
프로젝트 제안서

방구: 방언 구분기



팀 Guess

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

- 변 경 이 력 -

일자	버전	변경 내역	작 성 자
2022.03.31	1.0.0	프로젝트 계획서 초안	김현빈, 윤성은, 김 유민, 김부용

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

- 목 차 -

1.	프로젝트 개요	- 3 -
1.1	소개.....	- 3 -
1.2	주요 일정	- 18 -
1.3	조직.....	- 19 -
1.3.1	조직도	- 19 -
1.3.2	역할 및 책임.....	- 19 -
1.4	생명주기 모델.....	- 20 -
1.5	도구.....	- 21 -
2.	규모 산정.....	- 22 -
2.1	WBS(WORK BREAKDOWN STRUCTURE)	- 22 -
2.2	작업에 소요되는 기간	- 22 -
3.	일저	- 23 -
4.	산출물 관리	- 24 -
5.	위험 관리 계획	- 25 -

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

1. 프로젝트 개요

1.1 소개

1.1.1. 기간

2022.03.23 ~ 2022.06.23

1.1.2. 팀명

Guess

1.1.3. 프로젝트 명칭

방구(방언 구분기)

1.1.4. 프로젝트 목적

많은 나라에서 수도 지역의 방언을 표준어로 삼지만, 이는 수도 방언의 우월성을 의미하지는 않을뿐더러 방언 간에는 우열이 존재하지 않는다. 이처럼 방언은 표준어의 변화와 발전에 영향을 미친다.

이러한 방언은 언어적 다양성과 함께 문화적 다양성을 일구어 내는 토대로써 국어가 다양성을 유지한 채 건강하게 발전할 수 있게 하는 지역의 특색과 국어의 역사를 간직하고 있는 소중한 언어 문화의 유산이다.

그러나 박용식 국어문화원장은 한 인터뷰에서 이러한 방언은 15년 이내에 소멸할 것이라 전망했다. 실제로, 지난 2010년에는 유네스코에서 제주 방언을 ‘사라지는 언어’ 5단계 중 4단계인 ‘아주 심각하게 위기에 처한 언어’로 분류하였다. 다른 나라와 달리 표준어가 강력한 언어정책으로 자리 잡히면서 표준어와 비표준어로 구분하는 경향이 커, 교육이나 방송에서 사투리를 지양하게 되었다. 특히, 정승철 교수는 70~80년대 정부에서 시행한 ‘바른말 고운말 쓰기 운동’등과 가은 표준어 사용 정책을 펼치면서 방송이나 학교 교육 등에서 사투리를 지양하면서 파생된 문제라고 지적했다.

이처럼 교통과 이동통신의 발달 등 사람들 간 소통을 원활하게 해주는 수단이 늘어남에 따라 방언의 분류가 점차 희미해지고 있다. 이에 따라 국립국어원에서는 현재 방언의 보전, 연구를 위한 지역어 조사 및 지역어 정보 서비스 시스템을 구축하고 있다. 하지만 우선적으로 이루어져야 할 것은 방언이 표준어보다 열등하다는 인식일 것이다.

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

더하여, 통신업계에서는 이미 다양한 분야에 음성 기반 AI를 적용하면서 서비스를 확대해 나가고 있지만 표준어가 아닌 지역 방언까지 모두 인식하기에는 한계가 있었다. 그러나 과학기술부에 따르면 경상, 전라, 충청, 강원, 제주 등 다섯 개 도의 방언이 수집된 발화 데이터의 활용성을 검토한 결과 기존 서비스의 인식률이 12%나 확대되었다고 한다. 이는 국내 음성인식 기능이 표준어를 기준으로 만들어져 발생한 문제라고 설명했다. 이처럼 인공지능 기반의 언어 모델 개발이 표준어 데이터셋을 기반으로 개발되고 있으나 다양한 지역민과 연령의 구분 없는 언어의 소통 학습을 위하여 지역 방언 중심의 데이터셋을 구축해야 함은 분명하다.

위 프로젝트를 통하여 음성 파일을 입력 받아 지역 방언을 분류해줌으로써 방언의 바이럴 마케팅 효과를 끌어내어 방언에 대한 인식을 제고시킬 수 있는 지역 방언 분류 AI 모델을 이용한 체험형 웹사이트를 구현하고자 한다.

1.1.5. 수집할 데이터

표준어가 아닌 말은 모두 방언이라 한다. 방언은 크게 지역적 요인에 의한 지역 방언과 사회적 요인에 의한 사회 방언으로 구분한다. 위 프로젝트에서의 방언은 좁은 의미에서의 지역 방언을 의미한다. 현재 국어의 지역 방언은 동북 방언, 서북 방언, 중부 방언, 동남 방언, 서남 방언, 제주 방언의 여섯 개로 크게 나뉘어 있으나 자체 데이터셋에는 강원(0), 경상(1), 전라(2), 제주(3), 충청(4), 비사투리(5)로 분류하여 구축할 예정이다.

이렇듯 지역 방언 분류 AI 모델을 효과적으로 학습시키기 위해 AI model life cycle 및 Data lifecycle 과정을 거치며 각 강원(0), 경상(1), 전라(2), 제주(3), 충청(4), 비사투리(5) 여섯 개의 지역 방언으로 labeling 되어있는 음성 기반의 비정형 데이터로 이루어진 데이터셋을 구축할 예정이다.

1.1.6. 데이터 수집 방법



1.1.6.1. 데이터 collection

1. 오픈 데이터셋을 통해서 수집한 데이터

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문서명	프로젝트 계획서	버전	1.0.0

AI Hub 한국어 방언 발화 오픈 데이터 셋에서 사투리 단어가 확실히 포함된 문장을 골라 사용한다.

