데이터 셋 6 일상 대화

세 종 대 왕

오 성 균 · 양 민 석 · 이 오 연

DACON 훈민정음에 스며들다 대화 요약 역량 평가 데이터 셋을 활용한 일상 대화 요약 프로젝트

데이터 개요

◇◇◇◇

데이터셋 살펴보기

개요

SNS 일상 대화 원문 및 요약 원문 데이터 셋
총 16가지 주제의 35만건 데이터

데이터 선택

행사 데이터 선택 / 위치적 데이터 같이 유의미한 데이터를 얻기에 용이
화자 2명, 4턴 이상으로 동일하게 제한한 대화 원문 1만건을 활용해 결과 비교

최종 목표

명사 빈도수를 기준으로 다양한 방법으로 예측한 결과를 비교

세부 목표

비교 대상 선정

Okt

MeCab

분석 예측 1

분석 예측 2

분석 예측 2는 앞선 조건에서 4턴인 경우에 숫자 데이터를 살려 분석 진행 (약 1000개)

역할 분담

개인 역할 분담

양민석

데이터 추출
Mecab 활용 작업

이오연

데이터 전처리
데이터 분석

오성균

OKt 활용 작업

데이터톤 데이터셋 6

데이터 전처리

데이터 전처리

대화 원문 데이터 확인하기 1

1차 정제 데이터 제공

민감한 정보에 대한 정제가 진행된 데이터

이름	장소	기타
실명 → 이름 공인명, 별명은 유지	거주 주소 → 주소 상호명, 역은 유지	비범주 영역 → 기타
		소속
온라인	**각종 번호**	출신 및 소속 → 소속
아이디, 이메일주소 URL → 계정	전화번호 → 전번 금융번호 → 금융	**숫자 및 특수기호**
		이미 전체 제거

데이터 전처리

대화 원문 데이터 확인하기 2

1차 정제 데이터 제공

정제 과정에서 덜 정제되거나 깨진 문장 다수 존재

데이터 제거 대신 대체하는 방식으로 진행했으면 좋았을 것 같음 (예: 1시 → -시)

(*정규표현식 진행 후 가져오는 과정에서 누락 : 원본 데이터와의 대조 및 데이터 확인 필수 진행 필요성)

이모티콘 등 규모가 크고 정보성이 없는 기타 데이터 존재

실제 데이터 분석에서 혼동이 생기지 않게 안내 필요성 느낌

" 박 일 시스템 사진 행복해" → -박 -일 + 사진 데이터

"생각도 안햇슴 셋이 놀지머 이모티콘" → 이모티콘 데이터

데이터 전처리

◇◇◇◇

대화 원문 데이터 확인하기 3

긍정적 부분

대다수가 20-30대인 대화 원문으로 해당 연령대의 행사 데이터 수집에 용이
반말 80%, 존댓말20% 정도 비율 , 화자의 생활 환경(지역, 직업 등)이 다양한 편

"연차냈다" / "학원때문에 늦게 만나서 좋은거같다"
"결혼하면 자주못보니깐 그런파티하면 좋지"
"교회 청년부 동생 지금" / "퍼뜩 말해봐라마"
"가고 싶은겨" / "약속이 취소되었슴다"

데이터 전처리

◇◇◇◇

대화 원문 데이터 확인하기 4

부정적 부분

구어적 속성이 강해 문법적 오류가 많음 (오타, 줄임말, 띄어쓰기 생략 등)
한명의 화자가 일방적으로 대화를 이끄는 경우가 많아 행사 데이터 수집에 불리

"흑흑 알았써" / "셤보고 바로가" / "오키" / "알써" / "오오"
"옹옹 담주 토요일 워땨" / "언니들두가치오나여"
"보고띠포" / "백숙먹구 삼겹살꿔먹구 금요일날 계곡갈까"
"친구네댕댕세마리랑친해져잇어야돼" / "댕댕사진점보여줘"

데이터 전처리

◇◇◇◇

요약 원문 데이터 확인하기 1

긍정적 부분

본문의 내용을 잘 요약
대화 원문에 비해 문장 구성이 매끄럽게 이어짐

너싫다며, 몇일?, -ㅅ-, 안싫어돈이없어, 몇일로잡았는데, 나안정했지,
그냥가고싶다는거였는데, 아!!너운전면허있지?, 대박대박, 타지에서
죽고싶으면, 나한테운전하라고해, ㅋㅋㅋㅋㅋㄱㄱㅋㄱ, 장롱면허, 훠이훠이

→ **제주도에 가자고 하며 운전면허에 대해 이야기하고 있다.**

데이터 전처리

◇◇◇◇

요약 원문 데이터 확인하기 2

부정적 부분

요약문 길이가 절반 이상인 경우가 많음 (키워드를 제공하는 요약 방식 고려)
반말, 존댓말 상관 없이 -한다 형식의 문어체로 요약이 진행
각주 같은 부분이 있는 것으로 보아 사람이 직접 수기로 작성한 느낌이 듦

아르바이트(알바) 끝나고 퀴캔을 하자고 말하였다.
대만을 2박 3일로 가기엔 힘든데 예스허지 버스투어를 가면 하루 종일
걸리고 허우통에 오래 있고 싶지만 기차 배차가 한 시간이라고 말하고
작아서 볼게 많지 않고 고양이만 보러 가는 거면 괜찮다고 이야기한다.

