경영공학 종합설계

# 토픽모델링을 통한 항만 디지털 잠재 기술 분석

최종 발표

만능이조

# PROCESS

## 01 연구 배경

- 항만의 중요성
- 문제 정의

## 02 선행 연구

- 특허 분석 관련 연구
- 토픽모델링 관련 연구
- GTM 관련 연구

## 03 연구 방법

- 데이터 수집 및 전처리
- BERTopic
- LDA
- GTM

## 04 연구 결과

- BERTopic
- LDA
- GTM
- 결론

# 항만의 중요성 | 연구 배경

## 물동량 급증

2022년 세계 해상무역의 양
120억 톤을 기록

2024년에서 2028년까지 2% 이상 성장
할 것으로 예측(UNCTAD, 2023)

해상 운송을 통한 물동량 급증과 동시에
항만 터미널의 중요성이 대두

## 항만 기술의 동향

최근 운영 방식을 자동화하는 스마트 항만의
발전을 넘어 인간의 사고와 판단 능력을 갖춘
지능형 항만으로 거듭나는 추세

항만 기술이 더욱 발전할 필요성 부각

지속적인 혁신과 연구를 요구

# 문제 정의 | 연구 배경

항만 특허 연구의 부족

정량적 데이터 분석 부족

# 문제 정의 | 연구 배경

항만 특허 + 토픽모델링 + GTM

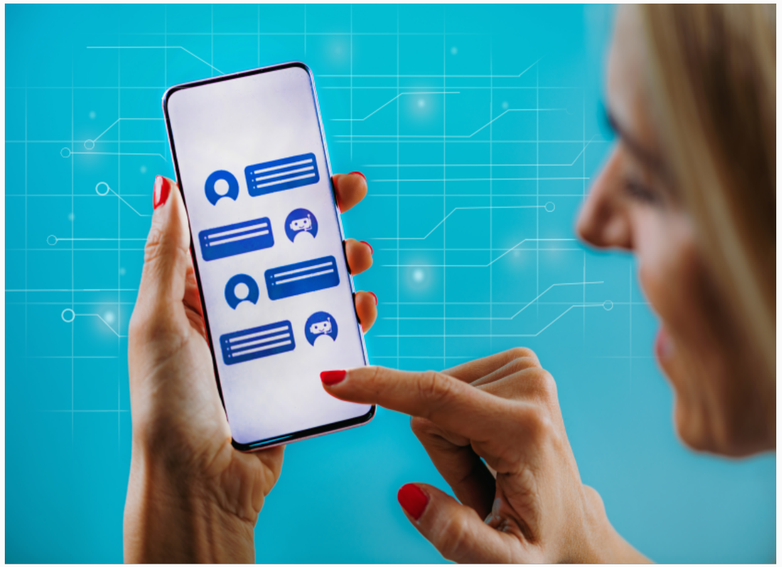
↓

미국 항만 특허 기반 데이터 분석을 통한 항만 기술 동향 분석 및 잠재 기술 도출

# 특허 분석 관련 연구 | 선행 연구

웰니스 케어

태양광



챗봇

# 토픽모델링 관련 연구 | 선행 연구

## LSA

고등 교육에서 기술 강화 학습(TEL)에 대한 연구 동향을 분석
(Chien-wen Shen & Jung-tsung Ho, 2020)

기술 기반 기업의 제품 전략 개발을 위한
신규성 탐지 통계 기법 기반 잠재적인 기술 기회 분석
(Juite Wang & Yi-Jing Chen, 2019)

# 토픽모델링 관련 연구 | 선행 연구

## LSA

**장점**

단순하고, 동의어 문제 해결

**단점**

결과에 대한 직관적인 해석이 어려움
문서별 토픽 분포 확인 불가

# 토픽모델링 관련 연구 | 선행 연구

## LDA

100명의 교육 전문가 대상 MP 기능 및 안전한 학습 환경에 대한 데이터 분석
(Alexandru Capatina et al., 2024)

지속가능한 재생 에너지 연구 분야의 주요 주제 도출
(Qiang Wang et al., 2024)

# 토픽모델링 관련 연구 | 선행 연구

## LDA

문서-주제 분포와 주제-단어 분포라는 두 가지 명확한 확률 분포 생성

예) 문서 A : '오늘 학교에서 수업을 듣고 과제를 한다.'

