텍스트 마이닝 및 토픽 모델링을 통한 산업 재해 사고 동향 연구

2021 대한산업공학회 | 제 17회 한국대학생 산업공학 프로젝트 경진대회

TEAM

박정선 심유경 이윤한 진승현

CONTENTS

산업 재해 현황

인구대비 산재 사고 사망자 영국의 10배?…한·영 통계상 사실로 확인
OECD 회원국 가운데 산재 1위?…1위는 아니나 '상위권'

올 상반기 중대재해 349건, 351명 사망… '후진국형 재해' 여전
정민혁 기자 | 입력 2021.07.29 10:58 | 수정 2021.07.29 11:03 | 댓글 0

7년간 산업재해 경제적 손실 159조원 추정

2022년도 **중대재해처벌법** 시행

산업재해 **사망 OECD 1위** 국가

산업재해 사고는 후진국형 사고

산업재해 은폐 비율 66.6%

산업 재해 현황의 문제점

- 다수가 누리는 이익을 위해 소수에게 희생을 떠안기는 구조 형성
- 산업의 지속 성장성에 영향을 끼침
- 산업 재해의 원인 규명 부족으로 2차 피해 예방 미흡

프로젝트 주제

텍스트 마이닝의 유사도 기반을 통한 산업 재해 사고 유형 모델링

프로젝트 목적

① 비정형 데이터를 통한 새로운 위험 요인 구체화

② 사회적 상황 반영으로 사회 및 정책의 사각지대에 있는 위험 유형에 대한 대책 제시

① 현재 산업 재해 통계의 인과관계 부재

(업종별)
건설업 사고 사망자는 485명(전체의 51.9%)로 전년 대비 30명 증가, 제조업은 201명(전체의 22.8%)로 전년 대비 5명 감소했다. 사고사망 만인율은 건설업 2.00퍼밀리어드, 제조업 0.50퍼밀리어드로 나타났다.

(규모별)
5~49인에서 402명(45.6%), 5인 미만에서 312명(35.4%)의 사고사망자가 발생하여 전년대비 각각 43명, 11명 증가했다. 사망만인율도 5~49인은 0.05퍼밀리어드p, 5인 미만은 0.04퍼밀리어드p 상승했다. 50~299인에서는 131명(14.9%), 300인 이상에서는 37명(4.2%)의 사고사망자가 발생하여 전년대비 각각 16명, 11명 감소했고, 사망만일율도 50~299인은 0.04퍼밀리어드p, 300인 이상은 0.03퍼밀리어드p 하락했다.

(1) 단순 통계치 형태

(2) 하락과 상승 이유의 부재

(3) 사고의 인과 관계 파악 불가

② 중대재해처벌법의 한계점

'중대재해처벌법' 충분한 역할을 다 할 수 있는가?

: *사업장에서 일하는 노동자가 다치거나 사망할 때 안전 조치를 소홀히 한 사업주에게 징역이나 벌금 등 강력한 형사처벌을 내리는 법*

50인 이상의 기업에는 2022년 1월부터 시행되며 50인 미만의 기업에는 2024년부터 시행될 예정

대상 사업장
- 5인 미만 사업장
- 50인 미만 사업장(3년후 시행)

처벌대상
- 안전보건 확보 의무를 다하지 않은 사업자, 경영책임자

기업

사업주, 경영책임자의 의무이행과 책임 범위 설정이 명확하지 않음.
→ 기업들의 대응을 위한 불확실성 커진 상태

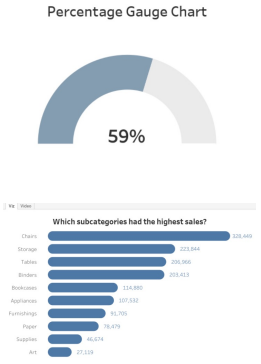
사고 발생시, 사고 원인과 인과관계를 두고 다툼 발생 가능성 다분
→ 실질적 사고 예방 위한 조치보다는 처벌 회피

