**데이터분석**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ** 다음 요구사항에 맞추어 주어진 Data Set을 분석하고,**  **결과물에 대하여 자유롭게 서술하시오.**  a. 요구사항   |  | | --- | | 미국 숙취해소식품 소비 트렌드 분석 |   b. Data Set 정의   |  |  |  | | --- | --- | --- | | 데이터량 | 숙취해소 관련 식품 댓글 3,895건 | | | Data 정의 | 미국에서 판매되는 숙취해소 관련 식품 Data | | | 엑셀 열 항목별 정의 | Name |  | | Type | 아마존 판매 제품 댓글 DB | | Country | 국가 | | Title | 댓글 제목 DB | | Content | 댓글 본문 DB | | Date | 댓글 입력일 DB | | URL | 관련 URL 정보 |    분석 진행 후, 결과물에 대하여 아래 양식에 맞추어 자유롭게 작성할 것 |

|  |
| --- |
| Q1. 요구사항에 대한 데이터 분석은 어떻게 진행하였고  이를 위한 전처리, EDA 작업은 어떻게 진행하였는가? |
| 먼저 숙취해소 식품에 관련된 정보를 읽어야 하기 때문에 distinct URL을 알아보았습니다. URL을 하나의 상품으로 인지하고 주어진 데이터가 6개 정도의 상품에 관한 후기들이 나열되어 있는 dataset이라는 것을 인지했습니다. 각 URL을 그룹화하여 각 상품별 text들에 대한 data를 필터링해 파일로 분리했습니다.  각 상품은 5가지의 숙취해소제와 특정한 상품을 알 수 없는 다수의 후기로 이루어져있었습니다.  다음 각 상품에 따른 텍스트 마이닝을 하기 위해 텍스트 전처리용 함수를 만들어 소문자 만들기 및 불용어 처리와 어간을 추출하였습니다. 다음 공백으로 구분된 문자열로 결합해 결과를 반환시킨 데이터로 워드 클라우드를 진행하였습니다.  또 주어진 데이터로 감성분석을 실시해 제품에 대한 평가를 확인해보았습니다.  분석진행은 데이터 핸들링은 R을 활용했고 전처리 된 데이터는 Python을 활용해서 자연어처리를 진행하였습니다. |
| Q2. 분석된 Excel 결과물에 대하여 간단히 요약 설명하시오. |
| **review\_word\_count50** = 명사추출 단어 빈도수 분석  **review\_adjective\_count50** = 후기의 긍정/부정 관찰 시 명사보다는 형용사가 더 유의미하다고 판단해 형용사 추출 단어 빈도수 분석  **sentiment\_tot\_result** = title과 content의 내용으로 감성분석 실시  부정문으로 적혀있지만 실제로는 긍정적인 후기를 남기고 있는 리뷰의 경우 제대로 인지하지 못해 신뢰도가 낮다고 판단했습니다.  따라서 title의 polarity와 content의 polarity의 값을 더해 감성분석 실시했습니다.  (긍정적인 후기 중 content가 부정으로 나와도 title이 더 높은 긍정의 값을 가지면 긍정적인 후기로 판단)  몇 개의 test를 해보았을 때 비교적 신뢰도가 향상했습니다.  **sentiment\_result** = 앞선 csv file에서 각 제품 마다 임의로 정한 5개 척도의 통계를 나타냈습니다. |
| Q3. 해당 과제 진행과 관련된 개선 의견에 대하여 자유롭게 서술하시오. |
| 날짜 데이터에 따른 제품 판매 추이를 확인해보고 싶었는데 모든 데이터가 동일한 날짜의 데이터여서 날짜 데이터를 활용할 수 없는 점이 아쉬웠습니다.  일단 대부분의 숙취해소제들은 긍정적인 평가가 많고 비율도 비슷했습니다.  만약 감성분석의 신뢰도가 더 높다면 몇 개의 부정적인 평가가 긍정적인 평가로 바뀌어 좀 더 높은 비율의 긍정적인 반응을 예상해 볼 수 있을 것 같습니다.  Naïve Bayes 분류기를 사용해 몇 개의 후기를 train, test set을 만들어 학습 시켜 모델을 만든다면 좀 더 제품에 최적화된 정확도 높은 분석이 가능할 것 같습니다. |