**T셀파 연수원과 아이스크림 연수원의 연수후기 비교분석**

빅데이터 4기

김성민 오한나 울지바야르 이예원

**[요약]**

경쟁업체인 아이스크림 연수원과의 리뷰 비교를 통하여 어떤 점이 다른지 파악하고자 T셀파와 아이스크림 연수원 홈페이지의 연수후기를 각각 크롤링 하였다. 긍정적인 후기 수가 98% 이상이었기 때문에, 별점을 기준으로 분석하는 것은 의미가 없다고 판단하였다. 그렇기에 감정 사전의 점수를 기준으로 긍정과 부정 단어를 분리하고 시각화하여 비교했다. 그 결과 긍정적인 단어는 ‘이해’가 가장 높았고, 부정적인 단어는 ‘부담’이 가장 높았다. 또한 T셀파가 아이스크림 연수원에 비해 긍정 단어의 비중이 더 높은것을 확인했다. 다만 전체 리뷰를 대상으로 진행한 것이 아니라 정확성이 높지 않다는 한계를 가지고 있다.

**[서론]**

경쟁 업체와의 비교를 통하여 서비스 개선 등 시장 점유율을 보다 높이기 위하여 어떤 점이 다른지 파악하고자 하였다. T셀파와 아이스크림 두 연수원의 리뷰를 비교하여 고객의 만족도가 어느 부분에서 어떻게 다른지 파악하여 강의의 질을 높이는데에 도움이 될 목적으로 진행하였다.

**[연구방법]**

1. **조사대상**

T셀파 연수원의 연수후기와 아이스크림 연수원의 연수후기를 2022년1월1일~2023년7월28일까지의 데이터를 사용했다.

1. **데이터 수집**

개발언어는 Python을 사용하였으며, webdriver, selenium, pandas 라이브러리와 time 모듈을 사용하여 T셀파와 아이스크림 연수원의 연수후기를 각각 크롤링하였다.

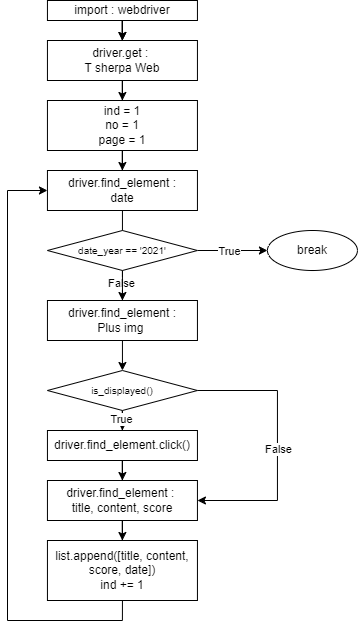
1. **분석 방법**

연수후기를 크롤링한 데이터는 리스트에 저장한 후 데이터프레임에 넣어 csv로 저장하였다. 아이스크림 연수원의 연수후기가 T셀파의 연수후기보다 수가 압도적으로 많아 샘플 수를 동일하게 1,000개로 지정하였다. 별점을 살펴본 결과, 4점과 5점의 비율이 98% 이상으로 별점 분석은 의미가 없다고 판단되어 전체 후기를 하나의 문장으로 합쳐 각각의 연수원 전체 후기 분석을 통하여 긍정 단어와 부정 단어의 수를 비교하고 시각화하고자 한다.

분석 전 전처리를 위해 합쳐진 후기 내용에서 한글을 제외한 문자와 기호 등을 제거하고 2글자 이상의 명사 형태소만 추출하였다. ranks.nl에서 제공하는 한국어 불용어 사전과 추가로 의미가 없다고 판단한 불용어를 추가하여 명사 리스트에서 불용어를 삭제하였다. 이후 KNU 한국어 감성 사전에 적힌 명사만 따로 추출하여 데이터 프레임에 저장하였다. 한국어 감정 사전의 감정 점수는 -2, -1, 0, 1, 2로 구성되어 있지만, 점수의 절대값과 중립인 0점은 고려하지 않았다. 양수는 긍정, 음수는 긍정으로 분류하고 긍정과 부정의 비율과 가장 많이 나온 단어를 확인하고 워드클라우드를 그려 비교하였다.