데이터 전처리

◇◇◇◇

불용어 사전 제작 기준 결정하기

포함 기준

1차 정제 안내 키워드 및 기타 데이터

어느정도 규모와 규칙성이 있는 1차 정제 과정 중 깨진 단어의 파편

불포함 기준

더 풍부한 대화를 구성하기 위해 줄임말 및 의성어 유지

(빈도수가 적은 의성어가 너무 많으면 노이즈가 발생하니 추후 정규화 작업으로 재처리)

자연스러운 대화 원문을 위해 나, 너 등 주어 삭제 미진행

데이터 전처리

텍스트 정규화 기준 결정하기

의의

일상 대화는 단어가 다양하고 빈도수가 적은 경향
빈도수가 적은 경우의 데이터를 정규화해 유의미한 데이터 추출

정규화 기준

빈도수 30 - 100 이하의 단어 정규화
의성어는 나름의 독자적인 의미가 있다면 수와 상관 없이 정규화 미진행

데이터 전처리

◇◇◇◇

텍스트 정규화 기준에 따른 근거 제시하기

예시 1

조아(355) / 죠아(52) 의 경우 크게 다르지 않은 의미와 형태를 가지고
빈도수가 작으므로 더 큰 빈도수를 가지는 '조아'로 정규화 진행

예시 2

(하하(45) / 히히(111) / 후후(36) / 허허허(36) / 허허(34) / 헤헤(81) / ㅋㅋ(51))
웃음을 표현하는 의성어 중 ㅋㅋ(51)의 경우 적은 빈도수를 가지고 있지만 형태가 다름
풍부한 표현력을 위해 정규화 미진행

데이터톤 데이터셋 6

데이터 분석

데이터 분석

요일 키워드에 따른 행사 분포 분석하기

월요일	화요일	수요일	목요일
233	157	207	258

금요일	토요일	일요일	평일	주말
472	585	406	155	327

토요일의 빈도수가 가장 높음

전체적으로 주말의 빈도수가 높음

데이터 분석

시간 키워드에 따른 행사 분포 분석하기

시간으로 따졌을 때 1-3시에 가장 빈도수가 높음

하지만 분류로 나눴을 때는 3분류 모두 저녁 시간대가 가장 높은 모습을 띔

예측 1. 행사의 종류와 성향이 다양 / 예측 2. 1-3시의 시간대가 새벽일 가능성

데이터 분석

기간 키워드에 따른 행사 분포 분석하기

오늘	내일	이번주	다음주	회사
1234	1688	328	576	143

하루	이틀	삼일	사일	학교
266	64	19	6	136

행사 약속을 잡는 경우 내일 약속을 잡거나 약속을 확인하는 빈도수가 가장 높음
여행과 같은 행사의 경우 하루의 빈도수가 가장 높음
대부분 학교나 회사에 다니는 20-30대의 대화 원문이기 때문에 하루가 가장 많다고 예상

데이터 분석

계절 키워드에 따른 행사 분포 분석하기

봄	여름	가을	겨울	날씨
157	92	30	84	143

봄의 빈도수가 가장 높음

날씨의 빈도수도 높은 것으로 미루어 보아 날씨를 중요하게 여길 가능성이 높음

데이터 분석

행사 종류 키워드에 따른 행사 분포 분석하기

여행의 빈도수가 압도적으로 높음

데이터 분석

여가 관련 키워드에 따른 행사 분포 분석하기

영화	수영	물놀이	캠핑	구경	관광	힐링
191	169	31	52	159	40	42

바다	등산	술	커피	카페	맛집	호텔
248	24	351	151	629	116	271

주로 식음료에 관한 빈도수가 높음

여행 이외에도 약속 등 간소한 행사도 포함되었기 때문이라고 추측

데이터 분석

여행 지역 키워드에 따른 행사 분포 분석하기

'비행기'의 빈도수가 높은데 비해 국내 지역의 빈도수가 해외 지역에 비해 월등히 높음
코로나의 빈도수도 높은 것을 보아 코로나로 인한 해외 여행 축소로 예상
하지만 지난 데이터와의 비교를 통해 해외 여행 빈도수 축소의 실제 근거가 필요

데이터 분석

인물 키워드에 따른 행사 분포 분석하기

주로 여자 가족이 함께 행사에 참여하는 빈도수가 높음
언니, 오빠 등 화자가 여성인 키워드의 빈도수가 높음
화자가 여성인 경우가 많다는 것을 추측 가능

데이터 분석

그 외 키워드 빈도수에 따른 행사 분포 분석하기

취소	예매	예약	검색
174	103	445	225
출발	**도착**	**시간**	**약속**
502	666	1412	368

주로 시간, 예약, 여행에 관련된 내용의 키워드를 중심으로 높은 빈도수를 보임

데이터톤 데이터셋 6

LDA / HEATMAP

데이터톤 데이터셋 6

추후 과제

추후 과제

추후 과제 계획하기

- 요약문에 존댓말과 반말을 반영할 수 있는 방법 찾아보기
- 코사인 유사도를 활용한 실제적인 성능 평가 비교
- 실제 숫자가 남아있는 더 큰 데이터로 예측 진행
- 실제 데이터에서 성별 및 연령, 지역, 데이터 수집일 등을 추출해 비교

데이터 셋 6 일상 대화

질 의 응 답

오 성 균 · 양 민 석 · 이 오 연

DACON 훈민정음에 스며들다 대화 요약 역량 평가 데이터 셋을 활용한 일상 대화 요약 프로젝트