**문서-주제 분포** : {교육 : 70%, 일상 : 30%}
**주제-단어 분포** : 교육 - {학교 : 0.3, 수업 : 0.2, 학생 : 0.1, 과제 : 0.15, ...}
일상 - {오늘 : 0.2, 과제 : 0.05, 학교 : 0.1, ...}

# 토픽모델링 관련 연구 | 선행 연구

## LDA

**장점**

결과 해석 가능성이 높음
결과를 쉽게 이해 및 활용 가능

**단점**

하이퍼파라미터 튜닝
주제 수 사전 지정이 필요

# 토픽모델링 관련 연구 | 선행 연구

## BERTopic

인간-건물 상호 작용(HBI) 기반 4가지 연구 주제 도출
(Hakpyeong Kim et al., 2023)

독일의 소비자용 및 규제 의료앱 리뷰를 분석하여 사용자 감정과 리뷰의 주요 주제를 비교 분석
(Marie Uncovska et al., 2023)

# 토픽모델링 관련 연구 | 선행 연구

## BERTopic

**장점**

초심자가 쓰기 좋음
주제 수를 자동으로 결정
대용량 데이터를 다루기에 적합

**단점**

사전학습 모델 기반으로 세밀한 조정의 한계

# GTM 관련 연구 | 선행 연구

## GTM

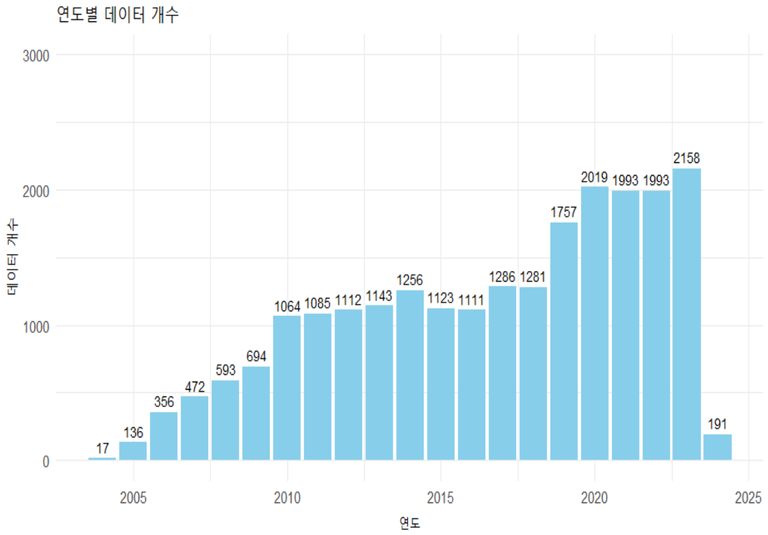
수소 공급망 유망기술 분석 및 특허 로드맵 개발
(J Yu et al., 2022)

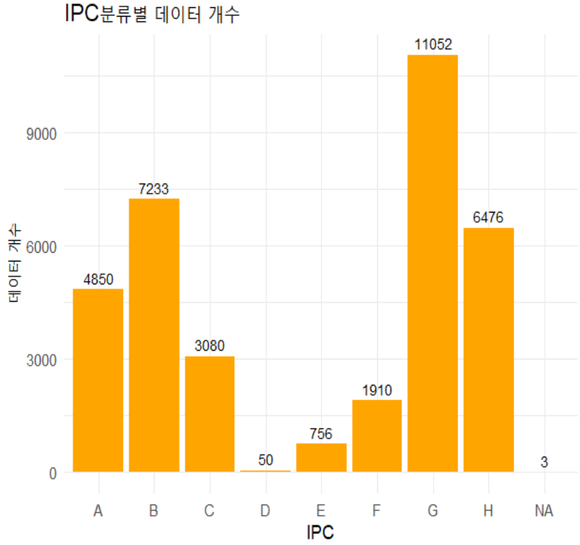
철도 주 변압기의 기술예측을 위한 특허 데이터 분석
(Y. Jeong et al., 2022)

GTM 기반의 특허 지도 개발을 통한 공백 특허 식별
(Changho Son et al., 2012)