음성/자연어 한국어 방언 발화(강원도) 텍스트 오디오 2020	음성/자연어 한국어 방언 발화(경상도) 텍스트 오디오 2020	음성/자연어 한국어 방언 발화(전라도) 텍스트 오디오 2020
음성/자연어 한국어 방언 발화(제주도) 텍스트 오디오 2020	음성/자연어 한국어 방언 발화(충청도) 텍스트 오디오 2020	

과제명	주요 내용	데이터 구축량	데이터 형식
한국어 방언 발화 데이터 (강원도)	방언(강원도)을 사용하는 일상 대화를 수집하여 음성을 문자로 변환한 방언 발화 데이터셋 구축	<ul style="list-style-type: none">조용한 환경에서 2,000명 이상의 화자가 발화한 3,000시간 이상의 음성 데이터셋원본 표준어 텍스트 및 방언 특성을 고려하여 전사한 텍스트 50만건	<ul style="list-style-type: none">원본형태 : 화자가 구분된 담화 텍스트 말뭉치학습용 데이터 형태 : 방언 발화된 음성 데이터가 맵핑된 텍스트와 음성 데이터셋
데이터 종류		포함 내용	제공 방식
음성 데이터셋		총 3,000 시간 정제된 음성데이터	wav 포맷 파일
텍스트 데이터셋		총 50건의 원본 표준어 텍스트 및 방언 특성을 고려한 이중전사 텍스트	JSON 포맷 파일

각 데이터는 원천 데이터인 오디오 데이터와 라벨링 데이터인 텍스트 데이터를 포함하고 있다. 라벨링 데이터에서는 두 명의 화자의 정보와 대화 데이터 등을 포함하고 있으며 ‘방언 전사’와 ‘표준어 대응상 전사’로 방언을 구분할 수 있다. 우리는 음절리스트에서 방언 여부 데이터인 isDialect가 true인 어절을 포함한 대화를 골라 선택적으로 데이터를 수집한다.

지역	보기
강원	이게 (다나?)/ (다니?) 나도 이쪽 동네 (출신이라.)/ (출신이야.) (이라)/ (이렇게)
경상	어제 어디 (갔었노?)/ (갔었니?) 미역 (줄거리)/ (줄기) (단디)/ (단단히)
전라	혼자 다 (묵어 분당께.)/ (먹어 버린다니까.) 아 (실땡키로)/ (실처럼) 가는 거 그거? (그랑께)/ (그러니까)

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

제주	아까 (집드레)/(집으로) (가라.)/(가더라.) 너 (하*구정 한)/(하고 싶은) 대로 (하*라.)/(해라.) (아매나)/(아무렇게나)
충청	동네 사람들은 (위떡헌다?)/(어떡한대?) 가만히 (두덜)/(두질) (못하.)/(못해.) (그려.)/(그래.)

```
{
  "id": 3,
  "eojeol": "감성이",
  "standard": "감성이",
  "isDialect": false
},
{
  "id": 4,
  "eojeol": "있다",
  "standard": "있지",
  "isDialect": true
},
{
  "id": 5,
  "eojeol": "아이가",
  "standard": "알아",
  "isDialect": true
}
```

2. 데이터 크롤링 후에 스크레이핑과 정제를 수행한 데이터

‘(지역명) 사투리’ 키워드로 유튜브 검색으로 음원을 추출한다. ‘(지역명) 사투리’ 키워드로 웹 크롤링을 통해 영상 주소 정보를 수집한다.

```
from selenium import webdriver
from bs4 import BeautifulSoup as bs
import pandas as pd
from selenium.webdriver.common.keys import Keys
import time

keyword = '경상도 사투리'

options = webdriver.ChromeOptions()
options.add_argument('--headless') # Head-less 설정
options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')
driver = webdriver.Chrome('chromedriver', options=options)

url = 'https://www.youtube.com/results?search_query={}'.format(keyword)
driver.get(url)
soup = bs(driver.page_source, 'html.parser')
driver.close()

video_list = soup.findAll('ytd-video-renderer', {'class': 'style-scope ytd-item-section-renderer'})
home_url = 'https://www.youtube.com'
video_url = []

for i in range(len(video_list)):
    result_url = home_url+video_list[i].find('a',{'id':'thumbnail'})['href']
    video_url.append(result_url)
    print(result_url)
```

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

```
https://www.youtube.com/watch?v=84dzzrt1OVM
https://www.youtube.com/watch?v=mRYFJWug-s
https://www.youtube.com/watch?v=DNDQ6zuKdv8
https://www.youtube.com/watch?v=BGSSlAPbl-M
https://www.youtube.com/watch?v=sPCY27BPQ64
```

크롤링을 통해 수집한 유튜브 주소와 pytube 유튜브 동영상 다운로드 라이브러리를 사용하여 음원을 추출한다.

```
from pytube import YouTube

#유튜브 전용 인스턴스 생성
yt = YouTube('https://www.youtube.com/watch?v=84dzzrt1OVM')

print(yt.streams.filter(only_audio=True).all())

# 특정영상 다운로드
yt.streams.filter(only_audio=True).first().download()
```

수집된 음성 데이터를 텍스트로 전사한 후 정제 및 가공하는 작업을 통해 최종 오디오 + 텍스트 데이터를 수집한다.

1.1.6.2. 데이터 labeling

데이터 라벨링은 in-house labeling 방법을 적용하여 팀 내에서 해당 데이터가 어느 지역 사투리인지에 대해 라벨링한다. 분류는 강원(0), 경상(1), 전라(2), 제주(3), 충청(4), 비사투리(5)로 총 6개이다.

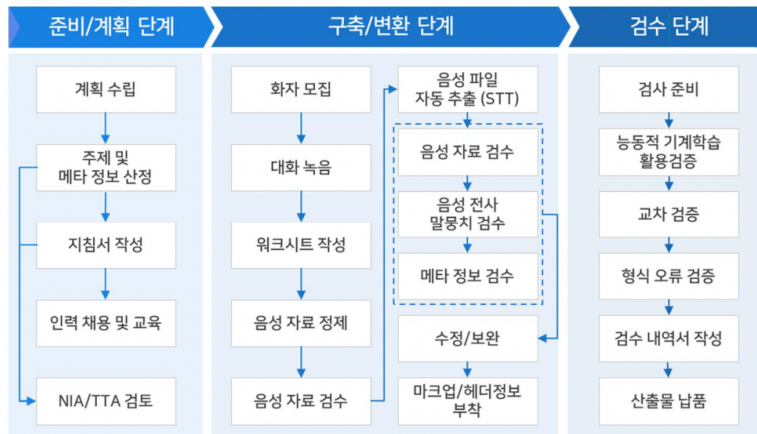
1.1.6.3. 데이터 verification

데이터는 AI Hub의 오픈 데이터셋과 유튜브에 전체 공개된 영상을 수집 대상으로 하며, 오픈 데이터셋을 이용할 땐 이용 신청을 하고 유튜브 영상에서 추출한 음원에 대해서는 원본 URL을 명시하는 페이지를 만들어 출처를 표기한다. 수집한 모든 데이터는 교육 목적의 프로젝트 진행에만 쓴다.

