이어서 '데이터 수집 및 분석' 파트 발표를 진행할 서석주라고 합니다. '데이터 수집 및 분석'은 '어플 리뷰 크롤링', '감성분석', '연관어분석' 순으로 진행했습니다.

먼저, '어플 리뷰 크롤링'에 대해 발표하겠습니다. 저희는 구글플레이, 앱스토어 별 쿠팡이츠 관련 어플 리뷰를 크롤링하였습니다. 쿠팡이츠 관련 뉴스 데이터나 SNS데이터에 비해 어플 리뷰가 양질의 데이터를 제공해서 어플 리뷰를 크롤링하기로 정했습니다. 또한 다양한 관점에서 기업의 이미지를 분석하기 위해 소비자의 관점을 대변하는 '쿠팡이츠', 자영업자의 관점을 대변하는 '쿠팡이츠스토어', 라이더의 관점을 대변하는 '쿠팡이츠배달파트너' 어플 리뷰를 크롤링했습니다.

크롤링은 크롬드라이버를 이용하여 날짜, 별점, 좋아요 수, 리뷰 내용을 추출하였습니다.

다음으로 '감성분석'에 대해 발표하겠습니다. 감성분석을 진행하기 위해 별점과 리뷰 칼럼만 추출했습니다. 다음으로 'STAR' 칼럼의 문자열에서 별점만 추출해서 나타냈습니다.

텍스트에 대한 전처리는 전처리 함수 및 불용어 사전을 직접 작성하여 진행했습니다. 정규 표현식 처리를 통해 띄어쓰기를 하나 이하로 포함한 단어를 추출했고, 거기서 한글자 키워드와 불용어는 제거했습니다.

다음으로 CountVectorizer 함수를 사용하여 BoW 벡터를 생성했고 TfidTransformer 함수를 사용하여 TF-IDF를 만들어 주었습니다.

다음으로 별점의 분포를 살펴보았습니다. 5점의 도수가 높은 것을 볼 수 있었습니다. 저희는 4, 5점은 긍정 리뷰로 1,2,3점은 부정 리뷰로 분류했습니다.

긍정 리뷰와 부정 리뷰의 클래스 불균형을 해소하기 위해 긍정 리뷰에서 부정 리뷰의 개수인 4501개를 샘플링 해 주었습니다. 그리고 모델 학습을 위해 train set과 test set을 분리해주었습니다.

모델을 학습하기 전 정확도를 기준으로 다양한 모델의 성능을 비교해보았습니다. LR(LogisticRegression), NaiveBayes, 선형 SVM, RBF_SVM, Polynomial_SVM을 비교했습니다.

모델의 정확도는 LR모델, 선형 SVM, RBF_SVM이 높게 나왔습니다. 그러나 단순 긍부정 분류가 아닌, 어플에 대한 고객의 평가를 알고자 하는 측면에서 LR모델을 선택했습니다. 오른쪽 아래 보이는 그래프는 LR 적합 모델 계수 그래프입니다. 가운데를 기준으로 위쪽은 긍정적인 의미를 나타내는 계수이고, 아래쪽은 부정적인 의미를 나타내는 계수입니다. 계수 그래프 상으로는 부정적인 의견이 더 많다고 보입니다.

다음으로 긍정 TOP 20 키워드와 부정 TOP 20 키워드를 추출했습니다. 긍정 키워드에는 최고, 실시간, 만족, 배송, 아주 등의 단어가 보이고, 부정 키워드에는 사람, 삭제, 기업, 최악, 쓰레기 등의 단어가 보입니다.

지금까지의 과정을 '쿠팡이츠배달파트너'에도 진행했고 결과는 위와 같습니다.

'쿠팡이츠스토어'에도 진행했고 결과는 위와 같습니다. 이 두 어플의 계수 그래프를 보아도 부정적인 의견이 더 많다고 유추할 수 있습니다.

좀 더 객관적으로 여론을 확인해 보기 위해, 세 어플의 여론 수치화를 통해 여론을 확인해보았습니다. Numpy.sign함수를 사용하여 전체 단어에서 상대적인 부정 키워드의 개수를 비교하여 수치화를 했습니다. 쿠팡이츠의 경우 -0.55, 쿠팡이츠배달파트너의 경우 -0.345, 쿠팡이츠스토어의 경우 -0.57의 값을 보였습니다. 쿠팡이츠 기업이미지에 대해 부정적인 여론이 형성되어 있음을 알 수 있었습니다.

더 나아가 구체적인 문제점을 알아내기 위해 연관어 분석을 진행했습니다. 전처리 및 토큰화를 진행해 주었고, Word2Vec 모델을 생성해주었습니다. Word2Vec 모델 생성 시 Skip-gram 모델을 사용했습니다. CBOW에 비해 더 좋은 성능을 나타낸다는 의견이 지배적이어서 Skip-gram 모델을 사용했습니다.

감성분석에서 얻은 긍부정 TOP 20 키워드와 연관된 단어들을 통해 문제점을 정리했습니다. 예를 들어, '최고'라는 키워드와 속도, 제일, 로켓, 장점, 아주 라는 단어들이 연관이 있다고 나왔습니다. 이를 통해 배달 속도에 강점이 있다고 추측할 수 있습니다.

부정 키워드의 경우에도 예를 들어, '사람'이라는 키워드와 손절, 사경, 사망, 분식집, 컨슈머 라는 단어들이 연관이 있다고 나왔습니다. 이를 통해, 최근 비상식적인 컨슈머의 갑질로 자영업자가 사망에 이르게 된 사건을 통해 사람보다 서비스를 우선시하는 쿠팡이츠의 모습이 부정적으로 비춰졌다고 추측할 수 있습니다.

좀 더 가시적으로 문제점을 확인하기 위해 별점이 4점 이하의 리뷰만으로 워드클라우드를 만들어 보았습니다. 쿠팡이츠의 경우 사람, 고객센터, 취소 등 감성분석에서 분석한 키워드가 나타나는 것을 볼 수 있습니다.

또한 고차원 벡터를 2차원으로 변환한 T-SNE 플롯을 만들어 보았습니다. 파란색 원부분은 긍정적인 단어들의 군집이고 빨간색 원 부분은 부정적인 단어들의 군집입니다. 긍정적인 군집에서 배달, 스피드 등의 장점을 다시 확인 할 수 있었고, 부정적인 군집에서는 사기, 수수료, 갑질, 문제점, 책임, 결제 등의 문제점을 다시 확인 할 수 있었습니다. 부정적인 군집이 더 많은 것을 확인할 수 있습니다.

다음으로 문제점 도출 및 해결전략 제시 파트를 인혁님께서 발표해주시겠습니다.