재난문자 유형별 분류 프로젝트

데이터마이닝 6조
산업정보시스템 20102036 이승규
산업정보시스템 20102045 임형빈
산업정보시스템 22101942 한다은

Github repo
https://github.com/daeun118/24_1_DataMining.git

Contents

분석 배경 및 필요성

- 과도한 재난문자로 시민들의 피로도 증가
- 재난문자에 대한 신뢰도 하락

‘영혼 없는’ 재난문자 좀 스마트해질 수 없나[논설위원의 뉴스 요리]

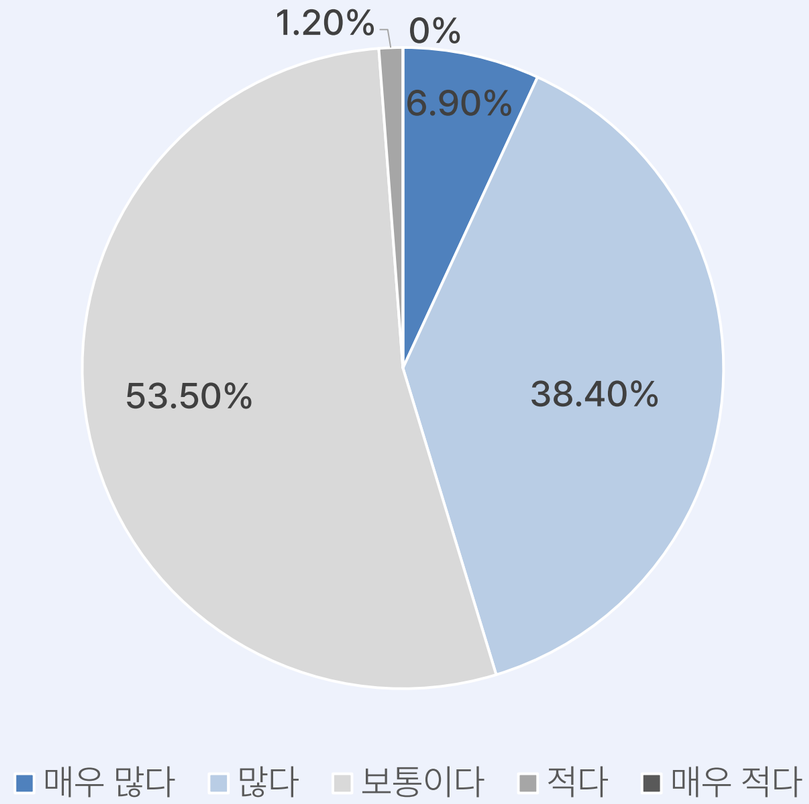
입력 : 2023-07-22 09:00:00 ㅣ 수정 : 2023-07-22 16:13:47

"시끄러워서 알람 꺼놨어요" 스팸 취급받는 재난문자

'스팸' 취급받는 재난문자, "면피용" 비판도…피로감에 알람 끄는 시민들

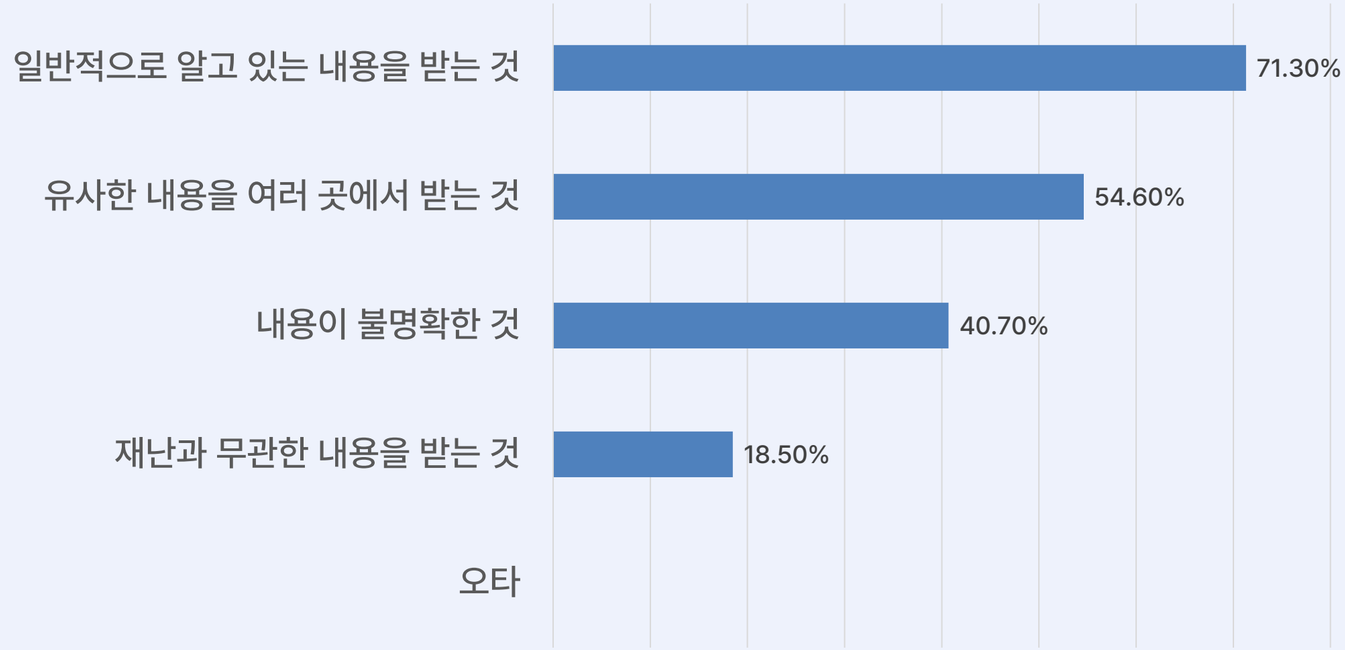
분석 배경 및 필요성

재난문자 수신량



8.9%가 재난문자 내용에 불만족

재난문자에 불만족하는 이유



이현지, 변윤관, 장석진, 최성종. 재난문자 이용자 인식조사. (2020).

