p1

네 안녕하세요, 땅콩버터팀의 발표를 맡은 장준하입니다.

2주간 고생하신 모든 팀들 고생하셨다는 말씀드리며

이번 세미프로젝트 땅콩버터팀의 Samsung Galaxy S24

SNS 댓글 텍스트 분석 발표를 시작하도록 하겠습니다.

p2

우선 발표에 앞서 목차를 소개드리겠습니다.

프로젝트 주제 선정 과정을 시작으로, 데이터 수집 및 전처리 과정,

데이터 탐색, 그리고 마지막에 결과 및 제안 부분 순으로 저희팀의 발표로 진행하겠습니다.

p3

네 먼저 저희 조원 및 역할 소개입니다.

저희 조 리더인 준홍님은 프로젝트 총괄 및 관리를 중심으로

데이터 분석 전략 수립, 데이터 전처리 및 탐색실행, 진행 상황 모니터링 등의 과정을 담당해주셨습니다.

진수님은 데이터 탐색 및 분석을 기반으로

데이터 탐색 및 시각화, 분석 결과 도출 및 인사이트 제안, 최종 발표 준비를 담당해주셨고

현진님은 보고서 및 발표 자료 제작을 기반으로

분석 결과 요약 및 문서화, 시각적 자료 준비, 발표 자료 제작을 담당해주셨습니다.

마지막으로 저는 데이터 수집 및 전처리를 중심으로 데이터 소스 확보,

데이터 클리닝 및 정제, 최종발표 준비 및 발표를 담당하였습니다.

p4

네 그럼, 본격적으로 프로젝트 과정에 대해 설명드리겠습니다.

첫 번째 목차, 프로젝트 주제 선정입니다.

p5

저희가 선택한 세미프로젝트 주제는 ‘sns 데이터를 활용한 브랜드 평판 분석’입니다.

오늘날 소비자들은 다양한 SNS 플랫폼을 통해 자신의 의견을 자유롭게 공유하고 있는 추세입니다.

이러한 의견들은 브랜드의 평판에 큰 영향을 미치기에,

기업은 이를 적극적으로 모니터링하고 분석할 필요가 있습니다.

p6

저희 땅콩버터팀은 많은 브랜드 중 " Samsung Galaxy S24 시리즈"를 선정하고,

해당 제품에 대한 sns 댓글 텍스트 분석 프로젝트를 진행하였습니다.

올해 1분기에만 약 1400만대가 팔린 '갤럭시 S24 시리즈'는 연말까지 3600만대 판매가 예상되며,

2016년 이후 8년 만에 가장 많이 팔린 '갤럭시 시리즈' 스마트폰이 될 것으로 전망되고 있으나,

인공지능 '삼성 가우스'가 탑재되어 기존 휴대폰과는 차원이 다른,

AI 기능에 만족하는 대중들과, 부족한 RAM 용량과 스피커 기능 문제를 주장하는 대중들 사이에서

제품에 대한 뜨거운 갑론을박이 이루어지고 있습니다.

p7

땅콩버터팀은 이러한 높은 실적에도 불구하고 소비자들 사이에서 논란이 되고 있는

'갤럭시 S24 시리즈'에 대한 대중 평판을 분석하여 제품의 장단점을 파악하고,

이를 바탕으로 마케팅 전략 수립하는 것을 목표로 설정하였습니다.

p8

저희 프로젝트는 다음과 같은 일정으로 진행하였습니다.

하늘색 부분이 사전에 계획한 일정이고 진한 파란색 부분이 실제로 진행한 프로젝트 일정입니다.

사전에 계획한 일정 중 빨리 마무리된 부분이 있을 경우, 이후의 일정을 당겨 미리 진행하기도 하였고,

시간이 부족한 부분은 시간을 늘려 유동적으로 진행하였습니다.

p9

이제 두 번째 목차, 데이터 수집 및 전처리 부분에 대해 말씀드리도록 하겠습니다.

P10

우선 저희는 sns 데이터 크롤링을 위해 ‘셀레니움’을 사용했습니다.

