연결 고리

# 경상도 사투리 번역기

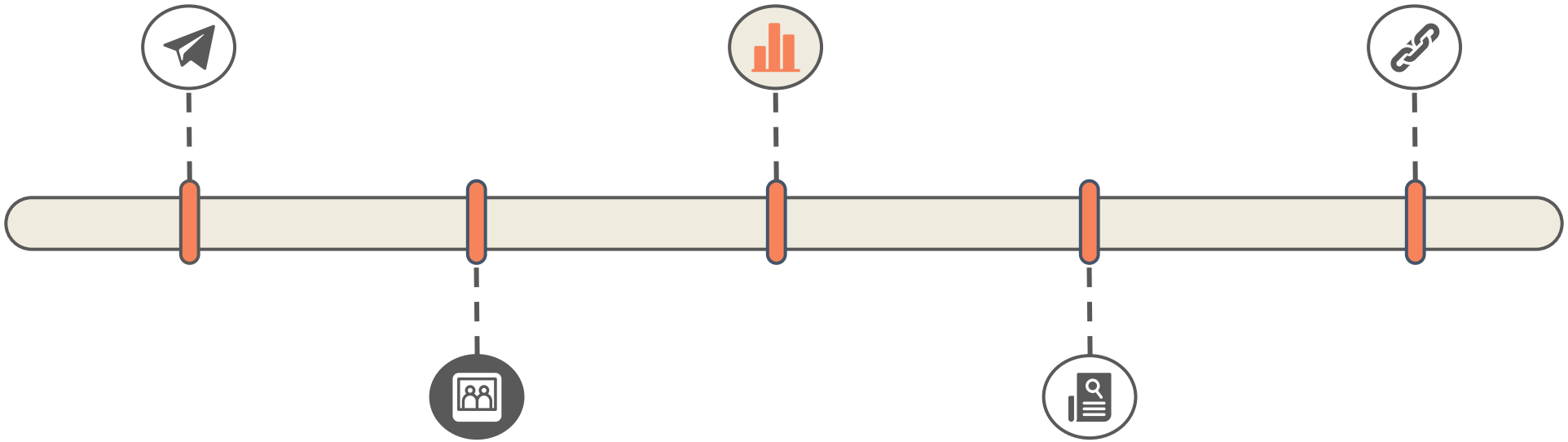
박시원

정소비

정연규

# Contents





# 프로젝트 배경

## 주제 선정 배경

선정배경

## 프로젝트 선정

문제인식
프로젝트 선정
프로젝트 목적

# 경상도 사투리 번역기

프로젝트 배경 | 수행 절차 및 방법 | 수행 결과

주제 선정 배경 프로젝트 선정

## 시도별 출생인구수

| 시군구별 | 2020 계 (명) | 2016 계 (명) | 2012 계 (명) | 2008 계 (명) | 2004 계 (명) | 2000 계 (명) |
|---|---|---|---|---|---|---|
| 전국 | 272,337 | 406,243 | 484,550 | 465,892 | 476,958 | 640,089 |
| 서울특별시 | 47,445 | 75,536 | 93,914 | 94,736 | 99,828 | 133,154 |
| 부산광역시 | 15,058 | 24,906 | 28,673 | 26,670 | 28,231 | 41,222 |
| 대구광역시 | 11,193 | 18,298 | 21,472 | 20,562 | 23,259 | 32,477 |
| 인천광역시 | 16,040 | 23,609 | 27,781 | 25,365 | 25,092 | 34,433 |
| 광주광역시 | 7,318 | 11,580 | 14,392 | 13,890 | 14,729 | 21,148 |
| 대전광역시 | 7,481 | 12,436 | 15,279 | 14,856 | 15,024 | 19,570 |
| 울산광역시 | 6,617 | 10,910 | 12,160 | 11,365 | 11,151 | 15,816 |
| 세종특별자치시 | 3,468 | 3,297 | 1,054 | - | - | - |
| 경기도 | 77,737 | 105,643 | 124,746 | 119,397 | 117,812 | 141,704 |
| 강원도 | 7,835 | 10,058 | 12,426 | 12,373 | 13,776 | 19,482 |
| 충청북도 | 8,607 | 12,742 | 15,139 | 14,064 | 14,331 | 19,628 |
| 충청남도 | 11,950 | 17,302 | 20,448 | 19,749 | 18,640 | 24,733 |
| 전라북도 | 8,165 | 12,698 | 16,238 | 15,878 | 17,257 | 25,173 |
| 전라남도 | 9,738 | 13,980 | 16,990 | 16,363 | 17,256 | 26,046 |
| 경상북도 | 12,873 | 20,616 | 24,635 | 23,538 | 23,553 | 35,190 |
| 경상남도 | 16,823 | 27,138 | 33,211 | 31,493 | 30,922 | 41,680 |
| 제주특별자치도 | 3,989 | 5,494 | 5,992 | 5,593 | 6,097 | 8,633 |

전체의 45.0%를 차지

# 경상도 사투리 번역기

프로젝트 배경 | 수행 절차 및 방법 | 수행 결과

주제 선정 배경 프로젝트 선정

## 선정 배경


**관심과 보전이 필요한 방언, 외국과 국내 몇몇 기업에선 보전하려 노력..**

# 경상도 사투리 번역기



문제 인식

## 방언의 중요성

- 방언은 과거 국어의 흔적을 발견할 수 있다는 데서 가치를 가집니다.
- 대부분의 언어 학습 데이터가 표준어로 만들어져 있고 그러한 데이터로 만든 서비스는 표준어로만 사용가능하기 때문에 노년층을 위한 AI 돌봄 서비스나 주로 농촌 지역에서 쓰이게 되는 스마트팜 서비스는 사투리를 인식하는 기능이 필요하다.

# 경상도 사투리 번역기



## 프로젝트 선정

방언의 역사적 가치를 이해하고 이 언어를 활용하여 방언을 사용하는 사람들을 위해 방언 – 표준어 번역기를 만들게 되었습니다.