분석 목적

수신자가 원하는 유형의 재난문자만
받아볼 수 있게 하기 위해,
재난문자를 각 유형별로 분류한다.

데이터 획득 방법

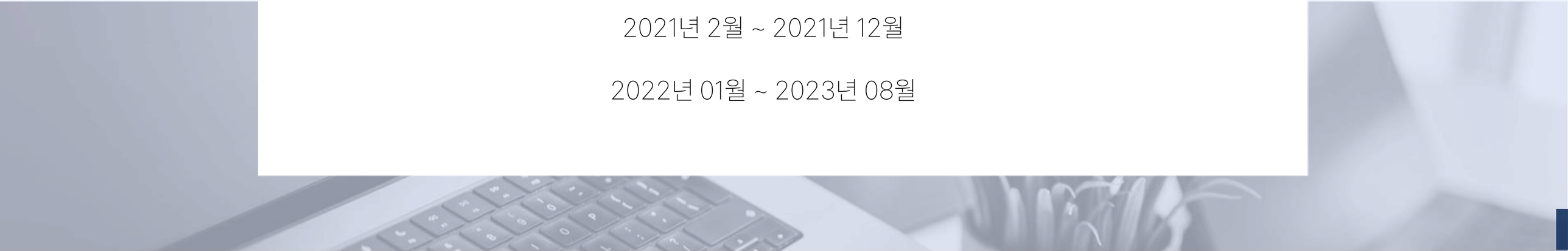
공공데이터포털

행정안전부_서울시 구청 재난문자 발송 현황

2020년 1월 ~ 2021년 01월

2021년 2월 ~ 2021년 12월

2022년 01월 ~ 2023년 08월



분석 과정

데이터 전처리

1. 토큰화 및 불용어 제거

2. TF-IDF

3. 데이터 라벨링

- 전처리 전 데이터

	연	월	일	시	분	수신지역	송출내용
0	2020	1	26	9	9	중구	[중구청] 오늘 07:00경 중구 장충동 앰버서더호텔 내 화재 발생. 이 지역을 우...
1	2020	1	28	16	59	광진구	[광진구청] 신종 코로나바이러스감염증 예방수칙 마스크 착용 흐르는물 30초이상 손씻...
2	2020	1	30	9	49	송파구	[송파구청]감염증 대응을 위해 구청장 중심으로 24시간 대책본부를 운영하고 있습니다...
3	2020	1	31	14	31	강남구	[강남구]중국 방문 후 14일 이내에 발열(37.5도 이상) 또는 호흡기증상(기침,...
4	2020	1	31	15	5	강남구	[강남구]중국 방문 후 14일이내에 발열(37.5도이상) 또는 호흡기증상(기침,인후...

- Okt 형태소 분석기로 토큰화(pos: 형태소에 품사를 붙여서 추출)
- 한글 불용어 사전으로 불용어 제거

	preprocessed_송출내용
0	오늘 경 중구 장충동 앰버 서다 호텔 내 화재 발생 지역 우회 하다 주시 인근 주민...
1	신종 코로나바이러스 감염증 예방 칙 마스크 착용 흐르다 물 초이 손씻기 중국 방문 ...
2	감염증 대응 위해 청장 중심 대책 본부 운영 확 진자 없다 발열 증상 시 보건소 문...
3	중국 방문 후 이내 발열 호흡기 증상 기침 후통 발 현시 강남구 재난 안전 대책 부...
4	중국 방문 후 내 발열 호흡기 증상 기침 후통 발현 시 강남구 재난 안전 대책 부로...

분석 과정

데이터 전처리

1. 토큰화 및 불용어 제거

2. TF-IDF

3. 데이터 라벨링

TF-IDF: TF와 IDF를 곱한 값

- TF: term frequency, 특정 단어가 특정 문서에서 사용된 횟수
- DF: document frequency, 특정 단어가 사용된 문서의 수
- IDF: inverse document frequency, 역문서 빈도

흔하지 않은 단어인데 특정 텍스트에서
자주 사용될수록 큰 값을 가진다.

- TF-IDF의 결과로 만들어진 행렬: (문서 수*단어 수)
- max_feature = 100으로 설정하여 (9348*100)의 행렬을 만든다.
- TF-IDF의 결과 희소 행렬이 생성되므로 toarray()를 사용해 밀집 행렬로 변환

