**중 간 보 고 서**

2023년 5월 26일 금요일

Team Winitech (임도영, 정도, 김도형)

본 프로젝트는 ㈜위니텍과 경북대학교 K-Digital Training의 협력으로 진행되었습니다.

**I. 서론**

**1. 기업 소개**

㈜위니텍은 기술로 이기는 강한 기업이 되겠다는 뜻을 담아 1997년 추교관 대표가 설립. 재난 사고로 인한 피해를 최소화하는 통합긴급사태관리시스템 (IEMS: Intergrated Emergency Management System)을 개발 및 구축하는 국내 1위의 독보적인 기업이다. 다양한 기술력을 바탕으로 119 긴급구조시스템의 핵심 역량을 보유한 기업으로 자리매김하였으며 2005년부터 해외 시장에 진출하여 동남 아시아, 아프리카, 중남미 등으로 영역을 넓혀가고 있다.

**2. 프로젝트 배경**

이번 프로젝트의 배경은 재난 관리 매뉴얼과 관련된 고객의 요청 사항을 수행 중 발생한 이슈를 해결하기 위함으로, 유사 항목 추출을 위한 자연어 처리 기반의 모델이 필요한 것으로 판단된다. 데이터는 국가법령정보센터 (https://www.law.go.kr)에서 수집하여 사용할 예정이다.

이번 프로젝트는 2023년 5월 2일부터 6월 15일까지 진행된다.

**II. 프로젝트 개요**

**1. 팀 구성 및 업무 분장**

|  |  |  |
| --- | --- | --- |
|  | **역할** | **주 업무** |
| **임도영** | **Project Manager** | **Web** |
| **정도** | **Team Member** | **Python** |
| **김도형** | **Team Member** | **Reports** |

**2. 개발 환경**

|  |  |
| --- | --- |
|  | **버전** |
| **Python** | **3.9.13** |
| **scikit-learn** | **1.2.1** |
| **Windows** | **11 Pro (64-bit)** |

**3. Work Breakdown Structure**

|  |  |
| --- | --- |
| **1.0.0** | **<착수 및 프로젝트 관리>** |
| **1.1.0** | **프로젝트 관리** |
| **1.1.1** | **일간 회의** |
| **1.1.2** | **최종 회의** |
| **2.0.0** | **<계획서 작성>** |
| **2.1.0** | **프로젝트 계획서** |
| **2.1.1** | **배경 및 목적** |
| **2.1.2** | **프로젝트 일정** |
| **2.1.3** | **예상 이슈** |
| **2.1.4** | **WBS** |
| **2.1.5** | **최종 계획서** |
| **3.0.0** | **<데이터>** |
| **3.1.0** | **수집 및 전처리** |
| **3.1.1** | **데이터 수집** |
| **3.1.2** | **데이터 전처리** |
| **4.0.0** | **<모델링 및 검증>** |
| **4.1.0** | **모델링** |
| **4.1.1** | **머신 러닝 기반 모델링** |
| **~~4.1.2~~** | **~~딥 러닝 기반 모델링~~** |
| **4.1.3** | **모델 검증** |
| **5.0.0** | **<디자인>** |
| **5.1.0** | **메인 디자인** |
| **5.1.1** | **디자인 구성** |
| **5.1.2** | **디자인 컨펌** |
| **5.1.3** | **디자인 수정·보완** |
| **6.0.0** | **<Web 구현>** |
| **6.1.0** | **Web 구현 및 수정·보완** |
| **6.1.1** | **Web 구현** |
| **6.1.2** | **Web 수정·보완** |
| **7.0.0** | **<구현 및 수정·보완>** |
| **7.1.0** | **테스트** |
| **7.1.1** | **구현** |
| **7.1.2** | **최종 수정·보완** |

**4. 일정**

|  |  |
| --- | --- |
| **1주차** | **분석 기획 (프로젝트 계획 수립) ~ 5/5** |
| **2주차** | **데이터 준비 (자료 수집) ~ 5/12** |
| **3주차** | **데이터 정제 (전처리) ~ 5/19** |
| **4주차** | **모델 구현 (모델 구축) ~ 5/26** |
| **5주차** | **모델 구현 (모델 평가 및 검증) ~ 6/2** |
| **6주차** | **모델 구현 (페이지 구축) ~ 6/9** |
| **7주차** | **평가 및 전개 & 프로젝트 평가 및 보고 ~ 6/16** |

