4 차 5 기 실무 Project

Cargle

2019 년 10 월 18 일

빅데이터(딥러닝) 활용 AI 설계 과정

투머치토카

임채명

이효정

김지현

목 차

1. 프로젝트 개요

1.1 프로젝트 기획 배경 및 목표

1. 프로젝트 기획 배경
   자동차 리뷰 관련 사이트로는 모터그래프, 오토뷰 등 다수의 사이트가 존재하지만 각 사이트에 게시된 글만 확인할 수 있어 리뷰 수가 부족함을 느꼈습니다. 따라서 흩어져 있는 자동차 리뷰를 통합하고, 수치화 하여 자동차에 관한 정보를 한번에 볼 수 있는 플랫폼을 개발하고자 하였습니다.

2. 프로젝트 목표
   웹 크롤링 기법, 자연어 처리 기법, 문서요약 기법을 이용하여 자동차 리뷰를 브랜드 또는 모델명으로 분류하고 해당 차종에 대해 긍정 리뷰와 부정 리뷰 중 어느 것이 많은 지, 각 리뷰가 평균적으로 긍정적인지 부정적인지를 알고자 합니다.

1.2 구성원 및 역할

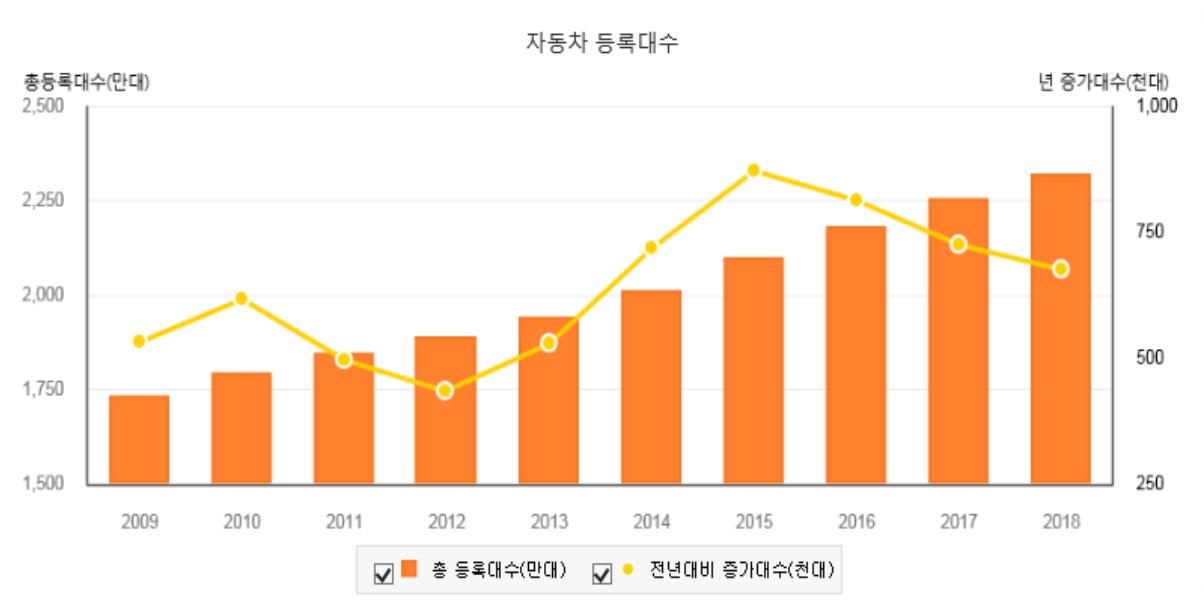
이름	전공	역할	구현 부분
임채명	자동차공학과	팀장	프로젝트 관리, 리뷰 데이터 수집 및 전처리, 감성사전 구축, 감성분석
이효정	컴퓨터공학과	팀원	리뷰 데이터 수집 및 전처리, 가격 데이터 수집, 전체 데이터 확인 및 마무리
김지현	컴퓨터공학과	팀원	리뷰 데이터 수집 및 전처리, 데이터 통합 및 정규화, 감성사전 구축, 검색 기능 구현

