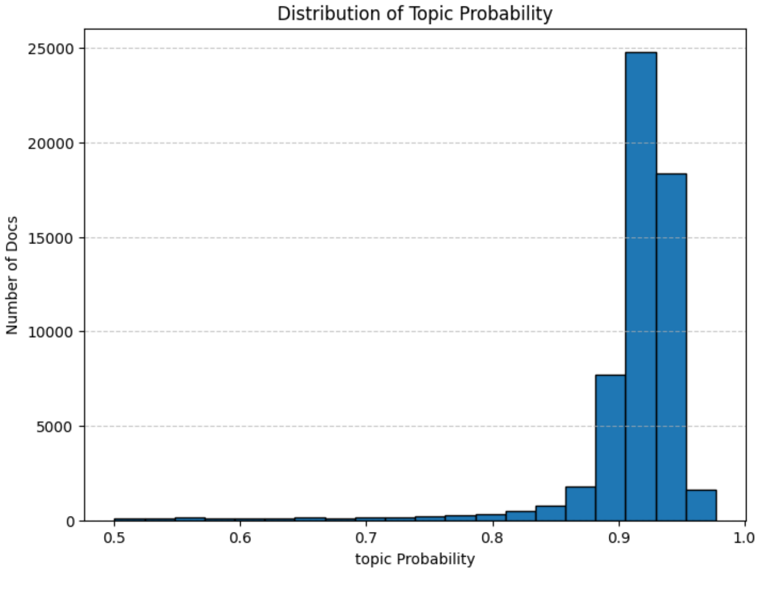
## 3. LDA Result

### Extract top keywords and assign topic labels.

| Cluster_No | Top_Keywords | Cluter_Label |
|---|---|---|
| 0 | 제품, 식약청, 관리, 업체, 위생, 지역, 소비자, 사업, 수입, 위해 | 식품,<br>제품 안전 관리 |
| 1 | 과장, 팀장, 정책, 본부, 승진, 부장, 인사, 실장, 기획, 국장 | 기관 정책,<br>조직 관리 |
| 2 | 식물_줄기세포, 유기농_수면, 해독_다이어트, 다이어트_성공, 늘어난_뱃살, 다이어트_프로그램, 일본_모스버거, 모스버거_코리아, 기존_패스트푸드, 단기_유기농 | 건강, 유기농,<br>다이어트 트렌드 |
| 3 | 남부_출장소, 충북_내수면, 암컷_생산, 부흥기_도래, 민물고기_소비, 이용_메기, 남부_어업, 성화_방지, 붕어_자어, 위해_이스라엘잉어 | 지역 수산업,<br>내수면 어업 육성 |

The top keywords of each topic(cluster) and the manually assigned topic labels through LDA.

### The degree of association of each document with its top topic.



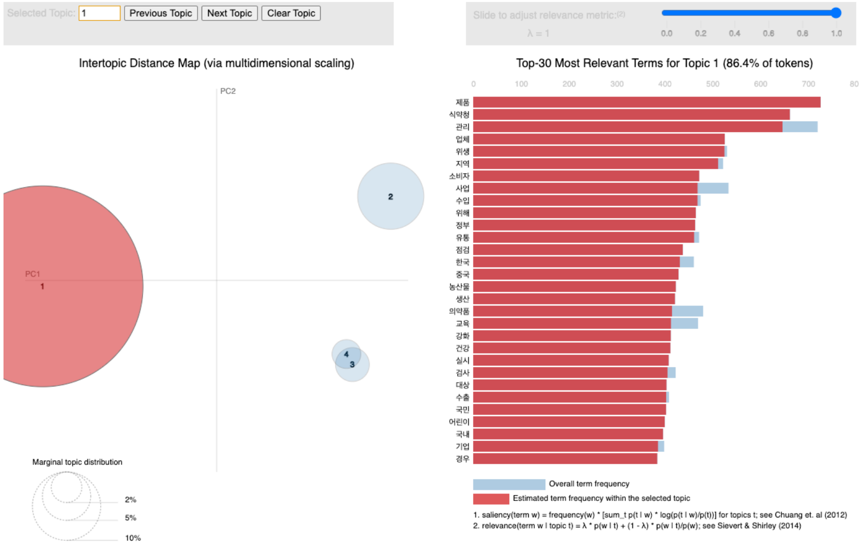
Most documents have a topic probability above 0.85, suggesting a strong confidence in topic assignment. Almost no document has a topic probability below 0.7. It means this model has strong skill of distinction.