## 프로젝트 목적

지역간 언어적 특성으로 인한 소통의 어려움을 해결하여 지역에 관계없이 원활한 소통 할 수 있도록 도와주는 프로그램을 개발

# 팀 구성 및 역할

## 박시원

- 자료조사
- 데이터 수집
- 알고리즘 설계
- 데이터 전처리
- 기능 구현
- 발표

## 정소비

- 자료조사
- 데이터 수집
- 알고리즘 설계
- 데이터 전처리
- 기능 구현
- 모델 구현

## 정연규

- 자료조사
- 데이터 수집
- 알고리즘 설계
- 데이터 전처리
- 기능 구현
- 문서 작업

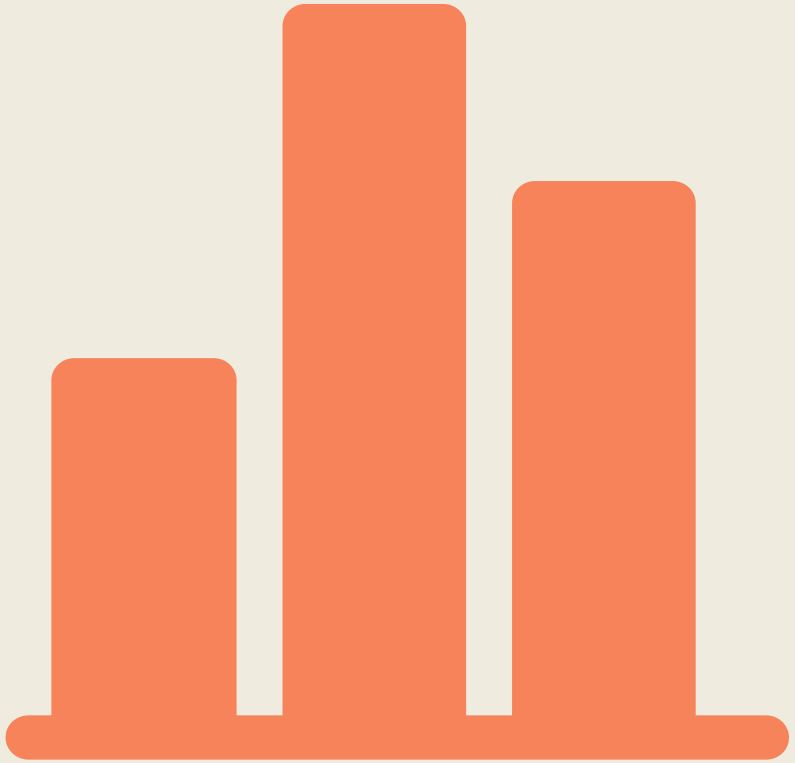
# 수행절차 및 방법

데이터 설명 및 전처리

EDA

모델링 과정

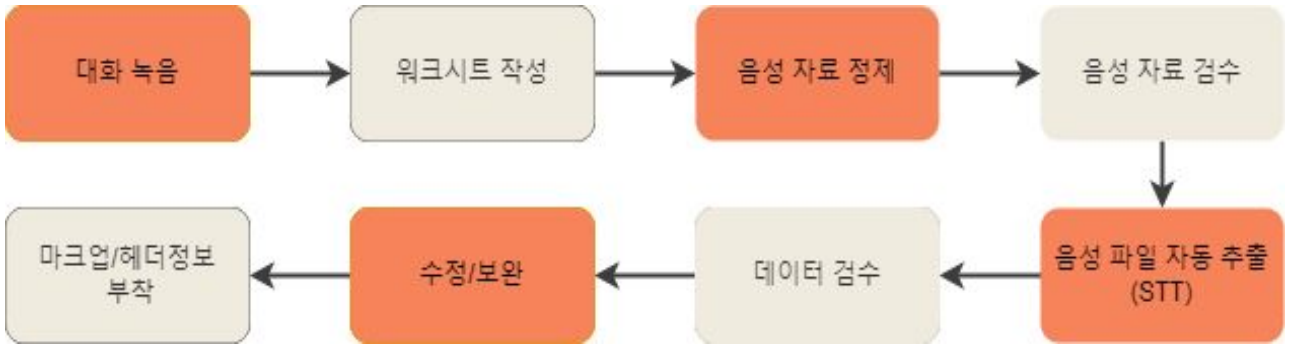
과정 및 개요



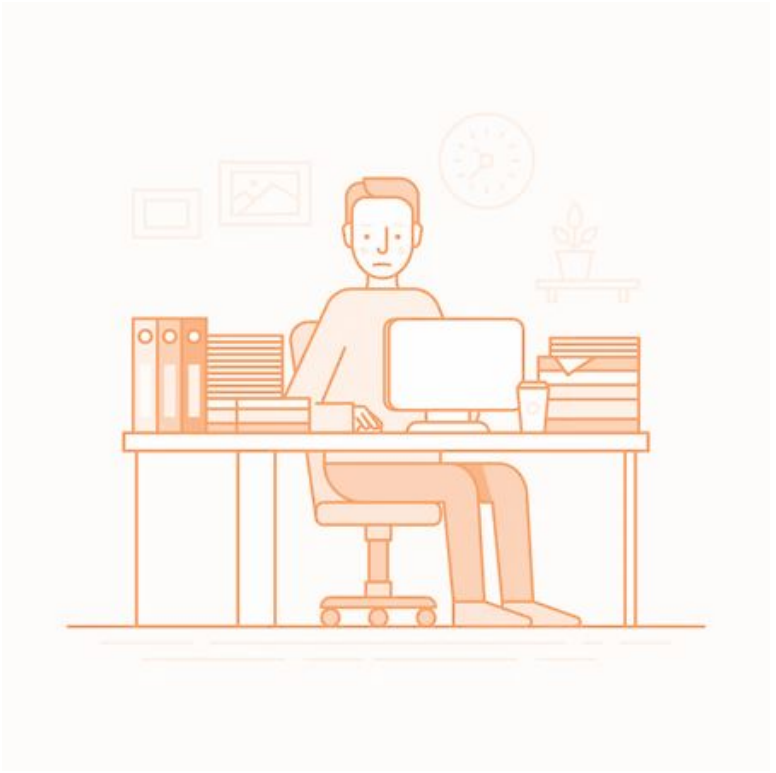
# 경상도 사투리 번역기



방언(경상도)을 사용하는 일상 대화를 수집하여 음성을 문자로 변환한 방언 발화 데이터셋 구축




조용한 환경에서 2,000명 이상의 화자가 발화한 3,000시간 이상의 음성 데이터셋의 원본 표준어 텍스트 및 방언 특성을 고려하여 전사한 텍스트 50만건



