데이터분석 보고서

주제 : 산막이옛길

– 사전 기반 빈도 및 감정 분석

1조

2022. 9. 23.

Ⅰ. 개요

1. 주제

빈도 및 감정 분석을 통해 산막이옛길에 방문했을 때 음식이나 방문하는 장소 등을 확인하고, 산막이옛길에 대한 방문자들의 감정이 긍정적인지 부정적인지, 어느 부분에서 그렇게 느꼈는지를 확인

2. 프로젝트 내용

데이터 내에서 내용이 없는 결측치를 제외하고, 1차적으로 글쓴이의 글 빈도수를 통해 관련이 없는 글들을 제거하고, 2차적으로 관련이 없다고 생각하는 키워드가 포함된 글들을 제거한 후, 형태소 분석기를 통해 데이터를 전처리 후, 빈도 및 감정 분석을 진행

Ⅱ. 데이터 분석

<table>
<tr><th colspan="2">1. 데이터 준비</th></tr>
<tr><td>데이터 정의</td><td>☐ 사용 데이터(명칭 : 산막이 옛길, 충북 관광 데이터)

1. 산막이 옛길에 관해 작성한 글의 정보
• 글제목
• 글쓴이
• 내용(사용할 데이터)
• 태그
• 공감수
• 댓글수

2. 데이터의 목적에 따른 분류
• 정상글(0) • 홍보글 및 뉴스(1) • 결측치 및 제거대상(9)</td></tr>
<tr><td>데이터 구성</td><td>☐ 사용 데이터 구성(명칭 : 산막이 옛길, 충북 관광 데이터)
1. 사이즈 (문장 수)
• total : 2,514 글
2. 크기(MB)
• total : 4.88 MB
3. 형식 : csv 파일

☐ 사용 데이터 구성(감정분석을 위한 감정사전)</td></tr>
</table>

	1. 사이즈 (단어 수) • positive = 124개 • negative = 280개 2. 크기(MB) • positive = 732Byte • negative = 1.43KB 3. 형식 : txt 파일
2. 분석 방법	
분석 방법	**□ 전처리 및 분석 방법** **1. 전처리 방법** 1. 글쓴이 • 한 글쓴이가 글을 6개 이상 쓰는 경우는, 광고 및 홍보에 관련된 글이 많음을 확인하고 그 글쓴이에 해당하는 행을 제거 2. 키워드 • 데이터를 보며 홍보성이 짙은 글들에 나타나는 공통적인 단어를 키워드로 지정하여 키워드가 들어간 글 제거 **2. 분석 방법** 1. 빈도 분석 • 형태소 분석기 : rhinoMorph==4.0.0.3 • pos = ['NNG','NNP','NNB','NP','MAG','IC'], eomi = True 2. 감정 분석 • 형태소 분석기 : rhinoMorph==4.0.0.3 • pos = ['NNG', 'NNP', 'VV', 'VA', 'VX', 'VCN', 'XR', 'IC', 'MM', 'MAG', 'MAJ'], eomi = True, combineN = True

3. 분석 결과

빈도 분석

- 산막이옛길을 찾아보니 충북 괴산군 칠성면 외사리 사오랑 마을의 산골마을인 산막이 마을까지 연결됐던 총 길이 10리의 옛길로서 흔적처럼 남아있는 옛길에 덧그림을 그리듯 그대로 복원된 산책로라고 한다.

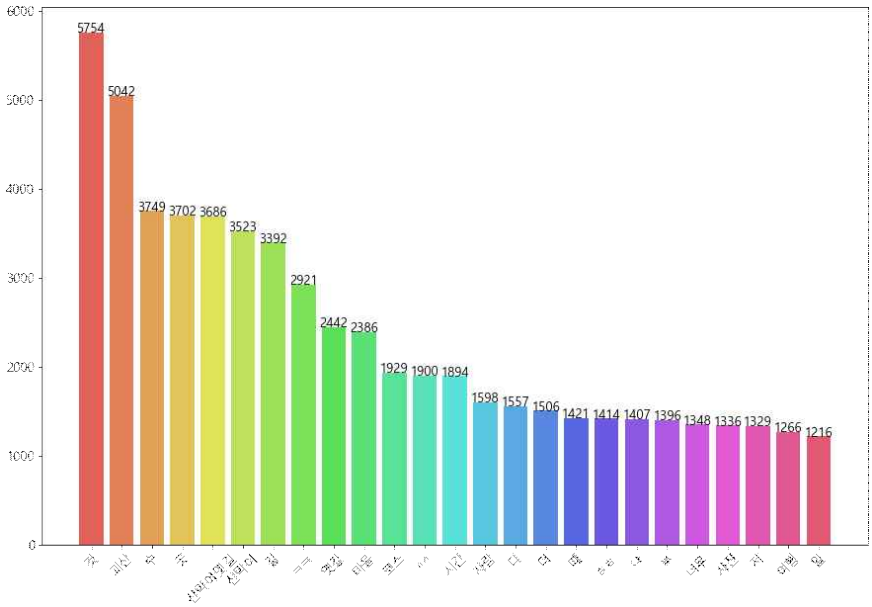


그림 1 불용어 처리 전 단어 빈도 그래프

- 불용어 처리를 안했을 때 ‘ㅋㅋ’ 등 여러 인터넷 용어들이 많이 나와 분석에 어려움이 있어 불용어 처리를 진행했다.
- 이때 산막이옛길에 관련된 글들만 가져온 데이터로 ‘괴산’, ‘산막이옛길’, ‘산막이’, ‘길’ 등 무조건 나올 수 밖에 없는 단어들은 분석에 방해된다고 판단하여 삭제하였다.