## Determine which themes are most dominant or frequently discussed.

| Cluster_Label | Doc_Count | Percentage |
|---|---|---|
| 식품, 제품 안전 관리 | 54,167 | 94.12% |
| 기관 정책, 조직 관리 | 3,258 | 5.66% |
| 건강, 유기농, 다이어트 트렌드 | 128 | 0.22% |

This table shows the document distribution to identified topics. Most documents (94.12%) fall under the topic '식품, 제품 안전 관리', means that this dataset majorly focus on 식품, 제품 안전 관리. A few documents(5.66%) focus on 기관 정책, 조직 관리, namely policy, management or something. And few documents (0.22%) focus on healthy and diet trend.

## Visualization

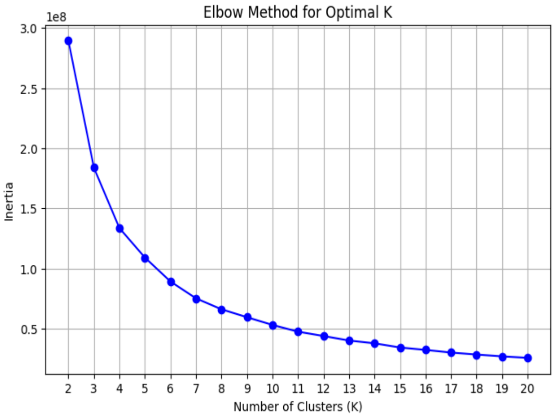

Selected Topic: 1

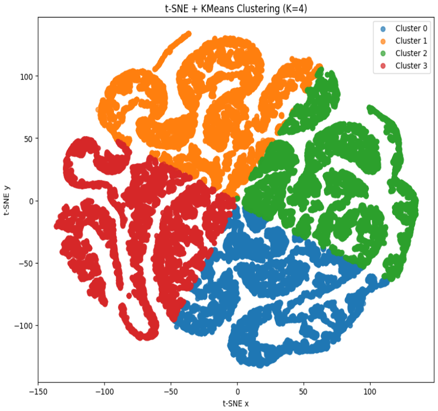
Slide to adjust relevance metric:(2)
λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

From this graph, we can see that Cluster 0 dominates in proportion. Meanwhile, Clusters 3 and 4 are positioned closely, suggesting they share some similarity. Since Cluster 3 relates to health and Cluster 4 to diet trends, it's reasonable to assume that a healthy diet contributes to overall health—hence the overlap.

# Identify the optimal number of clusters,and t-SNE + Kmeans visualization





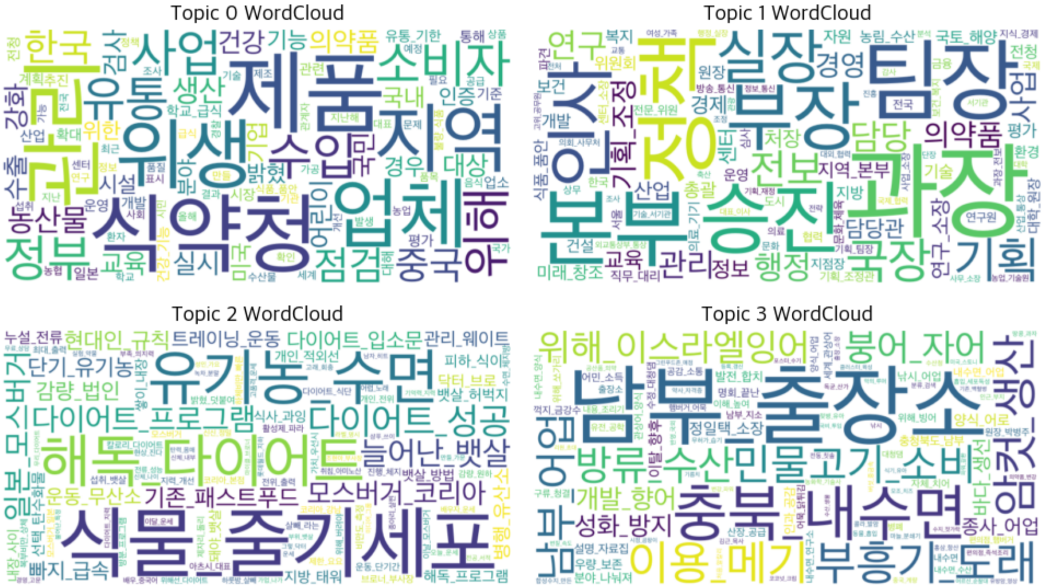
The first figure shows the Elbow Method for determining the optimal k. It shows a noticeable inflection point at k = 4, where the slope significantly decreases. It means when k increases, clustering result maybe not that good.