1.3 프로젝트 일정

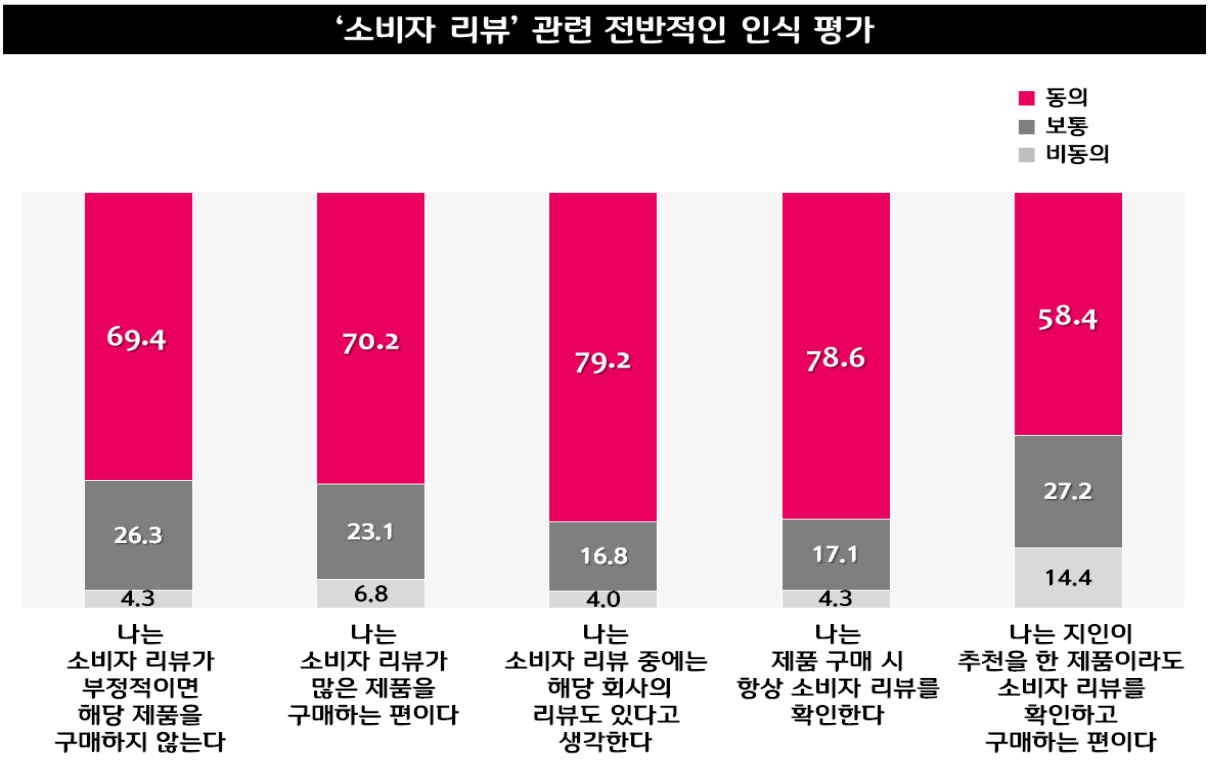
구분	기간	활동	비고
사전 기획	9/25	팀 구성	
	9/26~27	PJT 주제 선정, PJT 계획 수립	3~5 인/팀
	9/30	프로젝트 멘토링 [프로젝트 방향 설정 및 현업프로젝트 소개]	현업 멘토 참여
PJT 수행 / 완료	10/1~10/2	데이터 수집	
	10/3~10/12	데이터 전처리 및 라벨링	
	10/14	프로젝트 멘토링 [프로젝트 점검 및 기술자문]	현업 멘토 참여
	10/14 ~ 10/17	구현 및 보고서 작성	
	10/18	팀별 최종 발표 (구축 완료 보고)	최우수 한 팀 선발 멘토 평가

2. 프로젝트 현황

2.1 시장분석



<자동차 등록 현황> 출처: 국토 교통부(19.01.16)



