SAMSUNG CARD IDEA CHALLENGE

Track 1 모델 기획서

상담원 및 삼성카드 서비스에 대한 만족/불만족 피드백 분류 모델 개발

Team 람지

알고리즘에 입력될 토큰 & Feature의 구성 방법
(모델 입력까지 전처리)

ㅂ : 14　ㅈ : 13　ㄷ : 7　ㄱ : 5
ㅃ : 14　ㅉ : 13　ㄸ : 7　ㄲ : 5　· · ·
ㅍ : 15　ㅊ : 12　ㅌ : 6.5　ㅋ : 5.5

① 문장 내 불필요한 요소 제거
(. , ; ! 띄어쓰기 등)

안녕하세요, 삼성카드입니다!

② 문장을 2음절씩 자르기

통화대기시간이너무길다
-> 통화 화대 대기 기시 시간
간이 이너 너무 무길 길다

③ 쪼개진 2음절을 자음과 모음으로 나누기

삼성 -> ㅅ ㅏ ㅁ ㅅ ㅕ ㅇ

유니코드로 자음 모음 분리

ord 함수로 텍스트를 유니코드를 확인 후,
아래 식에 따라 초성, 중성, 종성으로 분리

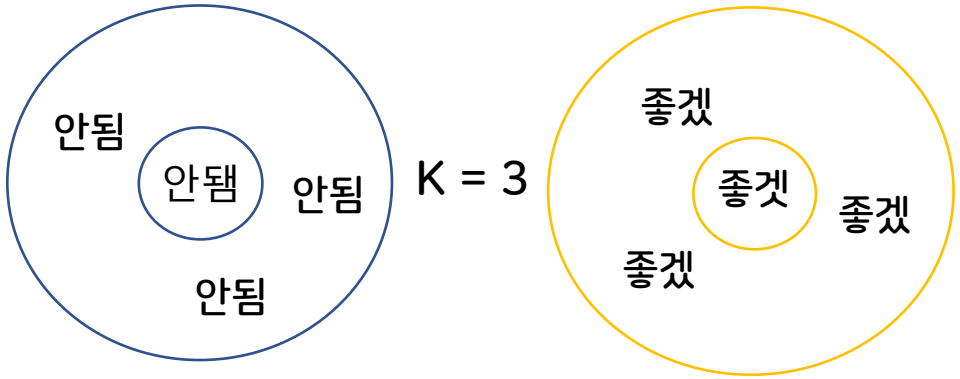
글자의 자모를 1차원 리스트에 저장 후,
2차원 리스트에 이를 순차적으로 추가

한글 코드 : (초성*21+중성)*28+종성+0xAC00
한글 코드 범위 : 44032 ~ 55199 (외 값은 예외처리)

오타 발생시 키보드의 근처에 위치한
자모음들로 바뀌었을 가능성이 높으므로,
쿼티와 천지인 키보드의 위치를
모두 고려한 좌표로 설정

2음절이므로 각 글자의 초성 중성 종성을
좌표로 설정하여 6차원의 좌표가 생성됨
ex. 상담 : [1, 9, 6, 7, 9, 10]

