더헬스 앱 리뷰 텍스트 데이터 분석

건강플랫폼사업팀 인턴 김영국

목 차

01

더헬스

앱 리뷰 개요

프로젝트 목적

분석 대상 데이터

프로젝트 목적

사용자 피드백 이해

- **긍정적 / 부정적 피드백 식별**
  - 앱의 강점과 약점 파악
- **사용자 요구 / 불만사항 분석**
  - 사용자가 좋아하는 기능
  - 사용자가 불만을 갖는 사항

토픽 모델링을 활용한 주요 주제 도출

- **주요 토픽 추출**
  - 사용자가 많이 언급하는 주제 파악
- **주제별 빈도 / 중요도 분석**
  - 각 주제의 빈도와 중요도 파악
  - 사용자에게 중요한 요소 식별

마케팅 전략 수립 + 앱 기능 개선

제품 개선 / 전략적 의사결정 지원

- **데이터 분석 기반 인사이트 제안**
- **사용자 행동 특성에 따른 마케팅 전략**
- **경쟁력 강화**
  - 다른 mHealth앱과 비교하여 경쟁력 강화 요소 식별

사용자 경험 향상

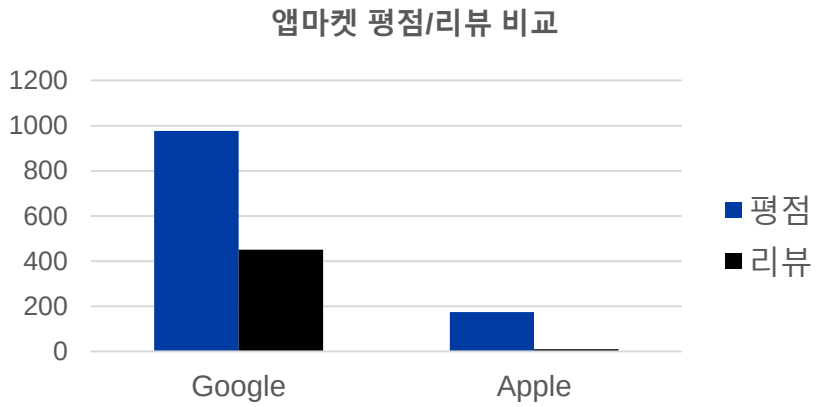
- **UI / UX 개선**
  - 불편사항, 개선요청 사항 반영
- **사용자 만족도 증대**
  - 사용자 피드백 적극 반영

분석대상 데이터

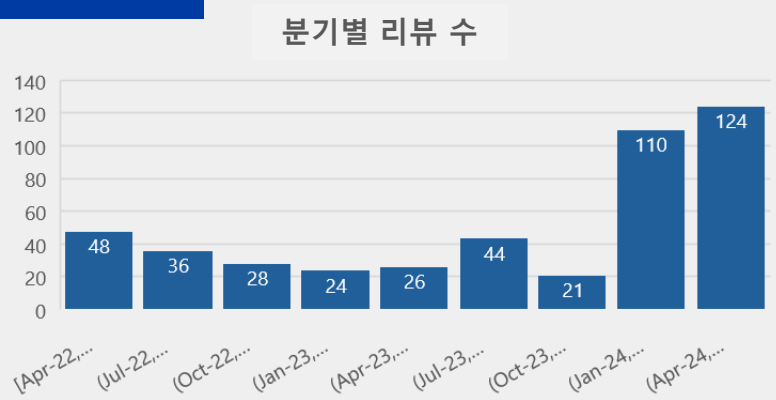
- **더헬스 구글 플레이 + 애플 앱스토어 평점 / 리뷰**
  - 평점 1,141개, 리뷰 461개, 평점 평균 3.8점
  - 기간 : 2022년 4월 ~ 2024년 7월