근로자

일부 이익집단의 시장만 창출
→ 실질적 인명 보호, 산업재해 감소 어려움

직업성 질병 범위 확대, 위험의 외주화 등 노동시민사회가 요구한 내용 미 반영
→ 노동자들이 실제 어려움을 겪고 있는 사항에 대한 개선 불가

기존 시스템



단순 보고서 형태
+
통계 수치

사고의 인과 관계 부족으로

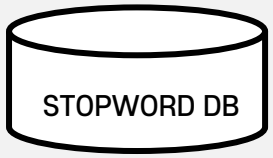
대안 마련 어려움

데이터 수집 및 전처리

텍스트 분석을 위한
데이터 정제 및 정규화 - 토큰화

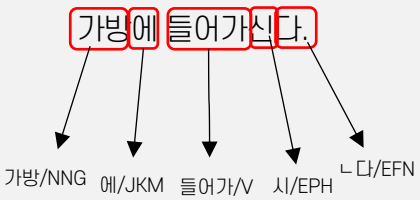
한국산업안전공단 & 산업재해사고 뉴스 크롤링

▼



불용어 제거

▼



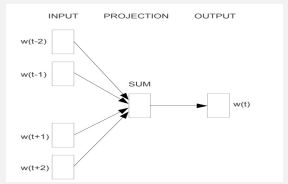
토큰화

유사도 기반 단어 벡터 산출

유사도 기반으로
단어의 벡터 값 도출

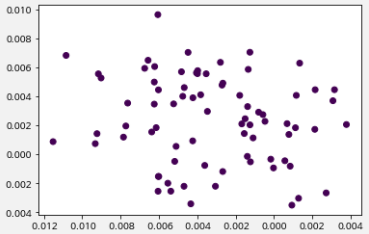
Wordcloud를 이용한 핵심 키워드 추출

▼



키워드 벡터화: Word2Vec
어휘 분석을 통한 유사도 추출

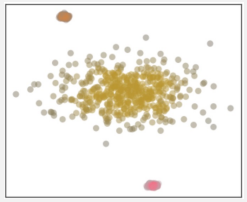
▼



벡터 차원 축소: t-SNE

각 산업별 단어 군집화

건설/화학/제조/기타업 별
단어 군집화 및 인사이트 도출



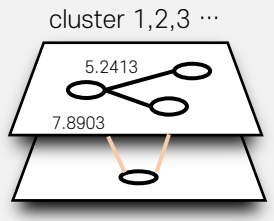
각 산업별 키워드 군집화: HDBSCAN

▼

$$DBCV(C) = \sum_{i=1}^{i=l} \frac{|C_i|}{|O|} V_c(C_i)$$

최적의 군집 수 도출 및 유효성 검증
: DBCV Score

▼



Topic Graph

사건 요인 별 위험도

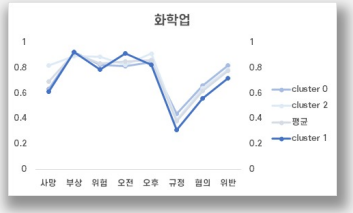
위험도 계산 및
고위험군 클러스터 선정

<특별 요인>

- 유사도
- Risk_score
- Law_score
- 오전
- 오후

수식을 통한 클러스터와 사건 요인 별 위험도 산출

▼

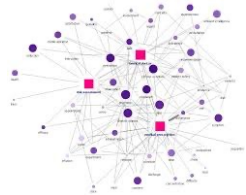


고위험군 클러스터 선정

개선 시스템

보고서
+
자연어 요인 분석
+
뉴스

Topic 추출



사고 유형별
- 피해 대상
- 법의 테두리/사각지대
- 국민적 공분
- ...

반영한 대책 제안

2 데이터 수집 및 전처리

❶ 특수문자/개행문자/영문/숫자 제거

[?.,;:|₩)*~`'!^₩-_+<>@₩#$%&-=#}※][a-zA-Z0-9]

정규표현식을 이용해 특수/개행/영문/숫자를
제거하는 함수를 만들어 적용

❷ 문장 분리 & 띄어쓰기

문장 분리: kss

정교한 패턴 기반의 문장 분리기
통계 기반에 비해 월등히 뛰어난 속도

띄어쓰기: PyKoSpacing

대용량 corpus를 학습하여 만들어진
딥러닝 모델로 뛰어난 성능 가짐

문장 분리: kss(korean sentence splitter) → 올바른 문장 간 분리를 시행

안전한 사업장을 만들기 위한 사업주의 … → 문장 내 띄어쓰기를 적용

❸ 유의미어 선별: 복합어 인식

형태소 분석기 Konlpy → Kkma

Kkma: 건설/업계/의/불만/이/팽배/하/고/이/ㅆ/다/.

Okt: 건설/업계/의/불만이/팽배/하고/있다/.

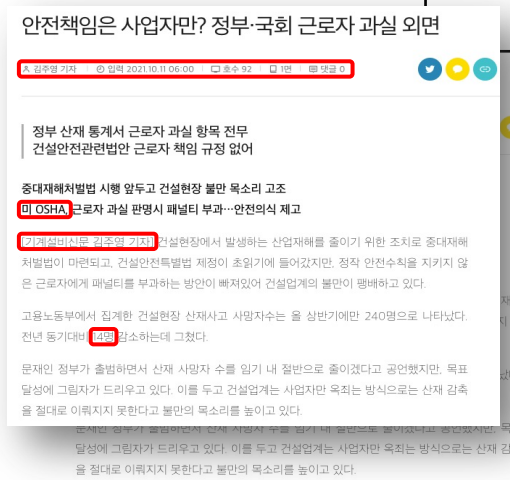
→ 속도는 느리지만 **Okt** 기준 품사 태그 비해 지원하는 품사 태그가 다양함

불용어 제거 프로세스

ID	WORD
1	기자
	입력
	…

STOPWORD DB

제거



키워드 추출

" 안전한 사업장을 만들기 위한 사업주의 투자와 노력도 중요하지만
작업장에서 스스로를 지키기 위한 근로자들의 의식 개선도 필수적이다."

