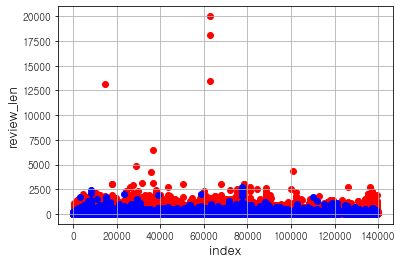
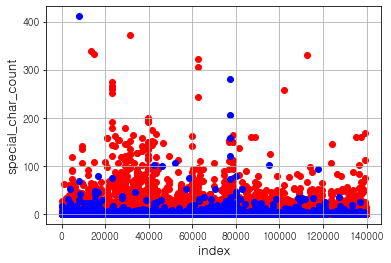
동시출현 단어분석 기반 스팸 문자 탐지기법 논문을 보면 스팸문자와 정상문자에는 특수문자 개수, 문장의 길이, 단어의 개수가 확연한 차이가 났다. 논문에서는 이를 특정한 의도 및 목적을 달성하기 위해 정보를 전달하려는 목적으로 문자를 사용했기 때문에 이러한 차이가 난다고 하였다. 쇼핑몰 스팸리뷰는 스팸문자와 비슷한 목적(광고)을 위해 작성되기 때문에 논문에서와 같이 언어기반기법으로 분석할 수 있을 것으로 판단되었다.



빨간점 : 5점 리뷰, 파란 점 : 1점리뷰

19만건의 리뷰 중 내용이 없는 리뷰를 뺀 15만건의 리뷰 중 1점과 5점의 리뷰를 분석 해본 결과 논문에서 본 것과 같이 특수문자 개수, 리뷰 길이 모두 5점 리뷰가 앞서는 것을 볼 수 있다. 따라서 1점 리뷰 보다 5점 리뷰에서 더 많은 스팸 리뷰들이 있을 것으로 판단되었다.

특수문자 개수와 리뷰 길이만으로는 스팸을 판단하기에 feature가 너무 적었기 때문에 추가 feature를 만들기 위해 머신러닝을 통해 1점과 5점의 단어들을 분석해 보고자 했다.

먼저 1점과 5점에서 공통적으로 나온 단어 690개, 1점에서 나온 단어 290개, 5점에서 나온 단어 290개를 단어 사용 빈도순으로 내림차순으로 정렬하여 가져왔다. 그런 다음 각각의 리뷰에 대해 1270개의 단어가 각각 몇 개가 나왔는지 카운트 하여 shape가 (1,1270)인 리스트에 저장하였다. 그리고 5점 리뷰가 1점 리뷰보다 많이 크롤링 되어 공통 단어와 1점 단어와 5점 단어의 총 개수가 불균형을 이루고 있기 때문에 정규화를 시켜 다시 저장하였다.

이렇게 각각의 리뷰는 길이가 1270인 리스트로 나타나게 되었고 아래의 모델로 머신러닝을 하였다.

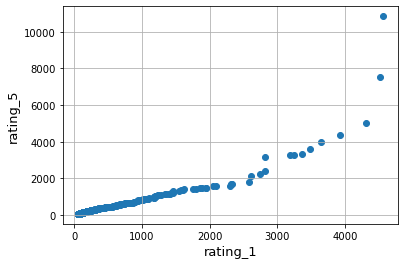
텍스트이(가) 표시된 사진

자동 생성된 설명

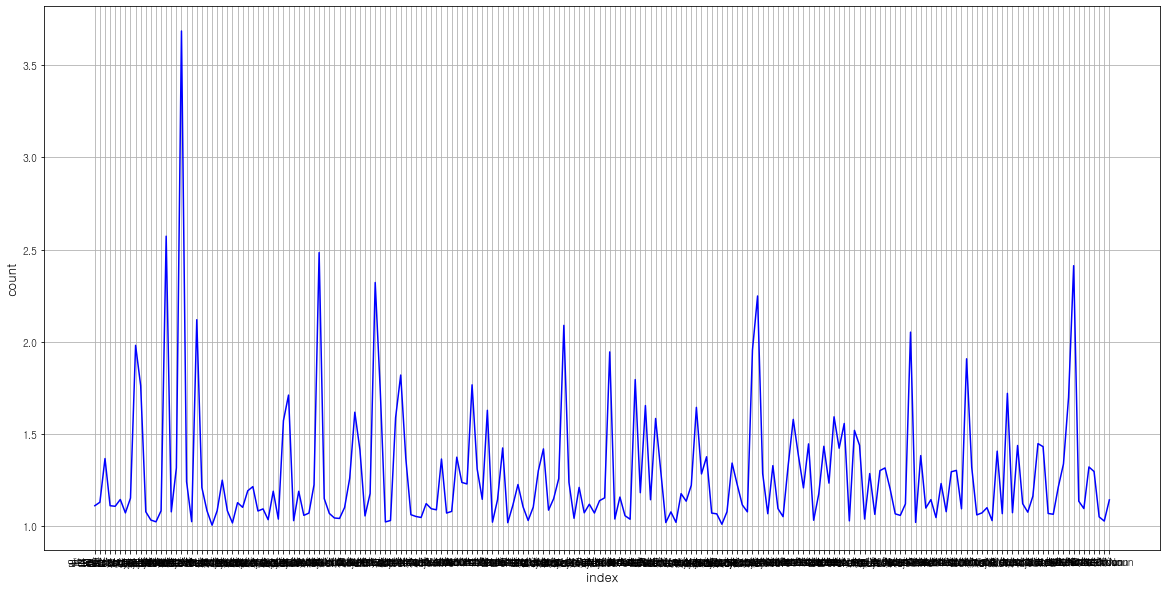
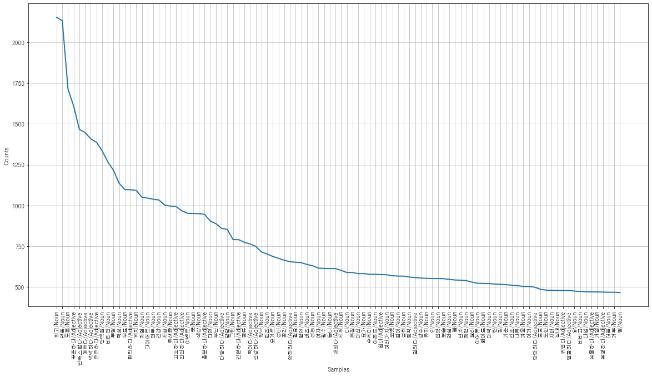
테이블이(가) 표시된 사진