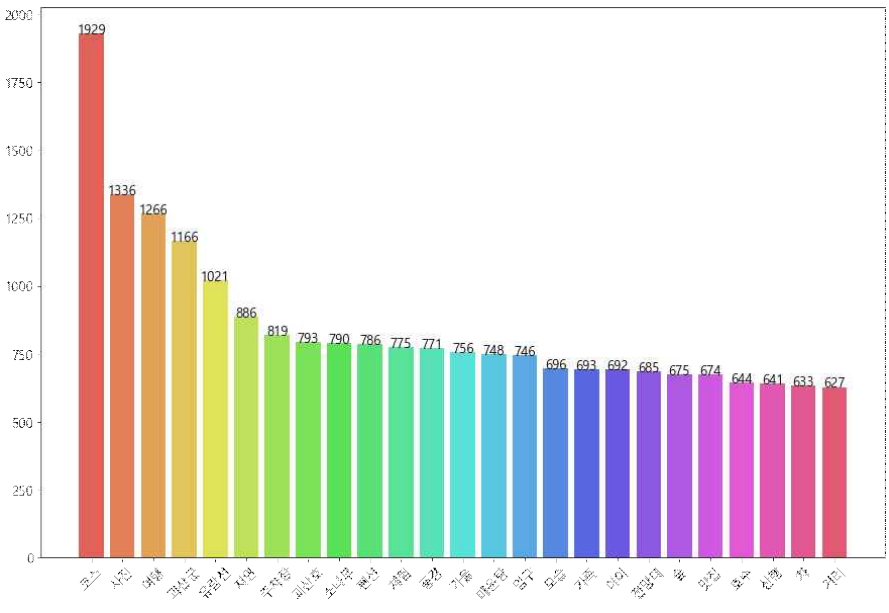


그림 2 불용어 처리 후 단어 빈도 그래프

- 길이가 10리이다 보니 코스들을 많이 확인하고 근처에 괴산호라는 호수가 있으며 여행을 많이 온다는 것을 알 수 있었으며 가족, 아이 등 이런 키워드들을 통해 가족 단위 여행이 많다는 것을 확인할 수 있었다.

- 그리고 유람선을 타고 관광을 즐길 수 있으며 매운탕을 파는 맛집이 많다는 것을 확인할 수 있었다.

그림 3 1음절 단어 및 불용어 처리 전 워드 클라우드

- 워드 클라우드로 보면 확실히 산막이옛길, 괴산과 같은 산막이옛길에 해당하는 단어들이 많이 보이는 것을 확인할 수 있었다.

- 이는 관련 주제에 대한 단어는 당연히 많이 나올 것이므로 산막이옛길과 같은 단어는 불용어 처리하여 진행했다.

<table>
<tr>
<td></td>
<td>

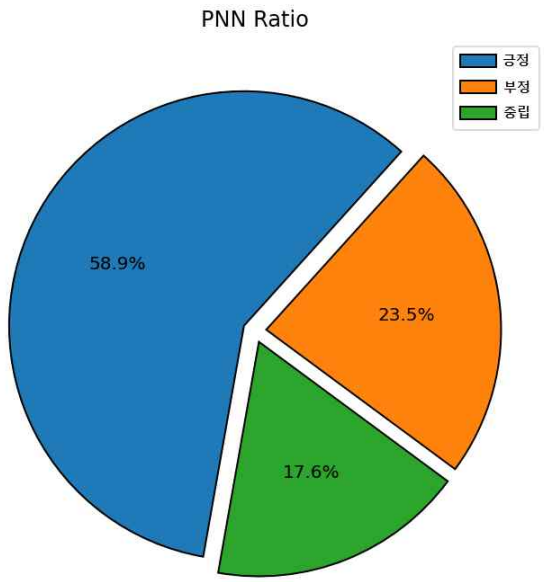

그림 4 불용어 처리 후 워드 클라우드
- 불용어 처리를 진행한 워드 클라우드를 보면 먹는 것과 관련된 단어들이 많이 보임을 확인할 수 있다.
</td>
</tr>
<tr>
<td>감정 분석</td>
<td>
- 팀에서 제작한 감정 사전 사용
- 데이터를 보며 그 외의 긍정, 부정 단어나 신조어, 축약어 추가