셀레니움은 동적인 웹페이지 크롤링에 최적화 된 라이브러리로,

웹브라우저를 자동으로 제어하고, 상호작용 할 수 있는 기능들을 제공합니다.

셀레니움은 평소 저희가 사용하던 함수 라이브러리와 같이 pip를 통해 설치 후 from과 import를 통해 호출하면 됩니다.

저희는 평판 분석에 사용할 데이터로 유튜브 댓글로 선정한 후 댓글을 크롤링하였습니다.

그럼 저희가 유튜브 댓글을 크롤링한 코드에 대해 간단히 설명드리겠습니다.

우선 저희는 두번째 주석의 driver = webdriver.chrome() 코드를 통해 크롬 웹드라이버를 실행시키고

유튜브 링크를 연 후, 페이지를 자동으로 스크롤 하여 댓글을 로드하도록 하였습니다.

p11

그리고 첫번째 주석의 첫줄에 있는 (body.send\_keys(Keys.END)) 코드를 통해 페이지 하단으로 스크롤하여

댓글을 로드하다가 두번째 주석에 있는 if문을 통해 더이상 새로운 댓글이 로드되지 않으면 종료 되도록 하였습니다.

이후 네번째 주석의 for문을 활용하여, 로드된 댓글 요소를 찾아 텍스트를 추출하고,

추출한 댓글을 리스트에 저장하도록 하였습니다.

p12

이후 첫번째 주석의 save\_to\_excel(comments, filename) 코드를 중심으로

새 엑셀 파일을 생성하고 시트 안에 추출한 댓글이 작성 가능하도록 하였습니다.

코드를 여러 번 실행해보니, 유튜브 url을 바꾼 후 이전과 동일한 저장명으로 실행할 경우

컴퓨터 내에 빈 엑셀 파일로 저장되는 유의사항도 확인할 수 있었습니다.

p13

그러면 저희가 셀레니움으로 직접 크롤링하는 과정을 담은 짧은 영상을 보고오시겠습니다.

p14

네, 이렇게 셀레니움을 이용한 유튜브 댓글 크롤링을 통해 저희가 수집한 유튜브 댓글 텍스트가

총 13782개인 것을 확인할 수 있었습니다.

p15

수집한 댓글 텍스트의 정확성과 신뢰성을 높이기 위해 ‘중복, 결측치 제거’/‘긍정,부정 label 출력’/ ‘문장 형태소 분해’ 순으로 데이터를 전처리하였는데요, 이 과정에 대해 설명드리도록 하겠습니다.

p16

우선 전처리의 첫 과정으로, 수집한 댓글 파일을 불러와 결측치 및 중복 데이터를 삭제하니

13484개의 데이터가 남은 것을 확인할 수 있었습니다.

p17

저희가 수집한 댓글 중 줄바꿈이 포함된, 여러문장으로 구성된 댓글을 찾아볼 수 있었습니다.

감성분석의 용이성을 위해 여러 문장으로 구성된 댓글은 각 문장으로 분할하는 것이 좋을 거 같아 줄바꿈과 마침표를 기준으로 분할해주었습니다.

분할 결과, 기존의 13484개의 댓글 데이터가 24641개로 늘어난 것을 확인할 수 있었습니다.

p18

이후 파이프라인 모델을 로드하여 감성 분석을 실행하였고,

감성 분석 결과를 데이터프레임으로 변환하고, 원본데이터와 합쳐질 수 있도록

보시는 것과 같이 코드를 구현하였습니다.

p19

그리고 value\_counts()함수를 통해 각 라벨의 빈도수 계산을 진행한 후,

데이터프레임으로 변환시켜 seaborn으로 시각화해보니,

p20

이렇게 중립, 부정, 긍정의 개수를 각각 확인할 수 있었습니다.

p21

이후 감성 분석을 마친 데이터 중

프로젝트에 사용하지 않을 중립 레이블 데이터를

보고 계신 코드를 활용하여 삭제하였습니다.

p22

왼쪽사진이 중립 레이블 데이터 삭제전,

오른쪽이 삭제 후의 모습으로,

확실히 데이터 개수가 차이나는 것을 확인하실 수 있습니다.