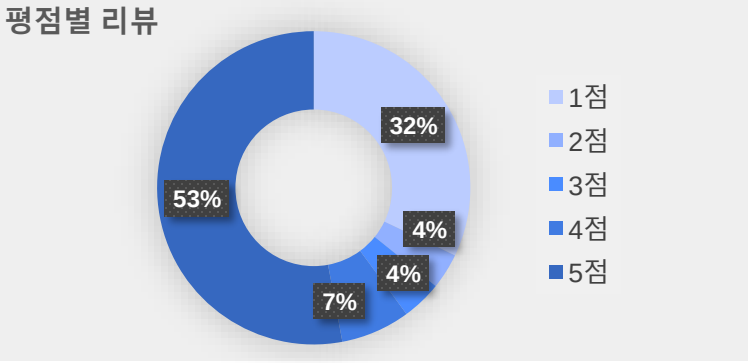
[구글 플레이]

[애플 앱스토어]



리뷰 분류





02

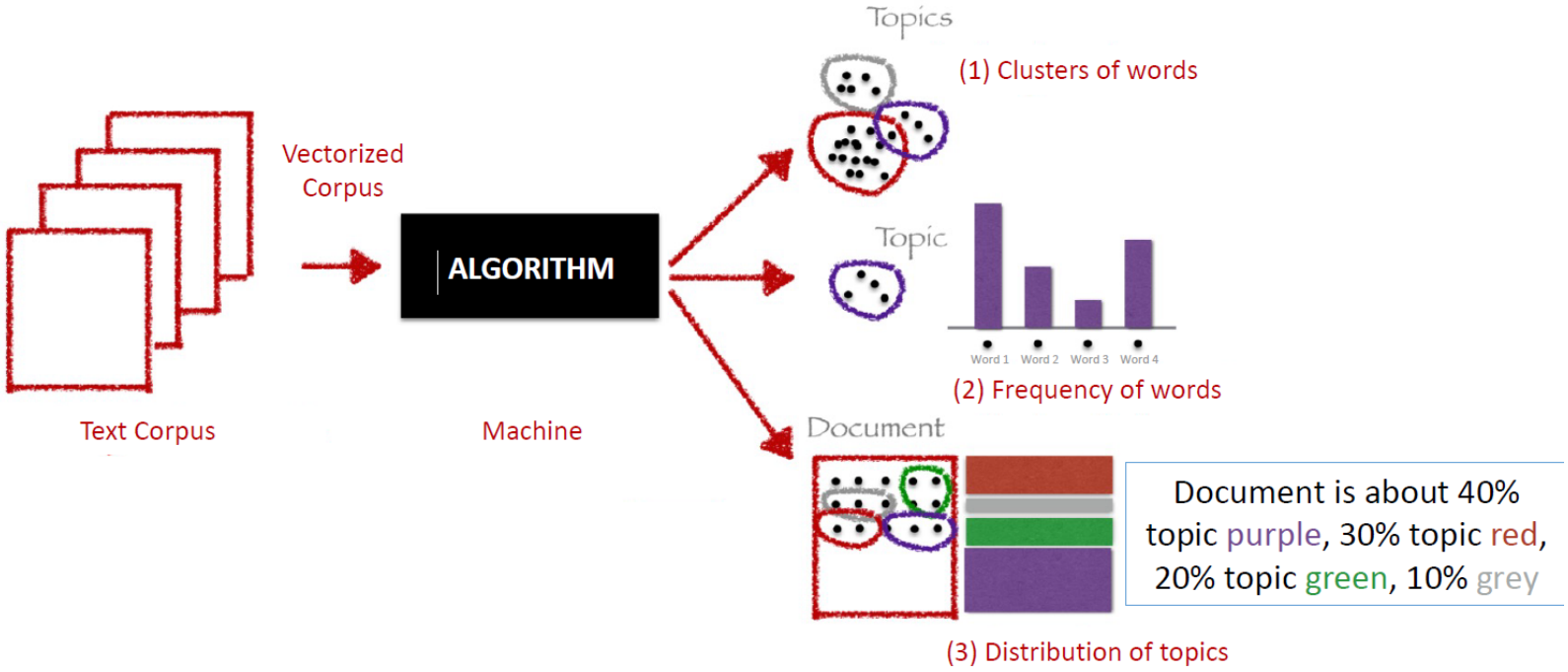
토픽 모델링

토픽 모델링 개요,
데이터 전처리,
하이퍼 파라미터 튜닝

토픽 모델링 개요

□ 문서(리뷰)를 머신러닝 알고리즘에 입력 후 1) 토픽, 2) 토픽별 빈출 단어, 3) 문서 내 토픽 비율 등 출력

□ 이번 프로젝트에서는 가장 흔히 쓰이는 LDA(Latent Dirichlet Allocation) 토픽 모델링 방식 활용



데이터 전처리

□ 데이터 정규화, 토큰화라는 텍스트 분석 사전 작업 수행: 전처리 작업 후 리뷰 총 279개 대상 분석 실시

데이터 정규화

- **결측치 확인: 평점 1,151개 중 리뷰 461개(40%)**
- **단어 치환: 같은 의미의 단어들을 하나로 통일**
  - 빈출 어휘 중심으로 분석하기 위해 단어 통일 필요
  - ex) 앱, 어플 → 어플리케이션
- **한국어 외 다른 문자형태 제거**
  - 숫자, 특수문자, 영문자 등 의미 추출이 어려운 문자 제외
- **불용어 (Stopword) 제거**
  - 토큰화 이후 토큰이 2글자 이상인 경우 불용어 확인
  - 토근이 1글자인데 1글자 키워드에 포함되는 경우 분석용 리뷰 데이터에 포함

