토픽 추출

1. min topic size 키우기   
   -> 불필요한 키워드를 -1로 많이 보내게 됨
2. UMAP 내의 파라미터 중에 n\_neighbors(기본값:15)를 키우기   
   -> 불필요한 토픽이 사라짐
3. hdbscan에서 min\_cluster\_size와 min\_samples 조절

+ topic의 개수를 엄청 늘렸다가 점차 줄여나가는 방식

* 위 시도들을 먼저 해보고, 최적에 가까운 결과인 상태에서 불용어 제거  
  (처음부터 불용어 넣고 제거하지 x)
* c-TF-IDF를 적용하여 불용어를 자동으로 처리해주는 기능 작동
* 너무 공통된 키워드만 나와서 토픽의 표현이 안된다면 top n word를 늘려서 확인
* 현재 토픽의 수는 적절한 데 불용어 때문에 토픽 표현이 되지 않는다면, 토픽 수 변경 없이 불용어 제거
* 중요한 키워드들이 나오는데 그게 하나의 토픽에서 나와서 설명이 어렵다면 토픽의 수 증가 시도

**현재 데이터**

- 산업안전 특허

- 901개 데이터

- ‘청구항’열 이용

토픽 완성 후

* 토픽의 단어들을 바탕으로 특허 데이터 찾기 & 타이틀 붙이기 (토픽명)
* GTM & 공백 확인
* 역맵핑  
  -> 각 patent vacuum 별로 keyword를 inverse해서   
  -> 그 결과로 각 공백 기술 별 명칭 생성 (=>> 인사이트 부여 시 용이)

### 좋은 결과?

* 토픽의 내용이 과도하게 겹치지 않고 적당히 잘 구분
* 각 토픽의 단어들을 통합하여 기술을 특정 (기술을 잘 설명할 수 있는지)
* 각 토픽에 대한 단어의 개수, 내용이 잘 포함되었는가

<문제상황>

* n\_neighbors 값을 늘리면, 설정한 토픽 수보다 적게(2개) 추출됨  
  -> 비교적 적은 input 데이터에서, 이웃 수를 크게 잡으니 묶이게 되는(주제 생성) 크기가 늘어나고, 개수는 줄어드는 듯  
  -> min\_topic\_size값도 크게 잡아 (10) 걸러지는 단어 & 하나의 주제를 형성하기 위해 수집하는 문서(데이터) 수가 많아지는 것으로 예상  
    
  => n\_neighbors 값을 줄이니까 정상적으로 작동 (15->8로 줄임)

# Vectorizer 모델 설정

vectorizer\_model = CountVectorizer(ngram\_range=(1, 1), stop\_words=list(stopword))

# UMAP 모델 설정

umap\_model = umap.UMAP(n\_neighbors=8, n\_components=3, min\_dist=0.0, metric='cosine')

# HDBSCAN 모델 설정

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=4, min\_samples=2,  metric='euclidean', cluster\_selection\_method='eom', prediction\_data=True)

**topic = 11 (Topic -1 제외 10개 토픽)**

**텍스트, 스크린샷, 메뉴, 흑백이(가) 표시된 사진

자동 생성된 설명**

**topic = 10**

**텍스트, 스크린샷, 메뉴, 흑백이(가) 표시된 사진

자동 생성된 설명**

**topic = 9**

**텍스트, 스크린샷, 메뉴, 번호이(가) 표시된 사진

자동 생성된 설명**

**topic = 8**

**텍스트, 스크린샷, 메뉴, 흑백이(가) 표시된 사진

자동 생성된 설명**

**topic = 7**

**텍스트, 스크린샷, 흑백, 메뉴이(가) 표시된 사진

자동 생성된 설명**

**topic = 6**

**텍스트, 스크린샷, 폰트, 흑백이(가) 표시된 사진

자동 생성된 설명**

**topic = 5**

**텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명**

* 여러 경우 중 가장 최적 (최소한의 토픽 개수로 다른 기술을 내포)
* 그러나, 아직 토픽들끼리 겹치는 내용 多
* 해결을 위해, top\_n\_words 수 늘려봄 (현재 6 -> 변경 8)

텍스트, 스크린샷, 폰트, 흑백이(가) 표시된 사진

자동 생성된 설명

* 구분이 더 모호해짐 (모니터링, 생체정보(자가진단), 센서 데이터 수집/분석)
* top\_n\_words 수 늘려봄 (변경 10)

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

* 크게 결과 향상 x
* top\_n\_words 수 늘려봄 (변경 12)

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

* 큰 차이 없음
* top\_n\_words = 10 & min\_topic\_size = 12

텍스트, 스크린샷, 폰트, 흑백이(가) 표시된 사진

자동 생성된 설명

* 이것도 별로,,,
* top\_n\_words = 10 & min\_topic\_size = 14 & UMAP의 n\_neighbors=10

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

* 조금 구체화 됨
* top\_n\_words = 10 & min\_topic\_size = 14 & UMAP의 n\_neighbors=15

텍스트, 스크린샷, 폰트, 흑백이(가) 표시된 사진

자동 생성된 설명

* 결과 동일
* top\_n\_words = 10 & min\_topic\_size = 14 & UMAP의 n\_neighbors=15  
  & HDBSCAN의 min\_cluster\_size= 4 -> 5으로, min\_samples= 2 -> 3으로
* 결과 동일
* top\_n\_words = 6 & min\_topic\_size = 10 , 나머지 초기상태