**III. 프로젝트 과정**

**1. 데이터**

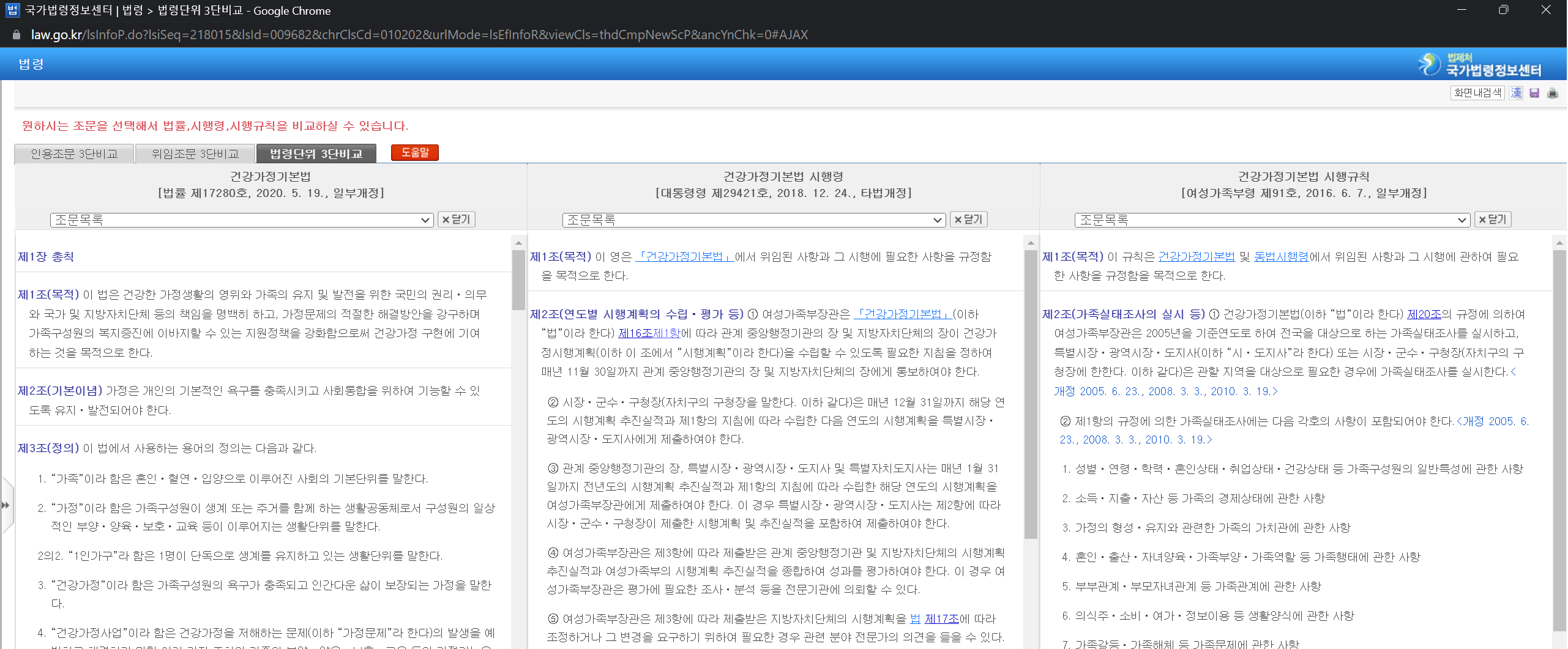
**(1) 데이터 출처**

데이터는 법제처에서 관리하는 국가법령정보센터 (https://www.law.go.kr)에서 수집하였다.