분석 과정

데이터 전처리

1. 토큰화 및 불용어 제거

2. TF-IDF

3. 데이터 라벨링

- TF-IDF, toarray()의 결과로 생성된 밀집 행렬

	가깝다	가족	강서구	거리	거주지	검사	격리	결과	경로	공개	...	채널	추가	추후	카카오	코로나	하다	해주다	현재	홈페이지	확인
0	-0.195993	-0.178289	-0.180621	-0.277417	-0.167665	-0.427924	-0.220432	-0.30342	-0.145053	-0.564277	...	-0.140914	-0.289634	-0.242714	-0.162418	-0.574206	1.009603	-0.137183	-0.195982	-0.925041	-0.446814
1	-0.195993	-0.178289	-0.180621	-0.277417	-0.167665	-0.427924	-0.220432	-0.30342	-0.145053	-0.564277	...	-0.140914	-0.289634	-0.242714	-0.162418	-0.574206	-0.533856	-0.137183	-0.195982	-0.925041	-0.446814
2	-0.195993	-0.178289	-0.180621	-0.277417	-0.167665	-0.427924	-0.220432	-0.30342	-0.145053	-0.564277	...	-0.140914	-0.289634	-0.242714	-0.162418	-0.574206	-0.533856	-0.137183	-0.195982	-0.925041	-0.446814
3	-0.195993	-0.178289	-0.180621	-0.277417	-0.167665	-0.427924	-0.220432	-0.30342	-0.145053	-0.564277	...	-0.140914	-0.289634	-0.242714	-0.162418	-0.574206	-0.533856	-0.137183	-0.195982	-0.925041	-0.446814
4	-0.195993	-0.178289	-0.180621	-0.277417	-0.167665	-0.427924	-0.220432	-0.30342	-0.145053	-0.564277	...	-0.140914	-0.289634	-0.242714	-0.162418	-0.574206	-0.533856	-0.137183	-0.195982	-0.925041	-0.446814
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9343	-0.195993	-0.178289	-0.180621	-0.277417	-0.167665	-0.427924	-0.220432	-0.30342	-0.145053	-0.564277	...	-0.140914	-0.289634	-0.242714	-0.162418	-0.574206	-0.533856	-0.137183	-0.195982	-0.925041	-0.446814
9344	-0.195993	-0.178289	-0.180621	-0.277417	-0.167665	-0.427924	-0.220432	-0.30342	-0.145053	-0.564277	...	-0.140914	-0.289634	-0.242714	-0.162418	-0.574206	2.219443	-0.137183	-0.195982	-0.925041	-0.446814
9345	-0.195993	-0.178289	-0.180621	-0.277417	-0.167665	-0.427924	-0.220432	-0.30342	-0.145053	-0.564277	...	-0.140914	-0.289634	-0.242714	-0.162418	-0.574206	2.137489	-0.137183	-0.195982	-0.925041	-0.446814
9346	-0.195993	-0.178289	-0.180621	-0.277417	-0.167665	-0.427924	-0.220432	-0.30342	-0.145053	-0.564277	...	-0.140914	-0.289634	-0.242714	-0.162418	-0.574206	0.930558	-0.137183	-0.195982	-0.925041	-0.446814
9347	-0.195993	-0.178289	-0.180621	-0.277417	-0.167665	-0.427924	-0.220432	-0.30342	-0.145053	-0.564277	...	-0.140914	-0.289634	-0.242714	-0.162418	-0.574206	2.478902	-0.137183	-0.195982	-0.925041	-0.446814

9348 rows × 100 columns

분석 과정

데이터 전처리

1. 토큰화 및 불용어 제거

2. TF-IDF

3. 데이터 라벨링

Table 1. Number of Disaster Alerts by Disaster Type

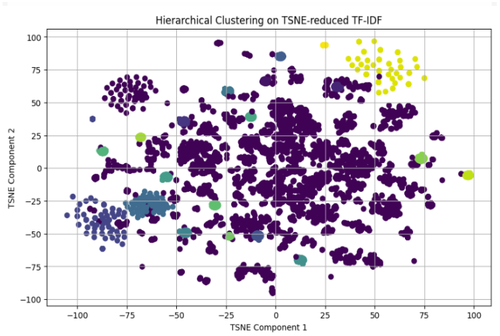
Disaster Type	Number
전염병	5,000 (67,984개 중)
태풍	2,584
호우	1,115
한파	816
교통	766
홍수	725
대설	574
폭염	435
산사태	408
기타	1,155
계	13,578

