**가족 사회문제 관련 특허 데이터에서 추출한 Problem 엔티티 의미기반 그룹핑 및 BERTopic 적용**

**요약**

7,406건의 원본 데이터에서 Problem 엔티티를 선별하고, 이를 2,461개의 고유 엔티티로 정제한 후, 1,668개의 대표 엔티티를 기반으로 74개의 클러스터를 도출했습니다. 그 중 73개의 유효 토픽이 식별되었습니다.

**1단계 – 의미 기반 그룹핑 (Semantic Grouping)**

1. 데이터 선별 및 전처리:
   1. 원본 데이터 **7,406건**에서 Problem 엔티티만 필터링하여 **3,582건**의 데이터를 확보.
   2. 중복 제거 및 전처리 후 **2,461개**의 고유 엔티티 도출.
2. 임베딩 및 유사도 계산:
   1. 각 고유 엔티티를 all-MiniLM-L6-v2 모델로 임베딩하여 384 차원의 벡터로 변환.
   2. 코사인 유사도 계산을 통해 의미적으로 유사한 표현들을 그룹핑.
3. 군집화 및 대표명 선정:
   1. 임계값(threshold)을 실루엣 점수 기준으로 설정하여 군집화 최적화.
   2. 그 결과 **1,668개** 고유 대표명으로 군집화 완료.
   3. 각 군집의 대표명은 해당 군집 내에서 가장 높은 TF-IDF 점수를 가진 엔티티로 선정.

**2단계 – BERTopic**

1. BERTopic 입력 데이터:
   1. 1단계에서 도출된 **1,668개**의 대표 엔티티를 BERTopic에 입력 데이터로 사용.
2. 모델링 및 군집화:
   1. UMAP을 통한 차원 축소와 HDBSCAN을 이용한 군집화 수행.
   2. c-TF-IDF를 통해 각 군집에서 주요 단어 추출.
3. 클러스터링 결과:
   1. **1,668개**의 대표 엔티티를 입력으로 사용하여 총 **74개**의 클러스터가 생성.
   2. 그 중 1개의 노이즈 토픽(-1)은 의미가 명확하지 않거나 분류하기 어려운 데이터로 처리됨.
   3. 유효 토픽 수는 73개로 나타남.