  * 1글자 키워드 예시: 컵, 방, 물, 돈, 꿈

데이터 토큰화

- **데이터 토큰화 예시**

토큰화 이전	"이 앱은 사용하기 매우 편리하고 유용합니다."
토큰화 이후	['앱', '사용', '편리', '유용']

- **한국어의 특수성 고려 전용 형태소 분석기 사용**
  - KoNLPy의 Mecab 형태소 분석기
  - 한국어는 명사가 문장 내 맥락을 파악하는데 핵심 형태소이며, 명사 중심 분석 시 빈출 어휘를 상대적으로 쉽게 파악할 수 있음
- **긍정 / 부정 리뷰 분류**
  - 긍정(평점 5~4점) / 부정 (1~2점)에 따른 차별적 토픽 가정
- **리뷰 내 토큰 수 2~20개 이내로 선별**
  - 지나치게 긴 리뷰는 주제 파악이나 리뷰 내 단어 간 조합을 활용한 특징 추출에 어려움 발생 → 적정 수준으로 조정

하이퍼파라미터 튜닝 (Hyperparameter Tuning)

□ 하이퍼파라미터는 모델 학습 과정에서 변경되지 않는 값 (cf. 파라미터: 모델 학습 과정에서 학습되는 매개변수)

□ 하이퍼파라미터 튜닝은 모델의 성능을 최적화하기 위해 학습 전 값을 조정하는 절차

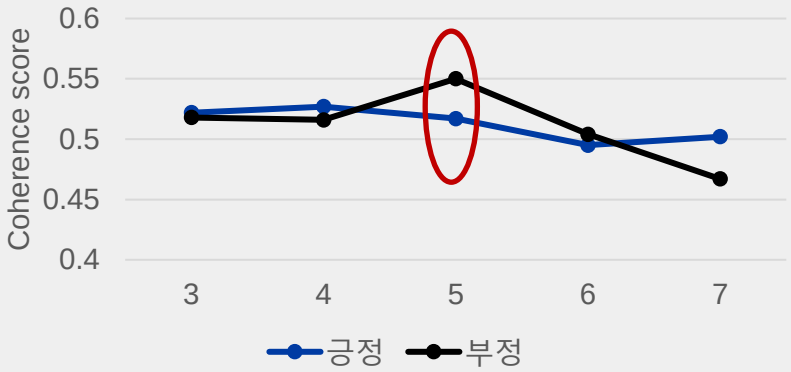
□ 토픽 모델링에서는 '토픽의 수(Number of Topics)와 '학습횟수(Passes)'가 하이퍼파라미터