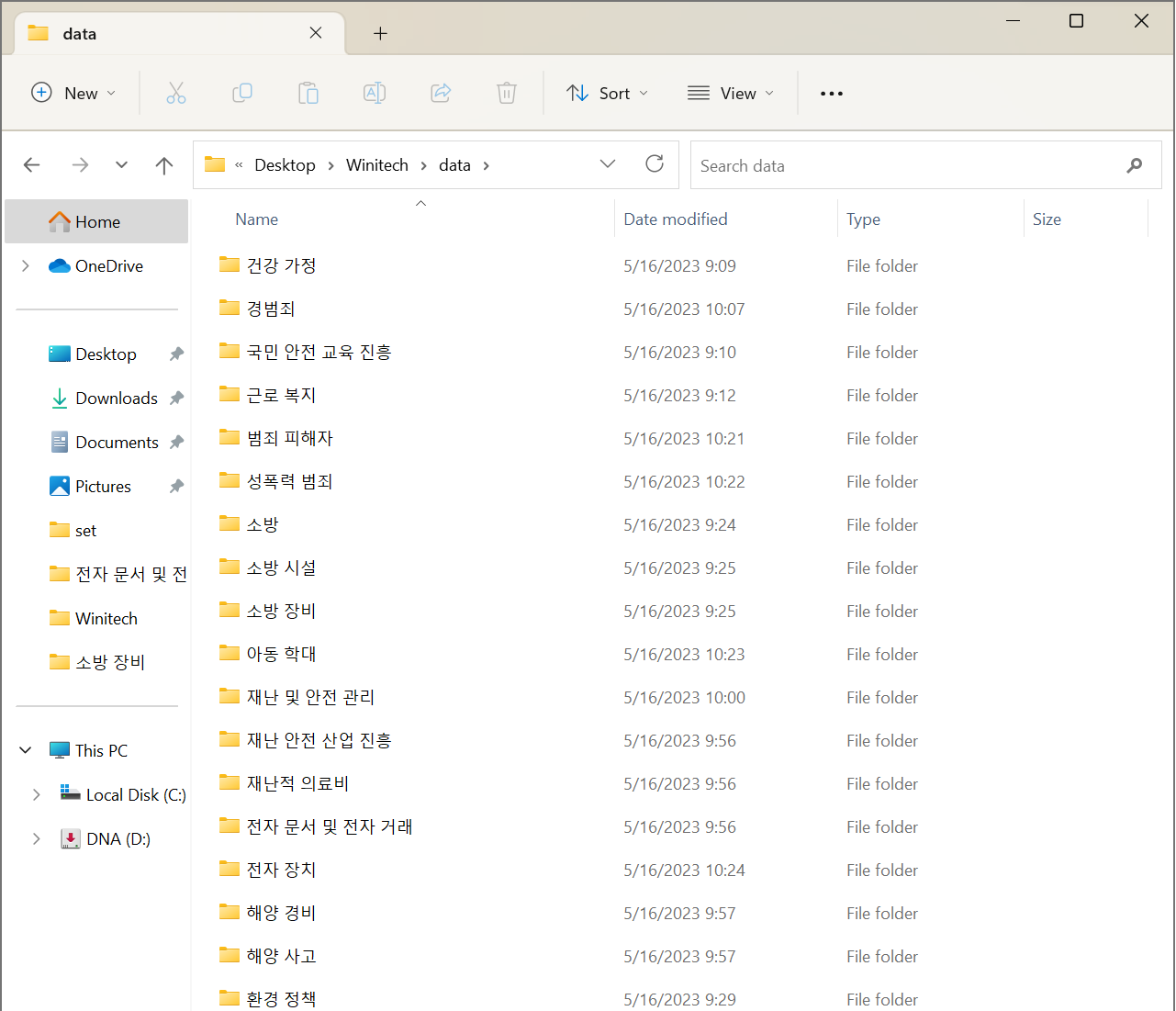
**(2) 데이터 수집 및 전처리**

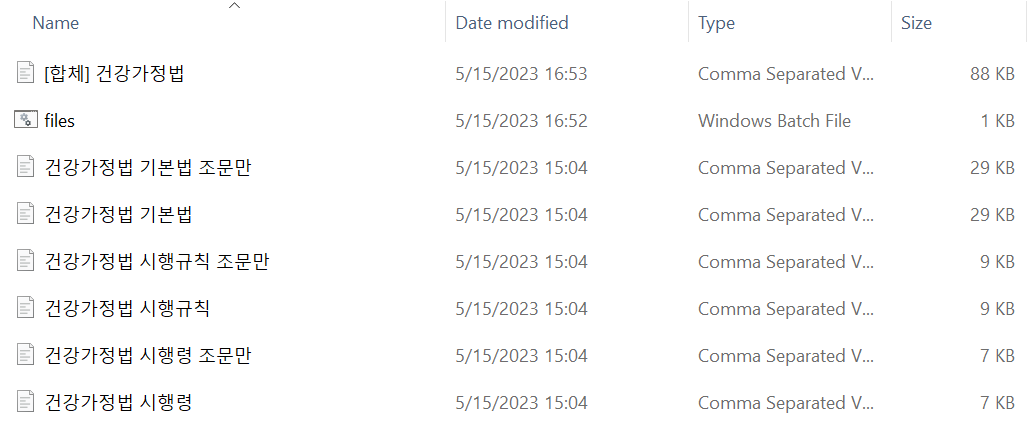
기본법, 시행령, 시행 규칙에 대한 비교가 필요하므로 크롤링 코드를 작성하여 각 법령 당 6개의 데이터를 수집하였다. 