The second figure shows the clustering results using t-SNE for dimensionality reduction combined with K-Means clustering. The four clusters are clearly separated in the 2-Dimension.

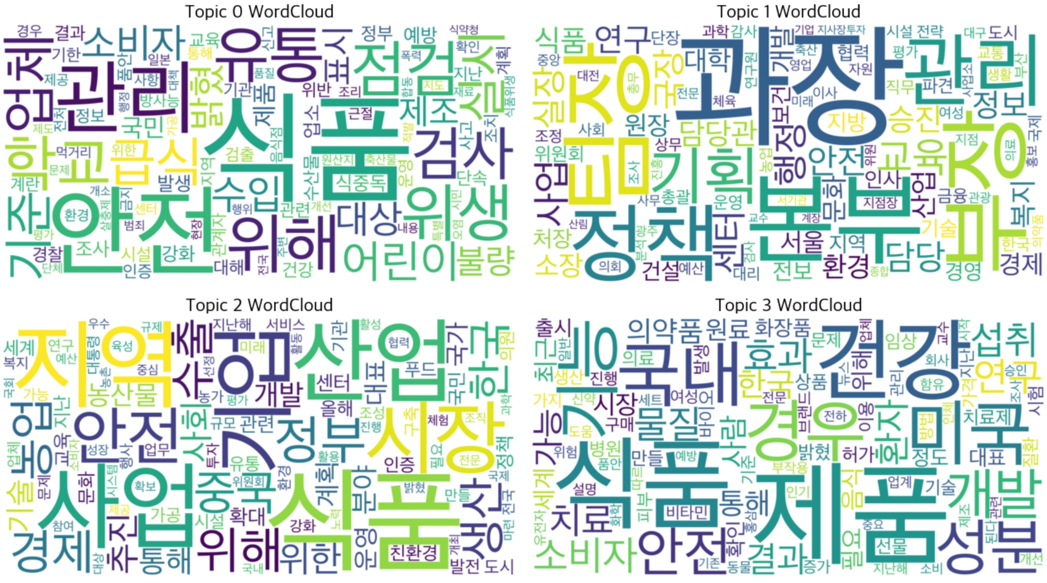
Together, these visualizations suggest that **k = 4** is an appropriate choice for clustering.

## 4. WordCloud or Top Keywords by Topic

### For WordCloud of LDA Model



### For WordCloud of MDR Model



Why are the topic of LDA/ MDR models different?

| Cluter_Label (LDA Model) | Cluster_Label (DMR Model) |
|---|---|
| 식품, 제품 안전 관리 | 식품 안전, 위생 관리 |
| 기관 정책, 조직 관리 | 기관 조직, 정책 행정 |
| 건강, 유기농, 다이어트 트렌드 | 지역 산업, 식품경제 |
| 지역 수산업, 내수면 어업 육성 | 건강 기능식품, 제품 개발 |

Topic 0 and Topic 1 show consistent results using both the LDA and DMR models, means this two topic are not sensetive to agency and date. Or both topics have some features of  agency and date.