1 형태소 분석기: Konlpy - KKMA

POS	WORD
Nouns	안전/사업/장/투자 …
Josa	한/을/위한/와 …
Verb	만들기/지키기 …

유의어 선별기 2

3 복합어 인식기

안전 사업장 사업주 투자 노력 작업장
스스로 근로자 의식 개선 필수

3 모델 설명 및 시각화 | (1) WordCloud

각 산업별 빈도 수 상위 100개 단어에 대한

WORD CLOUD

건설업	화학업	제조업	기타업
건설	폭발	작업	청소
사고	가스	설비	소속
위협	주민	지게차	차량
…	…	…	…

4개 산업별 사고 단어 분포가 다르기 때문에

산업별 모델링의 필요성을 확인

Tunning - Word2Vec

“ 4개의 모델로 업종별 특징 최적화 ”

Pretrained Model

: 사고의 특징 반영 불가

Fine Tunning Model

: 업종별 특징 세분화 불가

✓ **Tunning Model**

건설업 화학업

서비스업 제조업

Word2Vec

Parameter 설정

100차원으로 one-hot encoding 방식으로 embedding

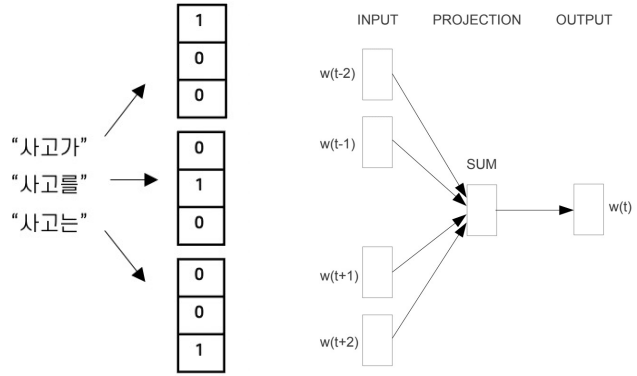
- size=100차원의 embedding vector)
- Window_size = 5
- Min_count = 5
- CBOW (크기가 작은 것에 적합)

❶ Tokenized sentence

1	'사고', '당시', '현장', '전', '관리자', '이씨', '씨', '안전', '모도', '지급'
2	'억', '미만', '소규모', '민간', '건설', '건설공', '공', '사장', '목숨', '조사'
3	'작업자', '개인', '개인보호구', '보호구', '미', '미착용', '착용', '비율', '건설업'
4	'안전', '안전시설물', '시설물', '불량', '보호구', '미', '미착용', '착용'
...	...