데이터 라벨링을 위한 클러스터링

- DBSCAN, Kmeans , Agglomerative Clustering 방법 시도

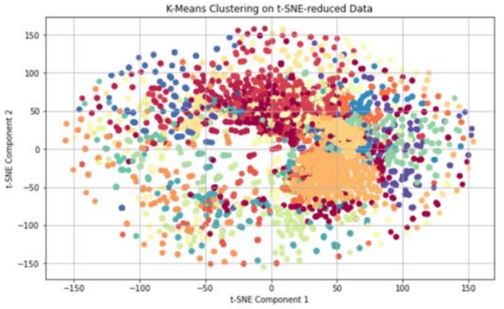
- DBSCAN

  거의 모든 데이터를 노이즈 처리



- Kmeans

  클러스터링이 적절하게 되지 않음



Agglomerative Clustering 방법으로 결정

분석 과정

데이터 전처리

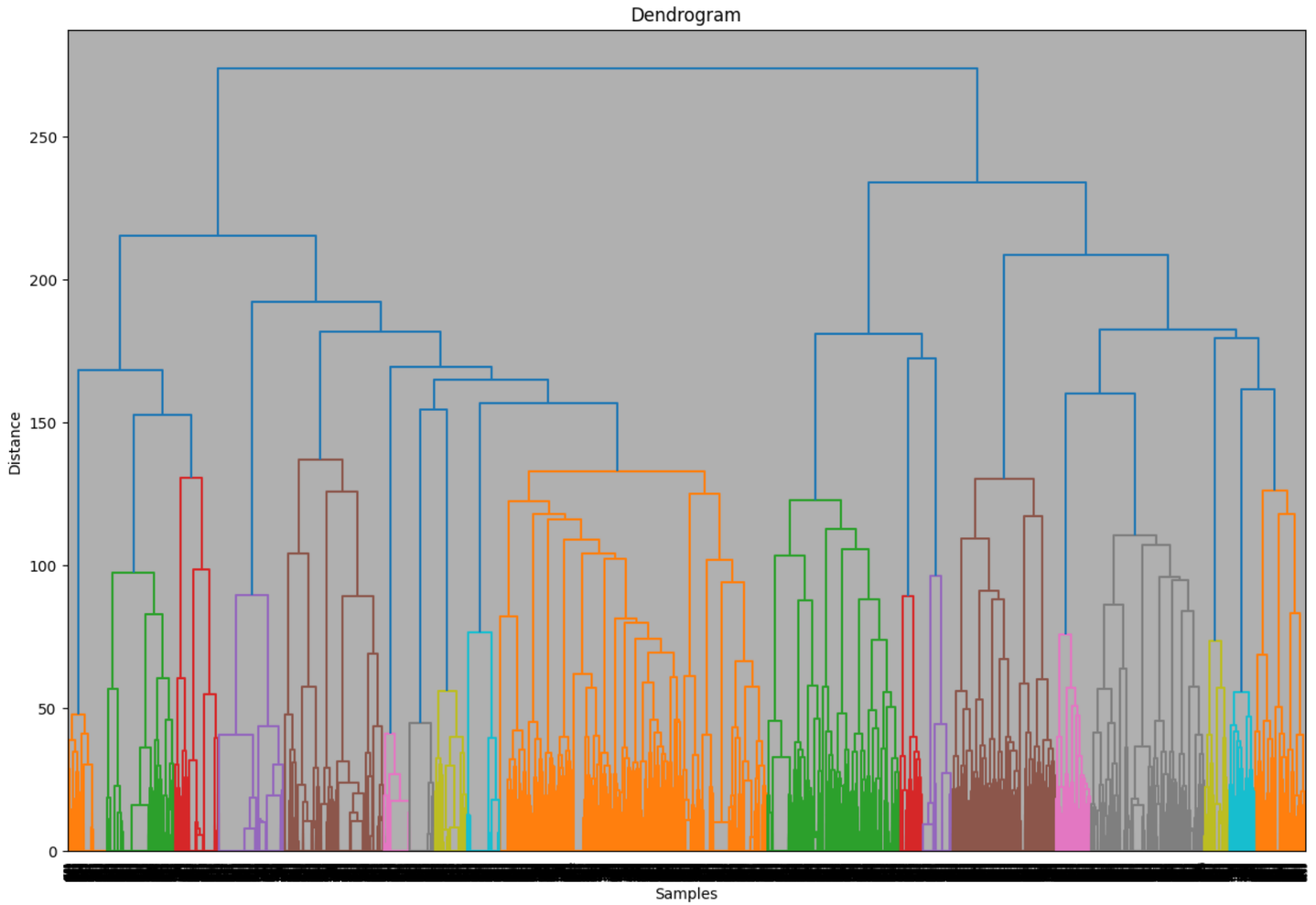
1. 토큰화 및 불용어 제거

2. TF-IDF

3. 데이터 라벨링

Agglomerative Clustering

- 거리 기반의 클러스터링 기법 -> StandardScaler 사용하여 스케일링
- 높이 150에서 군집화하기로 결정 -> 총 19개의 클러스터



분석 과정

데이터 전처리

1. 토큰화 및 불용어 제거

2. TF-IDF

3. 데이터 라벨링

Agglomerative Clustering

- 높이 150에서 군집화하기로 결정 -> 총 19개의 클러스터

- 0번(753): 확진자 발생안내
- 1번(329): 역학조사 진행중
- 2번(2020): 확진자 발생, 기상 관련
- 3번(1002): 확진자 발생, 기상 관련
- 4번(784): 기상 관련
- 5번(498): 확진자 동선공개
- 6번(385): 백신, 확진자발생
- 7번(855): 전염병 검사 권장
- 8번(515): 역학조사 진행중
- 9번(190): 확진자 동선 알림
- 10번(169): 거리두기 권장
- 11번(192): 방역완료
- 12번(298): 역학조사 결과 공개 예정
- 13번(227): 방역수칙 권장
- 14번(189): 자가격리 권장
- 15번(262): 임시선별소 운영
- 16번(246): 확진자 동선공개
- 17번(187): 역학조사 진행중
- 18번(247): 역학조사 결과 안내