법령 전체와 조문 제목만 추출한 내용에 해당한다. 실제로 비교하게 되는 것은 조문 제목만이다.



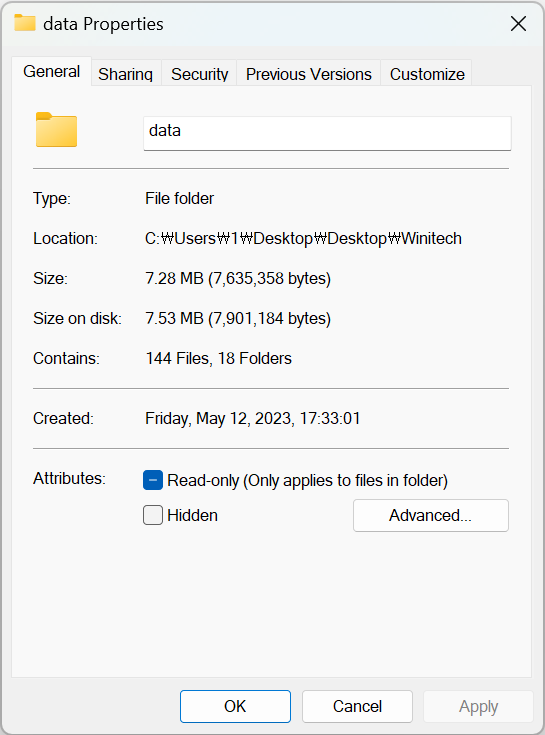


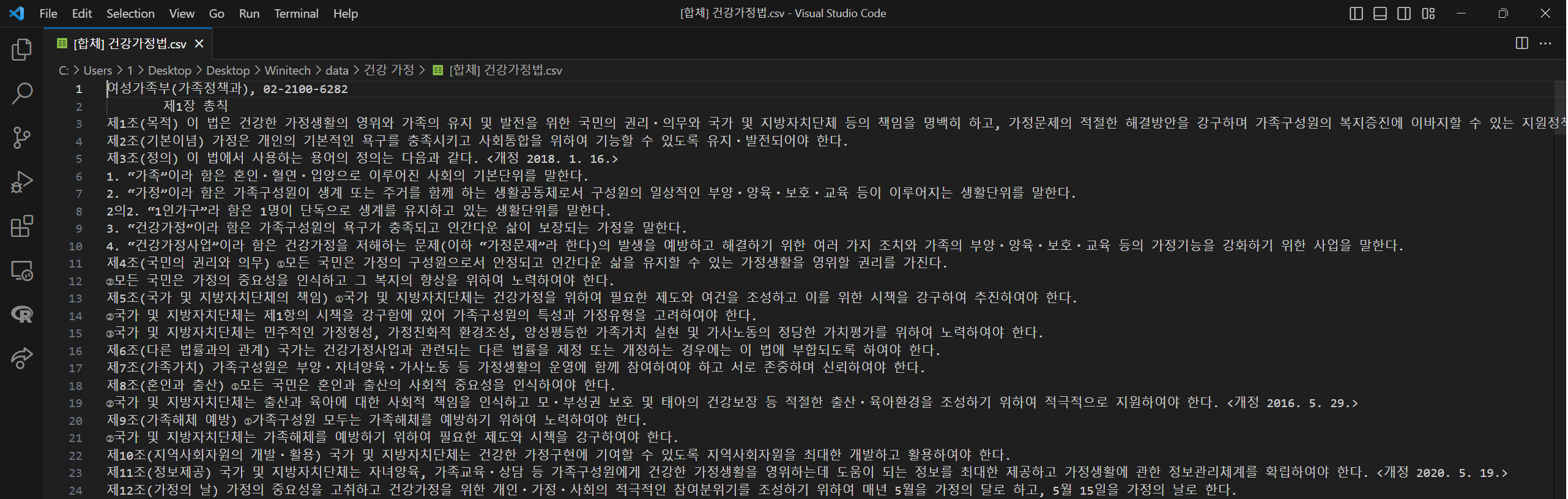
데이터는 소방, 재난, 안전 등 다양한 분야의 전체 18개 법령을 수집하였다.





