논문 추천 시스템 개발 및 분석

팀 명 JK　　팀 원 김동욱　　지도교수 이슬　　멘 토

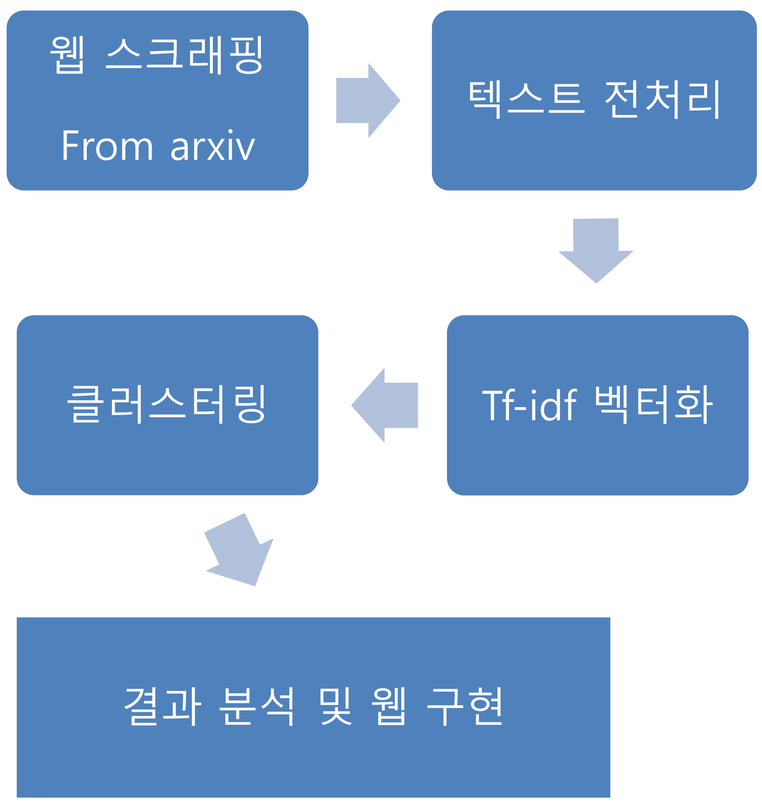
개발 동기 및 목적

최신 논문을 읽을 때 읽고 싶은 분야의 논문을 찾고, 흥미로워 보이는 논문들의 제목과 요약을 모두 다 읽는 것은 오래 걸리며, 힘든 일이다. 이런 문제를 해결하기 위해 사용자가 읽었던 논문 하나를 데이터로 사용하여 자동으로 논문을 추천하는 시스템을 구현하기로 하였다. 추천 시스템 중 하나인 콘텐츠 기반 추천 시스템의 경우 사용자가 이전에 읽었던 논문의 정보가 필요하다. 협업 필터링 바탕의 추천 시스템의 경우 다른 사용자가 읽은 논문 데이터가 필요하다. 하지만 클러스터링을 사용한다면 사용자 데이터 없이 추천 시스템을 구현할 수 있다. 또한 이 프로젝트는 완성도 높은 어플리케이션 구현보다 프로젝트를 진행하며 결과를 분석하여 텍스트 데이터에 대한 데이터 마이닝이 어떻게 진행되는지, 결과는 어떻게 나왔는지, 결과에 대한 분석을 목적으로 하고 있다.

개발 내용

논문의 텍스트 데이터를 얻기 위해 export.arxiv.org에서 request 라이브러리를 사용하여 최근 3달간 나온 컴퓨터 공학 분야의 논문의 html을 스크래핑 하였다. 이 프로젝트는 논문의 텍스트 중 논문의 요약(abstract)에 해당하는 텍스트를 데이터로 사용하기 때문에 수집된 html에서 논문의 요약 부분을 추출하였다. 얻어낸 텍스트 데이터에서 특수 기호를 제거하고, stop words와 max document frequency를 사용하여 간단한 전처리를 진행하였다. 전처리 된 텍스트에 tf-idf 벡터화를 하여 텍스트 벡터들을 얻었다. 마지막으로 minimum document frequency와 분산 기여도를 바탕으로 차원 축소를 진행하고 클러스터링을 진행하였다. 이때 파라미터로 min df와 분산 기여도, 클러스터링 방법의 종류에 따라 실험을 진행하였다.

