



# 중간고사 전 마지막 회합

2022.09.29. HAI 1팀

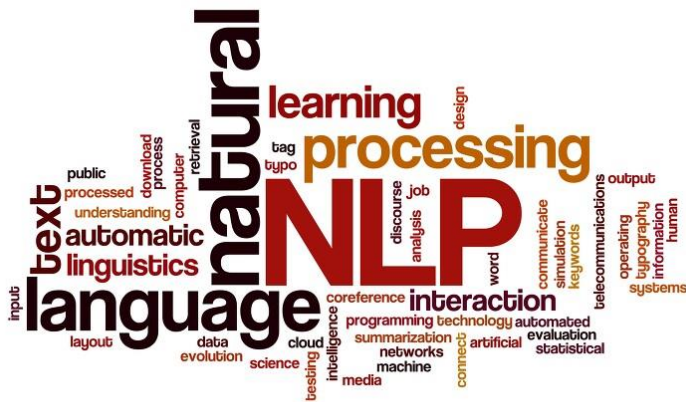


한양대학교

HANYANG UNIVERSITY

## 지금까지 해온 것들...

- NLP(자연어 처리)가 무엇이고, 어떤 응용 범위가 있는지 알아봤어요.
- 각 지역별 방언을 자동으로 인식하고 번역해주는 프로그램을 만들기로 했어요.
- AI hub에 업로드되어 있는 데이터를 살펴봤어요.
- Huggingface hub에 업로드되어 있는 GPT, T5 모델을 사용해 봤어요.