공공 데이터이어서 결측값이 존재하지는 않았다.





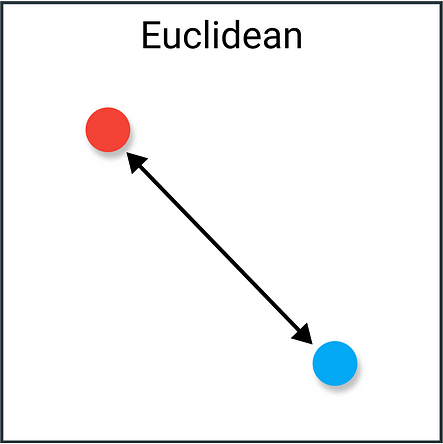
**2. 모델**

본 프로젝트의 주제는 자연어 처리 (Natural Language Processing, NLP) 기반 유사 항목 도출 모델 분석이다. 주제가 등장한 배경으로는 공공 기관의 재난 관련 솔루션 제공 기업으로서 재난 관리 매뉴얼과 관련된 프로젝트를 수행하였고 고객의 요청 사항 중 재난 관리 매뉴얼의 수직 비교에 대한 이슈가 있었다. 이에 대해 유사 항목 추출을 위한 자연어 처리 기반의 비교 모델이 필요한 것으로 판단하여 여러 가지 기법에 대한 정보가 필요하게 된 것이다.

그에 따라 Euclidean, Manhattan, Cosine 그리고 Jaccard의 4가지 방법으로의 접근을 모색하였다.

scikit-learn의 metrics 모듈의 euclidean\_distances, manhattan\_distances, cosine\_similarity 그리고 NLTK의 metrics 모듈의 jaccard\_distance를 활용하여 구현하였다. 먼저 비교하고자 하는 CSV 파일 2개에서 텍스트를 추출한다. 그런 다음 TfidfVectorizer를 통해 단어의 가중치를 조정한 BoW (Bag of Words) 벡터를 만든다. 그 후 해당하는 코드를 통해 유사도를 추출하게 된다.