자동 생성된 설명

정확도는 91.7%로 꽤나 정확하게 나왔고 이는 공통단어, 1점 단어, 5점 단어로 1점 리뷰와 5점 리뷰를 구분할 수 있을 정도로 의미 있는 차이가 있다는 뜻이다.

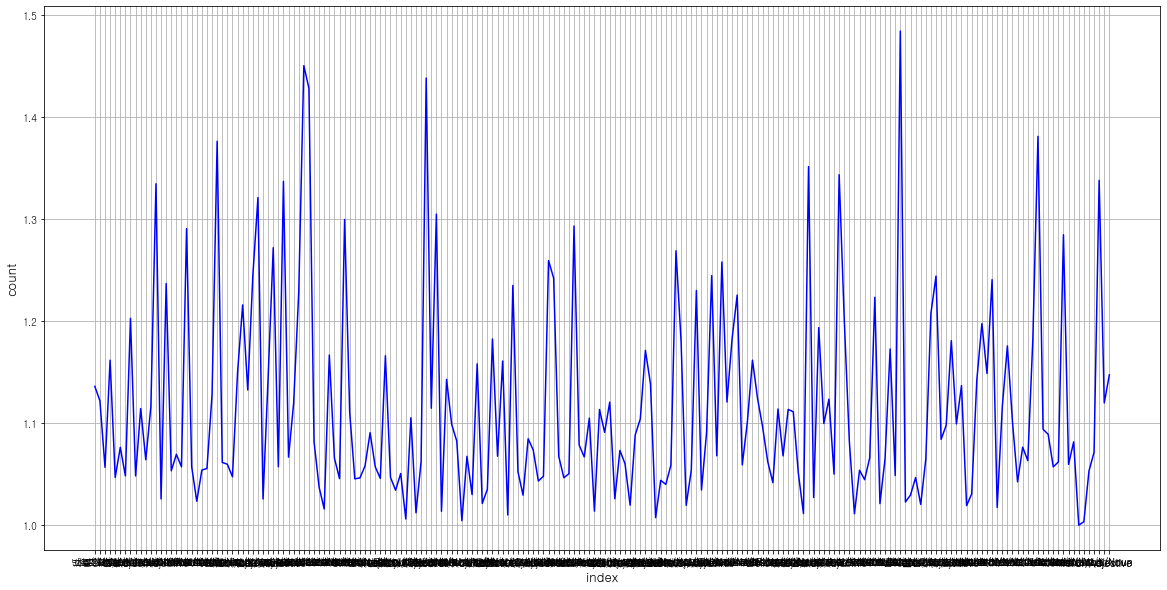
 옆의 사진은 1점 리뷰와 5점 리뷰에서의 공통 단어의 사용량을 계산한 것이다. 1점과 5점의 리뷰 개수의 차이가 있기 때문에 리뷰 개수의 비율과 리뷰 길이의 비율을 생각하여 5점 리뷰의 수치를 조정하였다. 그래프에서 보이듯이 공통단어의 경우에는 1점 리뷰와 5점 리뷰에서 비슷한 사용량을 보였다.

이는 5점리뷰와 1점리뷰를 구분하는데 있어서 가장 큰 역할을 한 것이 1점 단어와 5점 단어라는 뜻이다. 따라서 5점 리뷰에 스팸리뷰가 많을 것으로 예상하였기 때문에 5점 단어에서 리뷰를 스팸으로 만드는 단어들이 있을 것으로 생각하였다.