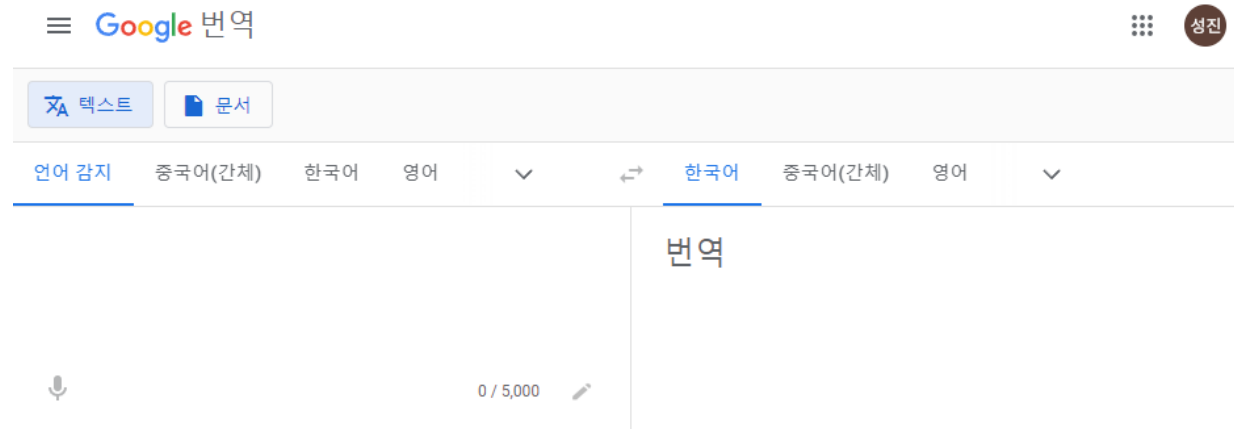
AI  Hub



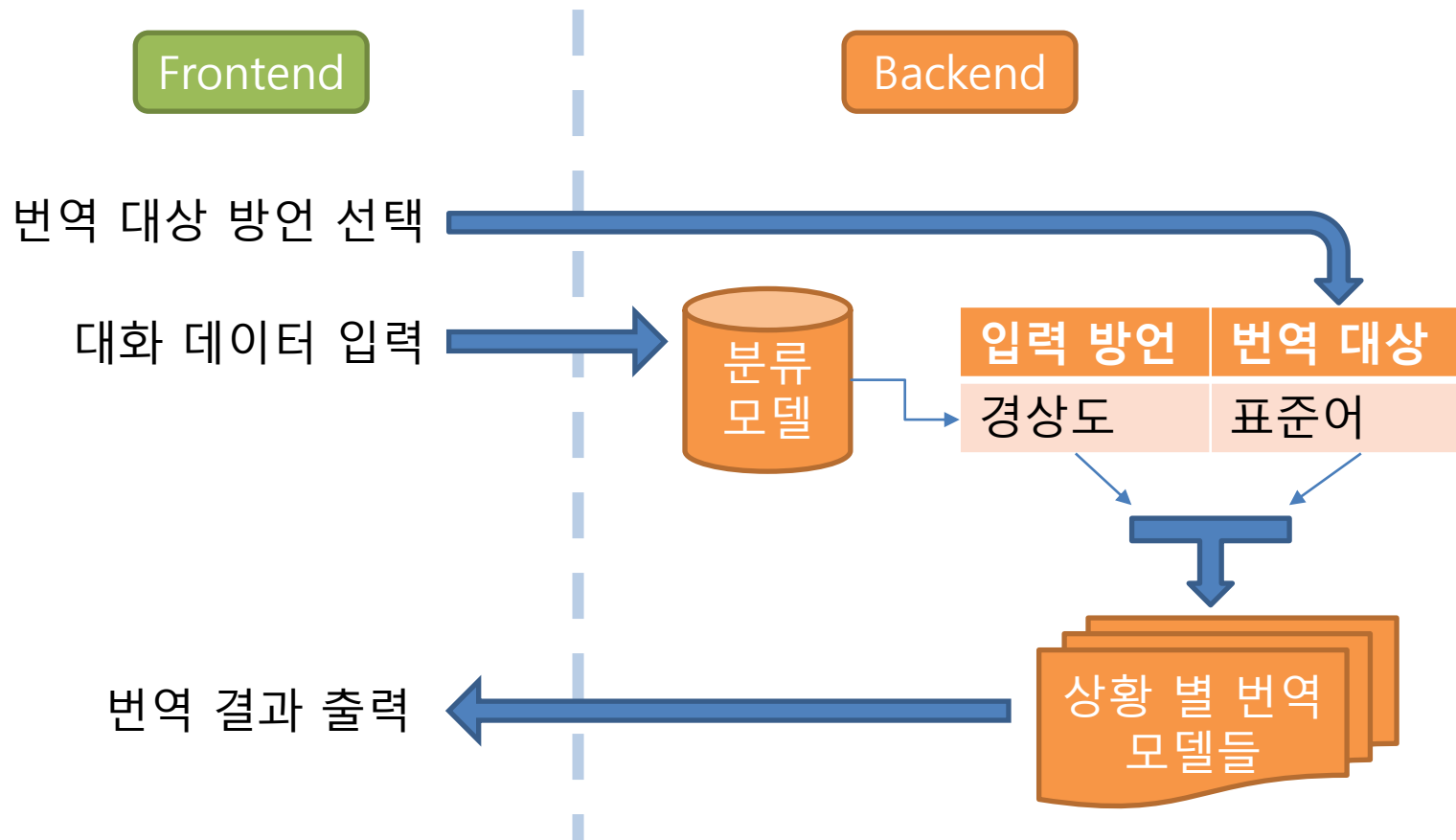
Transformers

## 우리가 만들 친구는 어떤 프로그램인가?

- 사람의 발화 텍스트를 입력하면, 표준어인지 아니면 특정 지역 방언인지 자동으로 분류해줘요(구글 번역기의 언어 자동 인식 기능).
- 현재 입력된 텍스트를 원하는 지역의 방언이나 표준어로 번역할 수 있어요.
- 웹 애플리케이션으로 구현하여 누구나 접근해서 사용 가능해요.