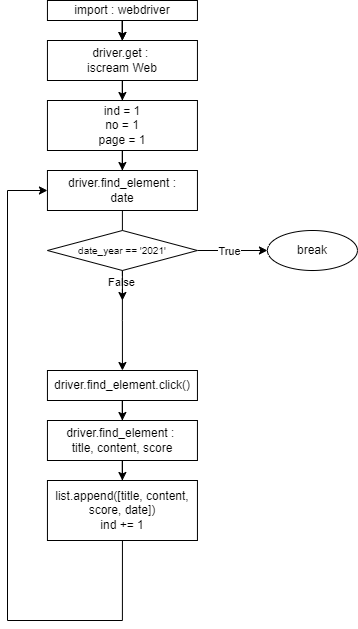
**[연수원 데이터 크롤링]**

그림1은 T셀파 연수원 크롤링 작업 흐름도이다. 크롤링은 Chrome Driver를 사용하였으며 등록일을 확인하여 2021년이면 반복문이 종료되도록 하였다. T셀파 홈페이지의 연수후기는 후기를 클릭하면 새 창이 뜨고 그 새 창에 들어가야 긴 후기를 읽을 수 있었다. 긴 후기를 크롤링하는 과정에서 시간이 많이 소요되므로 긴 후기와 짧은 후기로 나누어 크롤링 코드를 설계하였다. 긴 후기에는 플러스 이미지가 있는데 우선 이미지가 있는지 확인하고, 있으면 클릭하여 새 창으로 넘어가 제목, 내용, 별점을 크롤링하고 마치면 원 페이지로 넘어가 다음 후기에 이미지가 있는지 확인한다. 이미지가 없으면 새 창으로 넘어가지 않고 보이는 페이지에서 같은 내용을 크롤링한다. 하나의 후기에서 등록일, 제목, 내용, 별점까지 모두 수집하면 바로 리스트에 데이터를 저장한다.



<그림1. T셀파 연수원 크롤링 작업 흐름도>

다음 그림2는 아이스크림 연수원 크롤링 작업 흐름도이다. 크롤링은 Edge Driver를 사용하였으며 기본적인 작업 흐름도는 T셀파와 동일하다. 다만 아이스크림 연수원은 긴 후기와 짧은 후기를 구분할 필요가 없으므로 해당 작업만 제외하였다.