분석 과정

데이터 전처리

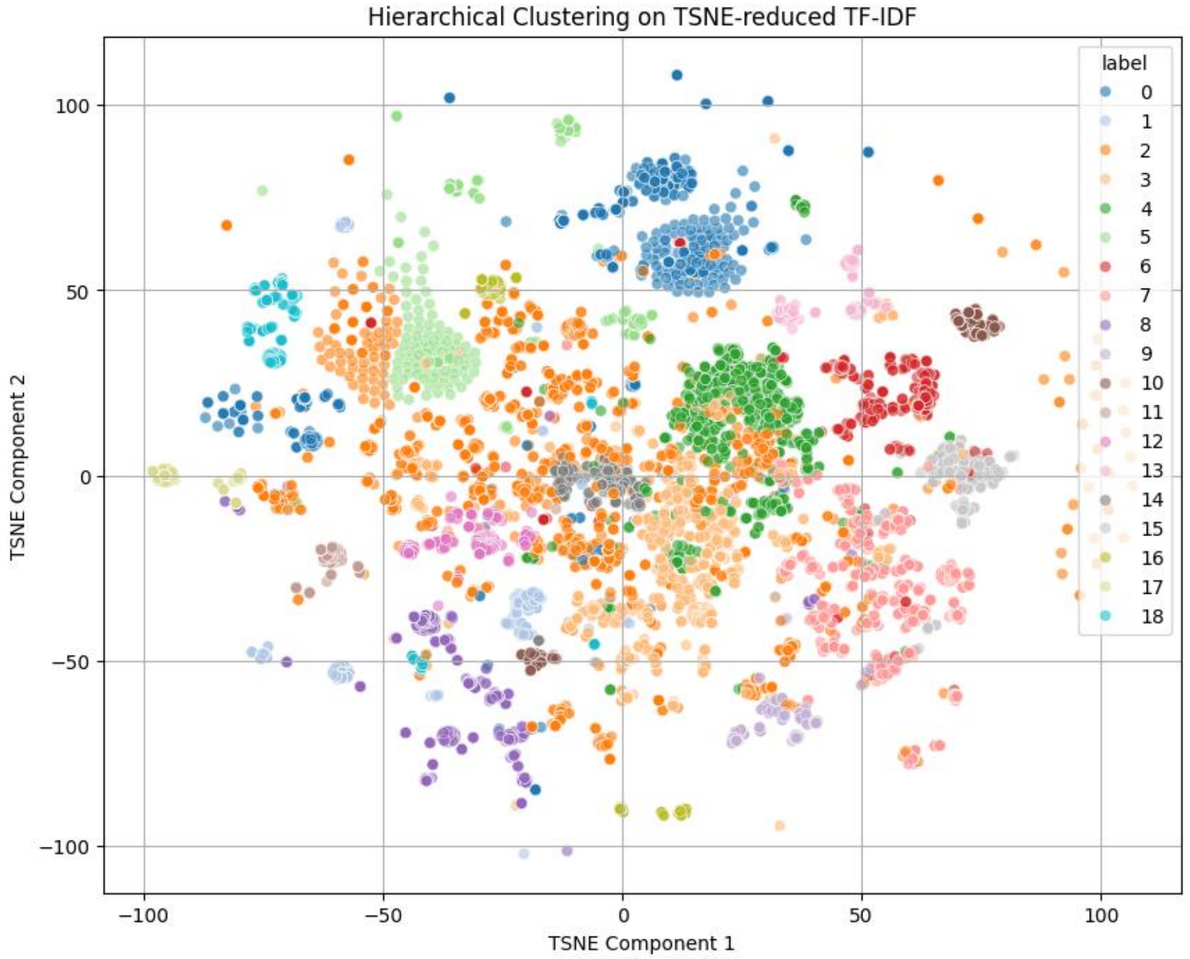
1. 토큰화 및 불용어 제거

2. TF-IDF

3. 데이터 라벨링

Agglomerative Clustering

- 높이 150에서 군집화하기로 결정 -> 총 19개의 클러스터



분석 과정

데이터 전처리

1. 토큰화 및 불용어 제거

2. TF-IDF

3. 데이터 라벨링

Agglomerative Clustering

- 19개의 클러스터 각각의 내용을 확인

유사한 클러스터끼리 병합 -> 총 8개의 클러스터

- 0번(753): 확진자 발생안내
- 1번(784): 기상 관련
- 2번(385): 백신, 교통, 확진자발생
- 3번(262): 임시선별소
- 4번(3022): 확진자 발생, 기상 관련
- 5번(2320): 역학조사
- 6번(1045): 전염병 검사 권장
- 7번(777): 방역