④ 키보드 위치를 고려하여 자음 모음의 좌표 설정하기



비슷한 단어들을 묶는 것만으로
실제 오타를 모두 발견하여
고치기에는 한계가 있음

추후 적용할 classification에서
오타를 다른 feature로 착각하는
것을 방지하며, 같은 feature로
통일 하는 것에 의의를 둠

⑤ 각 단어들의 좌표를 기준으로
k-Nearest-Neighbor(kNN) 알고리즘을 적용하기

현재음절로부터 거리가 가까운 3개의 음절 중
빈도수가 높은 음절로 수정함

전체 데이터 셋에서의 등장 횟수가 2 번 이하일 경우,
오타이거나 오타를 바로잡을 음절로 적합하지 않다고
판단하여, 해당 음절로 수정하지 않음

문장에서 앞 뒤에 위치한 음절들 중
빈도수가 더 높은 단어를 대상으로만 변환하여
해당구역의 의미가 바뀌지 않도록 함

ex. '~해택만족~' 에서 '택만'을 '객만'으로 수정할 때,
'객만'보다 '만족'의 빈도수가 높으므로 '택만' 수정 X

Feature Importance & Tokenizer

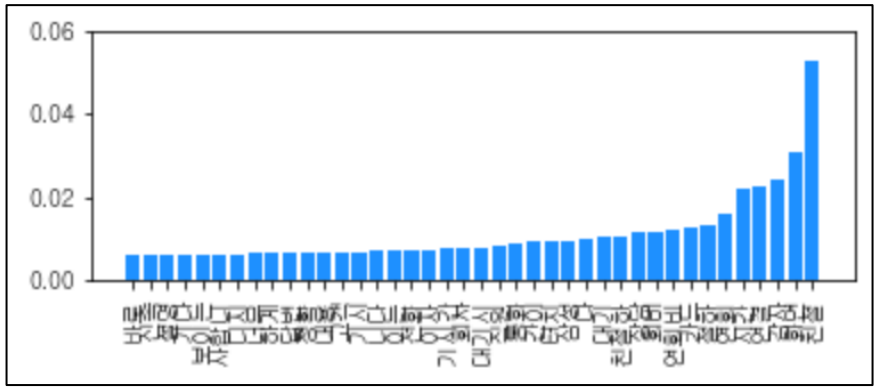
Tokenizer를 활용한 2음절 상위 n개 단어

상담	니다	친절	감사	합니
으면	절한	담감	하고	화면
면좋	기시	전화	좋겠	불편
네요	세요	원과	까지	해서
안내	보이	터치	화대	응대
성카	드사	지않	가너	는것
했습	주시	사하	절상	되었
무이	기다	결되	었으	질문
문의	어려	필요	했으	다는
되어	결시	지만	는상	절했

Tokenizer를 활용한 3음절 상위 n개 단어

합니다	감사합	사합니	친절하	습니다
기시간	셔서감	주셔서	해주셔	하게상
상담직	담직원	원연결	사드립	너무길
성카드	카드사	가너무	상담에	주세요
무이자	었으면	화연결	게해주	다감사
친절감	절감사	원과의	연결되	연결시
불편함	상담을	통화하	는상담	고객의
하게해	는화면	사했습	요감사	직원과
의통화	다른카	른카드	좋아요	설명해
드상담	었는데	화면으	면으로	해서감

Tokenizer를 활용해 빈도수에 따른
상위 n개 음절을 선별하여 1차 feature 설정



중요도 상위 40개에 대한 feature의 수치 그래프
(최상위 5개 음절 : 친절, 화면, 감사, 연결, 시간)

발화	이는a	다감사	하기	절한상	으면합	연결하	합니다	이길	rs	상담하	...	것같	대해	무친	원이	맡투	드사	드립	연결대	가너무	기분이
직원상담시간너무지연됨짜증유발	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
상담톡이부족	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
상담사와연결부분이어려웠어요	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
톡상담인원을늘려주세요요즘같은때에는전화상담보다톡상담이더신속하고일처리가빠른데인원이너무적은듯매번기다리다일처리못하고그냥지나간적이한두번이아닙니다	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0
ARS지동응답불필요한안내축소상담직원과빠른연결을원합니다주석연휴잘보내세요	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1상담원과의대화까지소요되는시간이길어서힘들었음2삼성카드사의직원과면대면하여질문+답변이되길수있으면좋겠습니다	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	1	0	0	0	0
문자상담이나온라인상담이가능한부분이많아졌으면좋겠습니다	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
상담직원과연결이좀신속했으면좋겟네요	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
상담번호를남겼는데너무늦게주심이넘불만입니다	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
상담직원대기시간이짧았으면한다	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