1.1.6.3. 데이터 verification

오픈데이터셋의 데이터는 다음과 같이 체계가 잡힌 검증단계를 거쳐 수집되었다.

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0



한국정보화진흥원의 데이터베이스 구축방법론(Ver.4)을 적용하여 음성 녹음, 이중 전사, 원시 말뭉치 구축에 대한 대상 자료별 공정 태스크와 주요 활동 절차를 표준화하여 효율적인 학습용 데이터셋 구축 체계를 확보하고 한국어 방언 시데이터 구축에 적합하도록 자료의 특성을 고려하여 준비/계획단계, 구축/변환단계, 검수단계의 3단계 공정을 거친다.

팀 내에서 직접 수집한 데이터를 포함한 전체 데이터에 대해서는 다음과 같은 기준으로 검증한다.

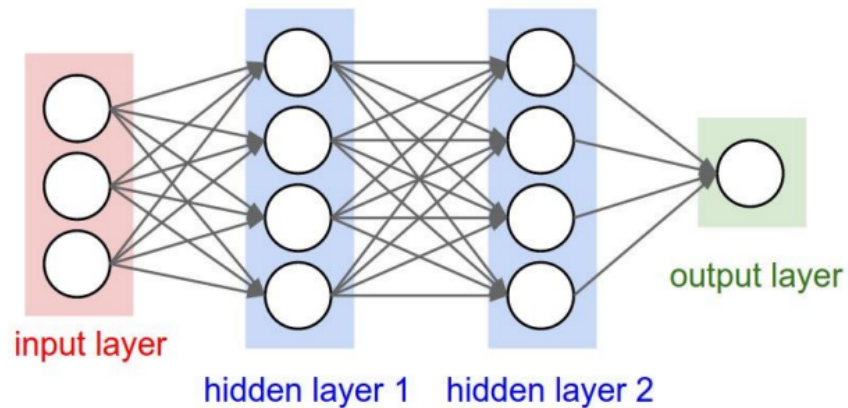
- 각 지역별로 데이터 수에 차이가 없다.
- 각 지역별로 오픈 데이터셋과 수집 데이터셋의 비율에 차이가 없다.
- 각 지역별로 발화자가 특정 성별에 치우치지 않는다.

1.1.7. 데이터 분석 방법

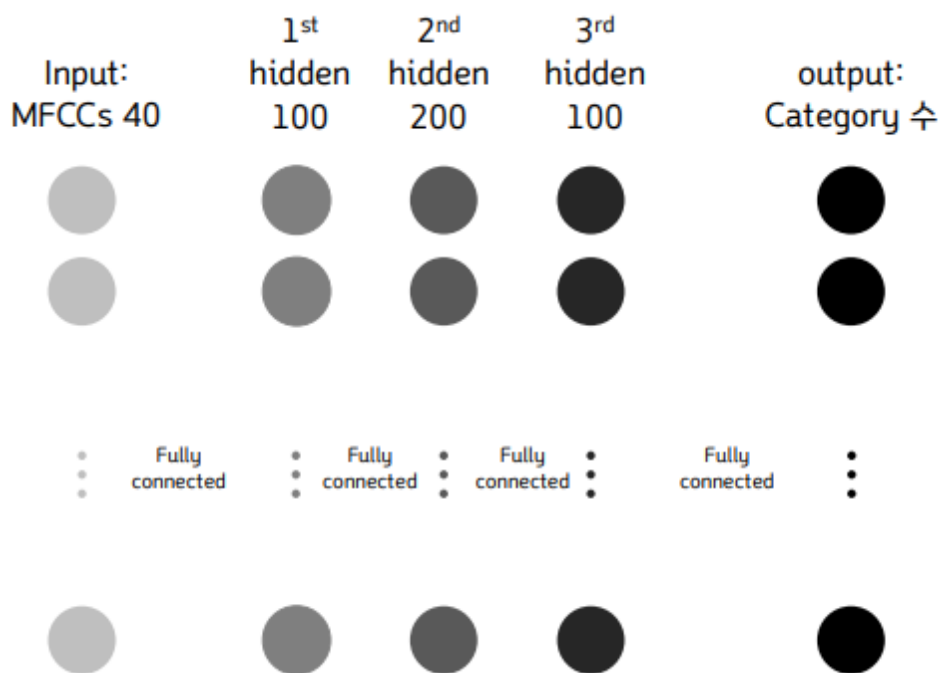
기존의 Model-centric AI는 알고리즘과 코드를 지속적으로 수정하는 과정에서 그 성능을 개선하는데 중점을 두고 있었다. 하지만 Model-centric AI의 문제점은 데이터 정리(보통 정상), 모델 교육, 유효성 검사, 배포, 저장, 공유 등 모든 것이 수동으로 수행된다는 점이었고, 이에 반해 Data-centric AI는 데이터 자체 개선만으로도 Model-centric AI보다 더 나은 AI의 성능을 보여준다는 점에서 그 효율성이 입증되었다. 이에 따라 프로젝트의 AI 방향으로 Data-centric AI를 선택하였다.

기계학습 중, 심층학습의 일종인 다층 퍼셉트론(Multi-Layer Perceptron)을 이용해 카테고리를 분류(classification)한다. 다층 퍼셉트론이란 인접한 층 사이의 모든 노드가 연결되어 있으며(fully-connected) 은닉층이 두 개 이상인 인공 신경망을 말한다. 기본적인 구조는 다음 그림과 같다.