□ 모델 평가 점수가 높은 하이퍼파라미터값 선택: Coherence Score를 평가 점수로 사용

□ 튜닝 기법: 'Grid Search', 'Random Search', 'Bayesian Optimization' → 'Random Search' 사용

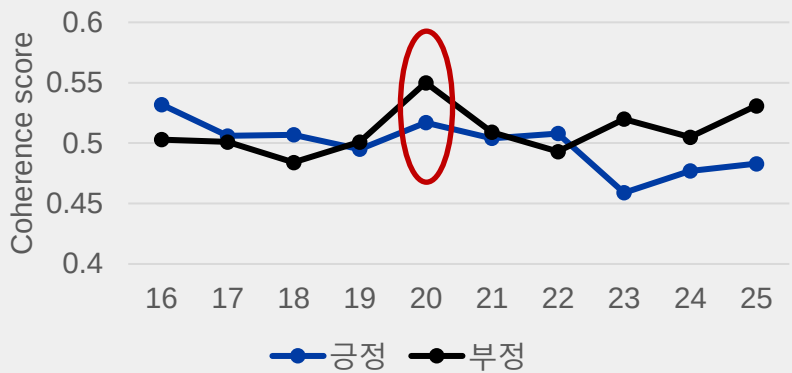
토픽의 수

- Topic = 5개일 때 긍정+부정 Coherence score 최대



학습횟수

- Passes=20회일 때 선택 (Topics=5개 고정)

