쏘카 - 차량 정비 메모

35조 - 산학지정과제

With SOCAR

20175168 조혜영
20185290 이하영

목차 소개

Step1

프로젝트 추진 배경

Step2

프로젝트 목표

Step3

프로젝트 주요 로직

001. 프로젝트 추진 배경

카 쉐어링 업체의
차량 정비 어려움은 무엇일까요?

이용 시간을
고려한 **점검**

동일한 클레임
지속적 발생

예기치 못한
점검 상황

→ TEXT 기반의 업무 처리

TEXT 기반 업무 처리의 한계

1. 배경지식에 따라 담당자마다 사용하는 언어 상이

정비 조치 내용을 범주화하기 어려움

2. 이용자의 신고 접수시, 고객의 표현을 그대로 작성

장애 요인에 대해서 명확한 인지가 어려워짐

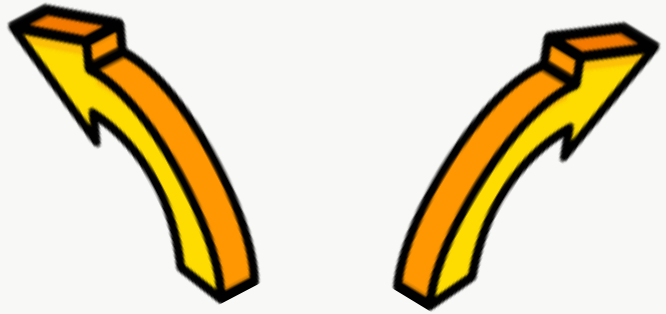
3. 10년간의 작성 방식과 사용 용어 테이블의 변화

처리 과정에서 기입이 누락된 단계가 발생

002. 프로젝트 목표

1. 카테고리 재정의

2. 정비 로직 일반화



정비 내용 범주화

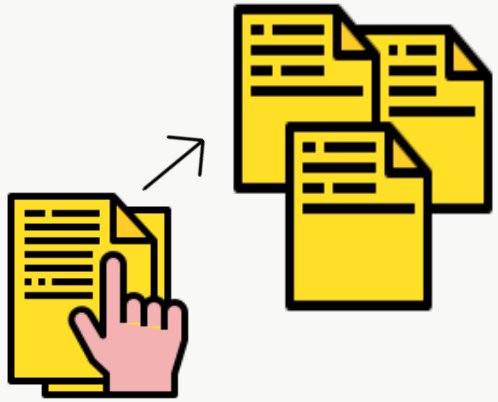
정비 메모 자연어 처리

정비 과정의 모호함 , 장애 요인의 오해 가능성

1. 카테고리 재정의

1.
각 수리 내역별
키워드 추출

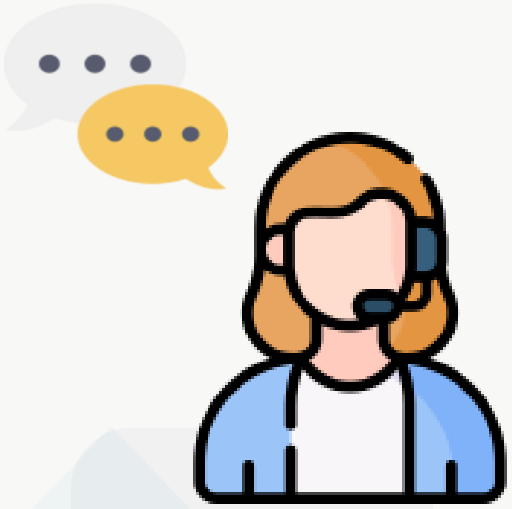
2.
혼합된 수리 내역
세분화



3.
새로운
카테고리 제안

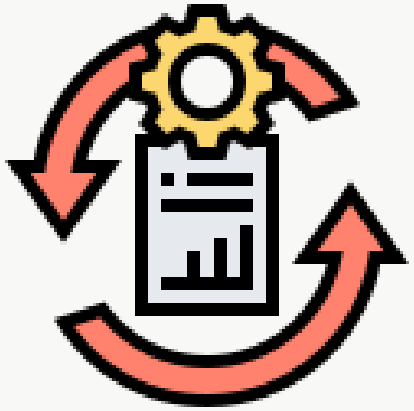
2. 정비 로직 일반화

1.
장애요인 추출



2.
누락된
진행 단계 복구

3.
정비 로직
일반화 도출



004.