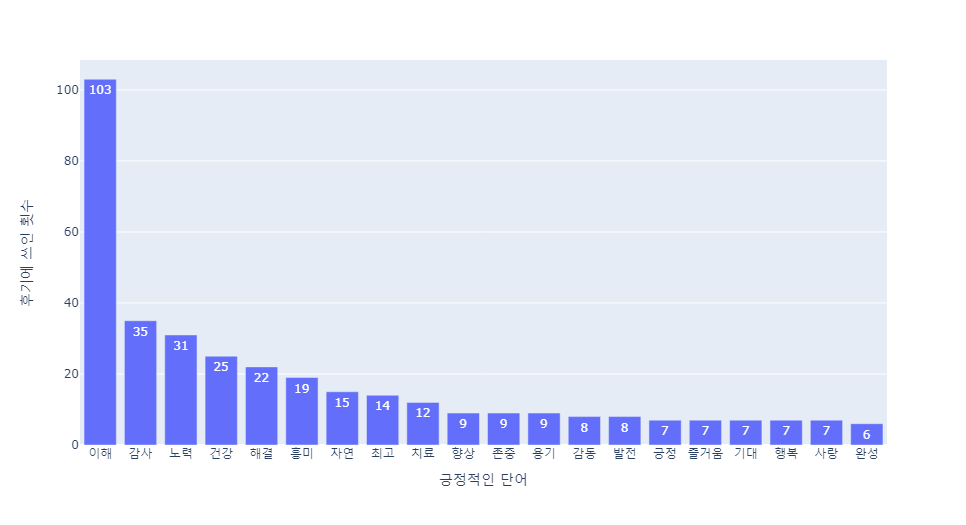


<그림2. 아이스크림 연수원 크롤링 작업 흐름도>

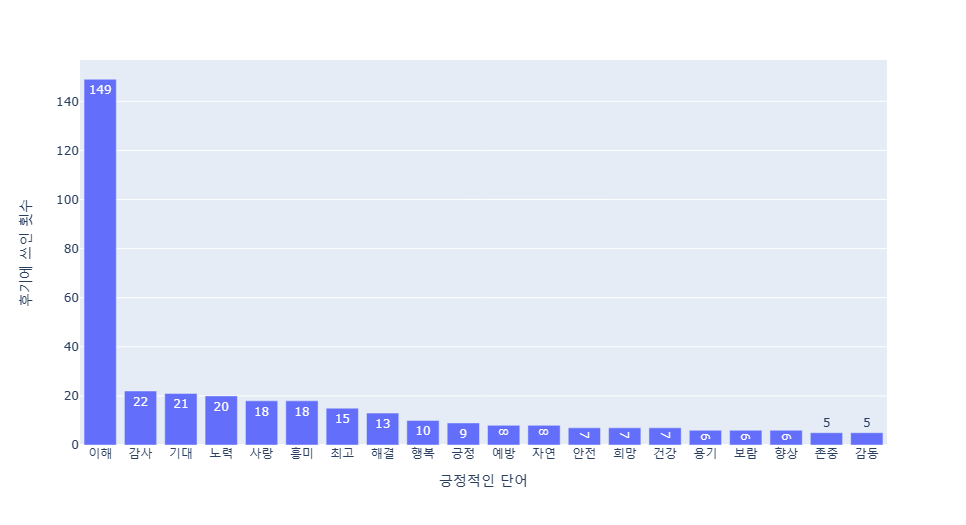
**[연수원 데이터 시각화]**

1. **T셀파와 아이스크림의 긍정적인 단어 빈도수**

T셀파와 아이스크림의 긍정적인 단어 빈도수를 각각 나타내어 확인한 결과 이해, 노력, 감사 등의 단어가 공통적으로 많이 나타났다.