# 데이터 수집 및 전처리 | 연구 방법

## 데이터 수집




**특허정보**

WIPSON의
2004년 ~ 2024년 1월
미국에 출원 및 등록된
항만 기술 특허들을
대상으로 추출

35,410개

# 데이터 수집 및 전처리 | 연구 방법

## 데이터 전처리

공백열 제거, 중복된 특허 제거

↓

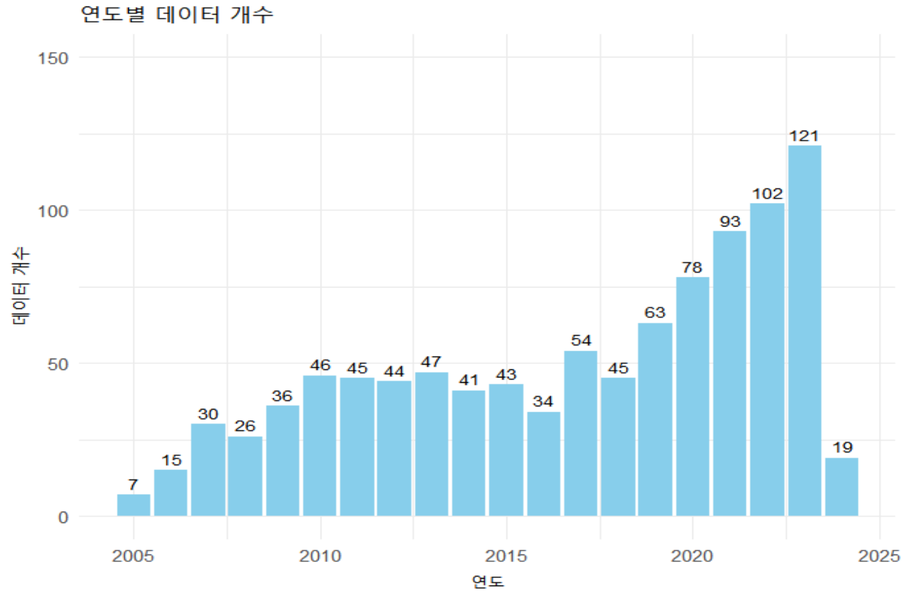
항만 디지털 관련 특허 선별

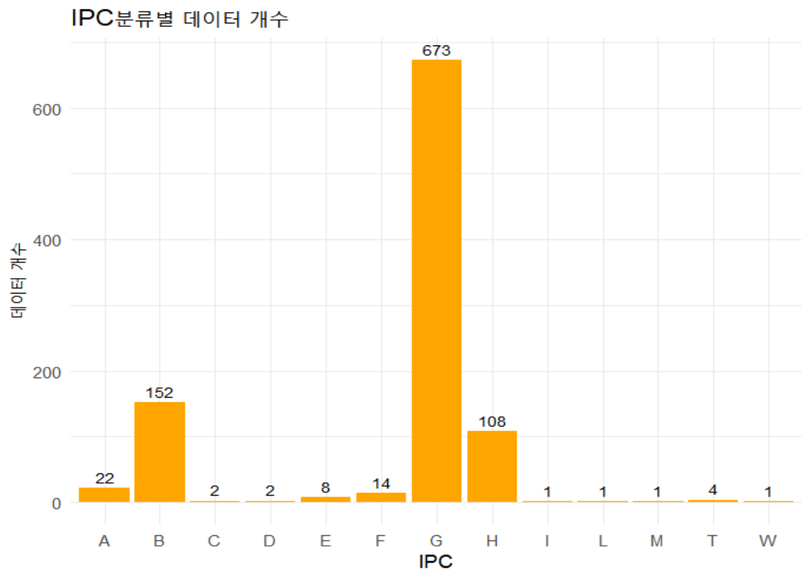
↓

최종 유효 특허 추출(989개)