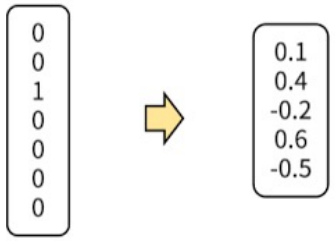
* Kkma를 이용한 형태소 분석 및 토큰화

❷ One hot encoding



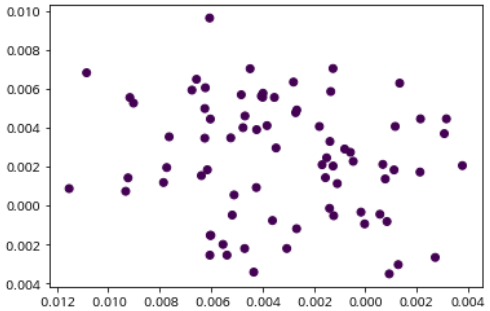
CBOW 적용

❸ Word Embedding

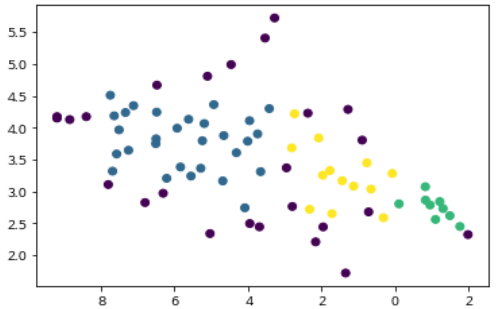


One-hot encoding 벡터
100차원의 분산 표현

차원 축소: t-SNE



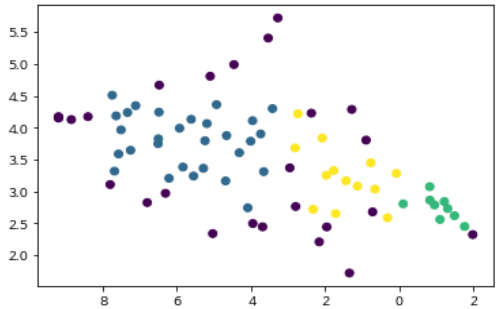
차원 축소가 없다면,
Clustering 불가능



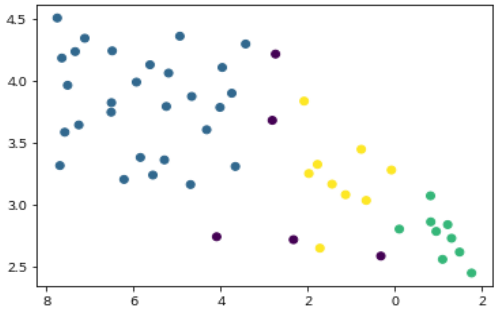
차원 축소 시,
Cluster 형성

* PCA는 고차원에 부적합 →t-SNE 이용하여 차원 축소

전처리: 이상치 제거

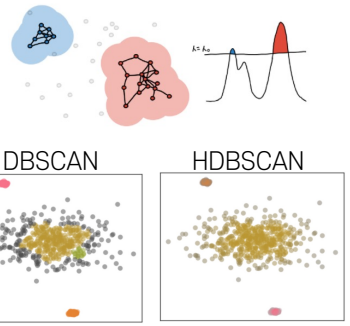


이상치 확인 가능



이상치 제거 후,
적은 이상치 확인

3 모델 설명 및 시각화 | (3) clustering



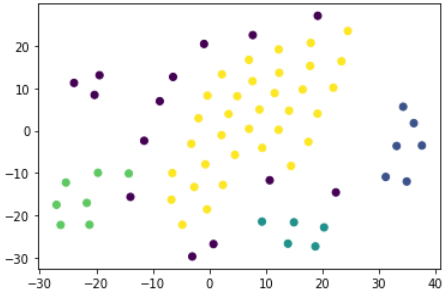
Variable dense 반영이 더 잘됨.

ε(클러스터포함거리) 파라미터는 필요하지 않으며 MinPts(최소이웃수)만 존재,

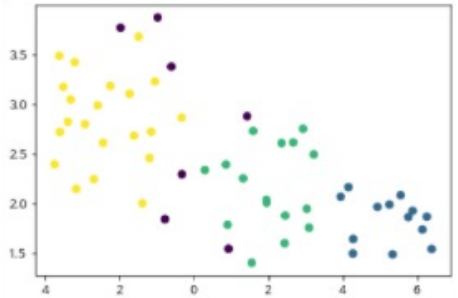
→ hyper-parameter에 대한 tuning 비용이 상당히 감소.

HDBSCAN

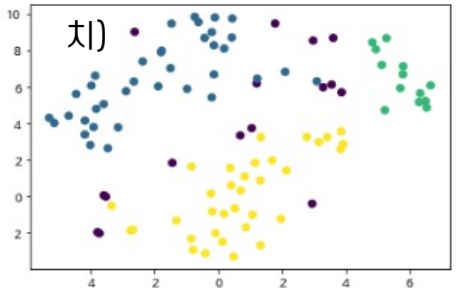
건설업, k = 3+1(이상치)



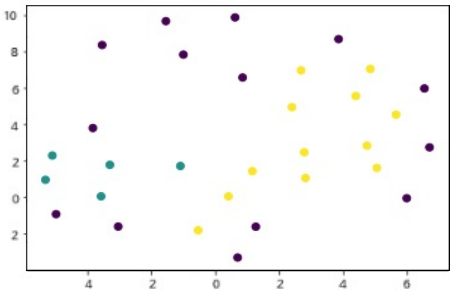
화학업, k = 3+1(이상



제조업, k = 4+1(이상치)



기타업, k = 2+1(이상



산업별 **다양한 유형**으로 사고가 발생한다는 점 고려

- ✓ 유형별 노출 키워드 **모여 있음**을 발견
- ✓ 데이터셋의 **밀도**를 기반으로 스스로 군집화

$$DBCV(C) = \sum_{i=1}^{i=l} \frac{|C_i|}{|O|} V_c(C_i)$$

$$C = \{C_i\}, 1, 1 \le i \le l$$

is defined as the weighted averaged of the Validity Index of all clusters in C

- ✓ 클러스터의 밀도와 속성 모양을 직접 고려
- ✓ 노이즈를 처리

→ silhouette, DaviesBouldin 에서는 고려되지 않음

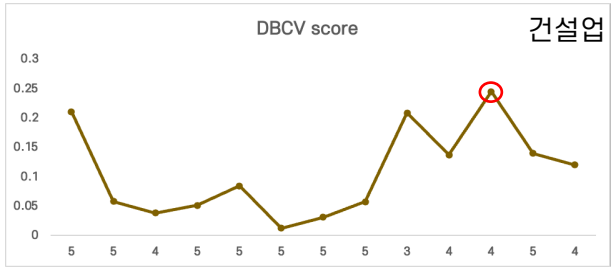
→ 노이즈 접근 방식은 클러스터링 검증에 대한 효율성 & 적용가능성을 확인