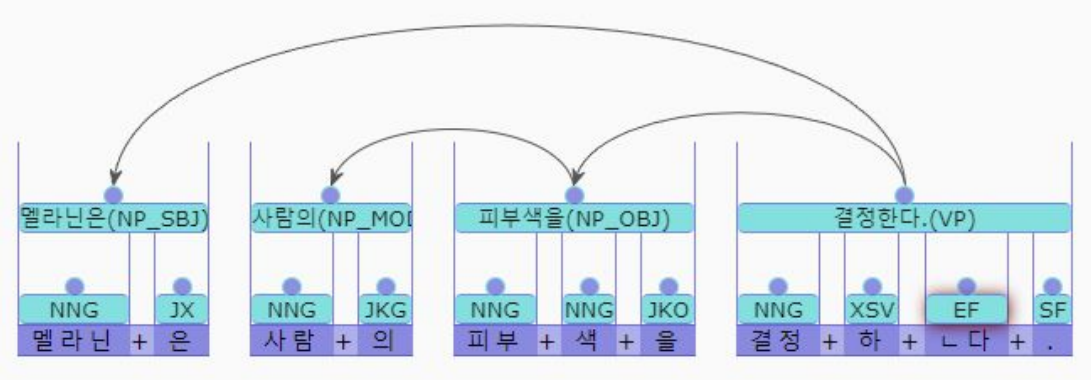
# 경상도 사투리 번역기

프로젝트 배경 | 수행 절차 및 방법 | 수행 결과

데이터 설명 및 전처리 EDA 과정 및 개요 모델링 과정

## ETRI 의존구문분석

의존 구문분석 API는 자연어 문장의 구조를 분석하는 기술로,
문장 내 각 어절에 대해서 지배소 어절을 인식하고, 주격,
목적격과 같은 세부 의존관계 유형을 인식하는 기술입니다.




## 구문태그셋

| 구문 태그 | 의미 |
|---|---|
| NP | 체언(명사, 대명사, 수사) |
| VP | 용언(동사, 형용사, 보조용언) |
| AP | 부사구 |
| ANP | 긍정 지정사구(명사+이다) |
| DP | 관형사구 |
| IP | 감탄사구(호칭 및 대답 등의 표현) |
| X | 의사구(pseudo phrase,조사 단독 어절 또는 기호 등) |
| L | 부호(왼쪽 괄호 및 따옴표) |
| R | 부호(오른쪽 괄호 및 따옴표) |

## 기능 태그셋

| 기능 태그 | 의미 |
|---|---|
| SBJ | 주어 |
| OBJ | 목적어 |
| MOD | 관형어 (체언 수식어) |
| AJT | 부사어 (용언 수식어) |
| CMP | 보어 |
| CNJ | 접속어(~와) |

# 경상도 사투리 번역기



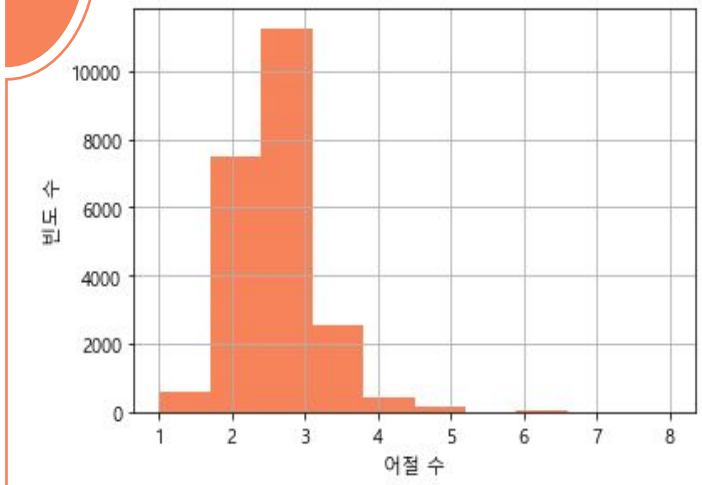
| 용도 | 이름 | 주요 내용 | 데이터 구축량 | 데이터 타입 |
|---|---|---|---|---|
| 번역 모델 | 표준어 사전 | {'이제': 161, '거기': 1, '조금': 3, '그렇게': 358, '하여튼': 128, | 389 | 딕셔너리 |
| | 사투리 사전 | {'인자': 0, '그기': 1, '인저': 2, '쫌': 3, '그래': 4, '하이튼': 5, | 389 | 딕셔너리 |
| 동의어 처리 모델 | 품사 사전 | {'NP_SBJ': 0, 'NP_OBJ': 1, 'NP_MOD': 2, 'NP_AJT': 3, 'NP_CMP': 4, 'NP_CNJ': 5, 'VP_SBJ': 6, | 68 | 딕셔너리 |
| | 전체 어절 사전 | {'이제': 88034, '인자0': 1, '거기': 914, '그기0': 3, '인저0': 5, '조금': 6, '쫌0': 7, '그렇게': 1264, '그래0': 9, | 31682 | 딕셔너리 |
| | 동의어 사전 | ['그래', '그니까', '땜에', '있나', '먹었다', '아이가', | 104 | 리스트 |
| | 어절별 의존 구문의 품사 사전 | pos: 3, word: 155743<br>pos: 8, word: 2194 | 6133 | 데이터 프레임 |