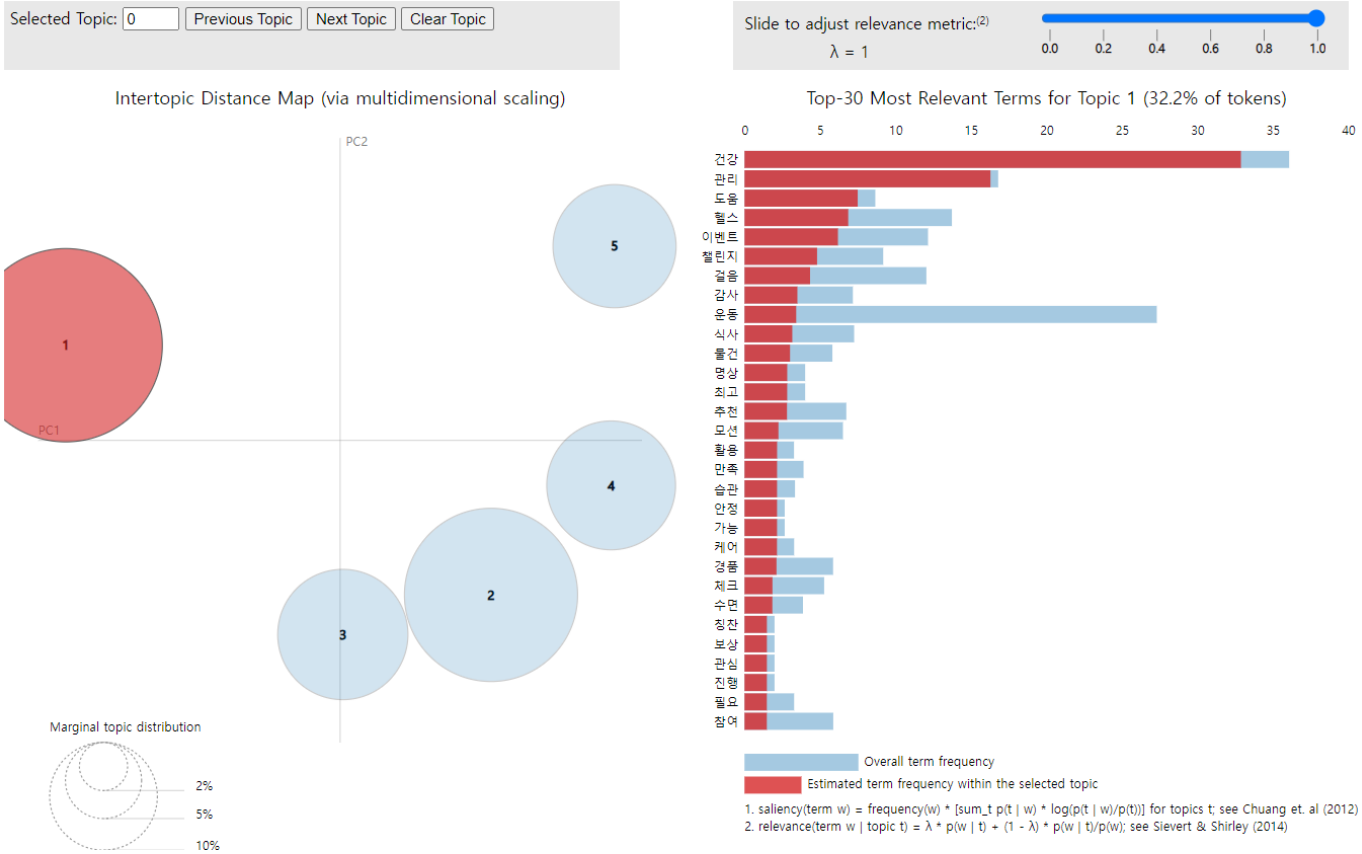


03

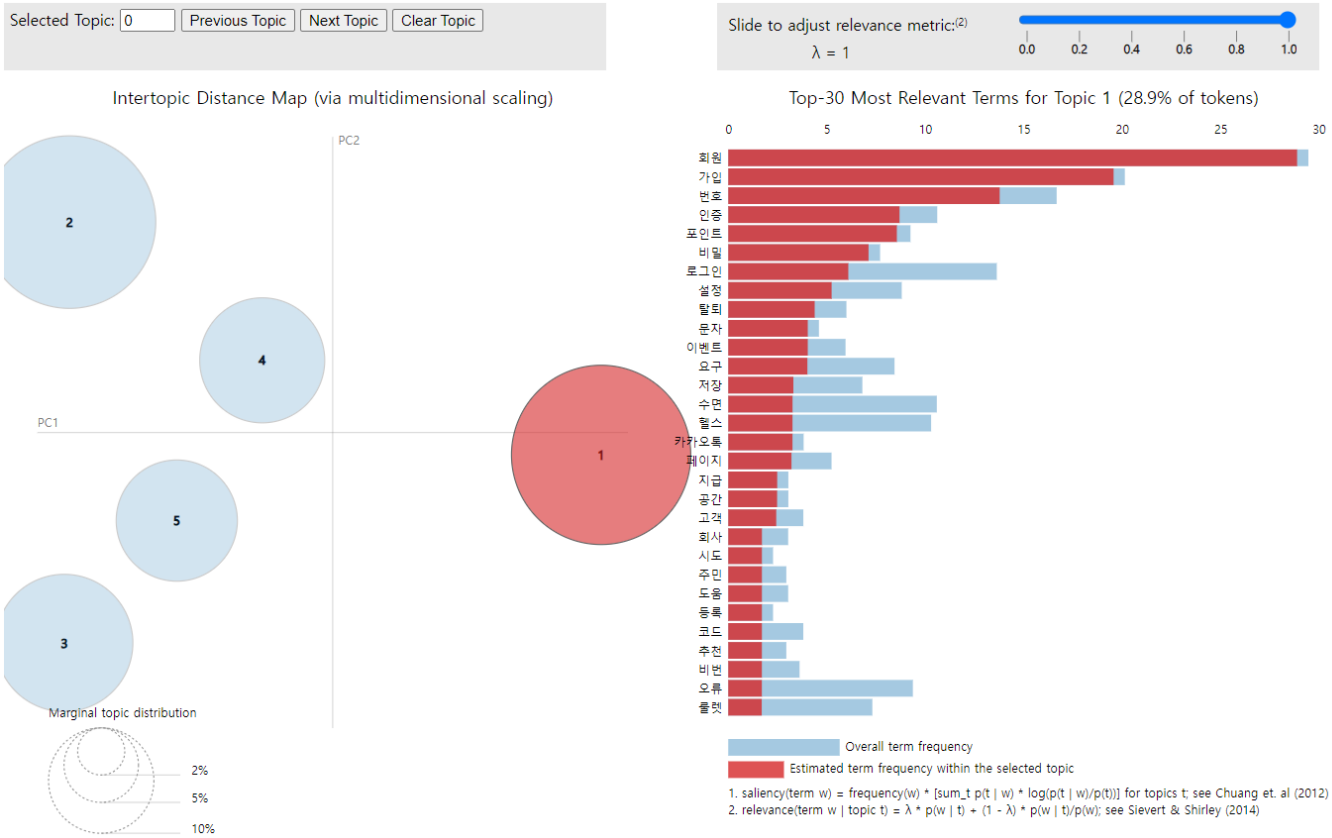
분석결과 및 토의

Inter-Topic Distance Map,
긍정 토픽 분석,
부정 토픽 분석,
N-gram을 통한 연관 단어 찾기

Inter-Topic Distance Map (긍정 답변)



Inter-Topic Distance Map (부정 답변)



긍정 토픽 분석

□ 1번 토픽: 일반적인 사용자 건강관리에 대한 만족도 확인 → 이벤트/챌린지 등을 통한 건강 프로그램 참여 촉진

□ 2번 토픽: 운동, 식단 관리, 걸음 등과 관련된 기능에 대한 만족도 확인 → 해당 서비스 고도화 검토 가능

□ 3번 토픽: 이벤트를 통한 건강 프로그램 참여 유도에 대한 긍정적 반응 → 앱 활성화 지속을 위한 이벤트 유지 검토

토픽명	토픽 내 주요 단어 10개	토큰 비율
일반적인 사용자 건강관리	0.153*"건강" + 0.076*"관리" + 0.035*"도움" + 0.032*"헬스" + 0.029*"이벤트" + 0.022*"챌린지" + 0.020*"걸음" + 0.016*"감사" + 0.016*"운동" + 0.015*"식사"	32.2%
운동, 식단 관리, 걸음과 관련된 기능	0.063*"운동" + 0.040*"식단" + 0.035*"걸음" + 0.033*"이벤트" + 0.025*"기록" + 0.021*"추천" + 0.017*"확인" + 0.017*"룰렛" + 0.013*"운동량" + 0.013*"선물"	25.8%