# 데이터 수집 및 전처리 | 연구 방법

## 전처리 후 데이터



# 데이터 수집 및 전처리 | 연구 방법

## 텍스트 전처리

특허의 핵심 요소를 집약적으로 기술한 특허 요약(Summary)을 데이터로 사용

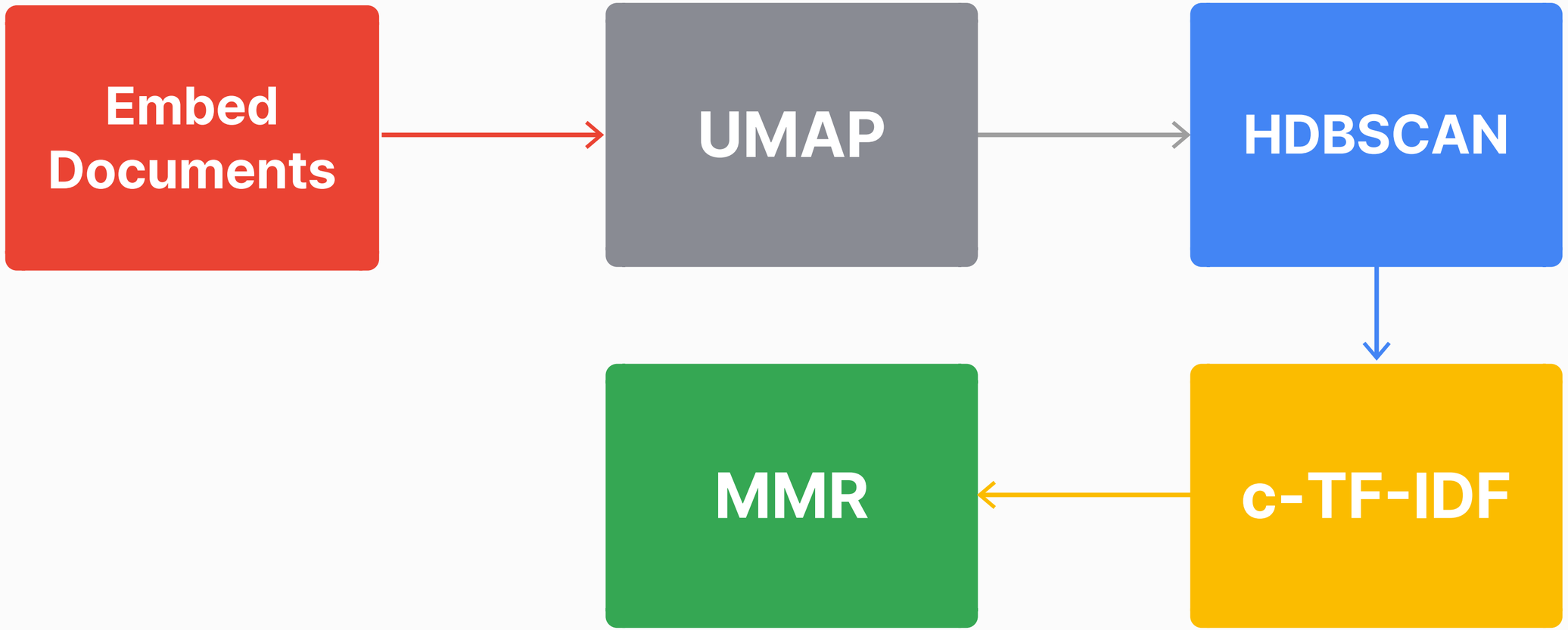
소문자 / 문장 부호 제거 / 숫자 제거 / 공백제거 / 띄워쓰기 시제 제거 / 표제어 추출 및 불용어 제거

# 토픽모델링 | 연구 방법

BERTopic

LDA

# BERTopic | 연구 방법



# BERTopic | 연구 방법

## 파라미터 설정

```
# 임베딩 모델
embedding_model = SentenceTransformer("all-MiniLM-L6-v2")
# UMAP 모델
umap_model = umap.UMAP( n_neighbors=15, n_components=2, min_dist=0.05, spread=1.0)
 #HDBSCAN 모델
hdbscan_model = hdbscan.HDBSCAN( min_cluster_size=5,  prediction_data=True, min_samples=5, alpha=1.0)
# CountVectorizer 설정
vectorizer_model = CountVectorizer(stop_words=stop_words, ngram_range=(1, 1))
# BERTopic 모델 초기화 및 훈련
topic_model = BERTopic( language="english", calculate_probabilities=True, nr_topics=7, top_n_words=10, min_topic_size=5, vectorizer_model
=vectorizer_model, embedding_model=embedding_model, umap_model=umap_model, hdbscan_model=hdbscan_model, ctfidf_model=None,
verbose=True)
```