10 rows × 400 columns

각 문장에서 400개의 1차 feature의 포함유무를 표기하는
One-Hot Encoding을 진행 후 Data set으로 저장

발화	담을	았으면	요감	는상담	절상	감사합	었습니	비가	활부	요감사	...	보다	대기	정확	보이	친절	디지	화면	시간	연회비	무이
직원상담시간너무지연됨짜증유발	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
상담톡이부족	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
상담사와연결부분이어려웠어요	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
톡상담인원을늘려주세요요즘같은때에는전화상담보다톡상담이더신속하고일처리가빠른데인원이너무적은듯매번기다리다일처리못하고그냥지나간적이한두번이아닙니다	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	0
ARS지동응답불필요한안내축소상담직원과빠른연결을원합니다주석연휴잘보내세요	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1상담원과의대화까지소요되는시간이길어서힘들었음2삼성카드사의직원과면대면하여질문+답변이되길수있으면좋겠습니다	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
문자상담이나온라인상담이가능한부분이많아졌으면좋겠습니다	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
상담직원과연결이좀신속했으면좋겟네요	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
상담번호를남겼는데너무늦게주심이넘불만입니다	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
상담직원대기시간이짧았으면한다	0	1	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	1	0	0

10 rows × 242 columns

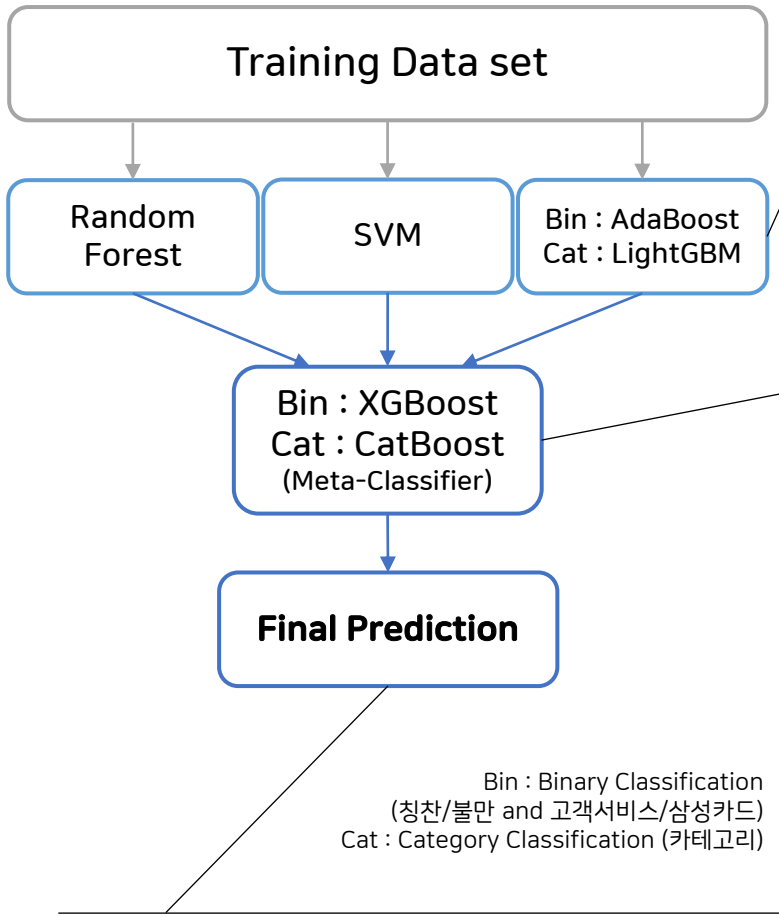
완성된 1차 feature와 label을
RandomForest와 XGB에 fit 후, feature importance를
상위 n개씩 뽑아 겹치는 feature 242개 최종 선별

선택한 알고리즘 - Stacking Ensemble

여러 개 모델의 예측 데이터를 통해
최종 메타 모델이 데이터셋을 재생성

장점

단일 모델 대비 잠재적인 결과를 이끌어 내어 성능 향상
데이터 특성상 적당한 overfit은 이로울 것이라 예측



사용된 개별 모델 & 활용 이유

RandomForest → 예측 변동성이 낮아 overfitting 방지
SVM(SupportVectorMachine) → 여러 카테고리를 다룰 때 좋은 성능
AdaBoost → 분류 쉬운 feature의 가중치를 낮춰 성능 상승
LightGBM → 카테고리 feature의 자동 변환 및 최적 분할 기능