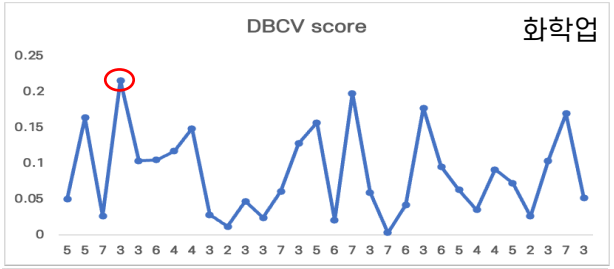
DBCV Score

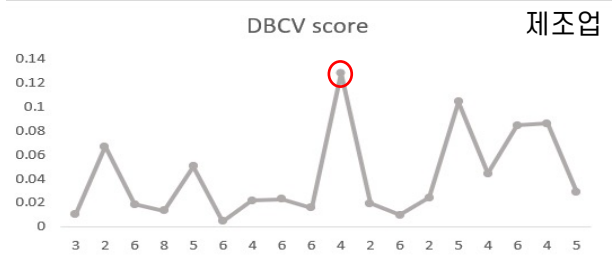
각 클러스터에 대한 DBCV score 계산

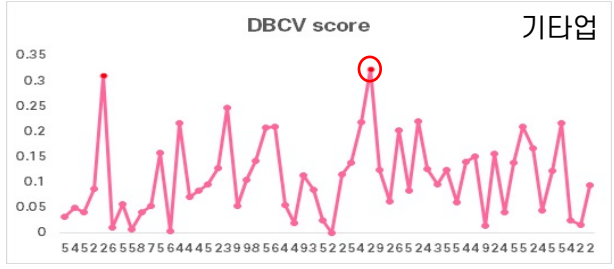
T-SNE 차원 축소 결과의 **변동성** 고려

→ **DBCV 값**이 가장 높았던 Cluster 개수를 **k**로 결정





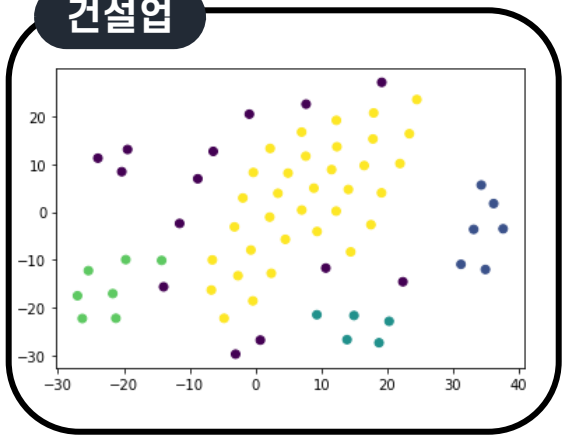


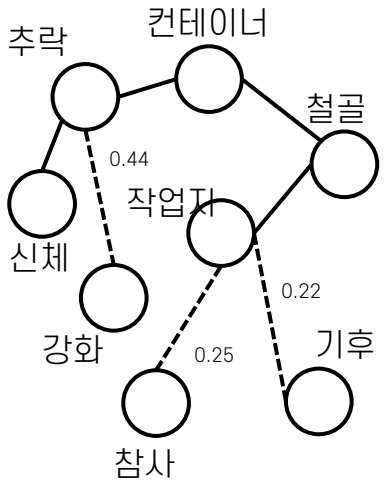


4 결과 해석 | (1) 산업별 클러스터 집합 해석

빈도 수 상위 100개 Cluster 진행 ▶ 각 클러스터 별 cosine similarity 계산 후, 유사도의 평균을 임계 값으로 설정하고 임계 값 이하의 단어 제거 & 유의미한 단어 추출 → 각 클러스터 별 Topic Graph 형성 ▶ 최종적으로 군집화 된 키워드 결과

건설업





1.00				
0.48	1.00			
0.88	0.78	1.00		
0.44	0.71	0.59	1.00	
0.25	0.22	0.66	0.69	1.00

cluster 0

["근로자/여성/외국인/남성" 등 "작업자"들은 "철골/컨테이너" 등에서 "신체/추락" 등의 사고 유형]

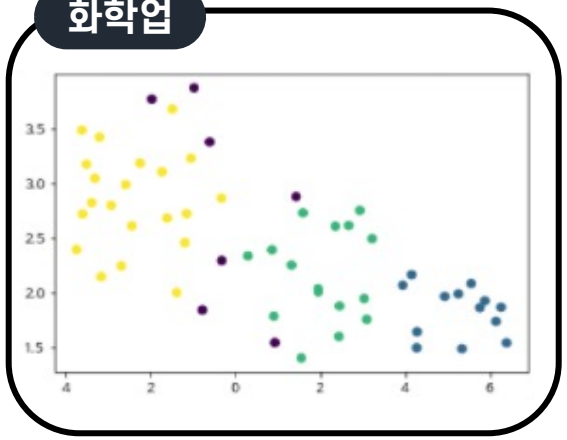
cluster 1

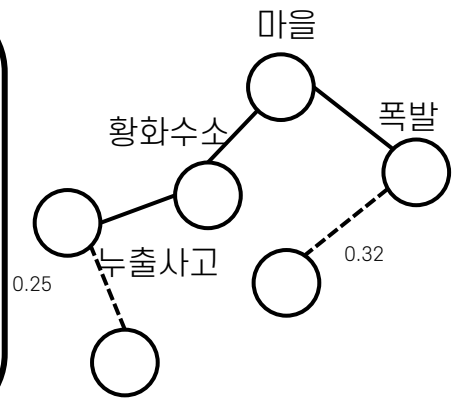
["폭염","열사병","집중호우" 등 기후이변으로 인한 "고령"작업자의사고유형]

cluster 2

["이윤"을 위한 "중소"기업의 "하도급" 으로 인한 사고유형]