In contrast, Topic 2 and Topic 3 differ significantly, root reason is the DMR model incorporates metadata such as agency and date, making topics sensetive to other feathers.

## 5. DMR Result

### Extract top keywords and assign topic labels.

| Topic_No | Top_Keywords (words, importance) | Cluster_Label |
|---|---|---|
| 0 | 식품(0.0378), 안전(0.0242), 관리(0.0148), 위해(0.0092), 위생(0.0088), 유통(0.0087), 점검(0.0084), 학교(0.0082), 검사(0.0074), 업체(0.0068) | 식품 안전, 위생 관리 |
| 1 | 과장(0.0244), 본부(0.0128), 팀장(0.0118), 부장(0.0109), 정책(0.0107), 관리(0.0098), 기획(0.0091), 교육(0.0089), 사업(0.0077), 행정(0.0071) | 기관 조직, 정책 행정 |
| 2 | 식품(0.0098), 지역(0.0069), 사업(0.0066), 산업(0.0063), 기업(0.0054), 시장(0.0051), 안전(0.0050), 정부(0.0050), 위해(0.0045), 경제(0.0042) | 지역 산업, 식품경제 |
| 3 | 제품(0.0128), 식품(0.0102), 건강(0.0077), 국내(0.0047), 미국(0.0047), 경우(0.0043), 기능(0.0043), 성분(0.0038), 개발(0.0037), 안전(0.0036) | 건강 기능식품, 제품 개발 |

When using the DMR model, metadata such as agencies and date enable the model to more sensitively capture the frequency and different topics across different agencies and temporal.

Continue naming the topic based on Top_Keywords by DMR Model. (Cluster_Label)

## Topic Distribution Over Time and Across Agencies

| Agency | Top_Topic | Top_Topic_Proportion |
|---|---|---|
| MBC | 3 | 0.543964 |
| OBS | 1 | 0.538462 |
| SBS | 3 | 0.425688 |
| YTN | 1 | 0.413592 |
| 강원도민일보 | 0 | 0.423788 |

**For Topic 1 - 기관 조직, 정책 행정**
OBS (53.85%), YTN(41.36%) show a strong focus on 식품 안전, 위생 관리.

**For Topic 3 - 건강 기능식품, 제품 개발**
MBC(56.04%), SBS(45.69%) show a strong focus on 건강 기능식품, 제품 개발.

**For Topic 0 - , 식품 안전, 위생 관리**
강원도민일보(46.30%) show a strong focus on 식품 안전, 위생 관리

It shows that different agencies have different reporting tendencies.

| Topic | Top_Agency | Doc_Count |
|---|---|---|
| 0 | 파이낸셜뉴스 | 1133 |
| 1 | 파이낸셜뉴스 | 1366 |
| 2 | 헤럴드경제 | 263 |
| 3 | 헤럴드경제 | 1863 |

Topic 0 and Topic 1 are most frequently reported by 파이낸셜뉴스 with 1452 and 1366 articles.
Topic 2 and Topic 3 are most frequently reported by 헤럴드경제 with 263 and 1863 articles.