# BERTopic | 연구 결과

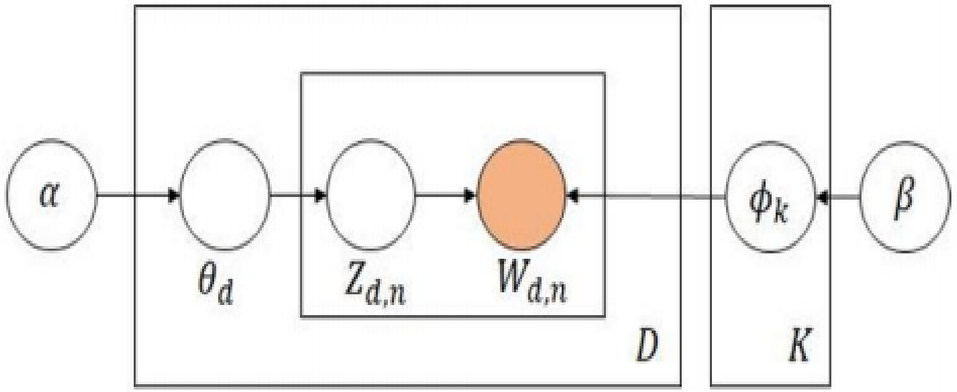
## 나온 결과 토픽

| | Topic | Count | Name |
|---|---|---|---|
| 0 | -1 | 23 | -1_tag_line_system_storage |
| 1 | 0 | 1447 | 0_system_container_include_data |
| 2 | 1 | 38 | 1_rfid_tag_reader_device |
| 3 | 2 | 19 | 2_refrigeration_temperature_control_container |

| | Topic | Count | Name |
|---|---|---|---|
| 0 | -1 | 506 | -1_system_container_method_network |
| 1 | 0 | 55 | 0_marine_vessel_underwater_ship |
| 2 | 1 | 36 | 1_portion_conveyor_rack_section |
| 3 | 2 | 35 | 2_rfid_tag_reader_antenna |
| 4 | 3 | 35 | 3_transaction_marketplace_good_sale |
| .. | ... | ... | ... |
| 64 | 63 | 6 | 63_line_drilling_traction_float |
| 65 | 64 | 5 | 64_semantic_deep_profile_enterprise |
| 66 | 65 | 5 | 65_time_base_receiver_particular |
| 67 | 66 | 5 | 66_tag_laser_contract_grid |
| 68 | 67 | 5 | 67_tilt_feedback_dolly_scan |

# LDA | 연구 방법

## LDA




$\theta_d$ : 문서 $d$에 대한 주제 분포

$\varphi_k$ : 주제 $k$에 대한 단어 분포

$\alpha$ : 디리클레 분포의 파라미터

$z_i$ : 문서 $d$의 $i$ 번째 단어에 대한 주제 할당

$\beta$ : 전체 주제에 대한 단어 분포의 파라미터

# LDA | 연구 방법

## 토픽 수

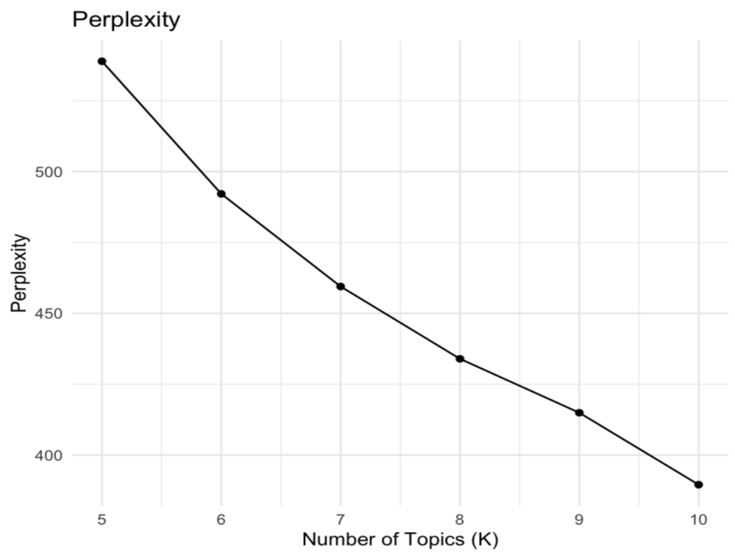
주제를 생성하기 전에 **몇 개의 주제를 생성할지**에 대한 기준이 필요