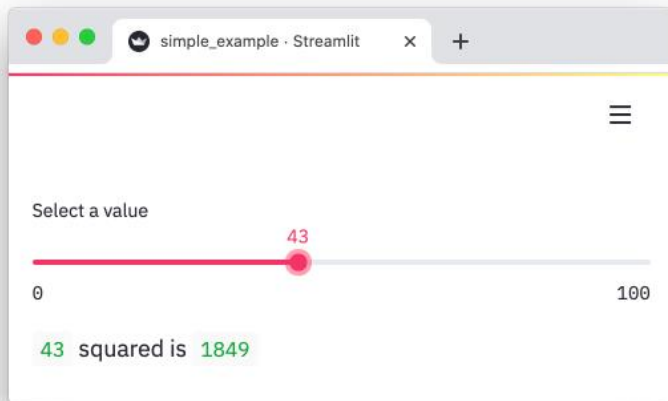


## 실제 개발해야 할 구조



## 프로토타입 웹 페이지 개발-Frontend

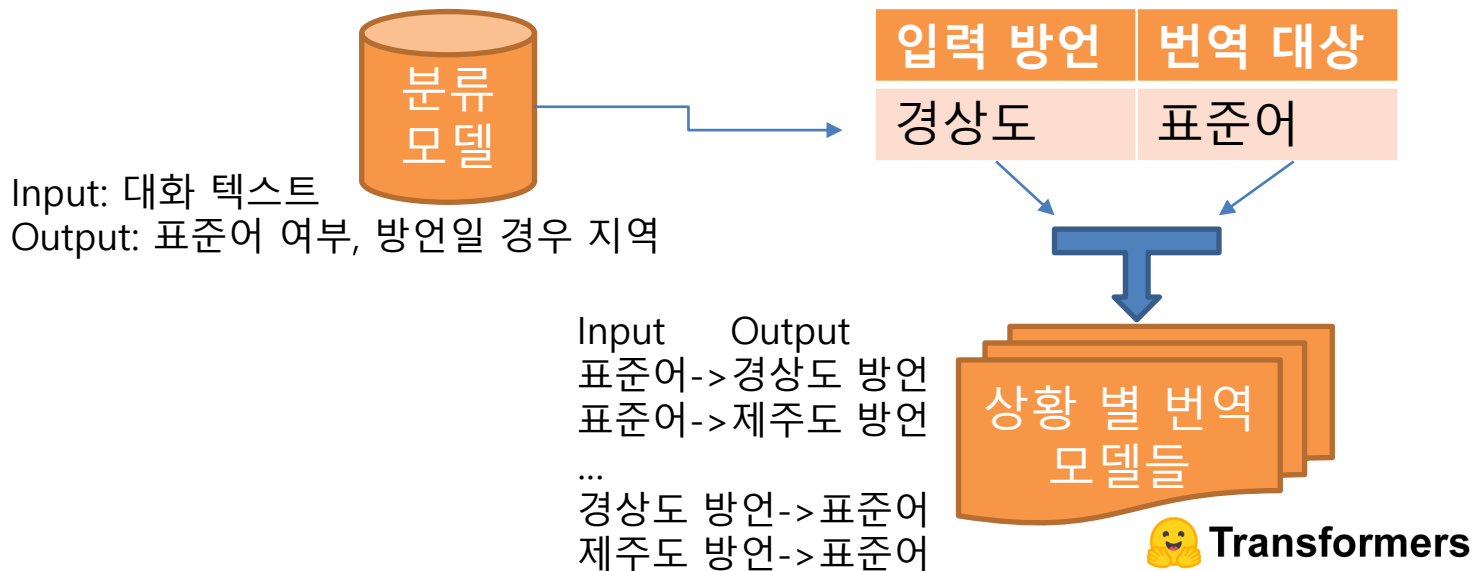
- 머신 러닝 모델을 학습하고 잘 동작하는지 확인하는 것 뿐만 아니라, 다른 사람들과 공유하기 위한 도구인 **Streamlit**을 활용하여 웹 페이지 형태로 구현할 예정입니다.
- 간단한 프로토타입을 웹 형식으로 빠르게 만들어 공유하기에 최적화된 도구로, 대화 텍스트를 입력하고 번역 타겟 방안을 선택하고 결과를 확인할 수 있는 UI를 구현해야 해요.