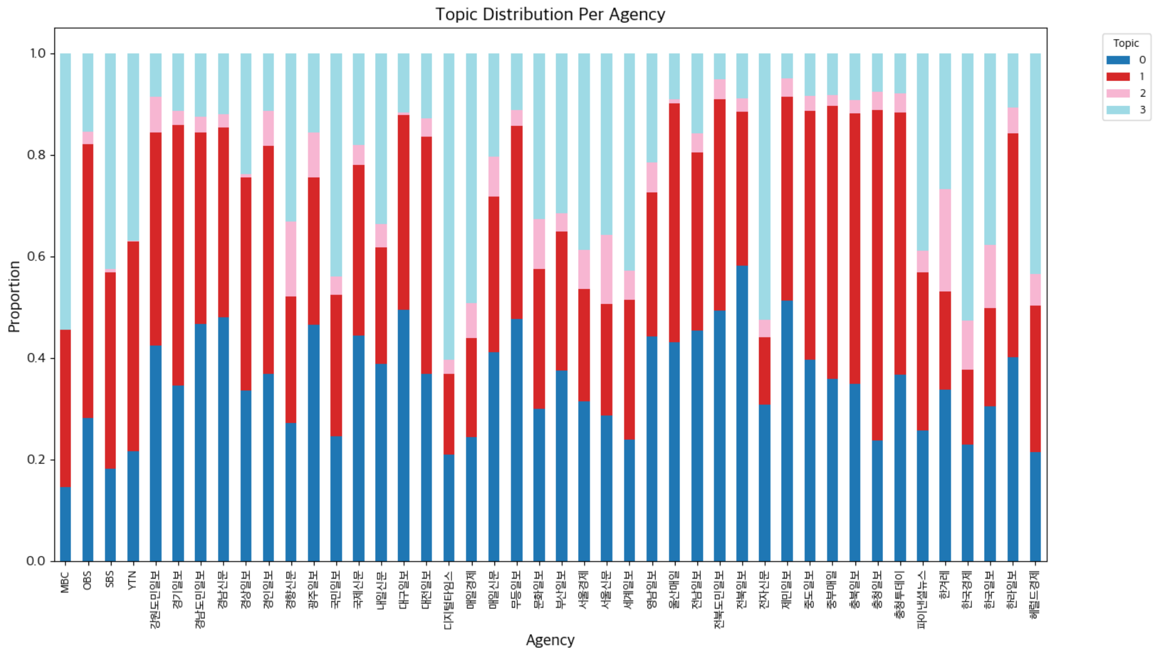
It still shows different agencies have different reporting tendencies.

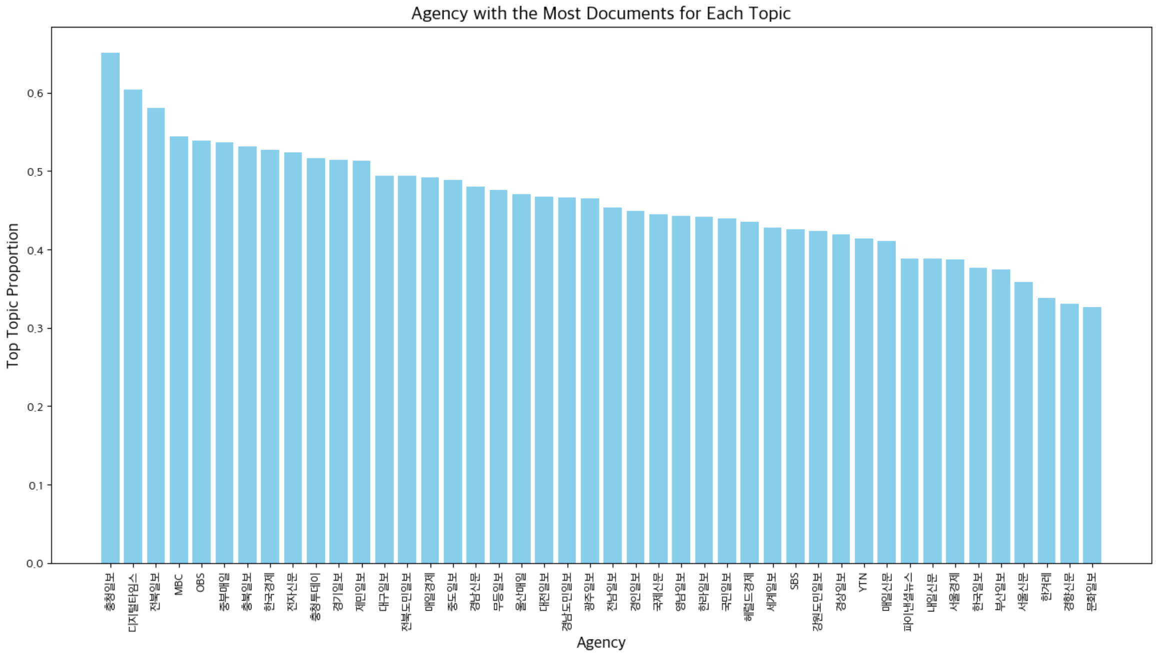
The total number: 57710

Topic_Prob > 0.7 the number of document: 40820

$$R_{Topic\text{Prob}>0.7} = \frac{40820}{57710} = 70.73\%$$

Out of 57710 documents, about 40820 of them (that's over 70.73%) have a topic probability higher than 0.7. That means for most of the documents, the model was pretty confident about which topic each one belongs to the topic.