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0



1출처: 2017 C231n lecture 4 p97

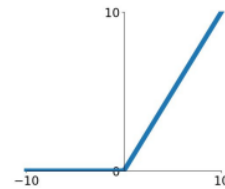


입력층(input layer)에 오디오로부터 추출된 mfcc값 40개를 특징(feature) 즉, 입력 데이터로 넣으면 노드가 각 100, 200, 100개씩인 은닉층(hidden layer) 3개를 거치며 데이터로부터 분류 작업에 의미 있는 특징들이 추출된다. 노드의 가중치(weight)와 곱해져 데이터를 잘 표현하는 값은 증폭되고 잘 표현하지 못하는 값은 작아진다. 가중치와 곱해진 특징 값은 각 은닉층마다 활성화 함수(activation function) ReLU(Rectified Linear Unit)에 의해 새로운 특징 공간에 비선형적으로 매핑된다. 여

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

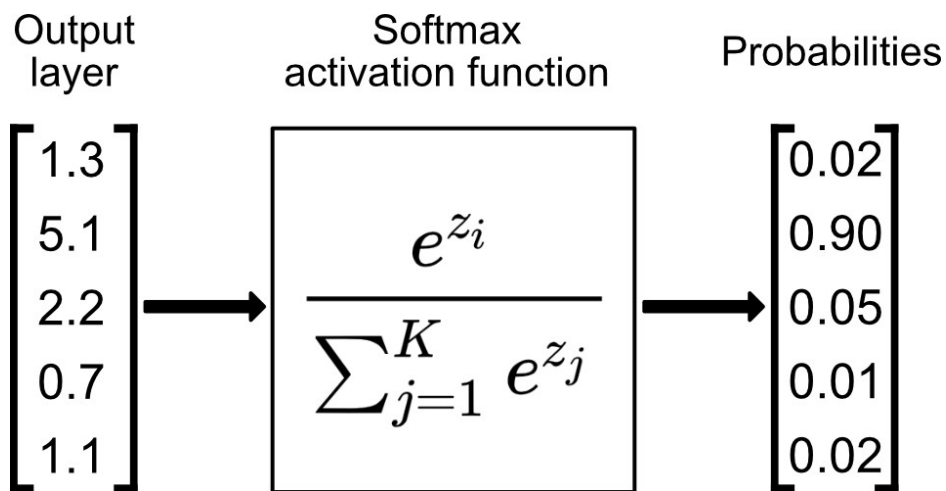
기서 활성화 함수란 특정 임계치에 미치지 못한 값의 탈락에 관여한다. ReLU는 그 일종이며 식은 다음 그림과 같다.

ReLU
 $\max(0, x)$



2출처: 2017 C231n lecture 4 p96

모든 은닉층을 거친 뒤, 출력층(output layer)에서 6개의 값이 출력으로 나오게 되는 데 이는 해당 오디오가 각 6개의 카테고리에 속할 확률을 나타낸다. 출력층에 Softmax를 적용하여 확률 값을 정규화하므로 최종적으로 출력되는 6개의 값은 총합이 1이 되는 0에서 1 사이의 값으로 나오게 된다. Softmax의 식은 다음과 같다.



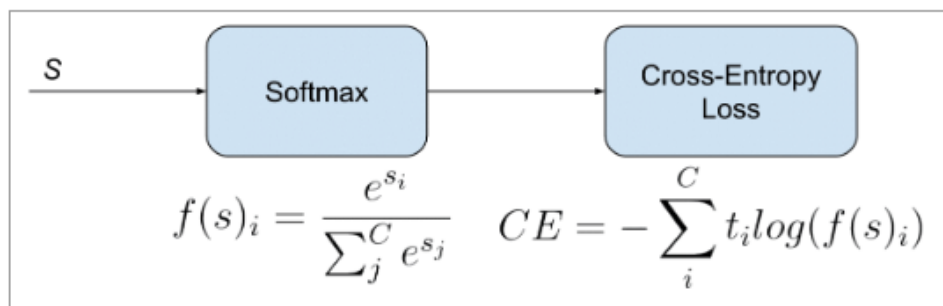
3출처: <https://towardsdatascience.com/softmax-activation-function-explained-a7e1bc3ad60>

위에 설명한 인공 신경망은 입력되는 특징 값에 곱해지는 좋은 가중치 즉 파라미터 (parameter) 값을 찾기 위해 학습 과정을 거친다. 학습 방법으로는 출력 값에서 구한 오차(loss)를 이용해 역방향으로 층의 파라미터를 갱신하는 오차 역전파법 (backpropagation)을 사용한다.

오차 또는 손실값이란 인공 신경망이 데이터에 대해 얼마나 성능을 내는지 알 수 있는

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

수치로 손실함수(loss function) 또는 목적함수(objective function)로 계산된다. 여기서는 손실함수로 CE loss(categorical cross-entropy loss)를 사용한다. 이는 데이터를 3개 이상의 카테고리로 분류하는 다중 분류(multiclass classification) 문제에서 MSE(mean squared error)보다 더 빨리 수렴한다는 이유로 많이 사용된다. 식은 다음 그림과 같으며, softmax로 정규화 된 출력층의 결과값에 로그를 취한 뒤 정답(true, t_i)과 곱한 것들의 총합에 -(minus)하여 구한다. Softmax를 사용하기 때문에 softmax loss라고도 불린다.



4출처: https://gombu.github.io/2018/05/23/cross_entropy_loss/

위에서 설명한 오차를 최소화하는 파라미터가 좋은 파라미터이며, 그것을 찾는 것이 인공지능망 학습의 목적이다. 인공지능망의 학습은 손실함수의 기울기를 이용해 파라미터가 최적화될 때까지 반복적으로 갱신함으로써 이뤄진다. 이러한 파라미터 최적화 방법을 일컬어 손실경사 하강법(Gradient descent)이라 하는데, 그 중에서 가장 널리 쓰이는 Adam(Adaptive Moment Estimation) 최적화를 이용한다. 파라미터 값을 갱신할 때, 이전까지의 손실값의 누적을 momentum값으로 이용해 학습 과정에 관성 내지는 탄력을 주며 특징 변수마다 고르게 학습이 진행되기 위해 스케일을 조정하고 forgetting factor를 이용해 오래 전 값일수록 덜 반영되도록 한다. 식은 다음과 같다.