# Streamlit

## 모델 구성-Backend

- 앞서 보았던 구조를 구현하려면, 꽤 많은 종류의 모델이 필요합니다.
- 입력된 자연어 텍스트가 어떤 지역의 방언 또는 표준어 발화인지 분류하는 **Classification task**를 수행할 모델이 필요해요.
- 그리고 입력 방언과 번역 대상 방언에 따라 양 방향으로 **Translation task**를 수행할 모델들이 필요해요.

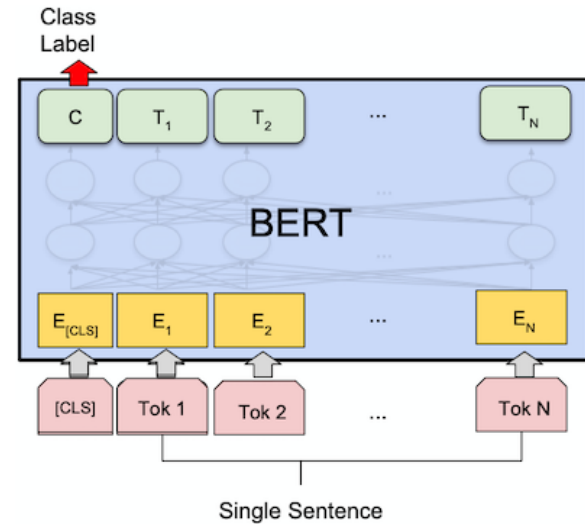


# 이 친구를 먼저 만들어 봅시다!

**Input:** 대화 텍스트  
Ex) 마! 니 밥 묵었나?



**Output:** 방언 라벨  
Ex) 표준어: 0, 제주도: 1, ...



## Step 1. 데이터 다운로드

- <https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=121>
- 구현을 쉽게 하기 위해, 입력된 대화 텍스트가 표준어로 이루어져 있는지, 아니면 제주도 방언으로 이루어져 있는지 **Binary classification**을 수행하는 모델을 만들기 위한 데이터를 준비할게요.
- 데이터를 다운로드할 때는 원천 데이터(음성 파일)의 용량이 불필요하게 크므로 제외하고 다운로드 해 주세요!
- 이 데이터를 우리가 학습시키려는 모델에 맞게 아래와 같이 변형시킬 예정입니다.

```
"form": "조금 애기도 나누고 (경)/(그렇게)(하믄)/(하면) 좋을 거 (땡다)/(같다)고 아",  
"standard_form": "조금 애기도 나누고 그렇게하면 좋을 거 같다고 아",  
"dialect_form": "조금 애기도 나누고 경하믄 좋을 거 땡다고 아",
```