<소비자들의 리뷰 인식 평가> 출처 : 트렌드모니터

매년 자동차를 구매하여 등록하는 수가 늘어남에 따라 자동차에 대한 정보도 늘어나고 있습니다. 게다가 그에 따른 리뷰가 소비자들의 인식에 점차적으로 크게 작용하고 있는 것을 볼 수 있습니다. 하지만 각각의 자동차 전문 사이트에는 그 사이트에 작성된 글만 확인할 수 있어 같은 차종에 대한 여러 리뷰를 보거나 비슷한 금액대의 다른 차를 비교하여 보기가 어려웠습니다. 그래서 서로 다른 사이트에 있는 다양하고 많은 데이터를 모아 같은 브랜드, 모델명, 금액별로 리뷰를 모아 분석하거나 비교할 수 있으면 편리할 것 같아 이 프로젝트를 시작하게 되었습니다.

2.2 경쟁 사이트 장단점 분석

네이버 자동차	장점	1. 해당 차량의 제원 및 가격 정보 제공, 댓글 기능, 모델 비교 기능 2. 자세한 차량 내부/외부 사진 3. 여러 사이트의 리뷰를 모두 볼 수 있음
	단점	1. 리뷰를 정리해서 한눈에 볼 수 없음 2. 금액이 아닌 모델명을 통한 검색 기능 3. 다양한 사이트의 리뷰를 확인할 수 없음
모터 그래프	장점	1. 출시예정차량, 정책, 부품 및 용품, 결함 관련 기사 제공
	단점	1. 타 사이트의 리뷰를 알 수 없음 2. 브랜드명을 통한 검색
보배 드림	장점	1. 해당 사이트를 통해 직접 차량 구입 및 판매 가능 2. 제조사, 차종, 연식, 가격, 주행거리, 연료, 변속기, 색상 등 상세 검색 3. 자세한 차량 내부/외부 사진
	단점	1. 판매자가 업로드한 매물 정보만 알 수 있음 2. 게시판 기능만 제공할 뿐 구체적인 리뷰를 제공하지 않음

2.3 차별화 핵심 전략 기술

네이버 자동차 리뷰 게시판이나 다른 자동차 관련 사이트들과는 다르게 16 개의 사이트에서 가져온 다양한 리뷰 내용들을 자연어 처리하여 감성분석을 진행합니다. 감성사전에 근거하여 각 차종에 대한 리뷰가 얼마나 긍정적인지 부정적인지를 수치화 합니다. 최종적으로, 원하는 금액대를 입력하면 미리 분석한 내용을 토대로 해당 금액대의 리뷰 분석 결과를 보여줍니다.