# 경상도 사투리 번역기

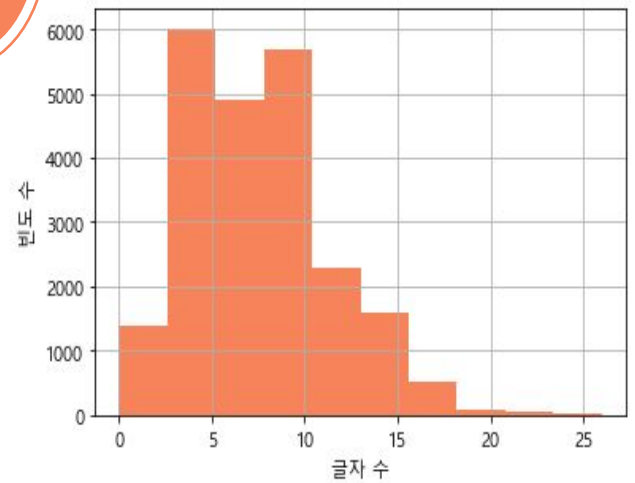
프로젝트 배경 | 수행 절차 및 방법 | 수행 결과

데이터 설명 및 전처리 EDA 과정 및 개요 모델링 과정

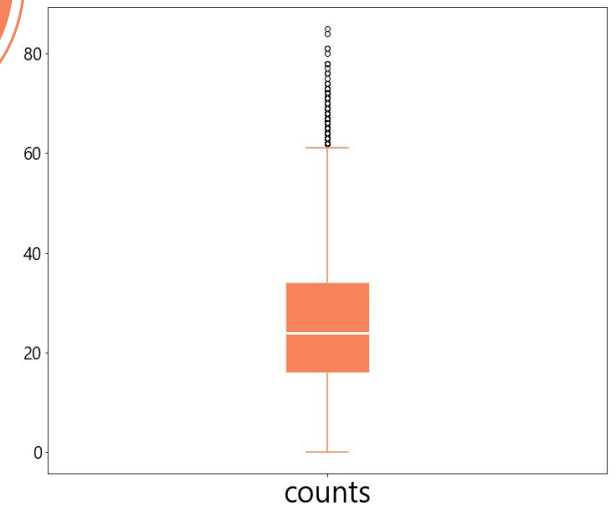
01




02




03




## 어절수 빈도

문장별 글자수는 2~3개가 제일 많음

## 글자수 빈도

어절별 글자수는 5~10개가 제일 많음

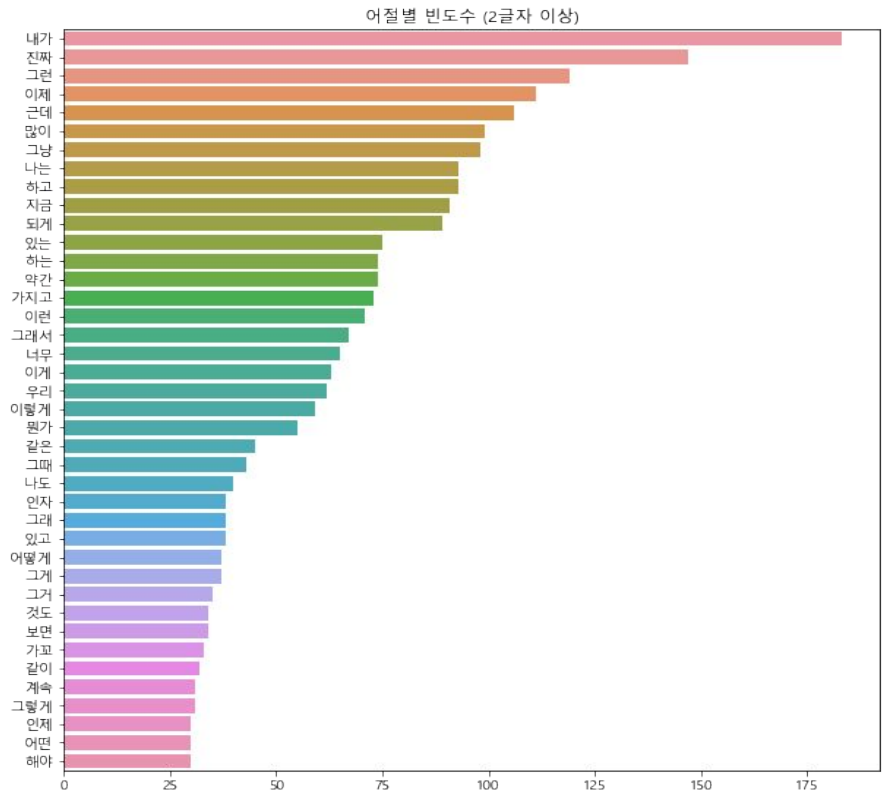
## 문장 길이

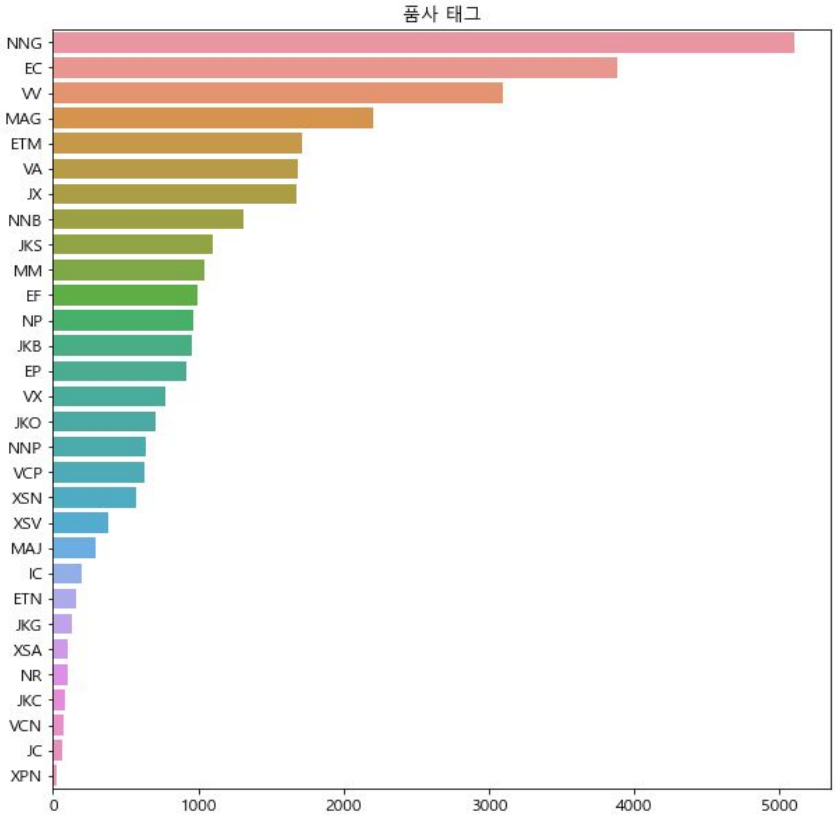
문장길이의 중앙값과 이상치 확인

# 경상도 사투리 번역기

프로젝트 배경 | 수행 절차 및 방법 | 수행 결과

데이터 설명 및 전처리 EDA 과정 및 개요 모델링 과정



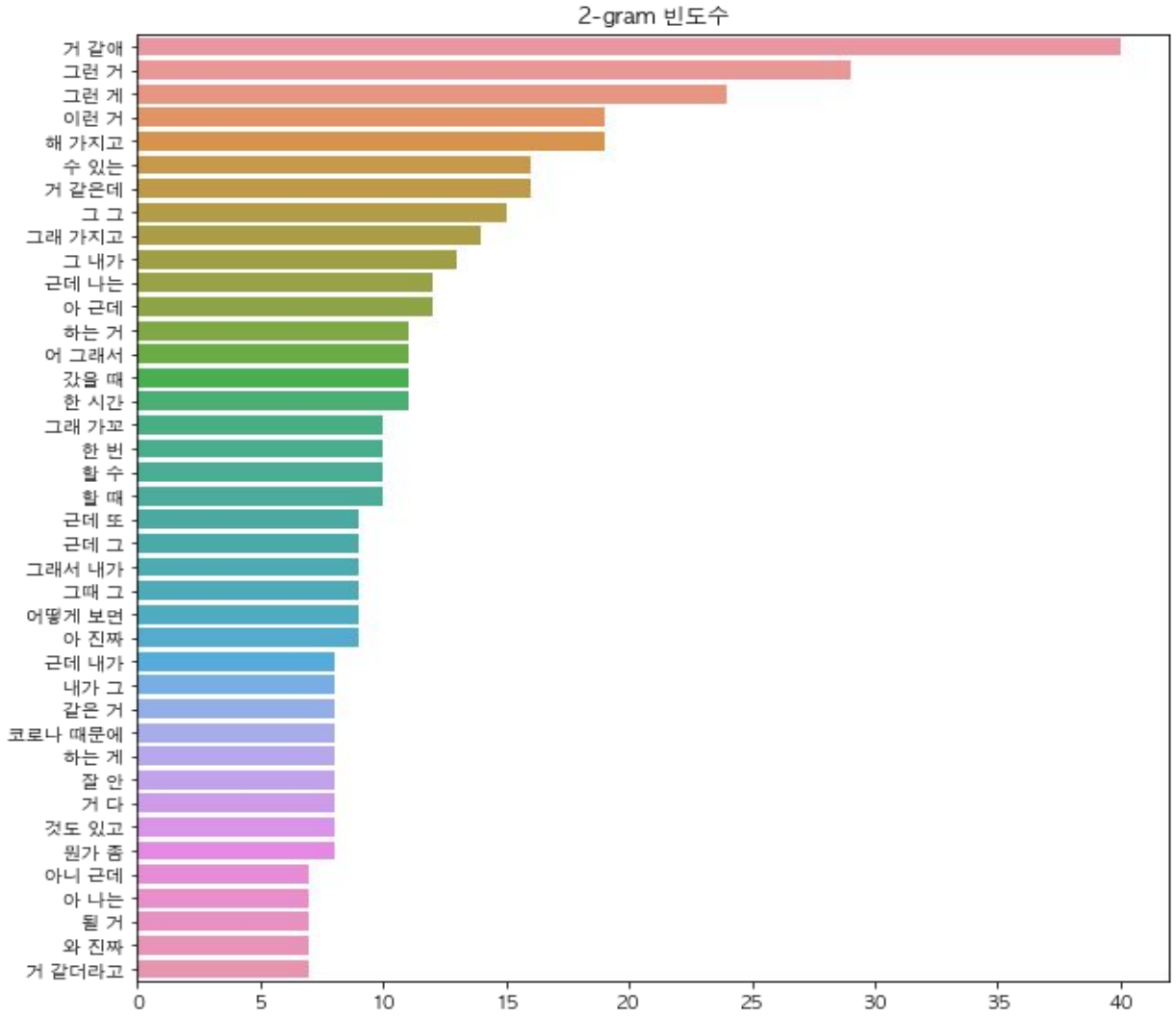




**KEY point**

- 2글자 이상인 어절별 빈도수
- pos_tag

# 경상도 사투리 번역기

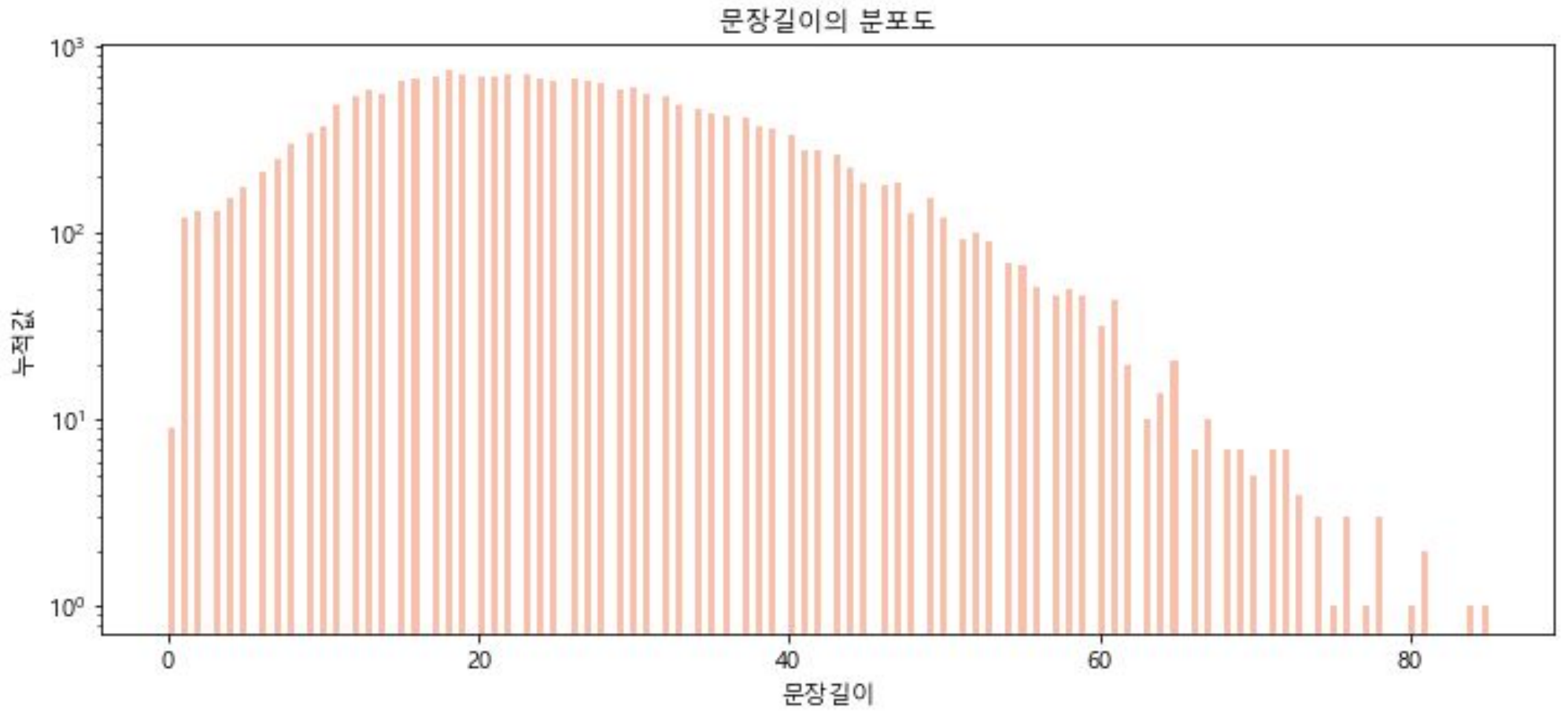