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

$$\begin{aligned}
m_w^{(t+1)} &\leftarrow \beta_1 m_w^{(t)} + (1 - \beta_1) \nabla_w L^{(t)} \\
v_w^{(t+1)} &\leftarrow \beta_2 v_w^{(t)} + (1 - \beta_2) (\nabla_w L^{(t)})^2 \\
\hat{m}_w &= \frac{m_w^{(t+1)}}{1 - \beta_1^t} \\
\hat{v}_w &= \frac{v_w^{(t+1)}}{1 - \beta_2^t} \\
w^{(t+1)} &\leftarrow w^{(t)} - \eta \frac{\hat{m}_w}{\sqrt{\hat{v}_w} + \epsilon}
\end{aligned}$$

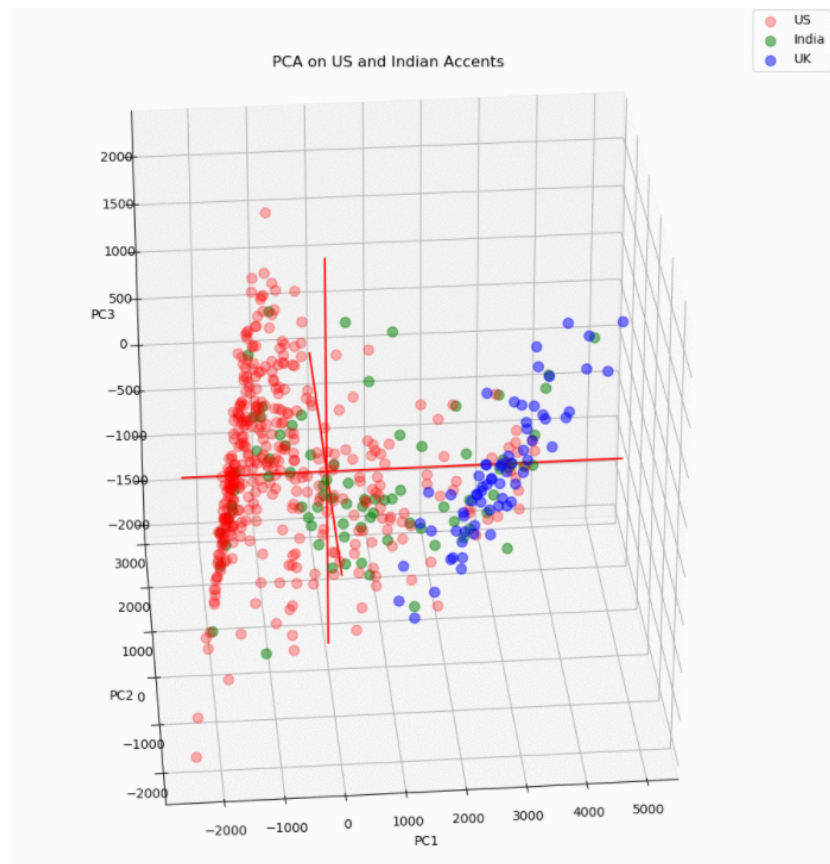
5출처: https://en.wikipedia.org/wiki/Stochastic_gradient_descent#Adam

1.1.8. 결과 시각화 방법

오디오 같은 비정형 데이터는 시각화하기 쉽지 않으나 세 가지 방법으로 시각화 하고자 한다. 그 방법 중 하나는 스펙트로그램이다. 스펙트로그램은 비주기신호를 시각화하는 방법으로 일정한 길이로 자른 시간대별 주파수 스펙트럼을 나타낸 것이다.

또 하나는 주성분분석(Principle Component Analysis)다. 오디오에서 추출한 mfcc 값을 시각화 가능한 저차원 값으로 축소시킨 뒤 그래프화 하고자 한다. 이 기법이 유사한 연구에서 활용된 전례가 있으므로 본 프로젝트에서도 유효하리라 기대한다.

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0



PCA를 통한 mfcc 시각화 예시: 미국 영국, 인도 억양 분류

출처: <https://github.com/nkrao220/accent-classification>

마지막으로 각 지역별 사투리에서만 자주 또는 고유하게 나타나는 단어의 빈도를 시각화 하고자 한다.

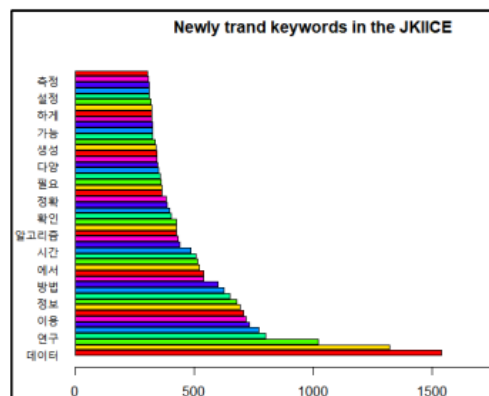


Fig. 3 Frequency analysis using word cloud

단어 빈도 분석 예시

출처: 반재훈; 하종수; 김동현. 빅데이터 분석도구 R 을 이용한 성경 데이터의 빈도와 소셜 네트워크 분석. 한국정보통신학회논문지, 2020, 24.2: 166-171.

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

1.1.9. 예상 결과

1.1.9.1. 오락 수단으로 이용되는 경우

방언 인식 AI 모델은 사용자들을 끌어 모으는 데 있어서 효과적일 것이다. 많은 대중들이 접근하기 쉬운, 웹 어플리케이션으로 제작하여 사용자들이 홈페이지를 통해 어플리케이션을 이용할 수 있는 환경을 조성하고, 사용자들이 녹음한 음성파일을 직접 업로드하여, 사용자들의 사투리 사용에 대한 결과를 제공한다면, 인터넷 사용자들의 흥미를 끌 수 있을 것이다.

<예상 홈페이지 UI>

상태	예상 홈페이지 UI화면
홈페이지 예상 메인 화면	
프로그램을 실행 후 결과 표출	

특히 사용자들에게 웹 어플리케이션 같은 손쉬운 접근성을 제공한다면, 인터넷 사용자의 큰 비중을 차지하는 MZ세대의 유입을 유도할 수 있다. 기존에도 많은 웹 어플리케이션들이 테스트라는 흥미유발 요소를 가지고 MZ세대의 관심을 끌고 있음을 밀의 자료를 통해 확인할 수 있다.

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0



Fig.1. 직장캐릭터 테스트 (출처: <https://poomang.com/detail/rqokm>)

색깔심리 테스트 (출처: <http://aiselftest.com/colortest/>)

동물상 테스트 (출처: <https://animalface.wookingwoo.com/>)

이중 AI를 활용한 동물상 테스트와 관상 테스트 같은 경우는 화제가 되어 뉴스 기사로도 올라왔다.