사용된 최종 모델

① 칭찬/불만 & ②고객서비스/삼성카드

XGBoost
Greedy를 활용하여 overfitting 방지
타 알고리즘과의 연계성 우수

③ 하위 카테고리

CatBoost
Categorical Boost로 카테고리
feature가 많을 때 성능 우수

각각의 모델에서 예측된 칭찬/불만, 고객서비스/삼성카드, 하위 카테고리
총 3가지 Label을 모두 합쳐 하나의 최종 Label이 예측됨

튜닝 예정 파라미터

Overfitting 예방을 위한 파라미터

learning_rate
학습 속도와 Stacking을 감안한 overfit을 고려하여 조정 ←

max_depth
overfit 방지를 위해 최소한의 분류 가능 수준에서 각 모델마다 조정 ←

여러 feature를 활용함에 따른 파라미터

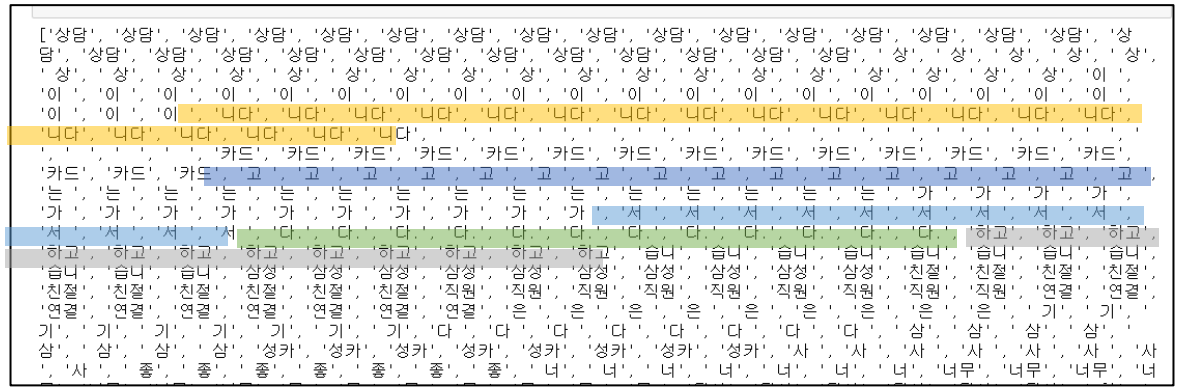
reg_lambda
많은 feature를 활용함에 있어 가중치의 L2 reg 미세한 조정 필수 ←

reg_alpha
마찬가지로 많은 feature에 대응하기 위한 L1 reg 미세한 조정 필수 ←

복문에 대한 대응 전략(Labeling과 Post-processing)

① 복문 찾아내기

데이터셋에서 발화가 2개이상인 문장을 복문으로 인식하고,
이 문장들 중 빈도수가 높은 순으로 2음절씩 뽑음
그 결과 명사와 조사를 제외한, 문장과 문장을 연결하는 음절로써
'니다' '고_' '서_' '다_' '하고' 가 가장 많이 등장함
따라서 위와 같은 음절이 나온다면 복문으로 판단함



② labeling + feature 구성하기

음절을 기준으로 발화 개수에 맞춰 문장을 최대 3개로 잘라
각각을 발화1, 2, 3에 Labeling 후 feature을 구성함

Ex.

1289	11930	정확하게설명이필요하고/상담원연결이잘안된다

이 문장에서는 '고_'를 기준으로 문장을 두개로 나눔
'정확하게설명이필요하고 ' : 발화1 불만>고객서비스>상담원
'상담원연결이잘안된다' : 발화2 불만>고객서비스>상담시스템

발화1	발화	발화2	발화
불만>고객서비스>상담원	4	불만>고객서비스>상담시스템	5

③ 새로운 데이터셋의 복문 처리하기 + 후처리

위에서 제시한 방법으로 복문인 문장을 찾은 후,
문장을 잘라 모델에 적용함

각 문장에 대한 카테고리를 제공된 우선순위에 따라 정렬 후
높은 우선순위를 가지는 카테고리를 최종 결과값으로 도출함

복문이 아닌 문장들은 카테고리가 하나만 도출되므로
우선순위를 고려할 필요없이 최종 결과값으로 확정됨

감사합니다 ☺