## 어절의 2-gram 빈도수

# 경상도 사투리 번역기






## 각 문장길이의 분포도

# 경상도 사투리 번역기

프로젝트 배경 | 수행 절차 및 방법 | 수행 결과

데이터 설명 및 전처리 | EDA | 과정 및 개요 | 모델링 과정



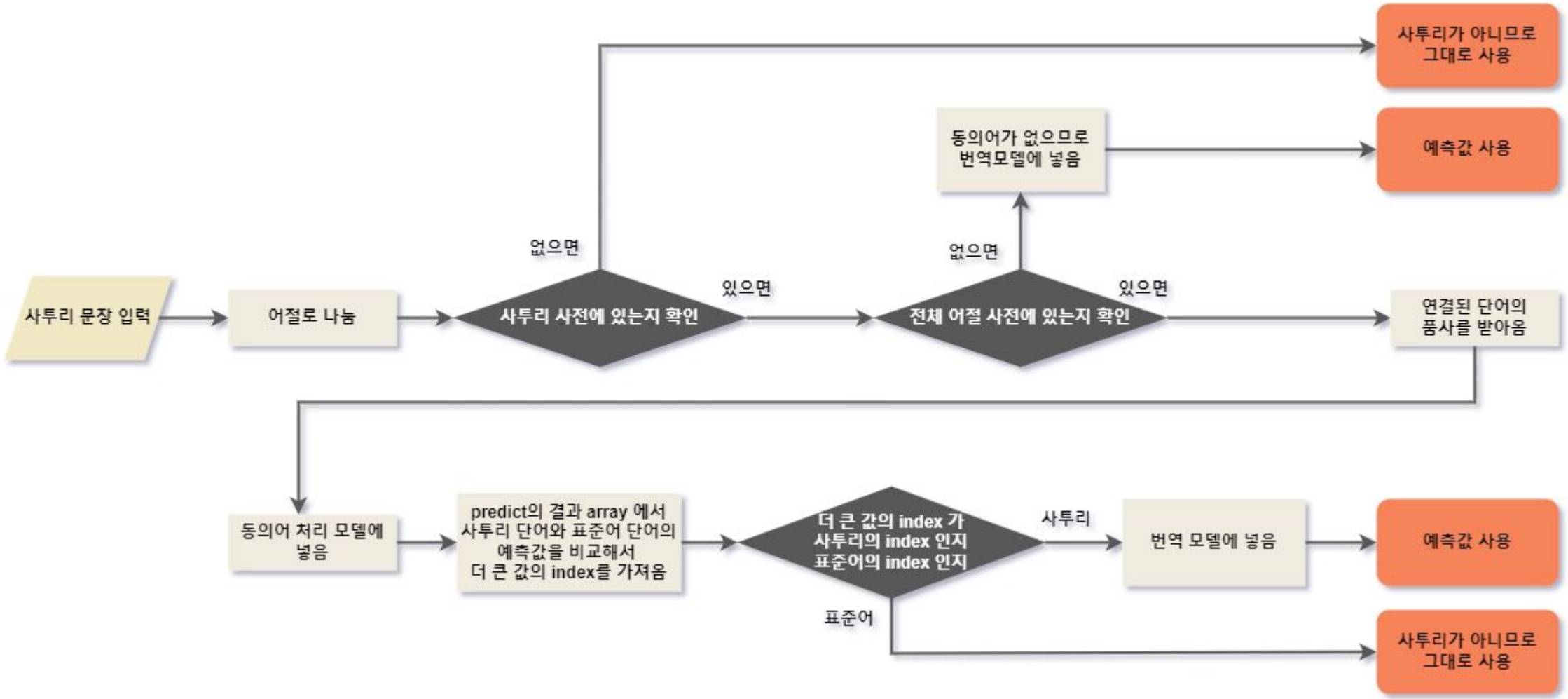
**KEY point**

- 표준어 사전의 워드 클라우드
- 사투리 사전의 워드클라우드

# 경상도 사투리 번역기



## 전체 프로세스



# 경상도 사투리 번역기

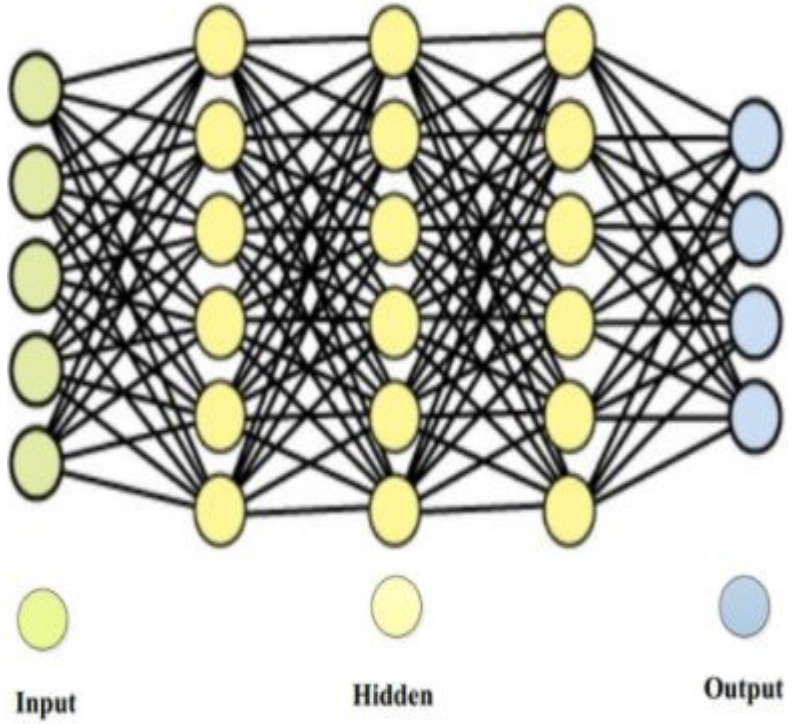


## DNN.

복잡한 데이터를 학습하기 위해 은닉 계층을 많이 쌓아서 만든 인공지능 기술
ANN은 주로 하나의 은닉 계층을 포함하나 DNN은 수십에서 수백의 은닉 계층으로 구성 가능
다수의 은닉 계층을 활용하면 은닉 계층 하나를 활용할 때보다 입력 신호를 더 정교하게 처리 가능

Purpose
분류 및 수치 예측. 영상처리, 음성인식, 자연어 처리와 같은 분야에서 자주 쓰임.