(개발 순서)

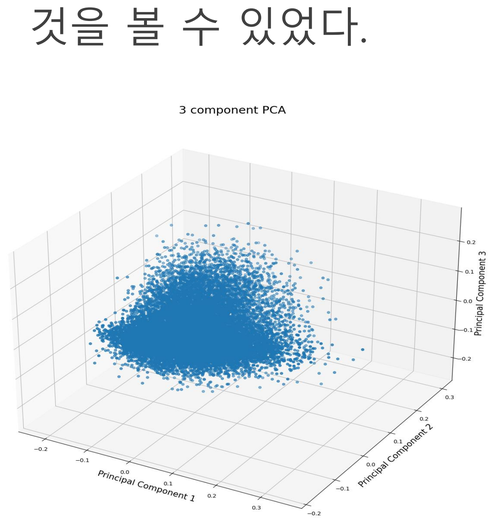


주요기술

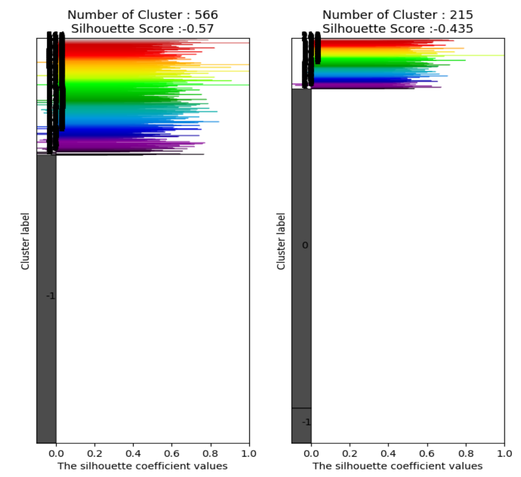
이 프로젝트는 Python 3.9.7에서 진행되었다. 웹을 통해 논문 데이터를 얻기 때문에 request 라이브러리를 사용하였고, 얻어낸 html 데이터 처리를 위해 bs4를 사용하였다. Html 데이터에서 얻은 텍스트 데이터를 처리하고 차원 축소와 클러스터링을 진행하기 위해 sklearn을 사용하였으며 웹으로 배포하기 위해 flask를 사용하였다. 차원 축소 기법으로는 분산 기여도를 계산하여 주성분 분석(PCA)과 특잇값 분해(SVD)를 사용하였다. 클러스터링으로는 k-최근접 이웃과 계층적 클러스터링, 밀도 기반 클러스터링을 사용하였다. 또한 클러스터 모델의 평가 지표로 실루엣 계수를 사용하였다.

결과 및 분석

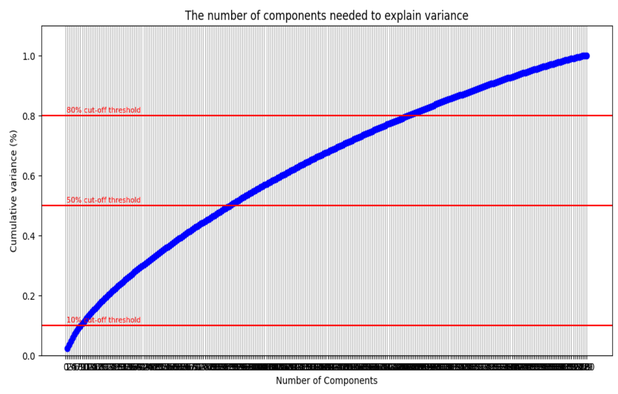
높은 차원의 tf-idf 벡터에 대한 클러스터링이라 대부분의 실험에서 직관적으로 좋은 결과를 나타내지 않았다. 특히 k-최근접 이웃과 계층적 클러스터링의 경우 직접 클러스터의 수를 정하는 문제가 있을 뿐 아니라 실루엣 계수 또한 상당히 낮게 나왔다. 밀도 기반 클러스터의 경우 클러스터의 수를 정하지 않아도 되지만, 차원 축소를 진행하여도 높은 차원의 벡터를 사용하여 클러스터가 잘 진행되지 않았으며, 특히 노이즈로 판별되는 클러스터가 절반 이상을 차지하였다. 또한 클러스터의 수가 보통 200개에서 많게는 1,000개가 나왔고, 대부분의 클러스터에는 2~3개의 데이터가 들어 있었다. 적은 양이라고 생각되는 최근 3개월의 논문을 사용하며, 논문의 특성상 이미 해결된 문제는 다시 짚지 않는 점에서 이러한 결과가 나왔다고 생각한다. 또한 모든 벡터의 분산 기여도가 낮게 나왔는데 데이터의 tf-idf 벡터가 차원이 너무 커서 분산 기여도가 전체적으로 낮게 나온 것으로 분석된다. 이는 3차원으로 PCA를 수행하여 데이터를 시각화 했을 때도 눈에 띄는 데이터 패턴이 보이지 않은 것으로 확인할 수 있었다. 하지만 클러스터에 성공된 데이터의 경우 실루엣 계수도 높게 나왔으며, 실제로 논문의 내용도 유사한 것을 보이며 클러스터가 가능한 것을 볼 수 있었다.



(데이터의 시각화)



(클러스터의 실루엣 계수 그래프)



(Tf-idf 벡터의 분산 기여도)