관상테스트 이틀째 화제...조코딩 '동물상 테스트'도 재조명

이승요 기자 | 입력 2020-12-03 08:48



Fig.2. 관상, 동물상 테스트 (출처: <https://www.ajunews.com/view/20201203080723752>)

동물상 테스트는 다운로드 횟수 10만 이상이라는 수치적 지표로도 높은 인기를 보여주는 것을 확인할 수 있으며, 각종 AI 기반 테스트 어플리케이션들이 MZ세대에게 큰 어필을 할 수 있다는 사례를 보여주고 있다. 이런 어플리케이션 시장의 경향성에서 우리 팀의 어플리케이션은 경쟁력을 충분히 갖추 수 있다는 점을 알 수 있다. 그 외에도 MZ세대들이 사투리에 대해 가지는 흥미나, 관심도 또한 큰데, 그 정도는 유튜브나 인스타그램 같은 대형 플랫폼을 통해서도 쉽게 확인

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문서명	프로젝트 계획서	버전	1.0.0

할 수 있다.



Fig.3. 유튜브,인스타그램 자료화면

이렇듯 MZ세대를 타겟으로, 사투리를 이용하는 많은 콘텐츠들이 존재하는 것을 확인할 수 있다. 하지만 막상 내부를 들여다보면, 콘텐츠 창작자들은 사투리를 객관적으로 판단할 만한 적절한 수단이 없어, 주관적인 판단에 의존하는 경향이 있다.

그렇기 때문에, AI를 활용한 사투리 판독 모델은 콘텐츠 제작자들에게 객관적 지표를 제공할 수 있고 또 매력적인 수단으로 작용할 수 있다. 만약 이들의 콘텐츠 제작 수단으로 우리 팀의 프로그램을 이용한다면, 포털 사이트나 유튜브 등을 통한 간접적 홍보 효과를 얻어 더 많은 사용자들을 창출할 수 있을 거라 기대 한다.

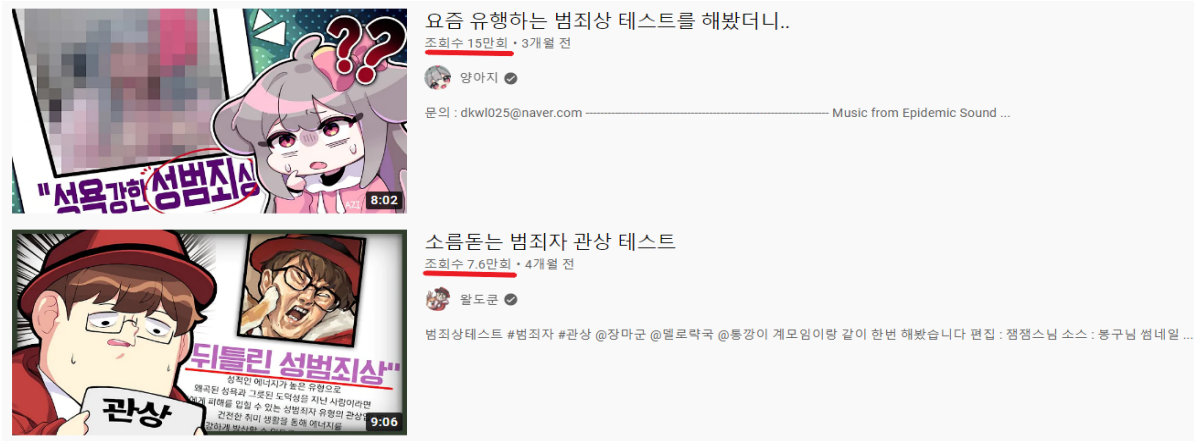


Fig.4.유튜브 자료화면

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문서명	프로젝트 계획서	버전	1.0.0

컨텐츠 크리에이터들을 통한 홍보효과와는 위의 예시로 드는 범죄자 관상 테스트만 봐도 높은 조회수를 기록하는 등, 높은 홍보효과를 기대할 수 있다.

더불어, 사투리에 흥미를 가진 사용자들이 홈페이지에서 어플리케이션을 여러 번 이용할수록 더 많은 화자들의 음성 데이터가 쌓인다면, 그걸 기반으로 사투리 식별 정확도를 더 높일 수 있을 점을 기대할 수 있는데, 기존 프로토타입의 DNN 모델을 학습시켜 샘플 데이터들을 이용해 정확도를 측정해 봤을 때 90퍼센트 이상의 준수한 정확도를 보여주는 것을 통해 확인할 수 있다.

```

[180] test_accuracy=model.evaluate(X_test,y_test,verbose=0)
print(test_accuracy[1])

0.9955456852912903

```

Fig.5.DNN 모델을 이용한 사투리 인식

이를 통해, 상용화를 통해 더 많은 데이터를 수집하게 된다면, 사용자에게 적절한 품질의 서비스를 제공할 수 있을 것이라 기대한다.

1.1.9.2. 학술적인 목적으로 이용되는 경우

관련 기사나 자료를 통해, 사투리 관련 데이터는 갈수록 가치가 중요해지는 걸 확인할 수 있고, 그에 따라 데이터 수집에 대한 필요성도 점점 높아져 가고 있다.



Fig.6.지역어 자료관 구축 (출처:

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

https://www.ytn.co.kr/_ln/0106_202202081622001283)

제주어를

지켜라

(출처:

<https://www.khan.co.kr/local/Jeju/article/202202181452011>)

AI

보이스피싱

(출처:

<http://mbiz.heraldcorp.com/view.php?ud=20200320000411>)

이렇듯 우리나라 고유의 문화를 유지하기 위한 목적으로, 화자가 갈수록 줄어드는 방언의 소멸을 막고 정통성을 지속하기 위해 우리나라의 지자체와 국립국어원 등에서 사투리 데이터 수집에 많은 노력을 하고 있다. 우리 팀의 데이터는 지역간 억양과 단어 차이를 구분하는데 초점을 둔 만큼, 지자체 및 여타 기관들이 요구하는 사투리 데이터의 성격에 잘 부합한다. 그러므로, 타기관들이 데이터 확보를 하는 데 있어서 우리 팀의 AI 모델은 많은 기여를 할 수 있다고 기대할 수 있다,

더불어 경찰도 보이스피싱 수사 등에 빅데이터와 AI를 활용한 출신 지역을 식별하는데 초점을 맞추고 있다. 이런 상황에서 우리 팀의 AI 모델을 통한 사투리 데이터 확보는, 갈수록 사투리 데이터의 가치와 중요성이 늘어나고 있는 상황에서 학술적으로도 효과적인 데이터 수집 모델로서 작용할 수 있을 것이다.