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

* top\_n\_words = 10 & min\_topic\_size = 10 , 나머지 초기상태

텍스트, 스크린샷, 폰트, 흑백이(가) 표시된 사진

자동 생성된 설명

* top\_n\_words = 10 & min\_topic\_size = 10 & UMAP의 n\_neighbors=10  
  & HDBSCAN의 min\_cluster\_size= 4, min\_samples= 2

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

**topic = 4**

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=10, n\_components=5,

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=30,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=8,  # 각 주제의 상위 단어 수

    min\_topic\_size=200,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=15, n\_components=5,

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=30,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=8,  # 각 주제의 상위 단어 수

    min\_topic\_size=200,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 흑백이(가) 표시된 사진

자동 생성된 설명

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=15, n\_components=10,

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=30,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=8,  # 각 주제의 상위 단어 수

    min\_topic\_size=200,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=15, n\_components=10,

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=30,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=8,  # 각 주제의 상위 단어 수

    min\_topic\_size=100,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=20, n\_components=10,

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=10,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=8,  # 각 주제의 상위 단어 수

    min\_topic\_size=100,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=20, n\_components=10,

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=30,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=8,  # 각 주제의 상위 단어 수

    min\_topic\_size=100,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

>> 데이터, 지표 불용어 추가

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=30, n\_components=10,

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=30,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=8,  # 각 주제의 상위 단어 수

    min\_topic\_size=100,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

# 시드값 고정

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=14, n\_components=10,

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=20,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=10,  # 각 주제의 상위 단어 수

    min\_topic\_size=20,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=14, n\_components=10, min\_dist=0.2

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=20,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=10,  # 각 주제의 상위 단어 수

    min\_topic\_size=20,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

* + 누수, 단말기, 데이터 불용어

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

* + 연료, 전지 불용어

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

위 상태에서

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=12, n\_components=10, min\_dist=0.2

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=20,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=10,  # 각 주제의 상위 단어 수

    min\_topic\_size=20,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

* + 보안, 이상, 장치 불용어

텍스트, 스크린샷, 폰트, 흑백이(가) 표시된 사진

자동 생성된 설명

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=11, n\_components=10, min\_dist=0.2

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=20,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=10,  # 각 주제의 상위 단어 수

    min\_topic\_size=22,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 흑백이(가) 표시된 사진

자동 생성된 설명

* + 센서, 상태 불용어

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=10, n\_components=5, min\_dist=0.0

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=20,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=10,  # 각 주제의 상위 단어 수

    min\_topic\_size=25,  # 주제의 최소 크기

텍스트, 스크린샷, 폰트, 흑백이(가) 표시된 사진

자동 생성된 설명

* + 변전소, 감지, 연동, 이탈, 리소스, 페이퍼, 레벨, 상점, 스트레스, 레이더, 재래, 시장, 표적 불용어

텍스트, 메뉴, 스크린샷, 흑백이(가) 표시된 사진

자동 생성된 설명

👍

불용어 목록 - (수정)

umap\_model = umap.UMAP(n\_neighbors=10, n\_components=7, min\_dist=0.0

hdbscan\_model = hdbscan.HDBSCAN(min\_cluster\_size=20,

nr\_topics='auto',  # 주제의 수 제한

    top\_n\_words=10,  # 각 주제의 상위 단어 수

    min\_topic\_size=25,  # 주제의 최소 크기

텍스트, 스크린샷, 흑백, 폰트이(가) 표시된 사진

자동 생성된 설명

교수님 피드백 후

* nr\_topics = ‘auto’ 고정
* top\_m\_word = 10 고정

min\_cluster\_size = 30 & min\_topic\_size = 50

1. umap\_model = umap.UMAP(n\_neighbors=10, n\_components=7, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=8, n\_components=7, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

min\_dist = 0.5

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

min\_dist = 1.0 -> 인덱스 에러

1. umap\_model = umap.UMAP(n\_neighbors=5, n\_components=7, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=10, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=20, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=20, n\_components=15, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=20, n\_components=5, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

0 : 작업현장 모니터링 및 위험 예측분석

1 : 설비 위험 분석 & 작업자 안전교육

2 : 설비 제어 (원격제어, 자동제어)

min\_cluster\_size = 30 & min\_topic\_size = 70

1. umap\_model = umap.UMAP(n\_neighbors=10, n\_components=7, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=10, n\_components=5, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=20, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=20, n\_components=5, min\_dist=0.0

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=20, n\_components=15, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=30, n\_components=15, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=30, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

min\_cluster\_size = 30 & min\_topic\_size = 100

1. umap\_model = umap.UMAP(n\_neighbors=10, n\_components=5, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=20, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=30, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=100, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

min\_dist = 0.5 & min\_dist = 1.0

인덱스 에러

n\_neighbors = 80

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

n\_neighbors = 105

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

n\_neighbors =110

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

n\_neighbors =130, min\_cluster\_size = 50

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

min\_cluster\_size = 40 & min\_topic\_size = 50

1. umap\_model = umap.UMAP(n\_neighbors=100, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=10, n\_components=5, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=20, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=30, n\_components=5, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=50, n\_components=50, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

min\_cluster\_size = 40 & min\_topic\_size = 70

1. umap\_model = umap.UMAP(n\_neighbors=10, n\_components=5, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=20, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=100, n\_components=5, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

min\_cluster\_size = 40 & min\_topic\_size = 100

1. umap\_model = umap.UMAP(n\_neighbors=100, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=300, n\_components=10, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

1. umap\_model = umap.UMAP(n\_neighbors=10, n\_components=5, min\_dist=0.0

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명