프로젝트 - 진행 상황

004. 프로젝트 진행 상황 - 데이터 전처리

1. 카테고리 재정의

수리 내역
리스트

2. 정비 로직 일반화

상담 내역
리스트

1. 띄어쓰기와 맞춤법 교정으로 정확도 향상

Ex) Py-Hanspell 네이버 맞춤법 라이브러리

2. 규칙이 있는 텍스트는 정규표현식으로 데이터 정제

Ex) 고객 개인 정보, 상담 용어, 엔지니어 이름 …

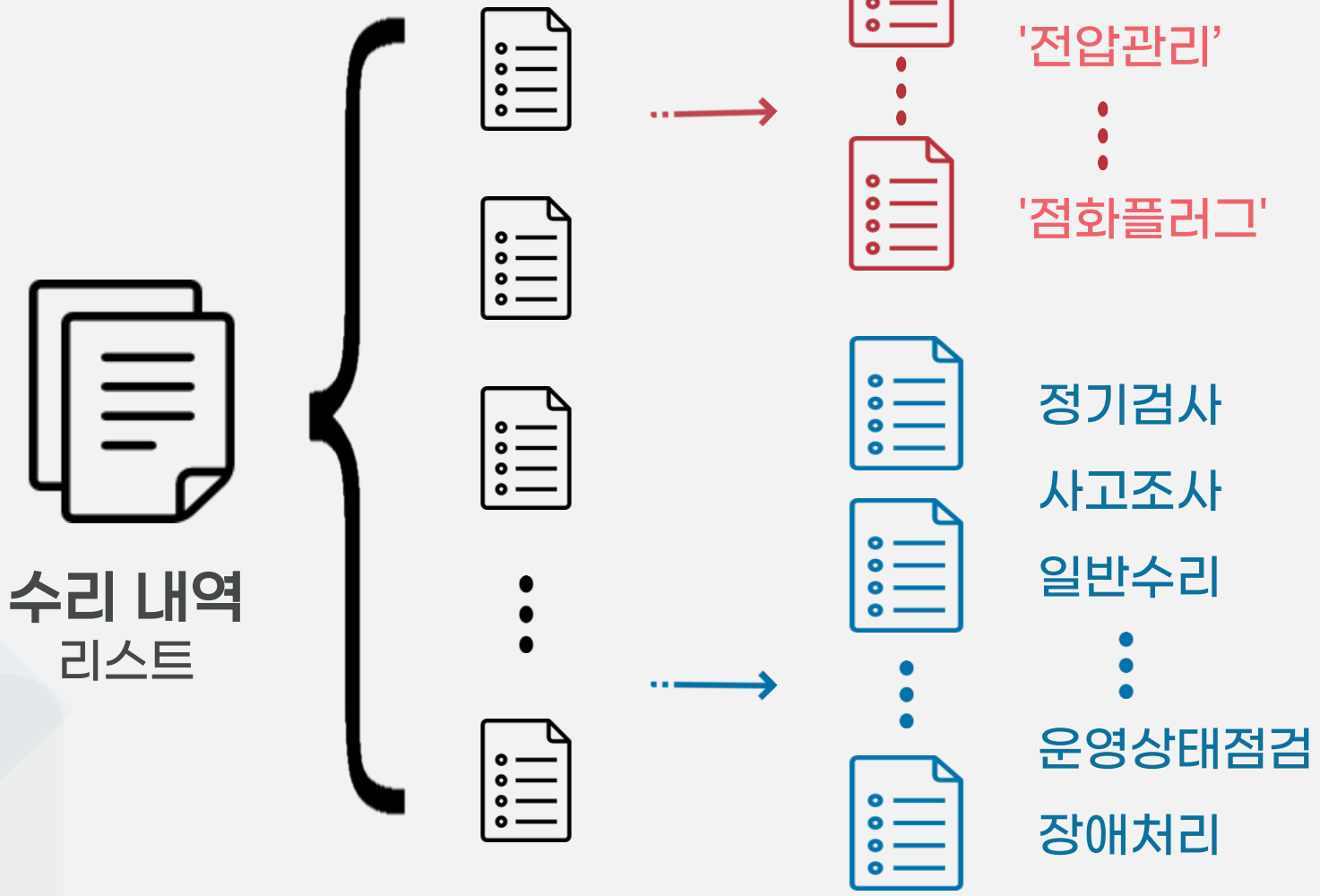
3. 규칙이 없는 데이터는 csv파일로 정제

정비 업체.csv 차종명.csv

Ex) 업체명, 차종명 …

001. 카테고리 재정의

1. 데이터 전처리 - 수리 내역 구획



001. 카테고리 재정의

2. 키워드 추출

▽ 1차원 수리 내역

'엔진오일'
'전압관리'
'점화플러그'