The **first graph** shows that topic distribution per agency.

For Topic 0, 1 (blue, red)
All agencies are very concerned about Topic 0, 1 (blue, red). It means '식품 안전, 위생 관리' and '기관 조직, 정책 행정' are the a universally emphasized topic.

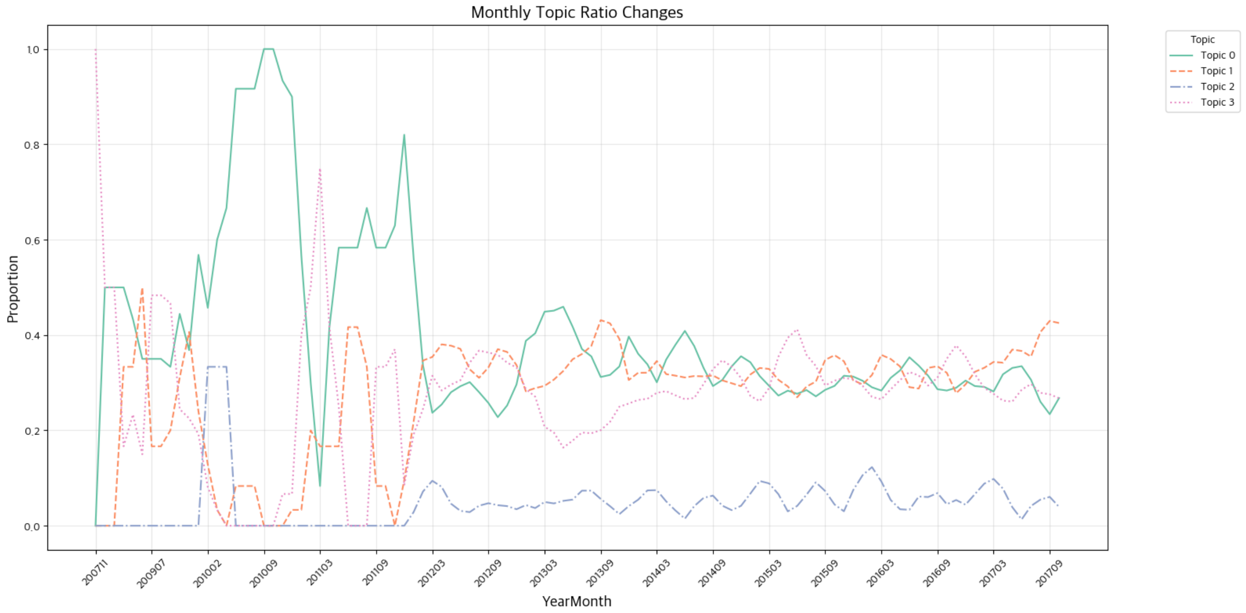
For Topic 2, 3 (pink, skyblue)
Almost all agencies do not pay attention to Topic 2 (pink), but a small portion of agencies such as '경향 신문', '한겨레' report Topic 2 (pink) frequency is high compared the other agencies

For Topic 3 (skyblue)
The proportion of agencies showing high vs low attention is roughly 1:1, indicating that agencies exhibit a clear tendency in whether or not they choose to report on Topic 3.

First graph show that agency's report have very **strongly tendency**.

The **seconed graph** shows different agencies have different reporting tendencies. Agencies like '충청일보', '디지털타임스', '전북 일보' have strong topic concentration. Agencies like '한겨레', '경향신문', '문화일보', topics covered in the report are more diverse compared to other agencies.

## Analysis of the dynamic changes in topic attention over the YearMonth.



**For Topic 0 (green) - 식품 안전, 위생 관리**
Maintaining the stable and high proportion all time. It means '식품 안전, 위생 관리' is a long-standing issue and media attention.