=>복잡도가 높은 비정형 빅데이터에 용이하지만
결국은 과적합을 얼마나 방지하느냐가 이를 제대로 활용하는 열쇠



# 경상도 사투리 번역기



## Adam

아담(Adam)은 Adaptive Moment Estimation의 약자입니다.
모멘텀과 RMSprop을 섞어놓은 최적화 알고리즘 이기 때문에,
딥러닝에서 가장 흔히 사용되는 최적화 알고리즘 입니다.

$$v_{dW} = 0, S_{dW} = 0, v_{db} = 0, S_{db} = 0$$

$$v_{dW} = \beta_1 v_{dW} + (1-\beta_1)dW$$
$$v_{db} = \beta_1 v_{db} + (1-\beta_1)db$$

$$S_{dW} = \beta_2 S_{dW} + (1-\beta_2)dW^2$$
$$S_{db} = \beta_2 S_{db} + (1-\beta_2)db^2$$

$$v_{dW}^{biascorr} = v_{dW}/(1-\beta^t)$$
$$v_{db}^{biascorr} = v_{db}/(1-\beta^t)$$
$$S_{dW}^{biascorr} = S_{dW}/(1-\beta^t)$$
$$S_{db}^{biascorr} = S_{db}/(1-\beta^t)$$

마지막으로 Momentum과 RMSprop의 가중치 업데이트 방식을
모두 사용하여 가중치 업데이트를 진행합니다.

## Softmax

소프트맥스 회귀는 가설 함수로 소프트맥스 함수를 사용하며, 비용 함수로는 CrossEntropy를 사용한다.

$$softmax(z) = [\frac{e^{z1}}{\sum_{j=1}^{3} e^{zj}} \frac{e^{z2}}{\sum_{j=1}^{3} e^{zj}} \frac{e^{z3}}{\sum_{j=1}^{3} e^{zj}}] = [p_1, p_2, p_3] = \hat{y} = \text{예측값}$$

분류해야하는 클래스의 총 개수가 k 개일 때, k 차원의 벡터 z = [z1, z2, z3]을 입력 받으면 소프트맥스 함수는, 위와 같이 각 클래스에 대한 확률 [p1, p2, p3]을 리턴한다.

p1, p2, p3 는 각각 1번 클래스가 정답일 확률, 2번 클래스가 정답일 확률, 3번 클래스가 정답일 확률이며, 각 확률은 0~1 사이의 값이고 확률들의 총합은 1 이다.