('엔진오일', 764)
('교환', 632)
('탈거', 263)
('점화', 332)
('교체', 135)

: 각 수리 내역마다 형태소 분석을 진행
중심 단어를 도출해 수리 내역의 특이성 확립

키워드 추출

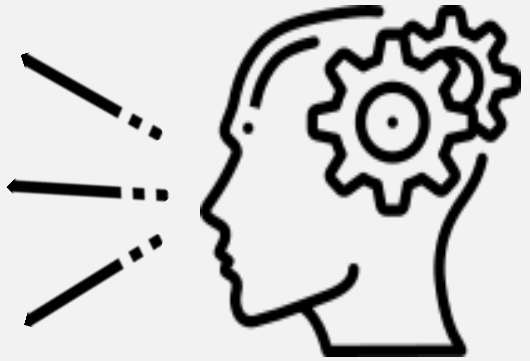
001. 카테고리 재정의

3. 분류 모델 생성

Transfer-Learning

('엔진오일', 764)
('교환', 632)
('탈거', 263)
⋮
('점화', 332)
('교체', 135)

카테고리 분류 모델



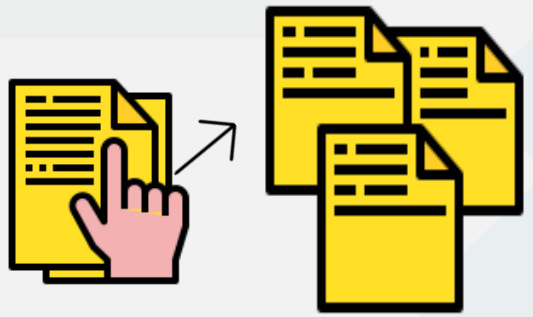
- 주어진 데이터가 충분하지 않기에 처음부터 모델을 생성하는 것이 아닌 Pre_trainning 모델을 이용하여 학습률을 높인다.
- 이때 구글에서 제공하는 BERT의 경우 multi-lingual이므로 보다 한국어에 특화된 수백만개의 한국어 말뭉치로 학습을 진행한 KoBERT로 전이학습을 한다.

001. 카테고리 재정의

4. 카테고리 분류

수리 내역 세분화

추출한 키워드를 바탕으로
어떤 수리 내역 카테고리에
속하는지 분류



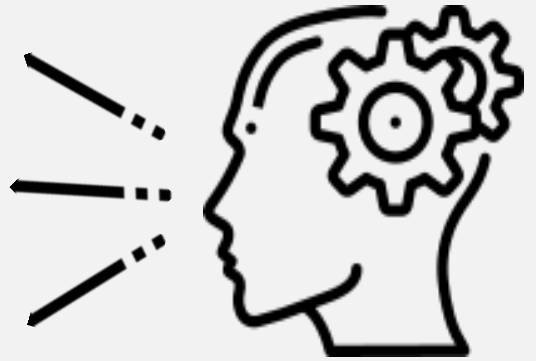
▽ 혼합된 수리 내역

정기검사
사고조사
일반수리
⋮
운영상태점검
장애처리

('엔진오일', 764)
('교환', 632)
('세트', 263)
⋮
('점화', 332)
('교체', 135)

형태소 분리

카테고리 분류 모델



002. 정비 로직 일반화

1. 데이터 전처리 - 단계 구획

상담 내역
리스트

확일화되지 않은 작성 방식
다양한 용어 사용

특정 장애 요인에 대한
프로세스를 도출하려면

업무 단계 구획 및
Missing Link 확인

'With 정규표현식'