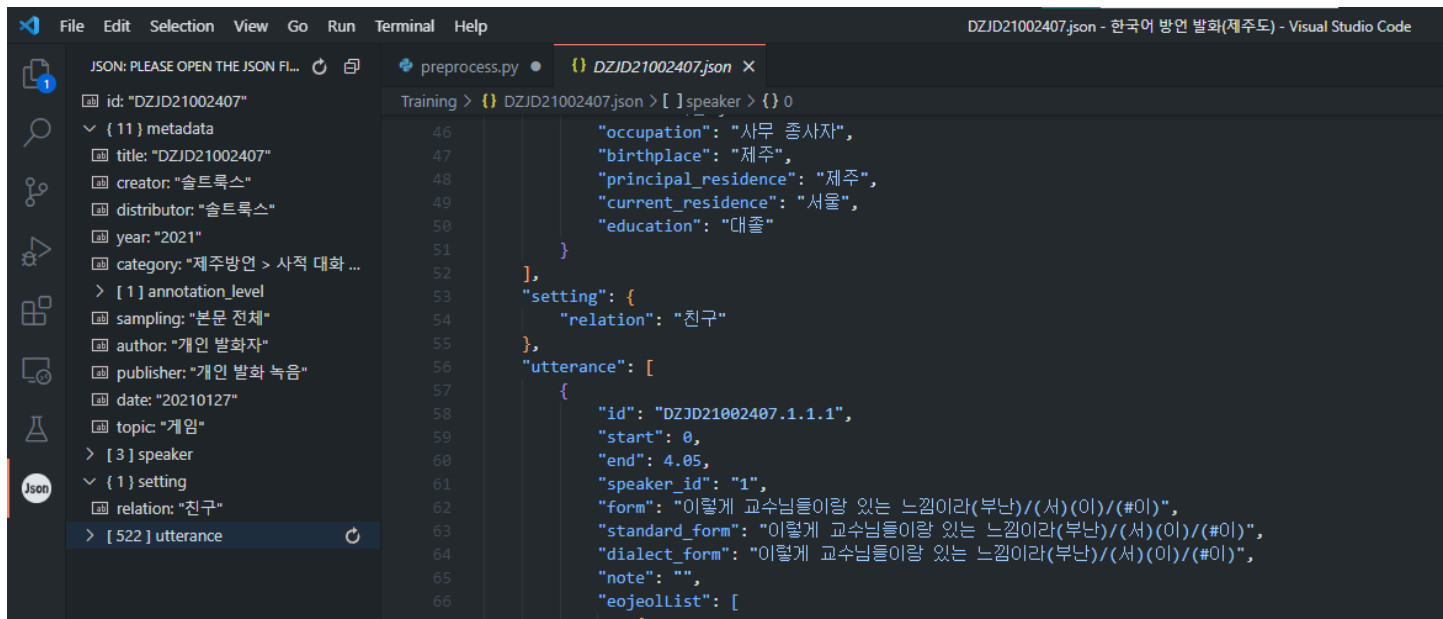


Text	Label
"조금 애기도 나누고 그렇게하면 좋을 거 같다고 아"	0 (표준어)
"조금 애기도 나누고 경하믄 좋을 거 땡다고 아"	1 (방언)



## Step 2. 데이터 살펴보기

- 다운로드 받은 json 형식의 파일을 열어보면 특정한 구조(Tree와 유사한)를 따르고 있다는 것을 알 수 있습니다.
- 여러 메타데이터들이 나타난 이후 utterance 부분에 실제 발화 텍스트가 저장되어 있네요.
- 각각의 텍스트에 (AAA)/(BBB) 꼴이 들어있는 경우 앞 부분이 방언, 뒷 부분이 표준어에 해당하는 어절이라고 할 수 있습니다.



```
File Edit Selection View Go Run Terminal Help
DZJD21002407.json - 한국어 방언 발화(제주도) - Visual Studio Code

JSON: PLEASE OPEN THE JSON FILE...
preprocess.py • DZJD21002407.json X
Training > {} DZJD21002407.json > [ ] speaker > {} 0

{
  "id": "DZJD21002407",
  "metadata": {
    "title": "DZJD21002407",
    "creator": "솔트룩스",
    "distributor": "솔트룩스",
    "year": "2021",
    "category": "제주방언 > 사적 대화 ...",
    "annotation_level": "본문 전체",
    "author": "개인 발화자",
    "publisher": "개인 발화 녹음",
    "date": "20210127",
    "topic": "게임"
  },
  "speaker": [
    {
      "setting": {
        "relation": "친구"
      },
      "utterance": [
        {
          "id": "DZJD21002407.1.1.1",
          "start": 0,
          "end": 4.05,
          "speaker_id": "1",
          "form": "이렇게 교수님들이랑 있는 느낌이라(부난)/(서)(이)/(#이)",
          "standard_form": "이렇게 교수님들이랑 있는 느낌이라(부난)/(서)(이)/(#이)",
          "dialect_form": "이렇게 교수님들이랑 있는 느낌이라(부난)/(서)(이)/(#이)",
          "note": "",
          "eojeollist": [
```