↓

언어모델의 복잡도(Perplexity) 알고리즘

↓

**최적의 토픽 수를 6으로 정함**



# LDA | 연구 방법

## LDA 코드

```
lda_model <- LDA(dtmLDA, k = 6, method = "Gibbs", control = list(
  seed = 26,
  alpha = 0.1,   # alpha 값을 0.1로 설정
  delta = 0.05,  # delta 값을 0.05로 설정
  iter = 5000,   # 반복 횟수를 5000으로 설정
  burnin = 1000, # burn-in 기간을 1000으로 설정
  thin = 500     # 샘플링 얇게 할 빈도 설정
))

lda_topics <- topics(lda_model, 1) # 각 문서에 대한 주제
lda_terms <- terms(lda_model, 9)   # 각 주제에 대한 상위 9개 단어
```

# LDA | 연구 결과

## 결과(6개)

| | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 |
|---|---|---|---|---|---|---|
| [1,] | "vessel" | "container" | "shipment" | "vehicle" | "container" | "network" |
| [2,] | "image" | "cargo" | "route" | "cargo" | "cargo" | "communication" |
| [3,] | "sensor" | "image" | "plan" | "robot" | "lock" | "wireless" |
| [4,] | "sail" | "sensor" | "package" | "autonomous" | "door" | "tag" |
| [5,] | "speed" | "apparatus" | "schedule" | "trailer" | "conveyor" | "track" |
| [6,] | "monitor" | "crane" | "logistic" | "truck" | "seal" | "sensor" |
| [7,] | "ais" | "radiation" | "cost" | "freight" | "electronic" | "monitor" |
| [8,] | "camera" | "security" | "network" | "warehouse" | "secure" | "rfid" |
| [9,] | "navigation" | "inspection" | "screen" | "unman" | "package" | "identification" |

# LDA | 연구결과

## Topic 1

"vessel", "image", "sensor", "sail", "speed", "monitor", "ais", "camera", "navigation"

System and method for monitoring vessel traffic information

System and method for collection of global AIS and radar track information

Apparatus, method and system for determining speeding of vessel based on artificial intelligence

지능형 선박 상태 모니터링 기술 (AIS, 센서, 이미지 데이터를 활용)

# LDA | 연구결과

## Topic 2

"container", "cargo", "image", "sensor", "apparatus", "crane", "radiation", "security", "inspection"

Cargo container X-ray apparatus
X-ray inspection system that integrates manifest data with imaging/detection processing
Fluoroscopic inspection method, device and storage medium for automatic classification and recognition of cargoes

AI 기반 X-ray 화물 검사 기술

# LDA | 연구결과

## Topic 3

"shipment", "route", "plan", "package", "schedule", "logistic", "cost", "network", "screen"

Application of a multi-objective optimization model for automatic vessel scheduling in liner shipping

Vector-based harbor scheduling

Ship assignment device, ship assignment method, program, and recording medium

운항 계획 최적화 기술

# LDA | 연구결과

## Topic 4

"vehicle", "cargo", "robot", "autonomous", "trailer", "truck", "freight", "warehouse", "unman"

Unmanned vehicle cargo handling and carrying system

General purpose robotics operating system with unmanned and autonomous vehicle extensions

Systems and methods for automated operation and handling of autonomous trucks and trailers hauled thereby

자율 주행 및 무인 차량 기반의 화물 운반 기술

# LDA | 연구결과

## Topic 5

"container", "cargo", "lock", "door", "conveyor", "seal", "electronic", "secure", "package"

Electronic security seal

Smart container monitoring system

Network-centric cargo security system

컨테이너의 스마트 보안 및 모니터링 기술

# LDA | 연구결과

## Topic 6

"network", "communication", "wireless", "tag", "track", "sensor", "monitor", "rfid", "identification"

