| **3팀** | | **KoBert를 활용한 감성분석 일기장** |
| --- | --- | --- |
| 2022.08.30 ~ 2022.10.17 |

o 팀원 : 강동현, 정영현, 방수진, 전아영, 성혜경, 지동주, 조혜리

| **프로젝트 개요(추후 변경 가능)** |
| --- |
| 1. 목적 설정(제공할 서비스 설정) 2. 목적 설정에 따른 데이터 수집(사용자 음성데이터, 텍스트데이터) 3. 데이터 전처리(음성데이터 > 텍스트데이터) 4. 수집 데이터 DB 연동 5. 텍스트 데이터 전처리(토큰화, 정체, 추출, 불용어, 정수 인코딩 등등) 6. 감성분석 모델 학습 및 성능 평가 7. 일기장 어플 구현 및 서비스 제공 |
| **프로젝트 목적(추후 변경 가능)** |
| 1. 사용자들이 음성으로 간편하게 작성 가능한 일기장 어플 구현 2. 사용자들의 일기장 내용을 감성분석하여 감정 파악 3. 파악된 감정으로 추천 여행지, 음식 기능 제공 4. 자주 사용하는 단어 시각화로 언어습관 분석 제공 |
| **프로젝트 방법(추후 변경 가능)** |
| 1. 목적 설정    1. 감성분석이 제공되는 일기장 어플 구현    2. 사용자들에게 감정분석과 힐링 문구 제공    3. 사용자들의 언어습관 분석 2. 필요한 모델 및 라이브러리 스터디    1. Speech\_Recognition(STT모델)    2. Kobert(감성분석 모델)    3. Kobart(문서요약 모델)    4. Django(HTML) 3. 데이터 분석    1. 모델링    2. 모델 평가 및 검증    3. 모델 적용 및 운영방안 수립 4. 시스템 구현    1. 설계 및 구현       1. 모델간 연계 프로세스 설계       2. Django 활용으로 HTML 구현    2. 시스템 테스트 및 운영 5. 평가 및 전개    1. 모델 발전계획 수립    2. 프로젝트 평가 및 보고 |
| **기대 효과(활용 또는 확장성)** |
| 1. 음성으로 저장 가능한 일기장 2. 저장된 데이터를 기반으로 감정분석 서비스 제공 3. 분석된 사용자의 감정에 맞는 여행지와 음악 추천 4. 사용자들의 언어습관 분석 제공 5. 배경화면으로 현재 날씨 정보 제공 |

# 궁금한거

<1주차>

멘토링 오시면 지원 받을 수 있는 부분, 작업 중 에러, 충돌시 메모해놓기(에러내용, 컴퓨터 사양 등등)

유튜브 영상을 mp3로 다운받아야 하나?, wav로 가져올 방법은?

wav를 그대로 써야하나, 시계열데이터(변환)로 바꿔서 쓸 수 있나?

→ 실시간 음성 분석과 녹음 파일을 기반으로 분석하는 모델은 다르기 때문에 실시간이 아니라는 것을 강조할 필요가 있음

→ [논문] MFCC와 LPC 특징 추출 방법을 이용한 음성 인식 오류 보정

MFCC와 사용함수 공부

warp-rnnt : 코랩 사용 추천

음성데이터의 경우 발음이 같은 단어로 나타날 수 있는 위험 제거(ex : 탯줄->대출로 인식한경우)

> 유튜브 원본 음성데이터가 들어가서 전처리가 안된 상태임(음성, 잡음 전처리 후 대량 데이터 학습 후 확인 > 중복, 단어빈도 확인하고 필요없는거 제거

→ 오류는 전처리 과정에서 그대로 두고

→ nlp 과정에서 오류 보정 가능 ( 학습 데이터가 정확한 카테고리로 분류되어있을 경우 오류를 보완할 수 있다 – 도메인 특화 )

\*) 고려 사항

→ 말뭉치데이터를 학습데이터로 사용할 경우 일반적으로 모델을 형성 가능하나, 위 오류는 생길 수 있음

→ 도메인 특화로 진행 시 학습데이터를 구하기 어려울 수 있음

용량제한

> (regonizer\_google() 이게 구글꺼 빌려쓰는거라서 제한 걸려있음 > 모델을 직접 만들어서 raw data를 살리자 > KoSpeech 모델(카카오 개발자가 만듦) > 가상환경 실행 해보고 뜯어보기, 수행시간 체크

→ 손실된 데이터 부분을 교집합으로 겹쳐서 txt 데이터로 변환 후 문맥에 따라 교집합된 데이터 제거

영상을 토대로 설명하는 내용이므로 화면상으로만 설명되어 음성데이터에는 남아있지 않는경우

→:[영상시간] 요약된 내용

→ 영상 요약은 난이도가 높음

kospeech 모델 전처리 -> txt 언어 품질이 안좋다? > 모델 전처리 부분 수정

사투리, 축약어같은 경우 처리를

→축약어는 처리 할 수 없음

[질문사항]

* NVIDIA사용을 위한 내장 노트북을 구할수 있는지?

- 코랩으로 돌리고 안되면 모델을 변경

* 구글 -> youtube to mp3 대안툴 고민

**<2주차>**

요약모델

>kobart모델을 이용하여 텍스트 요약 시 문장의 의미 그대로 추출이 아니라서 요약된 텍스트의 원본상 의미가 맞지 않음

-다른 모델을 사용하여 해야 하는지

-kobart를 사용하되 우리가 원하는대로 나타낼 수 있는 다른 요약방법이 있는지(우리가 모르는)

마이크

웹호스팅

코랩과 장고 연동