화학업





1.00				
0.48	1.00			
0.38	0.12	1.00		
0.44	0.32	0.25	1.00	
0.25	0.22	0.45	0.32	1.00

cluster 0

["폭발"로 "마을"에 "황화수소"등의 "누출사고"의 위험이 큰 사고 유형]

cluster 1

["가스" "유출"로 "염산", "메틸", "스티렌" 등으로 "주민"들이 피해를 본 사고 유형]

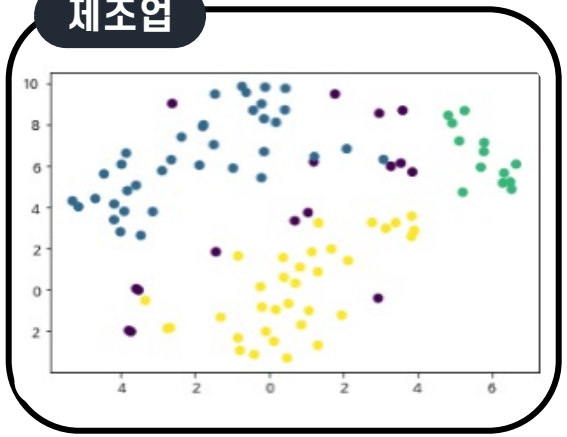
cluster 2

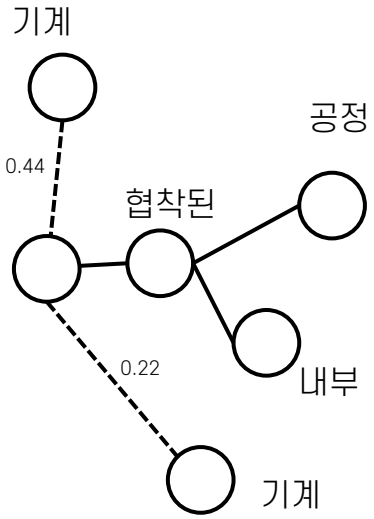
["근로자"또는 "노동자"가 "암모늄" ,"폭발사고"에 의해 "사망"한 사고 유형]

4 결과 해석 | (2) 산업별 클러스터 집합 해석

빈도 수 상위 100개 Cluster 진행 ▶ 각 클러스터 별 cosine similarity 계산 후, 유사도의 평균을 임계 값으로 설정하고 임계 값 이하의 단어 제거 & 유의미한 단어 추출 → 각 클러스터 별 Topic Graph 형성 ▶ 최종적으로 군집화 된 키워드 결과

제조업





1.00				
0.48	1.00			
0.38	0.58	1.00		
0.44	0.51	0.59	1.00	
0.25	0.22	0.66	0.69	1.00

cluster 0

["공정", "내부", "사이"에 "협착 된" 사고 유형]

cluster 1

["리프트", "승강기" 의 높은 "높이"에서 일하다가 "추락"한 사고 유형]

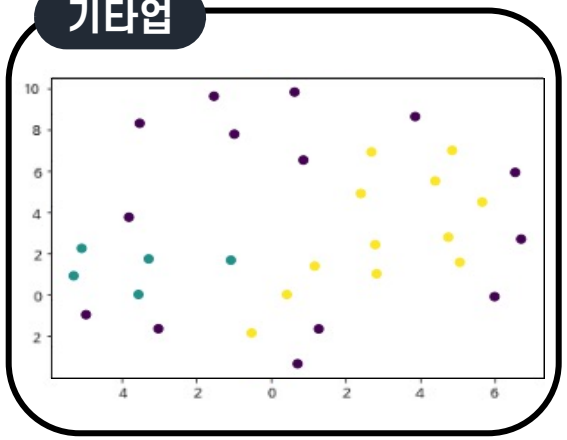
cluster 2

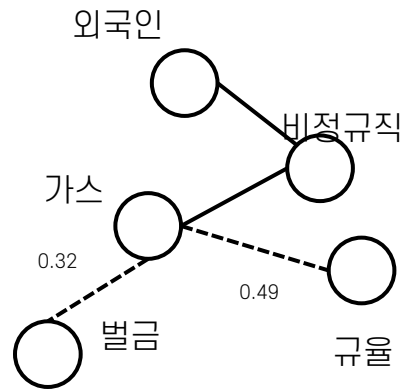
["설비"를 "정비"하는 과정에서 "기계"가 "작동"한 사고 유형]

cluster 3

["폭발"로 "화재" 가 발생한 사고 유형]

기타업





1.00				
0.48	1.00			
0.28	0.58	1.00		
0.14	0.51	0.49	1.00	
0.75	0.32	0.46	0.59	1.00

cluster 0

["외국인"과 "비정규직"을 주로 사용하는 해양에서의 "가스" 유출로 인한 사고 유형]

cluster 1

["위탁", "소속"에서 관리하는 "경비원"과 "미화원"의 "차량"에 의한 사고유형]