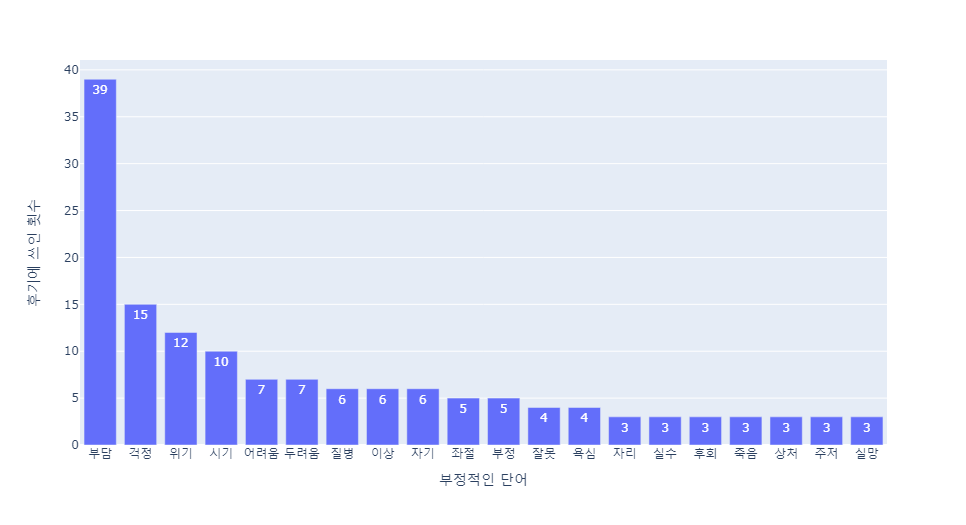


<표1. T셀파-긍정적인 단어 빈도수 순위>

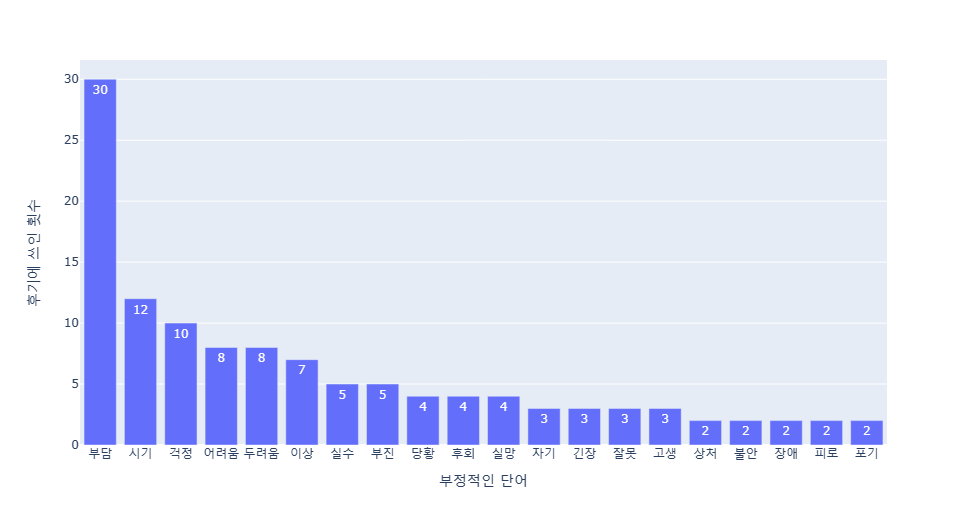
<표2. 아이스크림-긍정적인 단어 빈도수 순위>

1. **T셀파와 아이스크림의 부정적인 단어 빈도수**

T셀파와 아이스크림의 부정적인 단어 빈도수를 각각 나타내어 확인한 결과 부담, 시기, 걱정, 어려움, 두려움 등의 단어가 공통적으로 많이 나타났다**.**



<표3. T셀파-부정적인 단어 빈도수 순위>



<표4. 아이스크림-부정적인 단어 빈도수 순위>

1. **T셀파와 아이스크림의 긍정적인 단어 워드클라우드**

그림 3과 그림4는 T셀파와 아이스크림의 긍정적인 단어를 Wordcloud로 표현을 한 것이다. T셀파에서는 이해, 노력, 건강, 감사, 흥미 순이지만 워드 클라우드에서는 감정점수가 반영되어 크게 나온 글자로는 이해, 노력, 감사, 건강으로 나왔다. 아이스크림에서는 후기에 쓰인 빈도가 높은 단어로는 이해, 감사, 기대, 노력, 사랑 순이지만 워드 클라우드에는 위와 마찬가지 이유로 이해, 감사, 사랑, 노력 단어들이 크게 나왔다.