Wireless data communications network system for tracking container

Tracking location and realtime management system of a container using RF

Temperature tracking and monitoring system used for commodities transportation

해상 무선통신 네트워크 플랫폼 기술

# GTM | 연구 방법

## GTM 과정

EM 알고리즘을 통한 차원 축소

↓

특허 지도 생성

↓

가우시안 함수를 통한 역매핑

↓

공백영역 파악

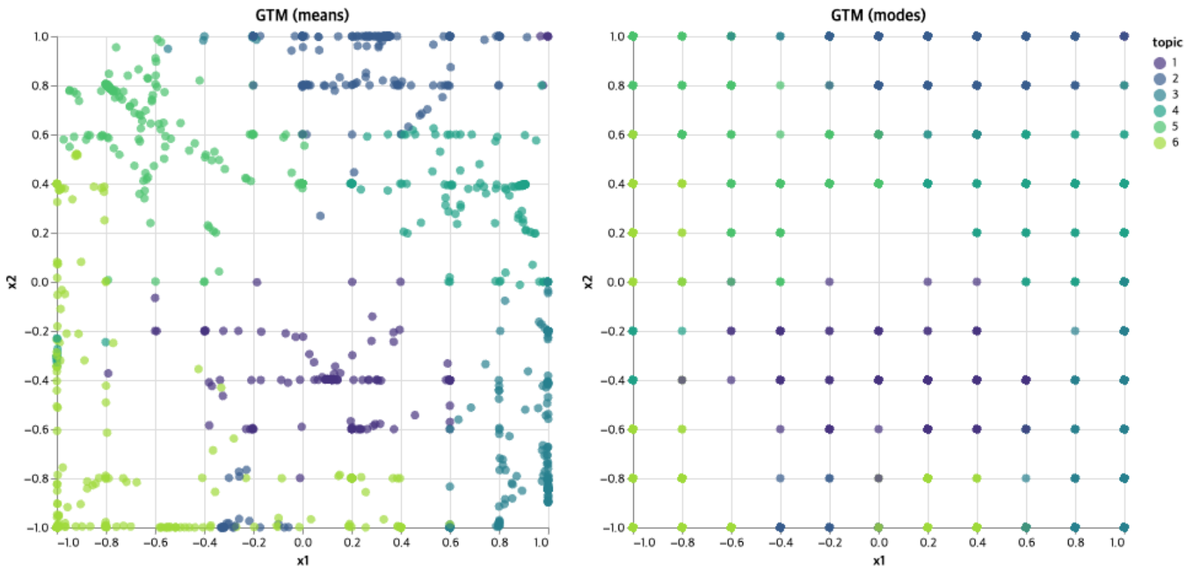
# GTM | 연구 방법

## GTM 코드

```
# CSV 파일 읽기
document_topic_assignments = pd.read_csv(path + "document_topic_assignments_k6.csv")
topic_distributions = pd.read_csv(path + "topic_distributions_k6.csv")

# 문서별 토픽 할당 추출
topic_assignments = document_topic_assignments['topic'].values

# LDA 토픽 분포 추출
topic_distributions = topic_distributions.drop(columns=['doc_id']).values

# GTM 모델 학습 및 변환
# 주어진 항만 특허 데이터의 주제 Scope를 고려하여 적합한 차원 수 11로 설정
gtm_model = eGTM(k=11, model="responsibilities").fit(topic_distributions)
gtm_means = eGTM(k=11, model="means").fit_transform(topic_distributions)
gtm_modes = eGTM(k=11, model="modes").fit_transform(topic_distributions)
```