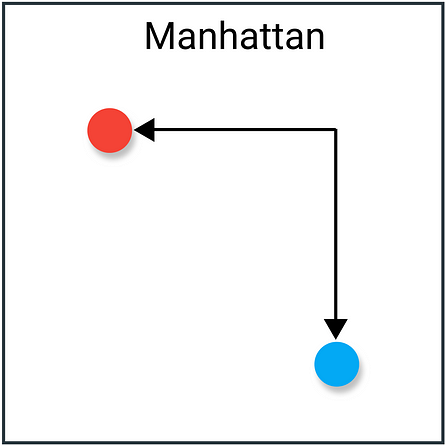
**(1) Euclidean**



**두 점 사이의 최단 거리를 구하는 방법**

**범위 없음 / 값이 작을수록 유사**

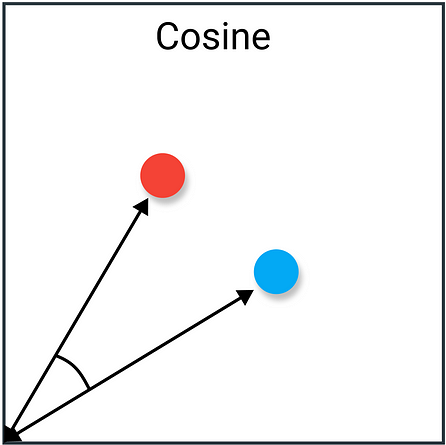
**(2) Manhattan**



**두 점 사이를 가로지르지 않고 갈 수 있는 최단 거리를 구하는 방법**

**값이 작을수록 유사**

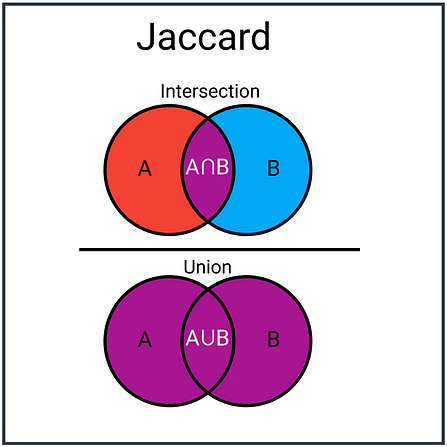
**(3) Cosine**



**두 벡터의 사잇각을 코사인으로 구하는 방법**

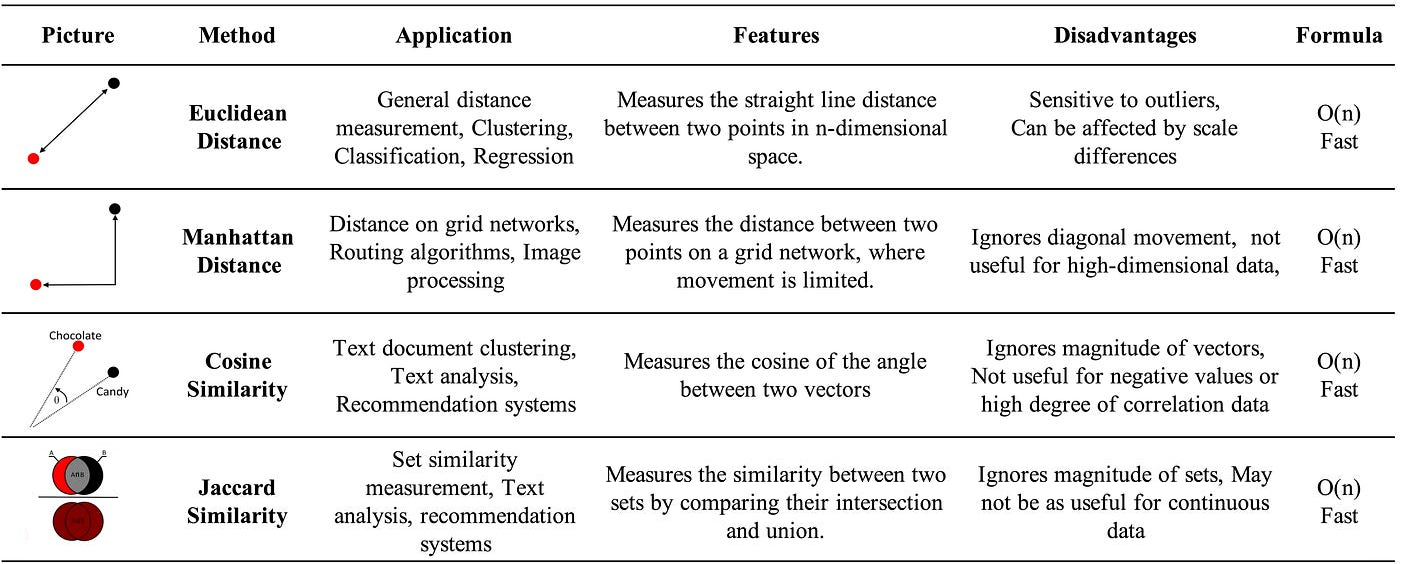
**-1 이상 1 이하 / 1에 가까울수록 유사**

**(4) Jaccard**



**두 집합 사이의 유사도를 측정하는 방법**

**0 이상 1 이하 / 1에 가까울수록 유사**



4가지 방법은 각각의 특징에 따른 장단점이 존재한다. 1) Euclidean은 계산하기가 비교적 간단하며 효율적이고 계산 속도가 빠르다. 수치형 데이터에 적합하다. 2) Manhattan은 이상치에 상대적으로 덜 민감하다. 범주형 데이터에 적합하다. 3) Cosine은 NLP 작업에서 텍스트 문서의 유사성을 비교하는 데에 널리 쓰인다. 그리고 4) Jaccard는 직관적인 해석과 성능 효율성, 이진 데이터 처리에 대한 장점이 있다.

이상 중간 발표 자료였습니다.