업무 4단계

- Trigger
- Figure
- Action
- Result

002. 정비 로직 일반화

2. 장애 요인 데이터 증폭

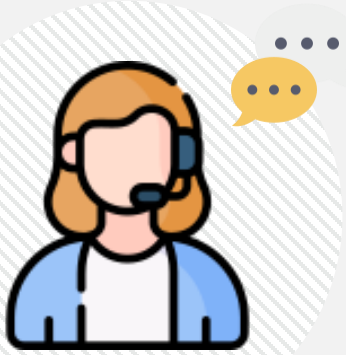
150가지의 장애 요인
But, 데이터 수 4920개

Trigger

Figure

Action

Result



장애 요인과 관련된
다양한 발화적 표현 수집

데이터 조합

핸들
핸들조작이
핸들링
핸들내 볼륨버튼

원래대로 안돌아옴
핸들조작이 잘 안됨
잘안된다고함
조작이 안돼
작동 불가

5, 핸들 돌리고난뒤 원래대로 안돌아오며 뻑뻑함
5, 핸들이 1km 운행시 핸들조작이 잘 안됨
5, 핸들링이 잘안된다고함
5, 핸들에 볼륨 조절 하는 버튼이 헐렁헐렁하며 조작이 안돼
5, 핸들내 볼륨버튼 내리는 것이 작동 불가

002. 정비 로직 일반화

3. 장애 분류 모델 생성

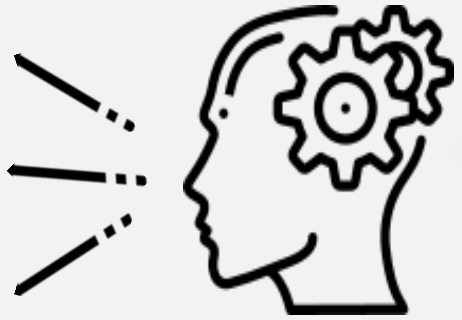
Transfer-Learning

KoBERT

추출된 **장애요인** 문장

장애 요인 판단

주행시 핸들떨림



장애 요인 분류 모델

5, 핸들 돌리고난뒤 원래대로 안돌아오며 뻑뻑함
5, 핸들이 1km 운행시 핸들조작이 잘 안됨
5, 핸들링이 잘안된다고함
5, 핸들에 볼륨 조절 하는 버튼이 헐렁헐렁하며 조작이 안돼
5, 핸들내 볼륨버튼 내리는 것이 작동 불가

005. 프로젝트 개선 사항

1. 데이터 수 증폭의 한계

- 현재 Trigger의 전체 문의 사항의 개수는 4920개임.
- '문의 사항' 내의 멘트들을 조합하여 나올 수 있는 가능성의 문장을 만들어도 데이터 셋이 10개 미만인 경우의 요소들은 데이터를 증폭시키는데 한계가 있음

부족한 양의 장애 요소는 데이터 확대를 위해 Oversampling 으로 가장 많은 샘플링의 개수에 맞춰 늘려준다.

'문의 사항' 데이터가 아니더라도 '수리 내역'이나 '문제 판단'의 데이터를 적극 활용

005. 프로젝트 개선 사항

2. 분류 범위 확장 

부족한 데이터 수로 인해 학습률이 떨어지는 것을 방지하고자
충분한 양의 데이터를 가진 요소로만 학습하여
- 장애요인 분석은 전체 150가지의 case중에서 60가지만 분류.
- 카테고리 분석은 전체 46가지의 case중에서 25가지만 분류.

Mecab을 이용하여 프로젝트에 특화된 vocab을 생성

보다 특화된 vocab을 생성하여 토크나이징을 진행한다면 데이터 양이 제한적인 상황에서도 모델의 학습률을 높일 수 있을 것이라 기대.

도움 받은 사항

1. 데이터 전처리

: 한국어는 띄어쓰기와 맞춤법에 의해 의도하지 않은 결과가 나오는 것을 방지하고자 맞춤법 교정기를 이용하여 학습에 앞서 텍스트의 문법적 오류를 제거하는 것을 권유받았습니다.
그 결과 높은 정확도로 형태소 분석이 진행되었습니다.

2. 키워드 추출기

형태소를 이용하여 단어의 빈도수만 확인하는 것이 아닌 KR-wordRank를 이용하여 키워드를 추출하는 방법을 얻게 되어 두가지 결과를 종합하여 보다 높은 수준의 대표성을 가진 키워드를 도출할 수 있었습니다.

앞으로 도움 받을 사항

모델별 Toknizing 차이에 의한 정확도

: 현재 사용하는 KoBERT 이외에도 BERT와 다른 Pretraining 된 모델을 이용하여 학습률이 가장 높게 나오는 모델을 선택할 필요성이 있고 이에 대해 멘토님에게 도움을 받을 예정

THANK YOU