**SentiWordNet을 이용한 감정분석**

**안진환1 정병기2 차경준3**

**1 2 3국민대학교 컴퓨터공학부**

**ajh4652@gmail.com, dennischa50@gmail.com, qudrl0284@gmail.com**

**Sentiment Analysis using SentiWordNet**

**JinHwan An1 ByeongGi Jung2 KyungJun Cha3**

**1 2 3School of Computer Science, Kookmin University**

**요 약**

사람의 감성을 분석하기 위한 알고리즘들이 많이 연구되고 있다. 그 중에서도 우리는 텍스트 데이터에서 사람들의 감성을 분석하는 알고리즘을 사용한다. 사람들이 작성한 상품 리뷰들의 텍스트 데이터를 긍정적인 것과 부정적인 것으로 점수를 매긴다. 그리고 그 결과와 상품의 판매량과의 상관관계를 분석해본다. 분석 결과 사람들의 상품 리뷰가 긍정적으로 분석된 상품은 판매량이 높게 나타났고 부정적으로 분석된 상품은 판매량이 상대적으로 저조하게 나타났다. 이 결과를 통해 상품의 질을 향상시키는데 도움이 될 수 있다고 판단된다.

**1. 서론**

우리는 텍스트 데이터에서 사람들의 감성을 분석하는 알고리즘을 사용하여 상품 리뷰들을 긍정적인 것과 부정적인 것으로 나누어 점수를 매긴다. 그리고 그 결과와 상품의 판매량과의 상관관계를 분석해본다. 상품 리뷰들을 분석하기 위해 필요한 데이터들은 아마존 상품의 데이터세트를 사용했다. 텍스트에서 감성을 분석하기 위해서는 감성 단어 사전이 필요한데, 대표적인 영어 감성 단어 사전인 SentiWordNet을 사용했다.

**2. 본론**

먼저 hadoop의 multipleInputs 라이브러리를 통하여 reviews와 metadata로부터 각각 {asin, reviewText}와 {asin, salesRank}를 받아온다. 이번 프로젝트의 사용자료는 books 카테고리의 상품들에 관한 데이터이므로 별도의 카테고리 분류 없이 salesRank의 값만을 뽑아왔다.

그 다음 파싱 된 데이터를 map에 넣은 뒤 reduce 작업에서 reviewText인지 salesRank인지를 분류하여 각각의 변수에 저장한다.

reviewText는 텍스트의 감정 분석 알고리즘을 적용하여 SentiWordNet 사전의 점수로 변환한 뒤 더하여 -10~10에 해당하는 점수로 변환한다. 이 때 하나의 상품에 대해 여러 개의 리뷰가 있으므로 평균값을 내어 각 상품별로 리뷰에 대한 감정점수 평균, 세일즈 랭크를 갖게 한다.

**3. 결론**

최종적으로 {asin. 최종감정점수 :: salesRank}의 형식으로 저장된다. 이 정보는 각 리뷰들에 대한 감정점수를 판매랭크를 비교하여 리뷰들과 판매랭크의 신뢰도를 구하는데 사용 할 수 있을 것이다.

**4. 참고문헌**

1. http://sentiwordnet.isti.cnr.it/

2. https://stanfordnlp.github.io/CoreNLP/

**3.** https://hadoop.apache.org/docs/r2.7.1/api/org/apache/hadoop/mapreduce/lib/input/MultipleInputs.html