NHTSA 고객 불만 데이터
토픽 이상 징후 예측 모델링

CONTENTS

01 데이터 요약

- NHTSA에서 수집한 북미지역 판매 전 자동차 모델 기준 1,005,512건의 고객불만 데이터

	CMPLID	ODINO	제조업체 이름	차량/정비 제조	차량/정비 모델	YEARTXT	...	DEALER_ ZIP	PROD_TY PE	REPAIAR ED_YN	MEDICAL _ATTN	VEHICLES _TOWED_ YN
1	1633288	11292384	Honda(Am erican Co...	HONDA	ACCORD	2018			V		N	N
2	1633289	112922384	Honda(Am erican Co...	HONDA	ACCORD	2018			V		N	N
3	1633291	11292386	Ford Motor Co..	FORD	EXPLORER	2020			V		N	N
⋮										⋮		
1005510	1089007	10615480	Ford Motor Co..	FORD	ESCAPE	2012			V		N	N
1005511	1089008	10610424	Ford Motor Co..	FORD	ESCAPE	2012			V		N	N
1005512	1089009	10610424	Ford Motor Co..	FORD	ESCAPE	2012			V		N	N

01 데이터 요약

✔ 기존 시스템의 문제점

- 불만 텍스트 데이터(500~1000자로 구성)를 활용하여 특정 키워드를 검색
  ex) not working, air bag
- 해당 키워드가 포함된 데이터를 단순 통계

순위	기능	불만 수
1	SERVICE BRAKES	90
2	EXTERIOR LIGHTING	69
3	FORWARD COLLISION AVOIDANCE	66
4	STEERING	60
	⋮	
13	EQUIPMENT	4
14	PARKING BRAKE	2
15	COMMUNICATIONS	1
	총합	472

- 불량 유형 매칭에 대한 정합성 및 신뢰도가 떨어짐
- 데이터를 활용하여 의사결정을 내릴 수 없음

결론

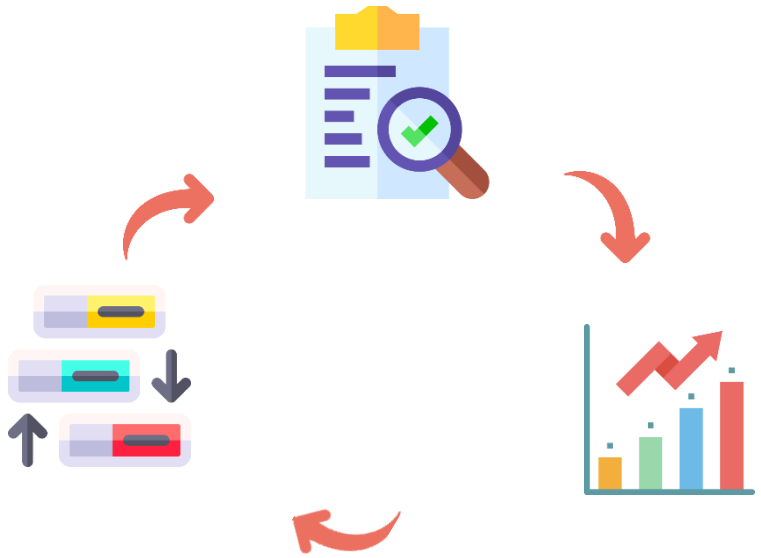
불만 텍스트 데이터를 적절하게 가공하여
예측 모델 구현을 원함

02 Baseline 구상

<기존 활용 방안>

- 토픽모델링(LDA)를 통한 자연어 처리 작업

문제점 : 특정 토픽을 특정 불만 유형으로 정의 할 신뢰성은 확보하기 어려움, 따라서 이를 종속 변수로한 모델은 구현할 수 없다고 판단



<새로운 활용 방안>

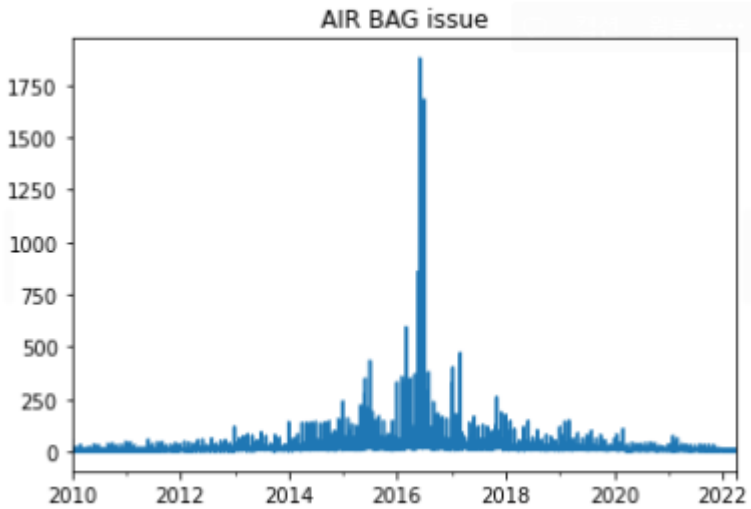
- 의사결정을 위해서는 텍스트 데이터 직접 확인은 필수적
- 부품 불량 위험이 높은 시기를 제안하여 텍스트를 직접 읽어볼 필요성이 높은 시기를 제안해주는 모델 구현

