

## 리뷰 데이터를 통한 리뷰 분석 - 리뷰 내용 토큰화

### 최종 키워드 산출

방법1과 방법2를 통해 산출된 키워드를 비교하면서  
겹치는 키워드를 우선적으로 최종 키워드 산출



최종 키워드 개수는 제품당 3-6개 사이

## 5. 데이터 군집화

## 제품 군집화를 위한 데이터 전처리

1. 최종 키워드 원핫 인코딩 수행
2. 가져올 속성 선택 -> 제품명, 원가, 별점
3. 원가 정규화(Min-Max 스케일링)

## KMeans 클러스터링

KMeans 클러스터링을 통해 군집 개수 **5개**로 설정



군집 개수 5개로 설정한 이유

1. elbow plot 그래프에서 군집수가 증가함에 따라 sse가 감소, 이때 급격히 줄어드는 지점이 최적의 군집
2. silhouette score 1에 가까울수록 좋음, 군집이 5개일때 silhouette 이 증가하므로 5개가 적절

최종 군집화

1의 개수 비율을 통해 각 클러스터 별 최종 키워드 선정

cluster\_0

cluster\_1

cluster\_2

최종 군집화

1의 개수 비율을 통해 각 클러스터 별 최종 키워드 선정

cluster\_3

cluster\_4

군집별 최종 키워드

cluster_0	수분감, 여름, 순함, 보습, 건성, 무난, 미백
cluster_1	각질, 자극, 부드러움, 모공, 블랙헤드
cluster_2	수분감, 순함, 진정, 밀착력, 자극, 트러블,보습
cluster_3	남성, 순함, 올인원, 민감성, 복합성, 수분감, 자극
cluster_4	순함, 세정력, 자극, 무난, 가성비, 대용량, 여름

## 6. 질문지 구성 및 데이터 시각화



# 질문지 구성

클러스터링 별 키워드를 기준으로 질문 구상  
→ 답변마다 가중치(0-3) 부여

## 데이터 시각화

설문 폼 제작 사이트 **스모어** 활용  
시각화 및 배포

**스모어 설문 테스트 링크**

<https://smore.im/quiz/nJe60YDNGH>