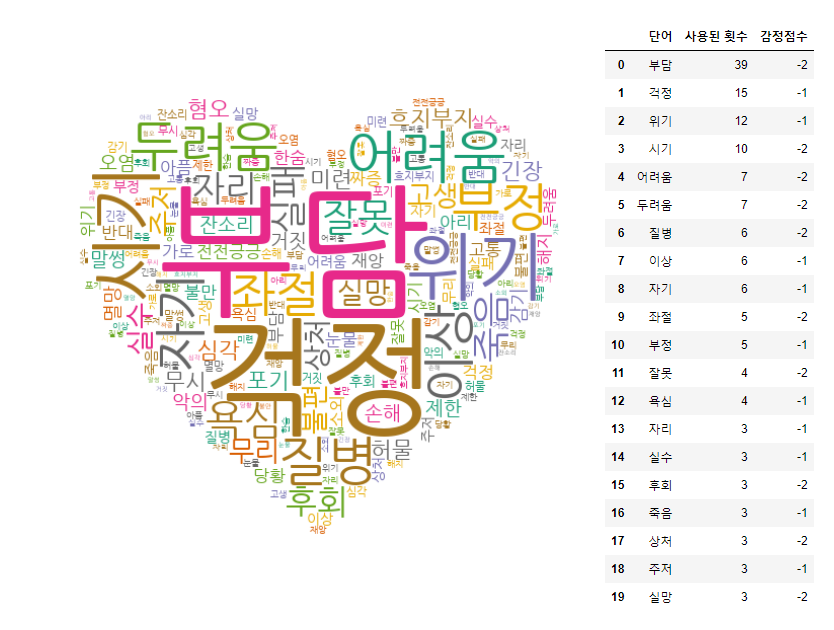
<그림3. T셀파-긍정적인 단어 Wordcloud>



<그림4. 아이스크림 긍정적인 단어 Wordcloud>

1. **T셀파와 아이스크림의 부정적인 단어 워드클라우드**

그림 5과 그림6은 T셀파와 아이스크림의 부정적인 단어를 Wordcloud로 표현을 한 것이다. T셀파에서는 부담, 위기, 걱정, 시기 순이지만 워드 클라우드에서는 감정점수가 반영되어 크게 나온 글자가 부담, 위기, 걱정, 좌절, 이상으로 나왔다. 아이스크림에서는 후기에 쓰인 빈도가 높은 단어로는 부담, 시기, 걱정, 어려움이지만 워드 클라우드에는 위와 마찬가지 이유로 부담, 부진, 어려움, 걱정 단어들이 크게 나왔다. 다만 긍정적인 단어와는 다르게 부정적인 단어는 T셀파와 아이스크림이 서로 유사한 단어가 더 많이 나왔다는 것을 알 수 있다.