p23

그리고 남은 긍정, 부정 데이터를 가지고 형태소 분석을 진행하였는데요,

konlpy 라이브러리에 속한, 한글로 ‘꼬꼬마 형태소 분석기’라고 불리는

kkma 객체를 생성하고 형태소 분석 함수를 정의하였으며,

형태소 분석의 진행률을 파악할 수 있도록 tqdm도 사용하였습니다.

p24

형태소 분석 이후에는 보고 계시는 코드를 활용하여 ,

"댓글 형태소” 열의 각 행을 리스트로 인식하여 데이터프레임으로 변환되도록 하였고,

p25

불용어 사전 파일을 불러와 ‘댓글 형태소 불용어 제거’ 열을 초기화 및 불용어 처리를 진행하였습니다.

처리한 결과, 오른쪽 하단에 보이시는 사진처럼

진심 / 확실 / 노트북 /고급 등과 같이 처리가 잘 된 것을 확인하실 수 있습니다.

p26

형태소 분석과 불용어 처리가 된 데이터는 추후 시각화 및 데이터 비교를 위해

다음 코드에서 “label”열을 기준으로 긍정 데이터와 부정 데이터로 분리하였습니다.

p27

상단의 사진이 긍정 데이터만 추출한 결과이고,

하단의 사진이 부정 데이터를 수집한 결과입니다.

p28

전처리 과정 이후 데이터 탐색 과정을 위해 형태소 빈도수 계산을 진행하였습니다.

“댓글\_형태소\_불용어제거”열에서

형태소 리스트 추출 및 빈도수를 확인한 후,

리스트 형식인지 확인하여 리스트가 아닌 행은 리스트로 변환하였습니다.

p29

이후 리스트 컴프리헨션으로 각 행에 있는 리스트를 하나의 리스트로 합친 후

데이터프레임으로 변환하여 긍정, 부정 형태소 빈도수를 각각 계산해보니

오른쪽 하단의 사진처럼 확인이 가능하였습니다.

좌측 사진이 긍정 데이터 형태소,

우측이 부정 데이터에 대한 형태소 빈도수 계산 결과입니다.

p30

네 이제 세번째 목차, 데이터 탐색에 대해 말씀드리도록 하겠습니다.

p31

전처리 과정을 마친 데이터를 다양한 그래픽 라이브러리를 통해 데이터 시각화를 진행하였습니다.

데이터 탐색에 있어 seaborn과 matplotlib를 이용해 키워드를 추출하고

wordcloud를 통해 텍스트 네트워크를 구축했습니다.

p32

seaborn을 이용하여 빈도수가 많이 나온 긍정 형태소 상위 20개를 추출한 바 그래프입니다.

이 그래프를 통해 긍정적 리뷰에서 ‘AI’, ‘기대’, ‘성능’, ‘디자인’ 등의 단어가 빈출했다는 점을 파악할 수 있었습니다.

p33

부정 단어 빈도수 바 그래프입니다. 해당 그래프를 통해서는,

‘문제’, ‘카메라’, ‘스피커’,‘불량’ 등의 키워드를 추출할 수 있었습니다.

p34

키워드 추출 후 긍정 단어와 부정단어의 비교를 통해 키워드의 중복을 확인하고,

중복 키워드들을 통해 소비자들의 니즈를 명확히 했습니다.

긍정 키워드를 기준으로 했을 때에는 AI 기능과 휴대폰 디자인에 대한 만족도가 높았고,

갑론을박이 오가던 RAM에 대해서도 부정보다는 긍정적인 반응이 높은 것을 확인할 수 있었습니다.

p35

이어 부정 리뷰 키워드 기준의 비교 그래프입니다.

부정 키워드에서는 대중들이 카메라와 스피커를 중심으로한

실망과 문제를 느끼고 있다는 결론을 내릴 수 있었습니다.

p36

다음은 파이차트입니다.