위험도 관계식

변수 정의 : k = 클러스터

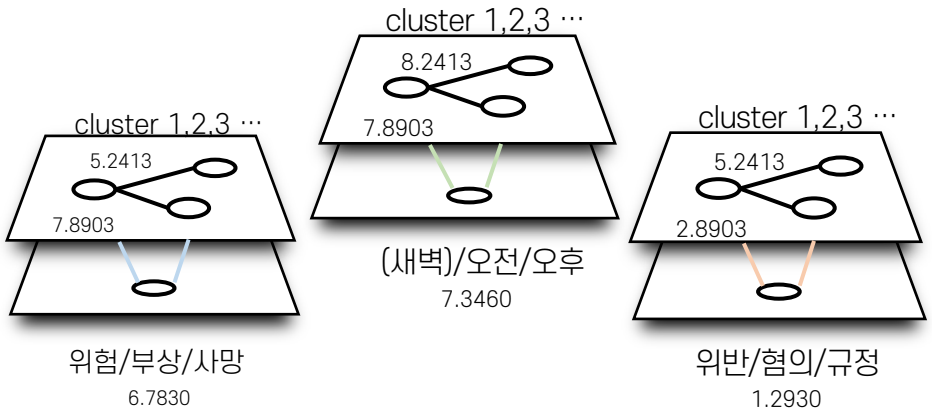
유사도 $f = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} Ai*Bi}{\sqrt{\sum_{i=1}^{n}(Ai)^2}*\sqrt{\sum_{i=1}^{n}(Bi)^2}}$

Risk_score = $\frac{\sum_{k}^{a}\{f_k(\text{사망})+(f_k(\text{부상})+f_k(\text{위험}))*0.5\}}{a}$

Law_score = $\frac{\sum_{k}^{a}\{f_k(\text{혐의})+f_k(\text{위반})+f_k(\text{규정})\}}{a}$

오전 = $\frac{\sum_{k}^{a} f_k(\text{오전})}{a}$

오후 = $\frac{\sum_{k}^{a} f_k(\text{오전})}{a}$



유사도 기반으로 묶인 단어 군집들에 대해,

건설업	사망	부상	위험	오전	오후	혐의	위반	규정	Risk_ score	Law_ score
0	0.675	0.916	0.867	0.915	0.918	0.916	0.915	0.915	0.522	0.915
1	0.665	0.882	0.829	0.878	0.880	0.880	0.881	0.880	0.507	0.880
2	0.653	0.895	0.869	0.898	0.897	0.894	0.897	0.896	0.512	0.896

화학업	사망	부상	위험	오전	오후	혐의	위반	규정	Risk_ score	Law_ score
0	0.635	0.926	0.827	0.815	0.848	0.660	0.820	0.440	0.504	0.640
1	0.675	0.872	0.829	0.788	0.699	0.560	0.720	0.310	0.509	0.530
2	0.823	0.895	0.889	0.828	0.917	0.640	0.790	0.390	0.572	0.607

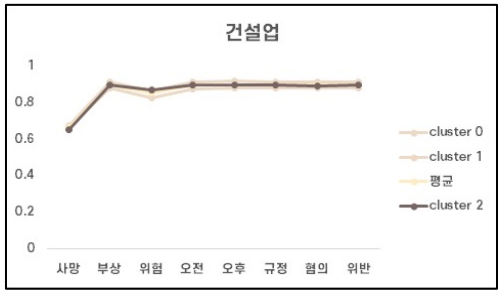
제조업	사망	부상	위험	오전	오후	혐의	위반	규정	Risk_ score	Law_ score
0	0.734	0.719	0.395	0.712	0.716	0.092	0.558	0.701	0.430	0.450
1	0.509	0.452	0.598	0.679	0.735	0.221	0.771	0.769	0.345	0.587
2	0.239	0.211	0.763	0.572	0.658	0.452	0.843	0.689	0.242	0.661
3	0.502	0.560	0.534	0.627	0.669	0.225	0.604	0.638	0.350	0.489

기타업	사망	부상	위험	오전	오후	혐의	위반	규정	Risk_ score	Law_ score
0	0.425	0.365	0.238	0.344	0.315	0.237	0.277	0.132	0.242	0.215
1	0.403	0.355	0.213	0.328	0.263	0.195	0.250	0.116	0.229	0.187

4 결과 해석 | 산업별 Topic Modeling 그래프

Cluster 2

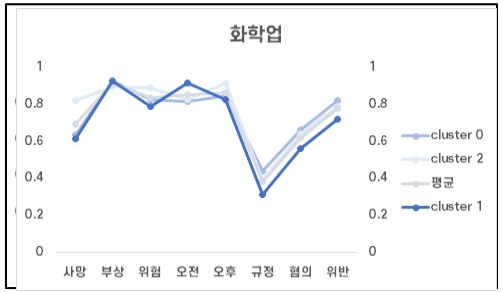
이윤"을 위한 "중소"
기업의 "하도급" 으로
인한 사고유형



건설업의 위험도(사망,부상,위험)값은 모든 클러스터에서 높다.
Cluster2는, 규정/협의/위반 수치가 높아 지속적인 관리 감독이 필요한 사고 유형으로 볼 수 있다.
건설업의 산재 사고는 불법 하청으로 인한 사건들이 많다. 실제 불법 하도급 금지 이슈로 떠오르고 있는 사건 중 하나가 광주 철거 건물 붕괴 참사이다. 이 사고는 다단계 하도급 계약 관행으로 다른 업체에서 공사를 맡았다가 대형 참사를 낸 것으로 보고있다.