## Step 3. 필요에 맞게 데이터 추출,정제하기

- 여러 가지 방법이 있지만, python을 이용하여 json 파일을 파싱하고, 원하는 부분만 추출해 봅시다.
- <https://choi-log-life.tistory.com/entry/python-json-parsing-to-dictionary> : json 모듈을 import하고 사용하여 json 파일을 파싱하고 원하는 데이터만 뽑아내는 과정 튜토리얼입니다. 관심이 있으면 도전해보세요!
- 주의해야할 점은, 대화 도중 표준어로만 이야기한 시점에는 standard\_form과 dialect\_form이 동일하므로 이 경우에는 표준어 데이터로 반영해야 합니다!

JSON	파이썬
오브젝트(object)	dict
배열(array)	list
문자열(string)	str
숫자 (정수)	int
숫자 (실수)	float
true	True
false	False
null	None

```
1 import json
2
3 ex = '{"a": "apple", "b": "banana", "c": "car"}'
4 dict = json.loads(ex)
5
6 for key in dict:
7     print(key)
```

a  
b  
c  
[Finished in 46ms]

## Step 4. 라벨링이 완료된 데이터 저장하기

- 아래와 같이 텍스트/라벨로 구성된 데이터셋이 완성되었다면, **pandas** 라이브러리를 이용해서 모델이 학습하는데 사용될 수 있도록 저장해 봅시다.
- Dataframe** 클래스를 이용하면 테이블 또는 시계열 형태의 데이터를 간단하게 다루고 csv 등의 파일 형식을 읽어오거나, 저장할 수 있습니다.
- to\_numpy** 메소드를 활용하면, numpy array로 즉시 변환하는 것도 가능합니다.
- 더 자세한 내용은 <https://rfriend.tistory.com/252>을 참고해주세요.

```
In [1]: import pandas as pd  
arr = [['조금 애기도 나누고 그렇게하면 좋을 거 같다고 아', 0], ['조금 애기도 나누고 경하민 좋을 거 났다고 아', 1]]  
  
In [2]: df = pd.DataFrame(arr, columns=['text', 'label'])  
  
In [3]: print(arr)  
print(df)  
[['조금 애기도 나누고 그렇게하면 좋을 거 같다고 아', 0], ['조금 애기도 나누고 경하민 좋을 거 났다고 아', 1]]  
      text label  
0  조금 애기도 나누고 그렇게하면 좋을 거 같다고 아      0  
1  조금 애기도 나누고 경하민 좋을 거 났다고 아      1  
  
In [4]: df.to_csv("dataset.csv")
```



dataset.csv

2022-09-29 오전 3:29

Microsoft Excel ...

1KB

## 다음 시간에 계속...

- 전처리한 데이터를 이용해서 BERT 모델을 학습시키고, validation data를 이용하여 모델의 성능을 테스트 해볼거예요.
- 제주도 사투리 이외에도 나머지 지역별 방언 데이터를 전처리해 볼거예요. Streamlit을 이용하여 만든 모델을 테스트할 수 있는 페이지를 만들어 볼 거예요.
- 완성된 분류 모델을 이용하여 알아낸 입력 텍스트의 방언 지역과 번역 대상 지역을 바탕으로 적절한 모델에 입력하여 번역 결과를 만들어낼 수 있도록 시스템을 구성해 볼 거예요.

## 추가로 시도해 보면 좋을 것들

- 데이터에서 잘못된 부분을 찾아 정확하게 고칠 수 있도록 해 봅시다.
- 지나치게 짧은 발화 파편들을 모아 일정 길이 이상이 되도록 배열해 봅시다.
- "표준어 or 특정 지역 방언"을 구분하는 binary classification이 아닌, 표준어, 경상도, 제주도 등등 여러 지역 방언을 구분하는 multi classification을 수행하는 모델을 학습시키기 위한 데이터셋을 만들어 봅시다.

즐거운 축제 기간 되시고  
중간고사 파이팅!  
10월 말에 다시 만나요!