분석 과정

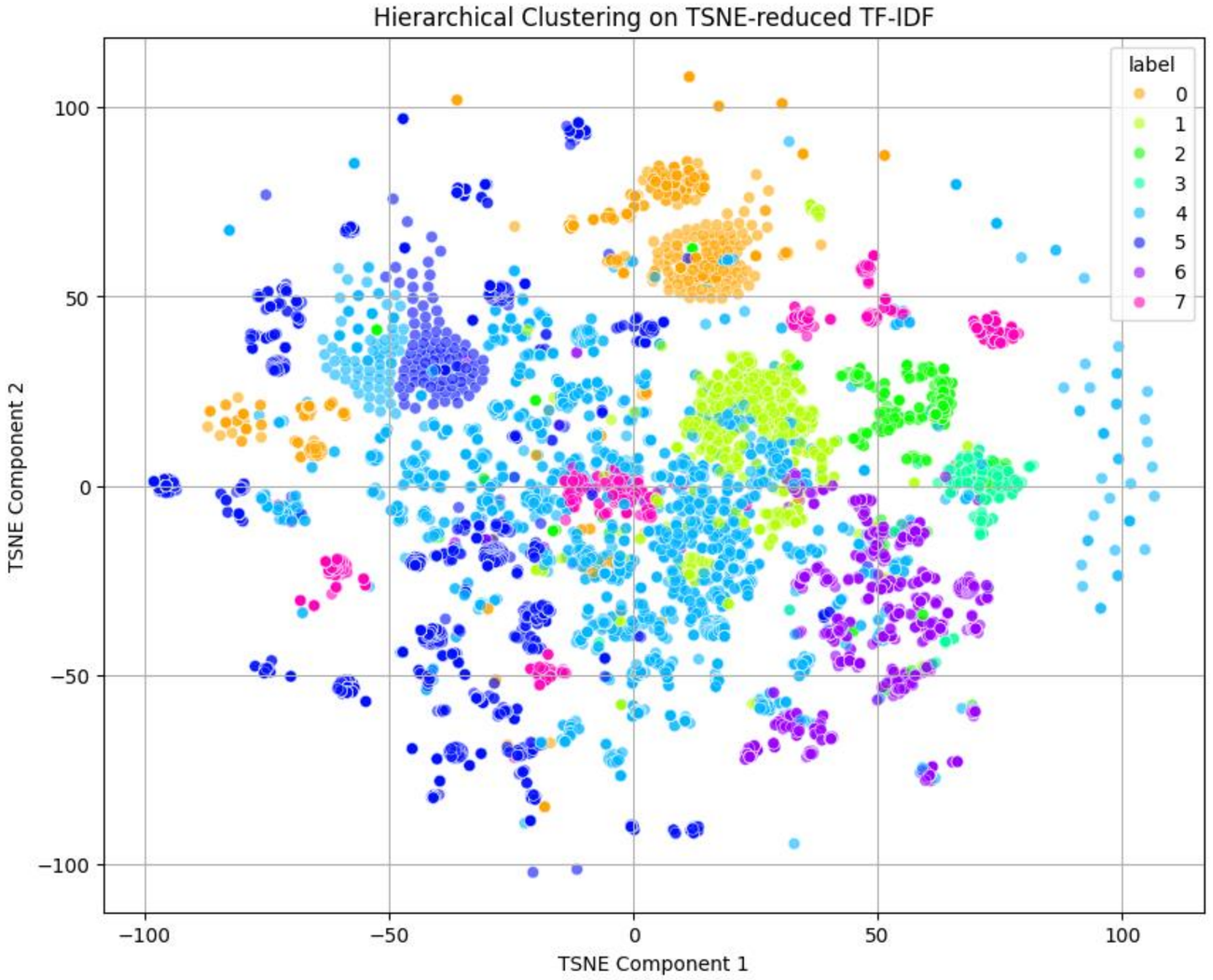
데이터 전처리

1. 토큰화 및 불용어 제거

2. TF-IDF

3. 데이터 라벨링

라벨링한 데이터를 T-SNE를 사용해 시각화



분석 과정

데이터 전처리

1. 토큰화 및 불용어 제거

2. TF-IDF

3. 데이터 라벨링

Agglomerative Clustering

- 라벨링한 데이터를 new_df.csv파일로 저장

	preprocessed_송출내용	label
0	오늘 경 중구 장충등 앰버 서다 호텔 내 화재 발생 지역 우회 하다 주시 인근 주민...	1
1	신종 코로나바이러스 감염증 예방 칙 마스크 착용 흐르다 물 초이 손씻기 중국 방문 ...	6
2	감염증 대응 위해 청장 중심 대책 본부 운영 확 진자 없다 발열 증상 시 보건소 문...	4
3	중국 방문 후 이내 발열 호흡기 증상 기침 후통 발 현시 강남구 재난 안전 대책 부...	6
4	중국 방문 후 내 발열 호흡기 증상 기침 후통 발현 시 강남구 재난 안전 대책 부로...	6
...	...	...
9343	태풍 북상 많다 비 강풍 예상 되다 안양천 수위 상승 산책로 자전거 도로 통제 오니...	1
9344	태풍 북상 중 많다 비 강풍 예상 되다 외출 자제 하다 주시 하천 출입 금지 배수 ...	1
9345	태풍 카눈 따르다 강풍 산림 내 수목 전도 우려 되다 국사봉 까치산 달산 출입 금 ...	1
9346	관악구 지역 집중호우 예상 되다 주민 별빛 린 위험 지역 접근 금지 하다 배수로 정...	4
9347	오늘 밤 내일 많다 비 예상 되다 따르다 피해 예방 위 하다 지하 주택 배수 시설 ...	1

분석 과정

모델링

- 데이터 분할 및 벡터화
- 스케일링

1. SVM
2. Random Forest
3. XGBoost
4. CatBoost
5. 최적 모델 선정

- train / valid / test를 6 : 2 : 2로 분할

X_train shape	(5608, 1)	y_train shape	(5608,)
X_valid shape	(1870, 1)	y_valid shape	(1870,)
X_test shape	(1870, 1)	y_test shape	(1870,)

- 분할된 각 데이터셋을 TF-IDF를 이용하여 벡터화(max_features = 30)
- 각각의 데이터셋을 StandardScaler를 사용하여 스케일링

분석 과정

모델링

- 데이터 분할 및 벡터화
- 스케일링

1. SVM
2. Random Forest
3. XGBoost
4. CatBoost
5. 최적 모델 선정

	C	gamma	train_accuracy	valid_accuracy
0	0.1	0.1000	0.809201	0.783957
1	0.1	0.0100	0.756419	0.731551
2	0.1	0.0010	0.600036	0.574866
3	0.1	0.0001	0.324358	0.324599
4	1.0	0.1000	0.901213	0.832086
5	1.0	0.0100	0.829886	0.792513
6	1.0	0.0010	0.748395	0.725668
7	1.0	0.0001	0.604315	0.577540
8	10.0	0.1000	0.925999	0.839037
9	10.0	0.0100	0.880350	0.817112
10	10.0	0.0010	0.803495	0.770053
11	10.0	0.0001	0.746969	0.721390
12	50.0	0.1000	0.926890	0.839037
13	50.0	0.0100	0.909058	0.824064
14	50.0	0.0010	0.827746	0.782888
15	50.0	0.0001	0.783167	0.759358

	precision	recall	f1-score	support
0	0.92	0.91	0.91	152
1	0.73	0.75	0.74	162
2	0.64	0.61	0.62	89
3	0.70	0.59	0.64	51
4	0.81	0.83	0.82	607
5	0.97	0.95	0.96	464
6	0.77	0.86	0.81	194
7	0.84	0.75	0.79	151
accuracy			0.84	1870
macro avg	0.80	0.78	0.79	1870
weighted avg	0.84	0.84	0.84	1870