Cluster 1

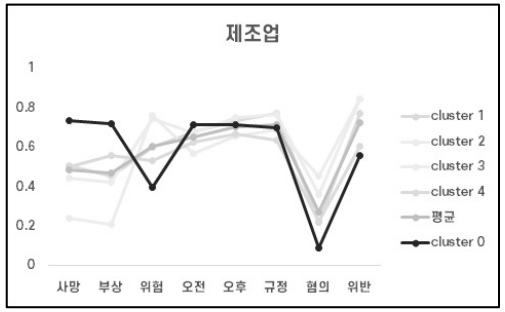
"가스" "유출"로 "염산", "메틸",
"스티렌" 등으로 "주민"들이
피해를 본 사고 유형



Cluster1의 토픽은 "주민", "가스", "유출" 등의 사고 유형으로, 사망, 위험 스코어는 낮은 것으로 간접적인 피해가 주를 이룬다. 또한 오전의 값이 상대적으로 오후 값 보다 높게 나온 것은, 근무 외 시간에 가스가 유출되는 경우가 많은 것으로 생각되고, 따라서 주민들의 피해에 대한 정책 또는 보상이 체계화가 필요한 사고유형으로 보인다

Cluster 0

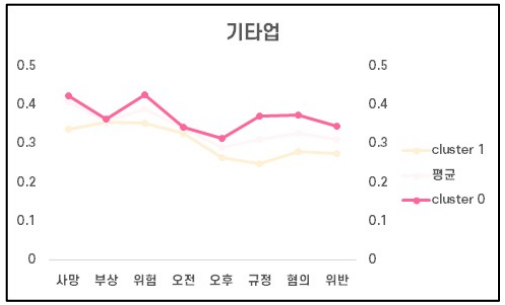
"공정", "내부"
"사이"에 "협착된" 사고 유형



사망, 부상 점수는 높지만, 위험 점수가 현저히 낮은 것과 오전, 오후에 고루 사고가 많이 발생하는 것으로 보아 산업현장 근로자들의 안전에 대한 인지가 낮아 안전수칙이 제대로 지켜지지 않아서 발생하는 사고라고 할 수 있다.

Cluster 0

"외국인"과 "비정규직"을 주로
사용하는 해양에서의
"가스" 유출로
인한 화재 사고 유형



해양 사고는 선박 내 가스 누출 사고로 인한 사고가 많다. 질식 재해자 중 절반 이상이 사망할 만큼 치명적으로, 위험도 수치를 확인한 결과 부상보다 사망까지 이르는 경우가 높다.
해양 업종은 외국인 노동자와 비정규직을 사용하며, 표준근로계약과 규정을 피하는 경우가 많다고 볼 수 있다.규정/협의/위반 수치를 확인한 결과 모두 score가 높은 것으로 보아 정책의 도입 강화가 필요하다는 것을 알 수 있다.

5 결론 및 기대효과

결론

제조업에 비해 화학업, 건설업 등이 상대적으로 인과성이 잘 드러났다.

시계열로 정리가 가능하여 사건 사고의 유형이나,
산업의 변화도 반영 가능하다

기존 통계에 결합하여 상세한 통계의 인과관계를 제안할 수 있다

HDBSCAN 군집화를 사용하여, 위계를 구성하여 유연성을 제공하였다

DBCV Score로 클러스터의 유효성을 검증하였다.

기대효과

사고의 감소로
산재비용을 줄일 수 있다

노동자의 정보에 따른
구체적인 안전 매뉴얼
설계가 가능하다

데이터를 실시간
업데이트를 하여 변화를 관측 가능하다

고찰

다른 자연어 알고리즘 Fasttext,
Transformer
등과 비교하여 최적의 모델을 찾을 수 있다

뉴스 기사의 특징 때문에 사고와 관련이 적은
말이 많아 이상치가 많이 발생한다

참고 문헌

강성식 외 3인. “자연어 처리 기법을 활용한 산업재해 위험요인 구조화”. 『Journal of the Korean Society of Safety 』, 2021, 36, pp. 56-63

이수상. “독후감 텍스트의 토픽모델링 적용에 관한 탐색적 연구”. 『한국도서관·정보학회』, 2016, 47(4), pp.1-18

박대영 외 2인. “토픽 모델링 기반의 국내외 공공데이터 연구 동향 비교 분석”. 『디지털융복합연구』, 2021, 19, pp.1-12

심재권. “국민청원글의 토픽 모델링을 통한 교육이슈 분석”. 『한국정보교육학회』, 2021, 25(4), pp.633-640

윤상훈 외 1인. “Word2Vec를 이용한 토픽모델링의 확장 및 분석사례”. 『한국정보시스템학회』, 2021, 30(1), pp.45-64

데이터 출처

- 한국산업안전보건공단
- 안전신문
- 매일경제
- YTN
- 연합뉴스
- 한겨레
- 매일노동뉴스
- …