<그림5. T셀파 부정적인 단어 Wordcloud>



<그림6. 아이스크림 부정적인 단어 Wordcloud>

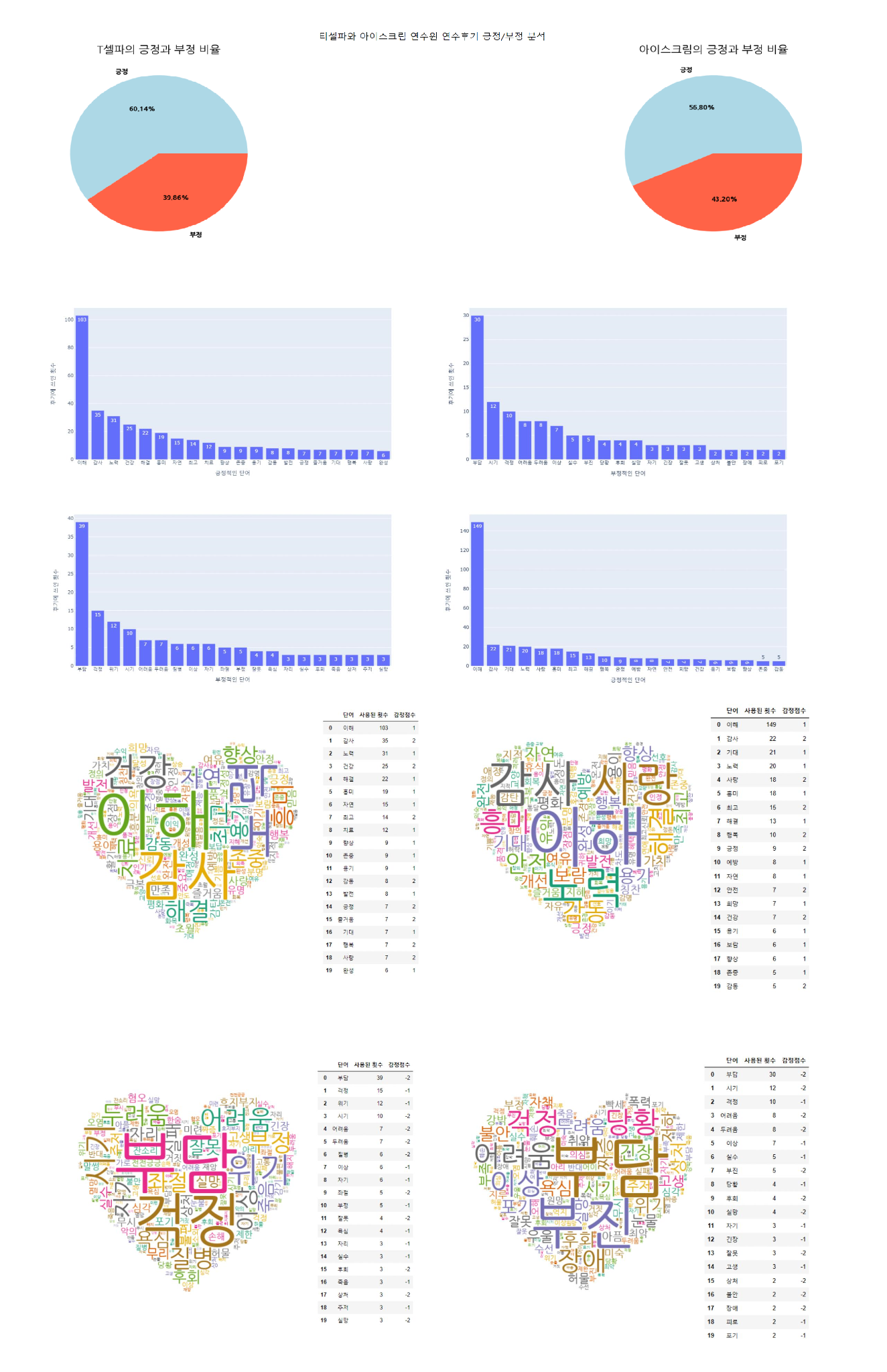
**[결론]**

T셀파 연수원과 아이스크림 연수원 후기의 단어를 분석한 결과, T셀파는 긍정이 60.14%, 부정이 39.86%로 나타났으며, 아이스크림은 긍정이 56.80%, 부정이 43.20%로 나타나며 T셀파의 긍정 단어가 더 많은 것으로 조사되었다.

| <그림7. T셀파 후기 긍정과 부정 비율> | <그림8. 아이스크림 후기 긍정과 부정 비율> |
| --- | --- |

**[대시보드]**

대시보드는 GitHub에서 코드를 다운로드 후 클론하여 브라우저에서 http://127.0.0.1:8050 주소로 접속하면 확인할 수 있다.

****

<그림9. 대시보드>

**[코드 제공]**

프로그래밍 코드는 아래 GitHub 링크에서 확인할 수 있습니다.

* <https://github.com/hannah-OH/Python/tree/master/project/team%20project2-data%20mining>

| <코드 파일 목록> |
| --- |
| * 티셀파\_홈페이지\_크롤링.ipynb * 아이스크림\_홈페이지\_크롤링.ipynb * 티셀파\_텍스트마이닝.ipynb * 아이스크림\_텍스트마이닝.ipynb * 대시보드.ipynb |

**[참고문헌]**

1. 정진구, 김자미, & 김홍래. (2009). 원격교육연수원의 교육과정 운영 실태 분석. 정보교육학회논문지, 13(2), 135-144.
2. 이종화.(2018).Python을 이용한 SNS 크롤링 시스템 구축1.한국산업정보학회논문지, 23(5), 61-76.
3. 이원주, 박봉균, 박문규.(2021).재난 문자 크롤링 및 시각화 애플리케이션 설계 및 구현.한국컴퓨터정보학회 학술발표논문집, 29(1), 89-90.
4. 박창현. (2017). 누리과정 정책 수요에 대한 빅데이터 분석: 워드 클라우드와 의미망 분석을 중심으로. 유아교육연구, 37(3), 73-91.