SVM 모델 성능

- Train accuracy: 0.927
- Valid accuracy: 0.839

분석 과정

모델링

- 데이터 분할 및 벡터화
- 스케일링

1. SVM
2. Random Forest
3. XGBoost
4. CatBoost
5. 최적 모델 선정

- GridSearchCV를 이용하여 hyperparameter 튜닝

max_depth	20
Max_features	auto
Min_samples_leaf	1
Min_samples_split	5
N_estimators	300

	precision	recall	f1-score	support
0	0.93	0.90	0.92	152
1	0.73	0.75	0.74	162
2	0.79	0.56	0.66	89
3	0.67	0.59	0.62	51
4	0.77	0.85	0.81	607
5	0.97	0.92	0.94	464
6	0.79	0.87	0.82	194
7	0.89	0.70	0.79	151
accuracy			0.83	1870
macro avg	0.82	0.77	0.79	1870
weighted avg	0.84	0.83	0.83	1870

Random Forest 모델 성능

- Train accuracy: 0.923
- Valid accuracy: 0.834

분석 과정

모델링

- 데이터 분할 및 벡터화
- 스케일링

1. SVM
2. Random Forest
3. XGBoost
4. CatBoost
5. 최적 모델 선정

- GridSearchCV를 이용하여 hyperparameter 튜닝

colsample_bytree	1
learning_rate	0.05
max_depth	7
n_estimators	300
subsample	0.6

	precision	recall	f1-score	support
0	0.92	0.90	0.91	152
1	0.71	0.74	0.73	162
2	0.74	0.56	0.64	89
3	0.60	0.55	0.57	51
4	0.79	0.83	0.80	607
5	0.96	0.93	0.95	464
6	0.77	0.87	0.82	194
7	0.85	0.73	0.78	151
accuracy			0.83	1870
macro avg	0.79	0.76	0.77	1870
weighted avg	0.83	0.83	0.83	1870

XGBoost 모델 성능

- Train accuracy: 0.923
- Valid accuracy: 0.828

분석 과정

모델링

- 데이터 분할 및 벡터화
- 스케일링

1. SVM
2. Random Forest
3. XGBoost
4. CatBoost
5. 최적 모델 선정

- GridSearchCV를 이용하여 hyperparameter 튜닝

depth	10
iterations	500
l2_leaf_reg	1
learning_rate	0.05

	precision	recall	f1-score	support
0	0.94	0.89	0.92	152
1	0.71	0.76	0.73	162
2	0.76	0.57	0.65	89
3	0.63	0.61	0.62	51
4	0.79	0.83	0.81	607
5	0.95	0.93	0.94	464
6	0.78	0.87	0.82	194
7	0.88	0.75	0.81	151
accuracy			0.83	1870
macro avg	0.80	0.78	0.79	1870
weighted avg	0.83	0.83	0.83	1870

CatBoost 모델 성능

- Train accuracy: 0.919
- Valid accuracy: 0.833

분석 과정

모델링

- 데이터 분할 및 벡터화
- 스케일링

1. SVM
2. Random Forest
3. XGBoost
4. CatBoost
5. 최적 모델 선정

- 선정된 최적 모델의 학습을 위해 train 데이터와 valid 데이터를 병합
  -> X_train_final, y_train_final

model	Cross-validation score(cv=5)
SVM	0.848
Random Forest	0.843
XGBoost	0.840
CatBoost	0.846

In [14]:

```
#bestmodel 생성

best_model = SVC(C=50, gamma=0.1)
best_model.fit(X_train_final, y_train_final)
```

분석 과정

모델링

- 데이터 분할 및 벡터화
- 스케일링

1. SVM
2. Random Forest
3. XGBoost
4. CatBoost
5. 최적 모델 선정

Test set에서의 분석 결과

cluster	precision	recall	F1-score	support
0 (확진자 발생)	0.91	0.95	0.93	143
1 (기상 관련)	0.68	0.79	0.73	157
2 (백신, 교통, 확진자발생)	0.6	0.65	0.62	77
3 (임시선별소)	0.68	0.67	0.67	57
4 (확진자 발생, 기상 관련)	0.82	0.81	0.82	596
5 (역학조사)	0.97	0.95	0.96	470
6 (전염병 검사 권장)	0.91	0.85	0.88	198
7 (방역)	0.86	0.8	0.83	172
accuracy		0.849		1870

분석 과정

모델링

- 데이터 분할 및 벡터화
- 스케일링

1. SVM
2. Random Forest
3. XGBoost
4. CatBoost
5. 최적 모델 선정

Test set에서의 분석 결과

- **2번, 4번 클러스터**는 라벨링 과정에서 깔끔하게 분리되지 않은 혼합된 클러스터이다.
- 2번 클러스터는 **precision = 0.6 , recall = 0.65**로 성능 개선이 필요하다.
- 3번 클러스터(임시선별소 관련)는 **precision = 0.68 , recall = 0.67**로 성능 개선이 필요하다.