03 가설 설정

✔ 데이터 특성을 활용한 가설 설정

① EDA 과정에서 과거 자동차 리콜 이슈 전후로 추세 파악
- 불만 데이터 급변 확인

② 시기별 불만 키워드 형성 / 토픽 차이 파악
- 시기에 따라 다양한 특징을 나타냄



가설 1 “ **불만 Text 토픽이 급변할 때 부품 리콜 등 불량이슈의 위험이 있다고 판단** ”

가설 2 “ **불만 Text 토픽이 급증할 때 부품 리콜 등 불량이슈의 위험이 있다고 판단** ”

04 모델설계

✔ 데이터 프레임 형성

- 각 데이터를 일주일 단위로 끊어 각각의 텍스트에 대한 LDA의 급변을 판단

  - Coherence, Perplexity값 + 불만 발생 횟수로 새로운 시계열 데이터 프레임 생성

✔ RNN(LSTM)모델 적용

- 새롭게 생성한 시계열 데이터 프레임에 RNN(LSTM)모델을 적용

  - 1, 2번 가설에 따른 불만 이슈 급변 시기를 제안

» 텍스트 데이터에 우선순위 부여

05 성능검증

✔ 모델의 일반성 검증

- 과거 데이터를 통한 검증
  - 과거 리콜 이슈에 대해 모델이 어느 시기에 이상 징후를 감지하는지 판단
  - 공식적으로 리콜이 진행되기 전에 미리 감지할 수 있는지 여부를 관찰

- 현재 데이터에 대해 직접 검증
  - 현재 발생하고 있는 데이터에 대해 실제로 모델을 통하여 불만 이슈의 위험성을 확인
  - 현재 데이터의 불만 텍스트 일부를 직접 번역 후 검증

06 최종모델 및 효과

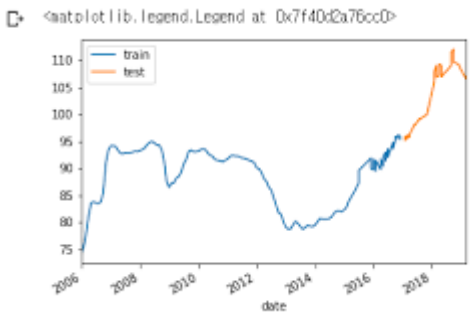
✔ 모델 파이프라인 정리

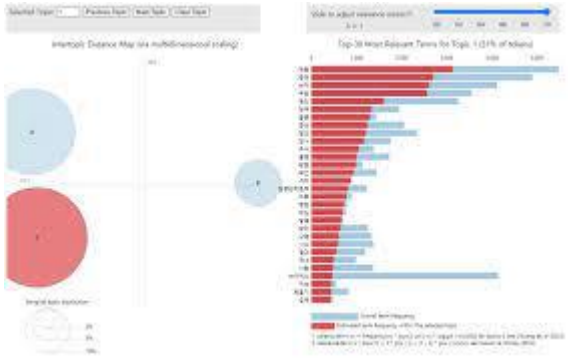
- 조회할 모델 명(input 변수)으로 데이터 전처리 ex) “SONATA”
- 학습할 구간 설정(input 변수) ex) 1년
- Window size설정(input 변수) ex) 1주일
- Window size마다의 텍스트에 대한 Dictionary 생성
- 각 Dictionary마다 LDA진행 후 Coherence, Perplexity 및 불만 횟수 전처리
- 학습할 구간에 대하여 빈 시계열 데이터 생성
- 각 시계열 데이터 포인트에 Coherence, Perplexity, 횟수 값 할당
- RNN 학습 데이터 생성 후 RNN 모델에 학습
- 토픽의 변화를 그래프로 시각화
- 변화가 심한 시점을 특정하여, 토픽 모델링 시각화

06 최종모델 및 효과

UI 제공

- 파이프 라인을 구현하여 최종적으로 UI로 그래프와 토픽모델링을 제공
- EDA → LDA → RNN 을 통해 불만 텍스트 시기의 우선순위를 제안
- 특정 자동차에 대한 토픽의 급변, 급증을 시각화





감사합니다