바 그래프와 같이 상위 10개의 키워드를 사용해서 시각화 했는데,

각 키워드들의 정확한 비율을 알고자 할 때 활용하면 좋을 것 같습니다.

p37

워드 클라우드를 이용해 단어 키워드를 추출하여, 텍스트 네트워크를 구성했습니다.

긍정 리뷰 분석 결과,

‘전작 대비 디자인이 이쁘고, AI 성능을 포함해 기능이 뛰어나다는 것’을

워드클라우드를 통해서도 알 수 있었습니다.

p38

부정 리뷰 분석 결과,

‘전작 대비 카메라 불량 문제가 있고

배터리 타임 감소, 스피커 문제 등의 이슈가 있다는 것’을 알 수 있었습니다.

p39

추출한 긍정적 키워드 중 상위 10개의 키워드를 포함하고 있는 문장을 데이터프레임을 통해 확인해봤습니다.

보시는 것처럼 댓글과 어떤 형태소가 포함되었는지 확인할 수 있습니다.

p40

부정적 키워드 상위 10개를 포함하고 있는 문장을 확인한 모습입니다.

이 역시, 댓글과 어떤 형태소가 포함되어있는지 한눈에 확인이 가능합니다.

p41

마지막으로 프로젝트의 최종 함의를 말씀드리겠습니다.

p42

키워드 추출의 결과를 통해 가져온 긍정 리뷰데이터는

AI 기능을 바탕으로한 이쁜 핸드폰이었습니다. 실제로 갤럭시 s24가

2016년 이후 갤럭시 시리즈 중 최대 판매량을 기록한 것과 맞물려

긍정 키워드 데이터는 어디서나 찾아 볼 수 있는 리뷰들을 대변했습니다.

한편, 전작보다 일부 성능이 좋지 않다는 갤럭시 s24의 카메라 문제,

빠르게 소모되는 배터리 수명 등의 이슈 또한 기사를 통해 찾아볼 수 있었습니다.

이렇게 갤럭시 s24 브랜드의

긍정, 부정 리뷰 데이터 키워드 추출을 통한 평판 분석을 수행했습니다.

그 결과로 “갤럭시 s24"라는 브랜드의 강점과 문제점을 도출했고,

실제 기사를 확인하여 비교도 해보았습니다.

이렇게 리뷰 데이터 분석을 바탕으로 마케팅 아이디어를 구성한다면,

고객의 구매 만족도가 상승하고 문제점을 개선할 수 있는 기회를 잡을 수 있으리라 생각합니다.

p43

분석 과정의 개선점과 프로젝트 소감 말씀드리며 발표를 마무리 하도록 하겠습니다.

전처리 과정 중 비정형 텍스트 데이터를 비지도학습 방식으로 분석을 해야 해서 어려움이 많았습니다.

이번 프로젝트에는 파인 튜닝 된 모델을 사용해서 감성분석을 진행했지만,

그 성능이 기대에 미치지 못해 아쉬움이 많이 남습니다.

기회가 된다면 직접 모델을 만드는 과정에 참여하여 이 부분의 성능을 끌어 올린다면,

프로젝트의 전체적인 퀄리티가 더욱 발전될 수 있으리라 생각합니다.

p44

세미프로젝트 후기도 나눠보았습니다. 짧게만 말씀드려보자면,

준홍님은 텍스트 분석의 전처리 과정과

변화하는 데이터의 형태를 파악하는 점에서 조금 어려움을 느끼셨고,

진수님은 결과를 도출했다는 점에서 만족하시면서도

데이터 수집과 전처리 과정에 어려움을 느끼셨습니다.

현진님은 프로젝트를 통해 데이터 분석의 흐름을 다시 한번 살펴보시면서

어떤 부분을 채워가야하는지 배우셨다고 합니다.

마지막으로 저는 짧은 시간내에 계획대로 프로젝트를 마무리했다는점과

저의 부족한 부분이 무엇인지 느낄 수 있어 만족했던 시간같습니다.

p45

이상으로 땅콩버터팀의 세미프로젝트 발표를 마치도록 하겠습니다.

경청해주셔서 감사합니다.

질문 있으신 분은 질문 남겨주시면 감사하겠습니다 !