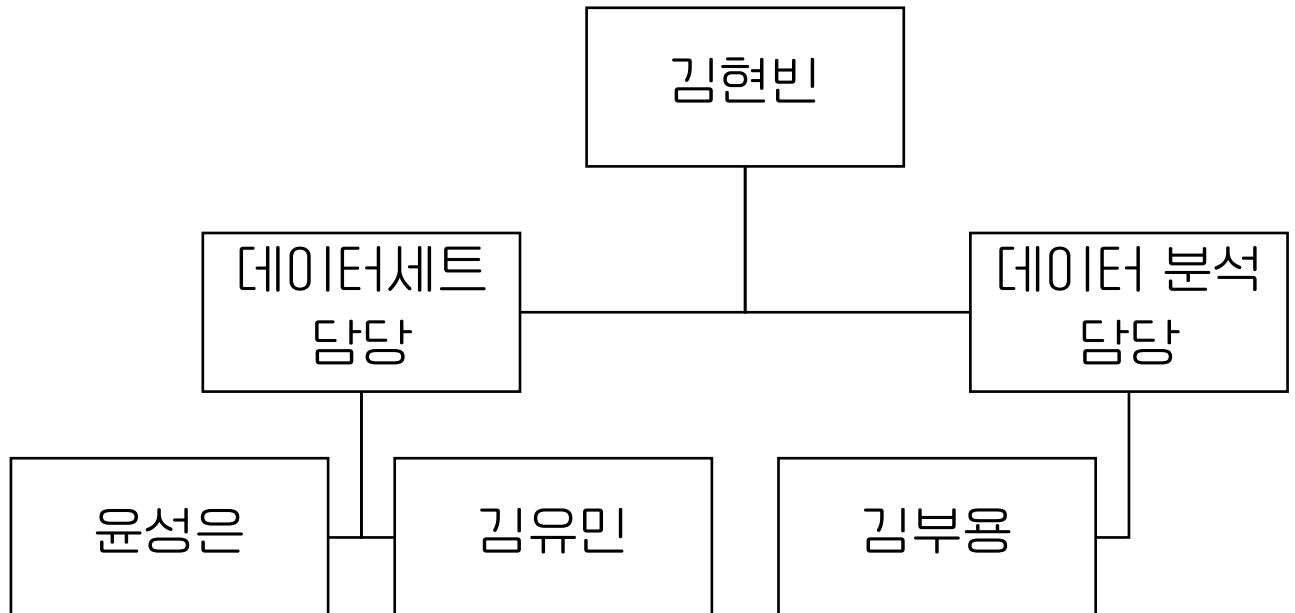
1.2 주요 일정

단계	일정	산출물
초기 요구사항 분석 및 계획 수립과 위험 분석	03.01 ~ 03.31	프로젝트 계획서
데이터세트 구축	04.01 ~ 04.14	데이터 확보 보고서
데이터 분석	04.15 ~ 05.19	데이터 분석 보고서
결과 시각화	05.20 ~ 06.02	최종 보고서 작성
최종 결과 발표	06.16	최종 결과 발표 자료

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

1.3 조직

1.3.1 조직도



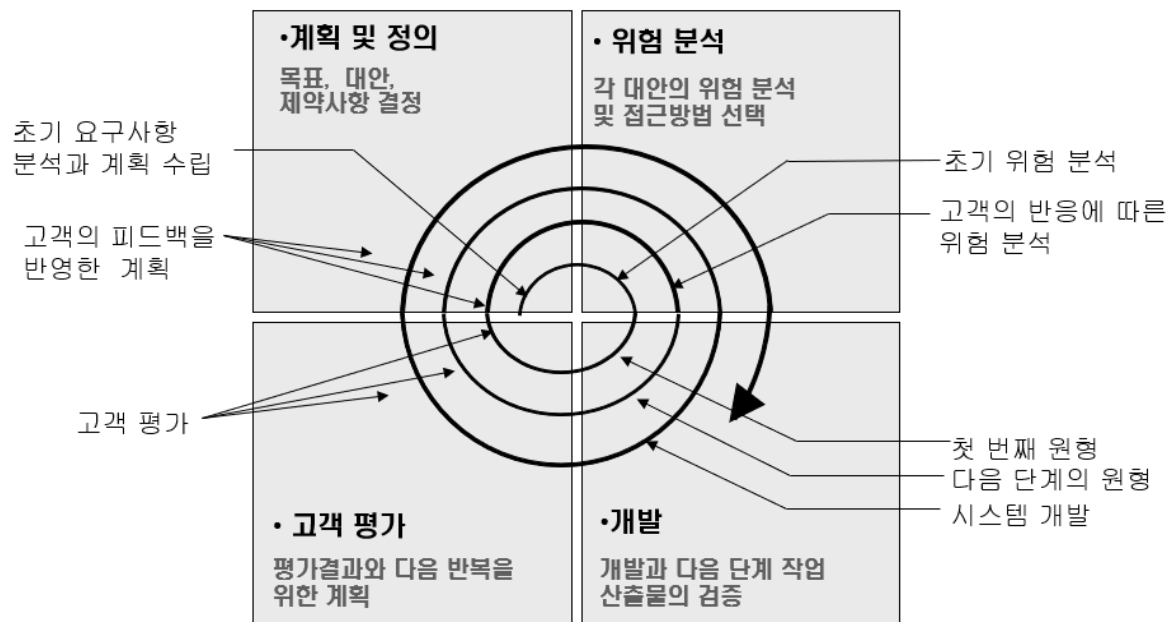
1.3.2 역할 및 책임

팀원	역할	책임
김현빈	팀장	<ul style="list-style-type: none"> 프로젝트의 종합적인 관리, 통제 및 웹 어플리케이션 기능 개발 데이터 수집 및 인공지능 학습
	웹 개발자	
	데이터 분석	
김부용	웹 개발	<ul style="list-style-type: none"> 어플리케이션 기능 개발
	데이터 분석	<ul style="list-style-type: none"> 데이터 수집 및 인공지능 학습
윤성은	데이터 수집 및 정제	<ul style="list-style-type: none"> 어플리케이션 화면 설계
	UI/UX 디자이너	<ul style="list-style-type: none"> 데이터 수집 및 정제
김유민	데이터 수집 및 정제	<ul style="list-style-type: none"> 어플리케이션 화면 설계
	UI/UX 디자이너	<ul style="list-style-type: none"> 데이터 수집 및 정제

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

1.4 생명주기 모델

Spiral Model (나선형 모델)