그림 5 글 내용의 긍정, 부정, 중립의 비율
- 긍정의 비율이 상대적으로 높음을 확인했다.
</td>
</tr>
</table>

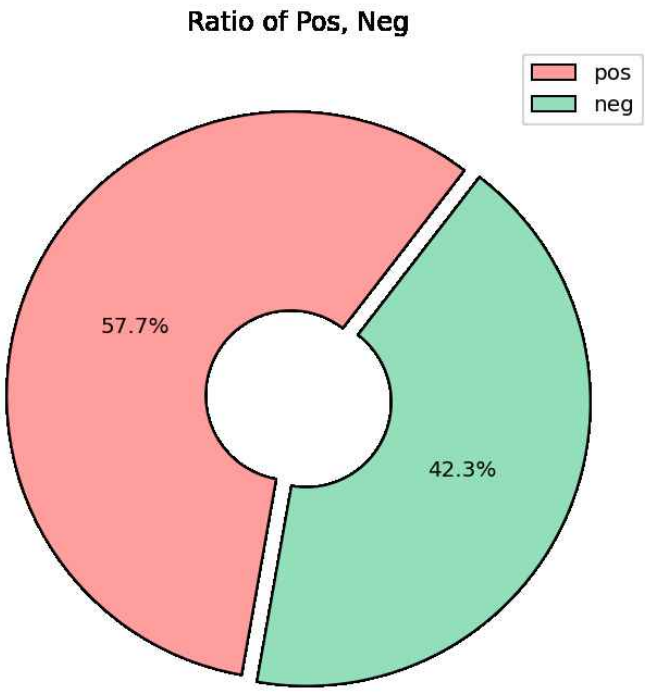


그림 6 전체 내용의 긍정, 부정 단어의 비율

- 긍정 단어가 더 많은 것을 확인했다.

-

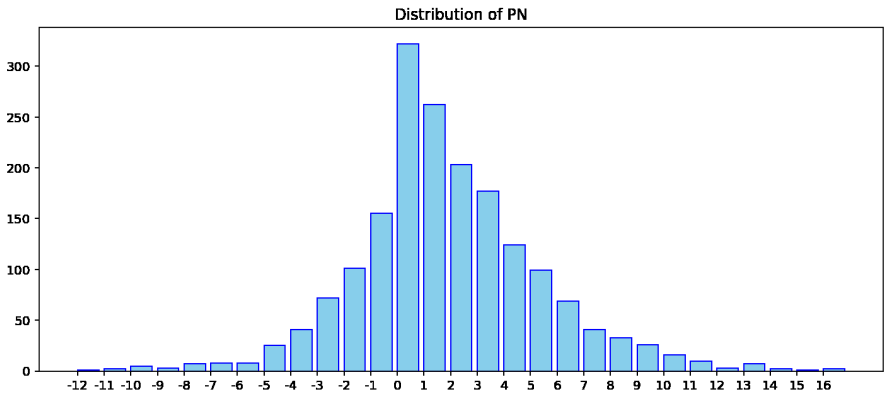


그림 7 전체 내용의 감정 점수 분포

- -1~5 정도에 분포되어 있음을 확인할 수 있다.

그림 8 부정단어 워드 클라우드

그림 9 긍정단어 워드 클라우드

- 부정단어는 주로 산막이옛길의 길이 많고, 험난한 길도 있어서 힘들고, 시간도 걸리고, 길도 잘 모르겠다는 단어들이 보인다.
- 긍정단어는 풍경이 아름답고 걷기 좋은 곳, 그리고 맛있는 음식에 관련된 단어들이 많이 보인다.

그림 10 전체 긍정 문장에 포함된 단어 워드 클라우드

그림 11 전체 부정 문장에 포함된 단어 워드 클라우드

그림 12 전체 중립 문장에 포함된 단어 워드 클라우드

- 전체적으로 단어가 중복되는 경우가 많이 보이는데, ‘걷기 좋다, 걷기 힘들다’, ‘좋아 보인다, 안 좋아 보인다’ 등 어떤 단어들이 같이 오느냐에 따라서 반응이 달라질 수 있기 때문인 것으로 보인다.
- 긍정 문장에서는 ‘아름답다’, 부정 문장에서는 ‘힘들다’라는 한 곳에서만 확인할 수 있는 단어들도 보인다.
- 글 안에서 긍정 단어와 부정 단어를 혼합해서 사용하기 때문에, 단어만으로 글의 긍정, 부정을 판단하기는 어려울 것으로 보인다.

4. 고찰	
고찰	- 어떠한 단어를 불용어 처리, 또는 긍정, 부정 단어의 구분을 지을 때, 많은 고민이 필요했다. - 불용어를 지속적으로 처리해도 불용어로 처리되지 않은 것이 분석할 때 방해가 되었다. - 형태소 분석에서 어떠한 품사를 사용할지에 대해서 고민을 많이 해야함을 느꼈다. - 신조어, 축약어 같은 단어를 반영하기 어려웠다. - 하나의 글에 주제가 여러 개일 경우, 해당 관광지에 대한 긍/부정인지, 관광지 근처 맛집이나 근처 장소, 먹은 음식 등에 대한 긍/부정인지 구분하기 어려웠다. - 사전 기반의 감정분석은 단지 개수를 비교하는 것이기 때문에 사용자의 언어습관에 따라 결과가 실제 감정과 얼마든지 불일치 할 수 있음을 깨달았다.