# 경상도 사투리 번역기



## Translate_model

**입력값**
x : 사투리 어절
y : 표준어 어절

↓

DNN layer

↓

비선형연산결과

↓

표준어 어절 예측

```
Epoch 50/50
12/12 [==============================]
- 0s 4ms/step - loss: 0.0292 - accuracy: 0.9975
```

## synonym_model

**입력값**
x : 어절과 연관있는 어절의 품사
y : 어절

↓

DNN layer

↓

비선형연산결과

↓

동의어인 어절이 표준어인지 사투리인지 예측

```
Epoch 50/50
13/13 [==============================]
- 0s 31ms/step - loss: 4.5753 - accuracy: 0.1958
```

# 프로젝트 수행 결과

결과

기대효과

# 경상도 사투리 번역기

프로젝트 배경 | 수행 절차 및 방법 | 수행 결과

결과 기대효과

| 입력문 | 번역 결과 |
| --- | --- |
| predict_standard("맞제 이거 약간 모의 투자 느낌.") | 맞지 이거 약간 모의 투자 느낌. |
| predict_standard("최고의 리더는 글을 쓴다.") | 최고의 리더는 글을 쓴다. |
| predict_standard("최고의 리더니까 최고의 글을 쓴다 아이가?") | 최고의 리더니까 최고의 글을 쓰지 않아? |
| predict_standard("너처럼 어린 아이가 무슨 운전 면허증을 따니?") | 너처럼 어린 아이가 무슨 운전 면허증을 따니? |

# 경상도 사투리 번역기

프로젝트 배경 | 수행 절차 및 방법 | 수행 결과

결과 기대효과

## 번역모델 잘 안된 케이스

predict_standard("게임을 그렇게 대충 해가지고 이기겠나?")

predict_standard("게임을 그렇게 대충 해가지고 이기긌나?")

→

게임을 그렇게 대충 해가지고 이길까?

게임을 그렇게 대충 해가지고 이기긌나?

## 동의어 처리 모델 잘 안된 케이스

predict_standard("어릴때 옆집 살던 가는 뭐하냐?")

→

어릴때 옆집 살던 가는 뭐하냐?

입력문

번역 결과

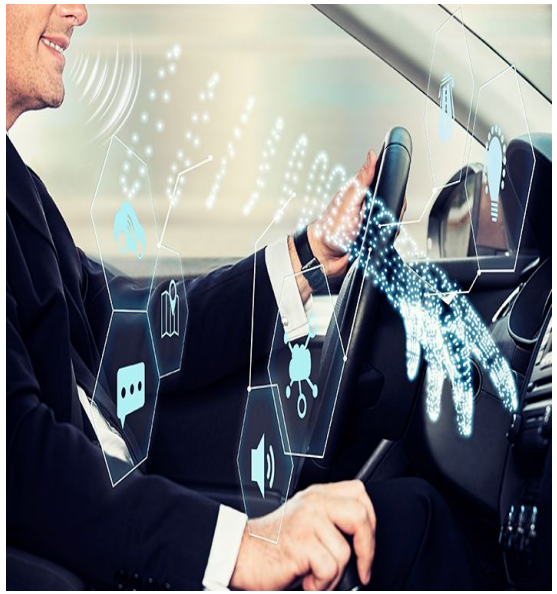
# 경상도 사투리 번역기

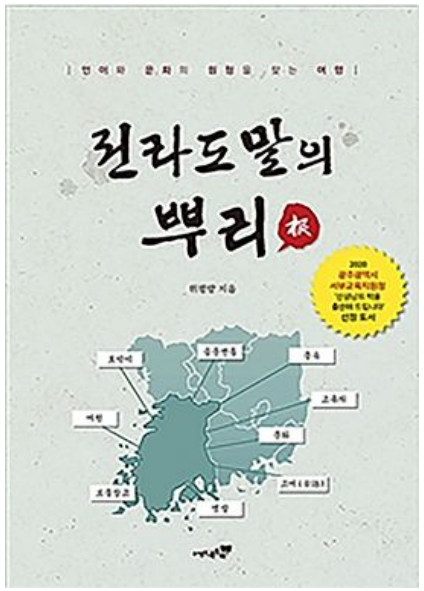


결과 기대효과

## 활용 방안

표준어 및 방언을 넘어선 소통 체제와 이를 기반으로한 각 분야별 활용 가능



# 느낀점

느낀점

앞으로의 개선 방향

# 경상도 사투리 번역기



느낀점 앞으로의 개선 방향

## 느낀점

- 제한된 리소스와 리서치의 부족으로 인해 만족하는 결과가 나오지 않아 아쉬웠다.
- 영어와 달리 한국어 자연어처리 특성상 어순이 정해져있지 않아 동의어 처리가 어려웠다.
- ETRI API의 일일 사용량 제한 때문에 많은 데이터를 사용하지 못하였기 때문에 아쉬웠다.

## 개선 방향

- 충분히 많은 데이터로 학습시켜 번역기의 성능을 개선
- 어절 대신 형태소로 학습시켜 번역기의 성능을 개선
- 카카오 API를 활용한 음성인식 및 음성합성을 구현
- 웹페이지로 구현

# Reference

## 사용 데이터 AI Hub

한국어 방언 발화(경상도) - 구축 기관(솔트룩스) https://aihub.or.kr/aidata/33981

## 참고 문헌


1. 임준호, 배용진, 김현기, 김윤정, 이규철, 의존 구문분석을 위한 한국어 의존관계 가이드라인 및 엑소브레인 언어분석 말뭉치, 제27회 한글 및 한국어 정보처리 학술대회 논문집, 2015.10.
2. Martha Palmer, Dan Gildea, Paul Kingsbury, The Proposition Bank: A Corpus Annotated with Semantic Roles Computational Linguistics Journal, 31:1, 2005
3. 임수종, 권민정, 김준수, 김현기, ExoBrain을 위한 의미역 가이드라인 및 언어처리 학습데이터 구축, 제27회 한글 및 한국어 정보처리 학술대회 논문집, 2015.10.


## 깃허브 https://github.com/hoofcas2/Dialect_translator

## Used tool

python™ TensorFlow draw.io GitHub Google Docs Microsoft Teams

# Q&A