이벤트 참여 및 경품	0.033*"헬스" + 0.033*"당첨" + 0.033*"운동" + 0.029*"건강" + 0.027*"룰렛" + 0.020*"경품" + 0.020*"참여" + 0.020*"반영" + 0.019*"재미" + 0.018*"물건"	14.6%
운동 기능의 유용성	0.088*"운동" + 0.040*"유용" + 0.028*"모션" + 0.021*"감사" + 0.021*"에러" + 0.021*"설치" + 0.015*"계산" + 0.015*"칼로리" + 0.014*"식단" + 0.014*"입력"	14.3%
칼로리 측정과 계산 기능	0.058*"칼로리" + 0.044*"측정" + 0.036*"사진" + 0.036*"계산" + 0.029*"편리" + 0.019*"식사" + 0.019*"수면" + 0.017*"운동" + 0.016*"업데이트" + 0.016*"데이터"	13.1%

부정 토픽 분석

□ 1번 토픽: 회원 가입 및 인증 시 불편이 발생하지 않도록 즉각적인 조치 필요 (특히 아이폰의 경우)

□ 2번 토픽: 특히 업데이트 이후 챌린지 참여나 수면 및 걸음 인식에 문제가 발생

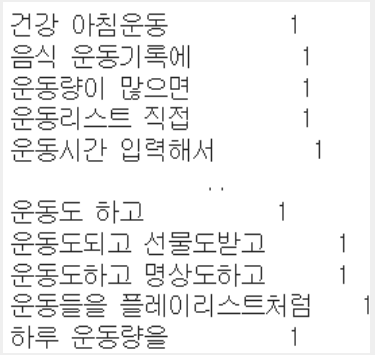
□ 3번 토픽: 사용자가 다른 기기나 앱과의 연동 문제를 겪고 있으며, 앱 실행과 설치 과정에서의 문제를 언급