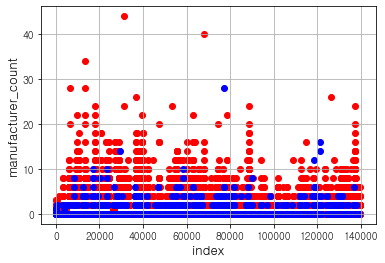
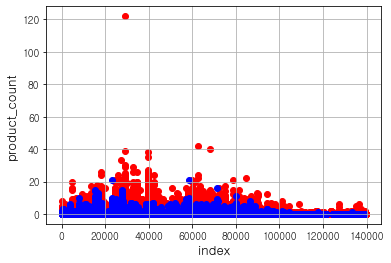
왼쪽의 그래프는 5점 단어들을 빈도순으로 나열한 그래프이고, 오른쪽 그래프는 빈도순으로 나열한 단어들이 리뷰에서 사용되었을 때의 평균 사용빈도이다. 한 번 사용되었을 때 자주 사용되는 단어들이 있다는 사실을 알 수 있다. 이를 1.5점을 기준으로 이보다 많이 사용된 단어들을 찾아본 결과 대부분의 단어들이 상품의 이름이나 제조사명, 상품의 정보에 관련된 단어였다. 

이는 1점 리뷰와는 대조되는데 1점 단어의 경우 튀어나온 5점 단어들의 기준선인 1.5회보다 아래에 머물러 있었고 1.2회를 기준으로 단어들을 추출해본 결과 5점 단어만큼 상품 관련 단어의 비율이 높지 않았고 오히려 부정적인 단어들이 많았다.

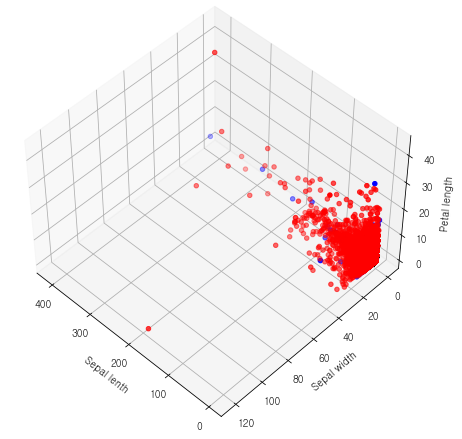


이를 통해 스팸 리뷰에는 상품에 관련된 단어들이 자주 사용되고, 특히 상품명과 제조사명이 사용된다는 것을 알 수 있었다.

따라서 스팸 리뷰를 필터링할 feature로 특수문자 사용 횟수에 더하여 제조사명 언급 횟수와 상품명 언급 횟수를 추가하였다.



위의 그래프를 보면 제조사명 언급 횟수와 상품명 언급 횟수 모두 5점 리뷰가 앞서고 있는 것을 알 수 있다. 상품명 언급 횟수의 경우 큰 차이는 나지 않았지만 그래도 일반적인 경우보다 많았음을 나타내는 지표라고 생각하여 그대로 사용하기로 하였다.



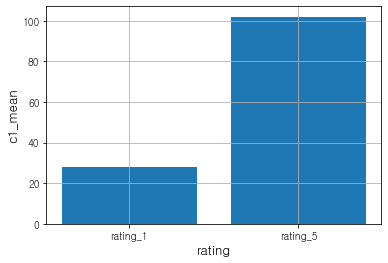
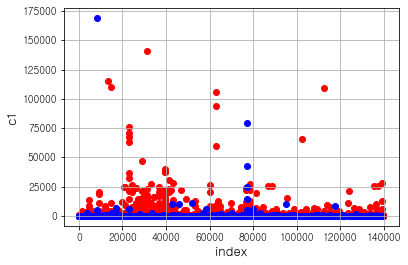
특수문자, 상품명, 제조사 세 feature를 3차원 공간에 나타내 보면 5점 리뷰에서도 정상적인 리뷰가 많아 이 세가지 feature로는 클러스터링 알고리즘을 실행하기에는 무리가 있을 것으로 생각되었다. 따라서 3차원 공간의 원점에서부터의 거리를 스팸을 나누는 기준으로 잡기로 하였다

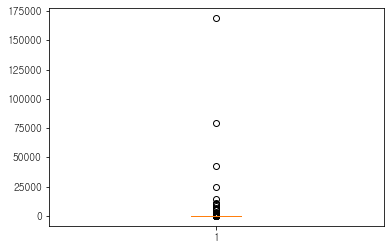
따라서 계산 식은

C1 = (특수문자 개수)^2 + (상품명 언급 횟수)^2 + (제조사명 언급 횟수)^2

가 된다

이를 그래프로 나타내 보면 1점 리뷰의 c1값은 대부분 아래에 깔려있는 반면 5점 리뷰의 c1값은 위로 퍼진 값도 많은 것을 알 수 있다





여기서 스팸을 정하는 c1값의 기준을 잡기 위해 생각해낸 방법이 스팸 리뷰가 상대적으로 적은 1점 리뷰의 c1값의 평균보다 높은 이상치의 최저값을 기준으로 잡는 것이다. 그래프에서는 잘 보이지 않지만 1점 리뷰의 평균보다 높은 이상치의 최저값은 4이다. 따라서 4보다 같거나 높은 c1값을 가지는 리뷰들을 필터링하였고 총 개수는 43188개였다.