웹 크롤링 및 머신러닝을 활용한 감성 분석 시스템 구현

송승엽, 박희재, 박래혁

서울과학기술대학교 컴퓨터공학과

{thdtmdduqdhk, prkhj98, lhpark}@seoultech.ac.kr

Implementation of Sentimental Analysis System using Web Crawling and Machine Learning

Seungyeop Song, Heejae Park, Laihyuk Park

Department of Computer Science and Engineering

Seoul National University of Science and Technology

요 약

현대 정보화 사회에서 데이터의 양은 기하급수적으로 증가하고 있으며, 이러한 대량의 데이터에서 유용한 정보를 추출하는 것에 대한 필요성이 증가하고 있다. 본 논문에서는 셀레니움을 사용한 웹 크롤링 기술과 로지스틱 회귀 분석을 활용한 머신러닝 기법을 결합하여, 웹에서 자동으로 정보를 수집하고 이를 분석하는 시스템을 구현하였다.

Ⅰ. 서 론

인터넷이 일상생활의 필수 요소가 되면서, 우리는 매일 방대한 양의 정보에 노출되고 있다. 하지만 이렇게 많은 정보 속에서 필요한 정보를 찾아내고 이해하는 것은 시간이 많이 소요되는 작업이다. 따라서, 자동으로 정보를 수집하고 분석하여 사용자에게 제공하는 시스템의 필요성이 증가하고 있다. 본 논문에서는 웹사이트에서 특정 장소에 대한 방대한 리뷰를 크롤링하고 이에 대한 감성을 분석하여 사용자가 장소에 대한 정보를 빠르게 습득할 수 있도록 만드는 시스템을 제안한다. 제안된 시스템은 웹 크롤링과 머신러닝 기법을 결합하여 효율적으로 정보를 수집하고 분석하는 과정을 수행한다.

Ⅱ. 본 론

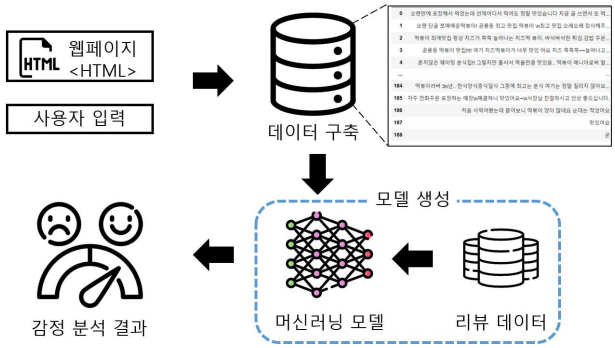

그림 1. 감성 분석 시스템 개념도

Fig 1. The conceptual diagram of information summary system

그림 1은 감성 분석 시스템 개념도를 보여준다. 시스템은 웹페이지를 크롤링하여 데이터를 구축하고 이를 머신러닝 모델에 적용하여 분석된 결과를 출력하는 과정을 수행한다. Python 언어로 개발했으며, 먼저 웹 크롤링 도구로 BeautifulSoup와 셀레니움을 사용하여 리뷰 정보를 추출 및 수집하였다 [1, 2]. 셀레니움은 웹 브라우저를 크롤링할 수 있는 도구로, 복잡한 동적 웹페이지에서도 효과적으로 데이터를 수집할 수 있다. 또한 셀레니움을 활용한다면 특정 장소에 대한 여러 사용자들의 리뷰, 방문평 등에 대한 정보를 포함하는 웹페이지를 탐색하고, 필요한 데이터를 추출하는 과정을 자동화할 수 있다. 그림 2는 셀레니움을 활용하여 예시로 추출한 데이터이며, 첫 번째 열에는 데이터의 순서를, 두 번째 열에는 리뷰 내용을 저장하여 데이터셋을 구축했다.