토픽명	토픽 내 주요 단어 10개	토큰 비율
회원 가입 및 인증 과정의 불편함	0.113*"회원" + 0.076*"가입" + 0.054*"번호" + 0.034*"인증" + 0.033*"포인트" + 0.028*"비밀" + 0.024*"로그인" + 0.020*"설정" + 0.017*"탈퇴" + 0.016*"문자"	28.9%
업데이트 후 기능 문제와 인식 오류	0.033*"업데이트" + 0.030*"챌린지" + 0.023*"수면" + 0.023*"걸음" + 0.020*"인식" + 0.020*"접속" + 0.017*"입력" + 0.017*"요구" + 0.017*"운동" + 0.017*"실행"	26.8%
앱 연동 및 실행 오류	0.057*"연동" + 0.044*"실행" + 0.042*"설치" + 0.037*"로딩" + 0.026*"오류" + 0.025*"무한" + 0.025*"기록" + 0.022*"헬스" + 0.020*"삭제" + 0.020*"권한"	17%
권한 문제와 기능 사용 제한	0.034*"미만" + 0.030*"권한" + 0.018*"룰렛" + 0.018*"삭제" + 0.018*"설치" + 0.018*"정보" + 0.018*"개인" + 0.018*"불가" + 0.013*"확인" + 0.013*"참여"	14.1%
로그인 및 화면 오류	0.057*"로그인" + 0.028*"화면" + 0.023*"운동" + 0.020*"번호" + 0.020*"설정" + 0.017*"수정" + 0.017*"오류" + 0.017*"자동" + 0.017*"챌린지" + 0.017*"전화"	13.2%

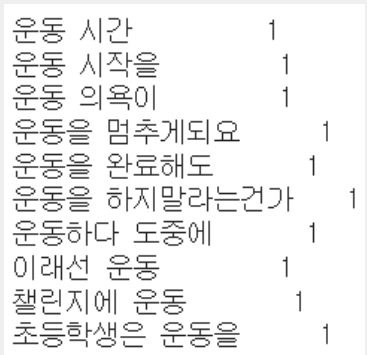
N-gram을 통한 연관 단어 찾기

□ 긍정 / 부정 리뷰 모두 높은 빈도로 출현하는 단어에 대해 연관 표현 확인

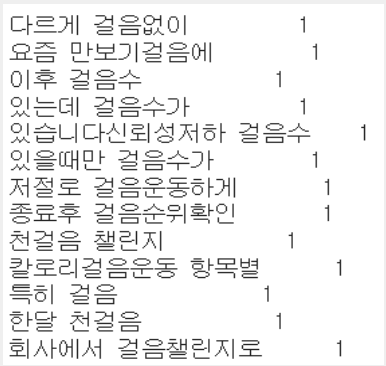
단어	긍정 리뷰	부정 리뷰
운동	전체 : 30회 / 토픽2 : 14회	전체 : 8회 / 토픽4 : 5회
걸음	토픽1 : 5회 / 토픽2 : 7회	토픽2 : 11회



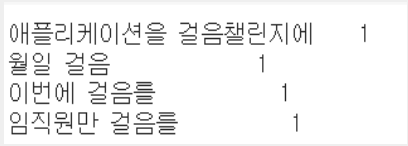
[운동 연관어-긍정리뷰]



[운동 연관어-부정리뷰]



[걸음 연관어-긍정리뷰]



[걸음 연관어-부정리뷰]

"행복한 가정은 모두 비슷하게 닮았지만, 불행한 가정은 저마다의 이유로 불행하다" (소설 안나 까레니나)
"행복한 사용자는 저마다의 이유로 행복하고, 불만족스러운 사용자도 저마다의 이유로 불행하다" (더헬스)

※ 한계점: Bi-gram 분석법은 영어에는 적합하지만 한국어의 경우 조사, 띄어쓰기 등의 제약으로 유의미한 인사이트를 얻기 어려움

04

결론

기술 및 마케팅 관련 제안사항

기능 / 마케팅 관련 제안사항