영화 리뷰 추천 시스템.

팀 명 리뷰 분석 팀 팀 원 박범수 지도교수 이슬 멘 토 이슬

개발 동기 및 목적

코로나 이후 비대면 시기와 인터넷 시장의 발달로 인해 리뷰의 수는 무수히 늘어나고 있고, 고객들이 리뷰에서 정보를 찾고자 하는 모습들이 많이 관측되지만, 현재 리뷰제공의 경우에는 단순히 시간순으로만 제공되고 있다. 이러한 제공 방식은 유저가 원하는 성격의 리뷰를 찾기 힘들뿐만 아니라, 리뷰이벤트와 같은 특수한 경우에는 리뷰의 질은 떨어지나 양은 늘어나게 되므로 이러한 경우를 고려하지 못한다. 이러한 부분들로 인해 현재 리뷰 제공 방식은 많은 리뷰데이터의 경우에 적합하지 않다고 판단하여, 이러한 부분들을 해결하고자 리뷰 추천시스템을 새로 구현하고자 하였다. 현재의 시간 순 정렬 리뷰는 많은 양의 리뷰가 있는 사이트의 경우 유용한 리뷰가 오래전 리뷰일 경우에는 유용한 리뷰들에 대해서 읽기가 힘들기 때문에, 이러한 부분을 고칠 수 있는 부분에서 강점이 있다고 생각된다. 또한 영화 리뷰뿐만이 아닌 다른 모든 리뷰에 대해서 확장할 수 있는 범용성이 장점이라고 생각된다.

.

주요기술

Content Based Recommendation

벡터화된 리뷰들을 content로 삼아 유저가 보고자 하는 리뷰에 대해서 모든 리뷰들에 대해 거리를 구한 뒤 거리순으로 나열하는 단순한 방식이다.

2. DBSCAN 알고리즘

밀도 기반 클러스터링으로 밀집도가 높은 부분을 활용하여 이러한 부분을 인식하고 클러스터링하는 방식이다. 클러스터의 수를 정하지 않아도 되며, 클러스터의 밀도에 따라서 클러스터를 서로 연결하기 때문에 기하학적인 모양을 갖는 군집도 잘 찾을 수 있다. 특히 Noise point를 통하여, outlier 검출이 용이하다는 장점이 존재한다.

개발 내용

네이버 영화 사이트를 크롤링하여 데이터를 수급하였다. 이러한 리뷰들은 문장형태의 데이터이며, 누구나 간단하게 쓸 수 있는 구조라, 체계적이고 차분하게 적혀진 글이 아니라 난잡하고, 이모티콘이나 자음 모음 그 자체로 쓰는 형태의 데이터, 아웃라이어가 굉장히 많고 난잡한 형태의 데이터이기 때문에 학습에 대한 고민 때문에 이모티콘, 자음, 모음, 불용어 등등을 모두 전처리 과정에서 잘라내고, 명사만 추출하여 순서를 고려하지 않는 단어 집합 형태로 가공하여 토큰화 하였다. 이후 단어의 빈도수와 가중치를 고려하여 tf-idf 벡터화 후 nomalization 과정을 거쳐 전처리를 마무리하였다. 이후 dbscan 알고리즘을 통해 각각의 리뷰를 유형별로 클러스터링 하였고, 비슷한 리뷰를 제공하거나, 다른 유형별로 다양성을 중심으로 리뷰를 제공하였다. 다음으로 content based recommendation 방식을 통해 유저가 원하는 리뷰를 고르면 유사한 리뷰별로 출력하는 방식을 구현하였다.

결과 및 분석

DBSCAN 알고리즘을 통해 구현한 경우 1500개 정도의 리뷰를 사용한 결과 16개의 클러스터로 구분되었고,각각의 개성이 뚜렷하게 구분되었다. 장단점을 설명하는 리뷰, 영화의 별점과는 다르게 재미있다는 역설적 성격의 리뷰, 제거해야 하는 노이즈 형태의 리뷰들 로 각각의 특성들이 확연하게 구분되었다. 노이즈가 자동적으로 제거된 이후에도 특히 단순한 형태의 칭찬식의 리뷰들이 다수 있어 가치가 떨어진다고 판단되어 제거해야 되는데, 이러한 형태의 리뷰들이 같은 클러스터로 뭉치기 때문에 처리하기 용이하였다.그러나 소규모의 인터넷 밈이나, 어떠한 공동체의 단어들이 자주 사용된 리뷰같은 경우에는 노이즈로 취급되는 단점이 존재함 그러나 평범한 인터넷 리뷰 환경에서 선호하는 리뷰에 유사한 리뷰들을 추천한다는 목적에 있어서는 단점을 뛰어넘는 장점이 존재한다고 판단되었다.
다음으로 Content Based Recommendation 방식으로 구현한 경우 비슷한 문장들을 추천하는 모습을 보이나 위의 DBSCAN 처럼 어떠한 경향으로 확실히 구분 되지는 않았다. 단순한 형식의 알고리즘이라, 위의 DBSCAN 보다 더 비슷하고 유용하게 제공한다고는 할 수는 없었으나 소규모의 크기를 가지는 사이트의 경우에는 적합하다고 판단되었다. 독특한 표현이나 같이 소규모의 인터넷 밈이나, 어떠한 공동체의 단어들이 사용된 리뷰같은 경우도 포함하여 검색할 수 있다는 장점 존재한다. 이러한 방식은 특정 밈이나, 소규모의 밈들이 자주 사용되는 리뷰들이 모여있는 사이트의 경우에 이러한 사이트의 목적이 정보보다는 웃음인 부분 또한 고려한다면 content-based-recommandation이 적합하다고 판단되었다.