0	맛있고 푸짐하고 소량씩 주문 가능해요. 멋모르고 혼 자가서 이것저것 많이 시켰는데,...
1	맛있고 친절하고 포장+현장식사 동일하게 대기는 30 분 기본. 애들 먹기에 매울갓 ...
2	지인소개로 근처갔다가 들렸네요 예상보다 손님도 많 고 전화예약등등..웨이팅후 포장해...
3	30-40분 기다려서 먹었어요□\n홀이랑 포장 손님 엄청 많아서 바쁘시던데\n친절히...
4	오랜만에 포장해서 먹었는데 언제어디서 먹어도 정말 맛있습니다 지금 글 쓰면서 또 먹...
...	...
225	맛있어요
226	다 먹어보진 못 했은데 먹어본 건 다 맛있고 친절합니다.\n줄 서 있을 때도 많대요.
227	저번에 사먹고 계속 생각나서 또 왔어요 파듬뿍에 어 묵도 많이 주고 양도 많아요 특...
228	늘 줄서서 이용하는 맛집 !! 사장님늘 친절하셔서 갈때마다 기분도 좋아져요!! 맵기...
229	맛있어요 친절해요

그림 2. 추출 데이터 예시

Fig 2. Extracted data example

다음으로 머신러닝을 이용하여 데이터셋을 분석했다. 추출된 데이터의 감성 분석을 위해 로지스틱 회귀 (Logistic Regression) 기반의 머신러닝 모델을 사용하였다. 로지스틱 회귀는 분류 문제에 널리 사용되는 방법으로, 본 연구에서는 이를 활용하여 수집된 데이터 중 사용자가 긍정의 감성을 가지고 있는지, 부정의 감성을 가지고 있는지에 대한 감성분석을 수행하였다 [3]. 특히, 텍스트 데이터의 특성을 벡터 형태로 변환하는 과정에서 TF-IDF (Term Frequency-Inverse Document Frequency) 방식을 적용하여, 정보의 중요도를 평가하는 기준을 마련하고 분석을 진행하였다 [4]. 머신러닝 모델에는 라벨링이 완료된 15000개의 리뷰 데이터셋을 학습데이터로 사용했다. 학습을 위해 데이터의 결측 값과 한글 외의 문자를 제거

하고 Train 용 데이터셋과 Test 용 데이터셋으로 분리한 후 토큰화와 TF-IDF 벡터화를 진행했다. 그리고 전처리한 데이터를 로지스틱 회귀모델에 학습시켜서 감성분석 모델을 구축했다. 그리고 학습된 모델을 이용하여 리뷰데이터를 분석하고 이에 대한 결과를 출력했다. 예시로 주어진 데이터에 대한 감성 분석 결과, 5%는 부정적인 리뷰였고 95%는 긍정적인 리뷰로 나타났다. 따라서 이 장소는 사용자에게 추천할 만하다라는 결론을 이끌어낼 수 있었다.

Ⅲ. 결 론

본 논문에서는 셀레니움과 로지스틱회귀를 활용해 웹 크롤링 기술과 머신러닝 기법을 결합한 감성 분석 시스템을 구현하였다. 이 시스템은 사용자가 필요로 하는 정보를 자동으로 수집하고 분석함으로써, 정보 검색과 이해 과정을 효율적으로 지원할 수 있다. 향후 연구에서는 더 다양한 머신러닝 알고리즘을 결합하는 앙상블 방식을 활용하여 시스템의 성능을 향상시키는 연구를 진행할 예정이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2024-RS-2022-00156353)

참 고 문 헌

[1] M. Rong, S. Zhang and E. Yi, “Big Data Collection and Building Material Price Estimation Based on Focused Web Crawler,” 2023 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Beijing, China, 2023, pp. 1-4.

[2] A. Pan and H. Pan, “Design and Implementation of Web Crawler System Based on Python,” 2023 8th International Conference on Information Systems Engineering (ICISE), Dalian, China, 2023, pp. 91-94.

[3] Tushar, R. K. Patel, E. Aggarwal, K. Solanki, O. Dahiya and S. A. Yadav, “A Logistic Regression and Decision Tree Based Hybrid Approach to Predict Alzheimer's Disease,” 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2023, pp. 722-726.

[4] P. Bafna, D. Pramod and A. Vaidya, “Document clustering: TF-IDF approach,” 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, India, 2016, pp. 61-66.