**For Topic 1 (orange) - 기관 조직, 정책 행정**
This graph shows that '기관 조직, 정책 행정' low attention for media attention. But during the 201002-201007, attention has increases, maybe one reason is '지역 산업, 식품경제' attention has increases.

**For Topic 2 (blue) - 지역 산업, 식품경제**
In the early year (200711 - 201203), attention is relative high, but the fluctuations are significant. It means that Topic 2 significantly affected by time, maybe has major events or policy.

Search for what events occurred in this period.
(1) 2007-2008 world food price crisis
During this period, world food prices increased dramatically, indicated the vulnerability of South Korea's food supply.

(2) 2008 US beef protest in South Korea
Public concerns over Bovine Spongiform Encephalopathy (광우병) led to widespread discussions on food safety policies and extensive media attention.

(3) Foot-and-Mouth disease during 2010-2011
A severe outbreak of foot-and-mouth disease in South Korea resulted in the mass culling of livestock and caused significant economic damage to the livestock industry. This incident heightened public attention to food safety and agricultural policies.

(4) Typhoon Bolaven (2012)
In 2012, Typhoon Bolaven struck South Korea, causing large-scale crop damage, especially to fruits such as pears and apples. This natural disaster disrupted agricultural production and impacted the food supply chain.

In Summary, Topic 0 and Topic 2 show a strong connected in MDR model , which differs from LDA result. One possible reason is that the DMR model incorporates metadata such as the publication date, which allows it to capture temporal variations in topic distribution. In contrast, the LDA model does not account for such features.

**For Topic 3 (purple) - 건강 기능식품, 제품 개발**
In the early year (200711), attention is the highest. During the 200711 – 201309, attention fluctuations are significant and then attention became stable. This means that early on, there was a high level of attention to health foods and health (with trends like those of the two curves).

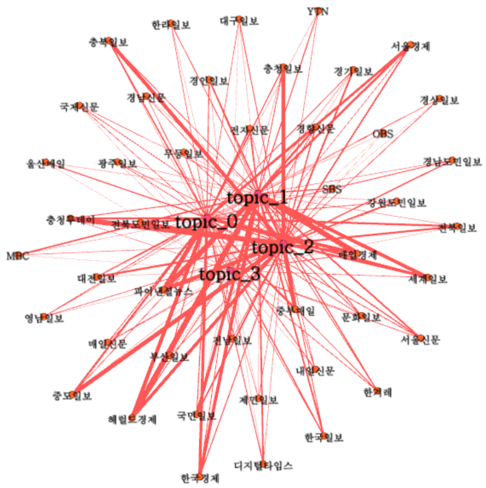
In summary, all of topics significantly effected temporal/date. As time changed, the reporting tendencies of agencies may change.

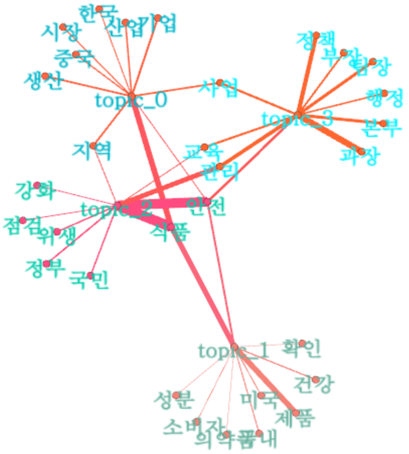
## 6. LDA vs. DMR

In summary, the DMR model outperforms the LDA model in this analysis.
By incorporating metadata such as agency and date, the DMR model captures variations in topic trends across different agencies and time periods more effectively.

## 7. Appendix

**Based on DMR Visualization with Gephi.**





The left graph shows the relationship between topics and agencies. The thickness of the edges represents the strength of association between each agency and topic, reflecting co-occurrence frequency or topic contribution strength. It shows that agencies exhibit a **tendency** in their reporting.

The right graph shows the relationship between topics and words. Through this graph can analyze different things such as '안전' is strongly related to all topics, '식품' is related to Topic 0, 1, 2.