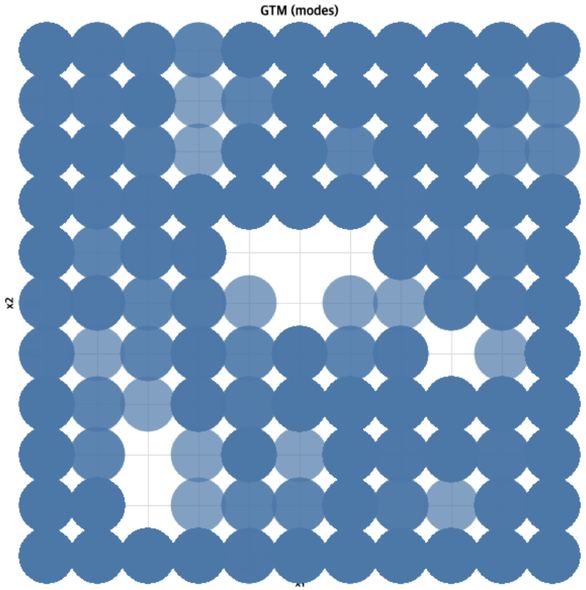
# GTM | 연구결과

## 특허 지도



# GTM | 연구결과

## 특허 지도



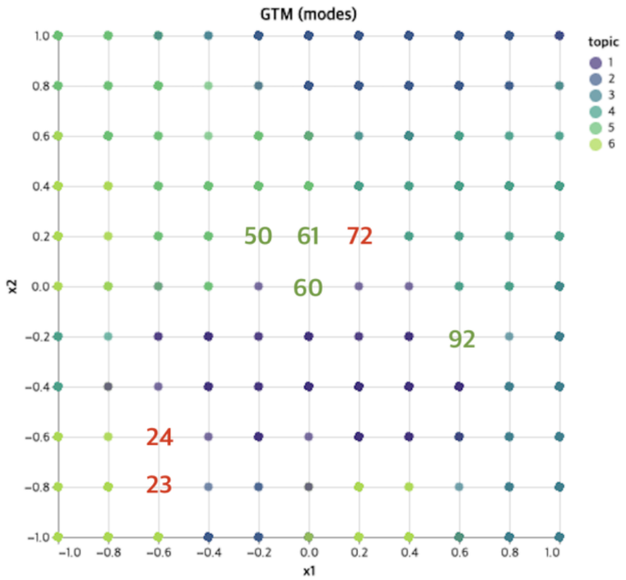
# GTM | 연구결과

## 공백 영역 식별

| | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 |
|---|---|---|---|---|---|---|
| 23 | 0.170887659 | 0.293915827 | 0.021151145 | 0.061388914 | 0.003552308 | 0.449125137 |
| 24 | 0.350367218 | 0.166781319 | 0.059173935 | 0.137995752 | 0.00532495 | 0.280374981 |
| 50 | 0.191558369 | -0.013036877 | -0.046113192 | 0.085094899 | 0.830632411 | -0.048146545 |
| 60 | 0.57199453 | -0.044883167 | -0.016225349 | 0.162134633 | 0.37658167 | -0.04962016 |
| 61 | 0.224487839 | -0.027538312 | -0.018564006 | 0.263426902 | 0.579996157 | -0.021825066 |
| 72 | 0.242324662 | -0.047645305 | 0.012244126 | 0.469411405 | 0.289724838 | 0.033925142 |
| 92 | 0.535837539 | -0.01391828 | 0.251636914 | 0.22170812 | 0.016046158 | -0.011309684 |

# GTM | 연구결과

## 공백 영역 식별




50, 61 - Topic 5

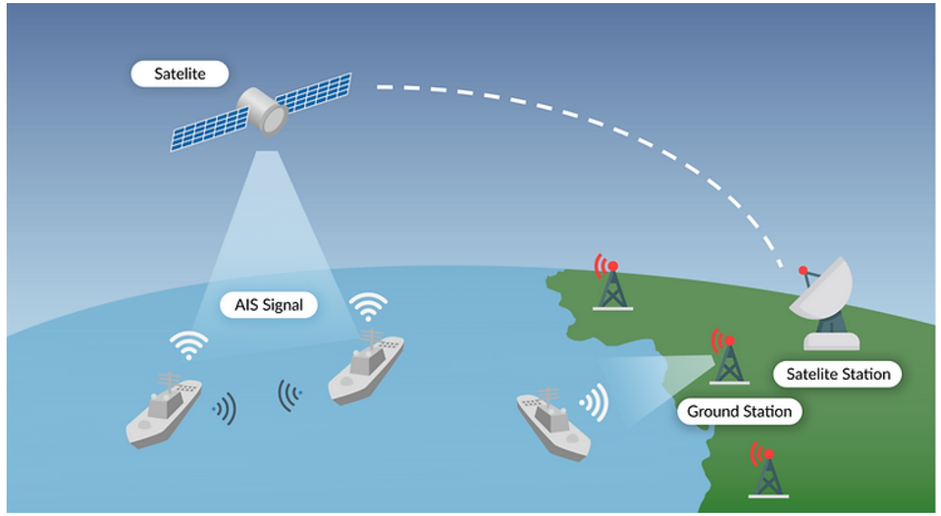
60, 92 - Topic 1

# 결론 | 연구결과

## Topic 1

"vessel", "image", "sensor", "sail",
"speed", "monitor", "ais",
"camera", "navigation"

지능형 선박 상태 모니터링 기술
(AIS, 센서, 이미지 데이터를 활용)



# 결론 | 연구결과

## Topic 5

"container", "cargo", "lock", "door", "conveyor", "seal", "electronic", "secure", "package"

컨테이너의 스마트 보안 및 모니터링 기술

경영공학 종합설계

# 감사합니다

## 최종 발표