시스템 개발 위험을 최소화하기 위해 나선으로 돌면서 계획, 위험분석, 개발, 평가의 단계를 반복하면서 점진적으로 소프트웨어를 완성하는 모델

단계	수행 Task
계획 및 정의	초기 요구사항 분석과 계획 수립 고객의 피드백을 반영한 계획
위험 분석	초기 위험 분석 고객의 반응에 따른 위험 분석
개발	구현 대상 기능에 대한 실제 구현 단위 테스트 수행
고객 평가	고객에 의한 시스템 평가 향후 목표 계획

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

1.5 도구

	▶ 디스코드 (Discord) 온라인 화상 회의 도구 팀 회의 및 산출물 작성
	▶ 카카오톡 (Kakao Talk) 모바일 메신저 팀 의사소통 기구
	▶ 엑셀 (Excel) 스프레드 시트 프로그램 테스트 케이스 등 문서 작성
	▶ 워드 (Word) 워드 프로세서 요구사항 명세서, 프로젝트 계획서 등 문서 작성
	▶ 깃 허브 (GitHub) 분산 버전 관리 시스템 소스코드의 버전 관리, 산출물 제출
	▶ 파워 포인트 (Power Point) 프레젠테이션 프로그램 발표자료 등 문서 작성
	▶ 구글 드라이브 (Google Drive) 산출물 버전관리 프로젝트 파일 문서 저장
	▶ Figma UI 디자인 도구
	▶ pycharm 개발 도구
	▶ 마리아 DB (Maria DB) DataBase

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

2. 규모 산정

2.1 WBS(Work Breakdown Structure)



2.2 작업에 소요되는 기간

프로젝트 계획 및 정의 – 1차(2주차), 2차(4주차)

데이터 확보 – 1차(5주차), 2차(6주차), 3차(7주차)







데이터 분석 – 1차(8주차), 2차(10주차), 3차(12주차)

결과 시각화 – 1차(14주차)

최종 결과 발표 – 1차(16주차)

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

3. 일정

항목	과업	3월				4월					5월				6월
		1주	2주	3주	4주	1주	2주	3주	4주	5주	1주	2주	3주	4주	1주
프로젝트 기획	팀 구성														
	프로젝트 계획서														
데이터 분석	데이터셋 구축														
	데이터 분석														
프로젝트 결과	결과 시각화														
	최종 결 과 발표														

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

4. 산출물 관리

분류	규칙	
산출물 명	1. Github와 Notion에 업로드 되는 산출물 명은 <조 이름.산출물 명_버전.확장자명>으로 기재한다. 2. 산출물을 수정을 하거나 업데이트를 할 때 소수점 이후의 버전 숫자를 증가시킨다. 3. Notion의 회의록 명에는 회의 당일 날짜를 기재한다. (예) <주간 회의(날짜)>	
산출물 저장	1. 프로젝트 일정 관리 및 다양한 자료 관리가 가능한 공유 문서 작업 툴 Notion을 통해 각자 맡은 작업물을 정해진 시간 전까지 공유한다. 2. Notion에 올라와 있는 산출물은 언제든지 팀원이 접근할 수 있으며 실시간으로 수정사항이나 피드백을 공유한다.	
산출물 세부 관리	프로젝트 계획서	프로젝트 계획서 구성에 맞게 계획서를 작성하고 진행되는 프로젝트 일정과 계획서를 비교하고 확인한다.
	회의록	1. 매주 수요일 Discode을 통해 비대면 회의를 진행하고 이때 회의는 Notion에 기록한다. 2. 추가적인 회의가 필요하다 느끼면 팀원들과 KakaoTalk 및 Discode를 통해 시간을 조율하여 회의를 진행한다.

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

5. 위험 관리 계획

5.1 프로젝트 측면 위험

위험 요소	가능성	영향도	대처 방안
휴학 및 코로나로 인한 참여 중단	중	하	<ul style="list-style-type: none"> - 팀원 역할 재분배 - 문서화 작업 습관화 - 문서 전달과 회의를 통한 빠른 인수인계
팀원 간의 불화	하	중	<ul style="list-style-type: none"> - 충분한 회의를 거쳐 업무를 분배 - 팀장 주도로 원만한 의견 중재안 도출
일정 초과	중	상	<ul style="list-style-type: none"> - 주기적인 회의와 프로젝트 일정을 참고해 효율적인 일정 관리
팀원의 연락 부재 및 담당업무 소홀	하	중	<ul style="list-style-type: none"> - 온/오프라인 연락처 공유 - 팀원 간 업무 진행 사항들을 지속적으로 공유하여 부재 시 바로 투입될 수 있도록 힘
작업 환경의 부재	하	하	<ul style="list-style-type: none"> - 담당자를 선정하여 미리 온/오프라인 회의실 예약
제품 품질 저하	중	상	<ul style="list-style-type: none"> - 각종 표준 및 품질 보증에 대한 충분한 교육 실시 - 동료 검토를 통해 지속적으로 품질 수준 체크

5.2 기술적 측면 위험

위험 요소	가능성	영향도	대처 방안
기술력 부족	중	상	<ul style="list-style-type: none"> - 개발 인력에 대한 충분한 사전 교육 실시 - 개발 분야에 대한 구체적인 분석
성능 미달	하	상	<ul style="list-style-type: none"> - 시스템 구축 전 충분한 능력, 용량 산정
산출물 관리 미숙	하	상	<ul style="list-style-type: none"> - 백업 및 복구 체제 마련 - 일일 백업 관리 철저
프로토 타입의 테스트 불가	중	상	<ul style="list-style-type: none"> - 개발 중 테스트 시점 미리 결정
찾은 오류 수정	상	중	<ul style="list-style-type: none"> - 구현하기 전, 구현할 때마다 일어날 수 있는 모든 상황을 테스트하여 기능들을 통합했을 때의 오류를 최소화

5.3 사용자 측면 위험

프로젝트 명	방구	프로젝트 기간	22.03.23 ~ 22.06.23
문 서 명	프로젝트 계획서	버전	1.0.0

위험 요소	가능성	영향도	대처 방안
요구사항의 누락	중	상	- 지속적 회의를 통한 체크리스트 확인
요구사항의 변경	하	상	- 변경사항 상시 파악, 적용 검토 최대 변경 상한선, 점증적 개발, 다음 버전까지 변경 연기
요구사항의 불만족	하	상	- 불만족 항목 조사 후 작업 과정이 신속히 복구되도록 대처