분석 과정

모델링

- 데이터 분할 및 벡터화
- 스케일링

1. SVM
2. Random Forest
3. XGBoost
4. CatBoost
5. 최적 모델 선정

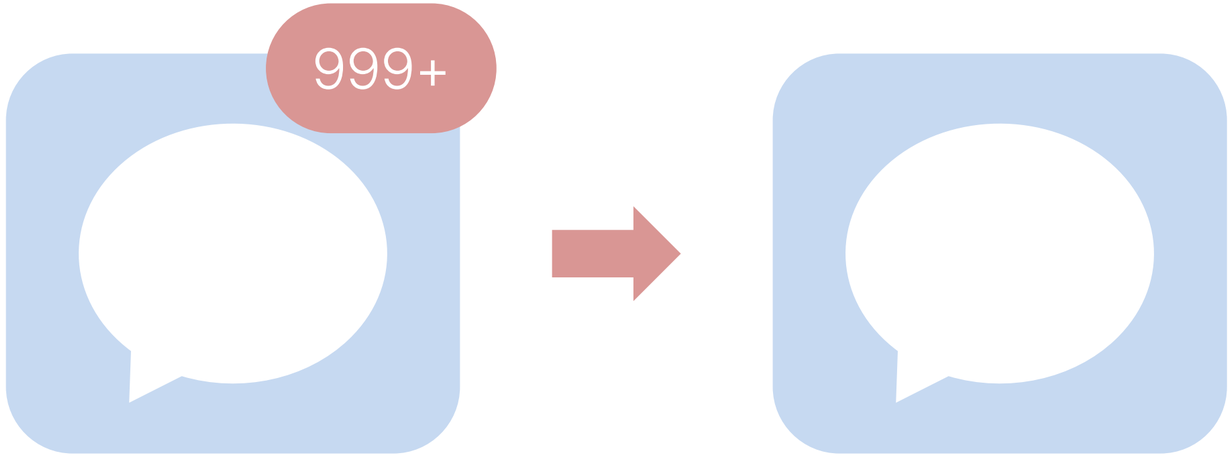
Test set에서의 분석 결과

- 전체 클래스 중 성능이 좋지 않은 2, 3번 클래스의 데이터는 각각 384개, 262개로 다른 클래스에 비해 적은 수이다. 따라서 **불균형 데이터셋이 모델의 분류 성능에 부정적인 영향**을 주었을 가능성이 높다.

- 전체적인 정확도의 경우, 최적 모델로 선정된 SVM 외의 다른 모델에서도 약 0.84의 비슷한 성능을 보였다.

기대 효과

- 수신자가 자신이 필요하다고 생각하는 유형의 재난문자만 수신할 수 있다.
- 재난문자에 대한 사람들의 피로도를 줄여서 재난문자에 대한 신뢰도를 회복시키고, 효용을 증가시킨다.



한계점 및 개선방안

한계점

1. 라벨링을 위한 클러스터링 과정에서 데이터가 완벽하게 분류되지 못함.
2. 전염병(코로나)관련 문자의 양이 매우 많아 일반 기상 경보 관련 문자들을 잘 분류하지 못함

- Train dataset을 벡터화 시킨 결과: 전염병 문자 관련 단어들만 feature로 선정되어 일반 기상 경보 문자(호우, 대설, 폭염 등)를 세세하게 분류하지 못함.

	검사	결과	공개	내용	동선	마스크	바라다	발생	방역	보건소	...	조사	준수	진자	참고	참조	추가	코로나	하다	홈페이지	확인
0	0.00000	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.213807	0.000000	0.0	...	0.709696	0.0	0.398862	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.418929
1	0.00000	0.0	0.000000	0.000000	0.0	0.0	0.255083	0.247325	0.000000	0.0	...	0.000000	0.0	0.230696	0.000000	0.563434	0.000000	0.0	0.0	0.000000	0.000000
2	0.00000	0.0	0.000000	0.000000	0.0	0.0	0.304473	0.000000	0.000000	0.0	...	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000
3	0.60247	0.0	0.000000	0.506698	0.0	0.0	0.000000	0.000000	0.000000	0.0	...	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000
4	0.00000	0.0	0.000000	0.345988	0.0	0.0	0.214147	0.207634	0.434480	0.0	...	0.000000	0.0	0.193674	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.406835
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5603	0.00000	0.0	0.000000	0.000000	0.0	0.0	0.184727	0.179109	0.374789	0.0	...	0.000000	0.0	0.167066	0.346138	0.000000	0.000000	0.0	0.0	0.223713	0.000000
5604	0.00000	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.731263	0.000000	0.0	...	0.000000	0.0	0.682095	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000
5605	0.00000	0.0	0.295847	0.000000	0.0	0.0	0.000000	0.172472	0.000000	0.0	...	0.286245	0.0	0.160875	0.000000	0.000000	0.435889	0.0	0.0	0.215424	0.000000
5606	0.00000	0.0	0.000000	0.000000	0.0	0.0	0.266094	0.258001	0.000000	0.0	...	0.000000	0.0	0.240654	0.000000	0.000000	0.000000	0.0	0.0	0.322253	0.505523
5607	0.00000	0.0	0.000000	0.000000	0.0	0.0	0.216278	0.209700	0.000000	0.0	...	0.348031	0.0	0.391200	0.000000	0.477719	0.000000	0.0	0.0	0.261923	0.000000

5608 rows × 30 columns

한계점 및 개선방안

개선 방안

- 서로 다른 유형의 문자에 자주 등장하는 단어가 한 문자에 동시에 존재할 수 있으므로 문장의 맥락 파악이 필요

(예)

[은평구청] 한파특보에 따른 은평구 임시선별검사소 운영시간 변경안내.
1.7.(목)~1.10.(일)까지 11시~15시로 단축운영됨을 알려드립니다.

➔ **딥러닝 기반 모델의 사용**

한계점 및 개선방안

개선 방안

- 데이터의 특성 상(2020년~2023년 데이터) 전염병 유형 문자의 수가 다른 유형의 문자보다 매우 많음

  ➔ **데이터 수집 확대 및 텍스트 데이터 증강(SR, RI, RS, RD, Back Translation)**

참고문헌

[1]이현지, 변윤관, 장석진, 최성종. 재난문자 이용자 인식조사. (2020). (https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10530859).

[2]Soonwook Park, Hyeyoon Jun, Yoonsoo Kim, Soowon Lee. A Deep Learning Model for Disaster Alerts Classification. (2021). (https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10818510).

[3]https://soojle.gitbook.io/project/undefined-2/nlp/tf-idf-term-frequency-inverse-document-frequency

[4]https://sooeun67.github.io/data%20science/text-data-augmentation/