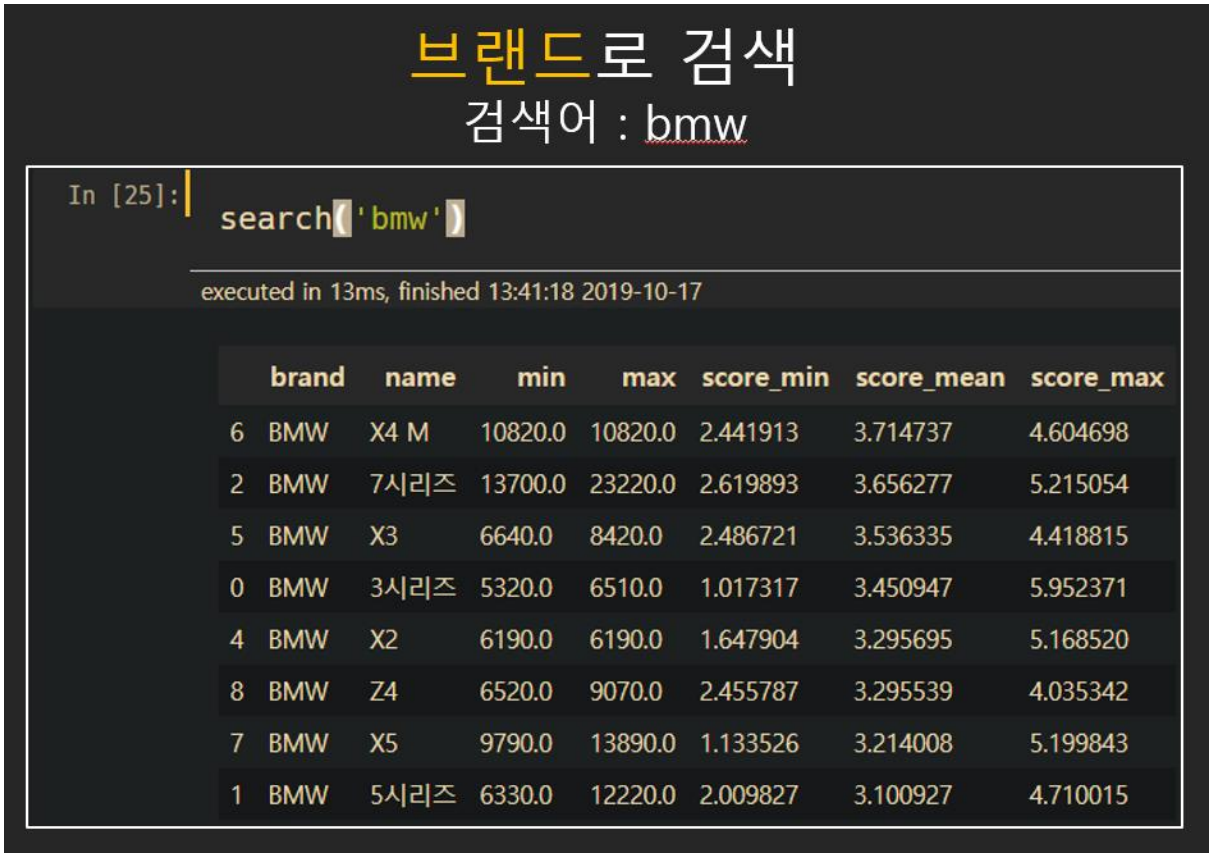
3. 프로젝트 개발 결과

3.1 프로젝트 수행 과정

데이터 수집	- Selenium, BeautifulSoup 라이브러리를 이용하여 자동차 관련 16 개 사이트에서 자동차 시승후기 및 리뷰를 수집 - 수집 시 정규표현식을 활용하여 1 차 전처리 <시승후기 출처> 모터그래프, 오토뷰, 글로벌오토뉴스, 카라이프, 엔카매거진, 카랩, 아이오토카, 오토다이어리, 오토트리뷴, 데일리카, 라이드매거진, rpm9, 모토야, 지피코리아, 탑라이더, 한경자동차
데이터 전처리	- 정규표현식을 활용하여 시승후기 본문에 포함된 기자 이름, 메일 주소, 이미지 캡션, html 코드 삭제 - Excel 및 Notepad++로 특수 기호, 띄어쓰기, 리뷰 내용과는 무관한 부분 추가 삭제
데이터 라벨링	- 브랜드와 모델명을 네이버에 등록된 명칭으로 라벨링 - 각 리뷰를 수집해온 사이트의 출처 라벨링 - 네이버에 차종 별 등록된 금액(최소금액과 최대금액) 크롤링하여 라벨링
감성사전 구축	- 군산대학교의 KNU 감성사전과 서울대학교의 KOSAC 감성사전을 통합 - 사전에 포함되지 않았으나 본문에 포함된 단어를 감성사전에 추가
자연어 처리	- Konlpy 라이브러리의 Komoran 형태소 분석기를 활용하여 리뷰 본문에 형태소 태그를 부착 - 1-gram, 2-gram, 3-gram 형태로 변환 - 한글, 알파벳, 문장 부호를 제외한 특수문자 제거

감성 점수 계산	- 자연어 처리한 데이터를 구축한 감성사전에 대입하여 수치화 - 글의 길이에 따른 영향력을 최소화하기 위해 감성 점수를 전체 단어 개수로 나누어 계산
검색 기능	- 원하는 금액대를 입력하면 해당 금액대에 속하는 브랜드, 모델명, 최소 가격, 최대 가격, 감성 점수를 출력 - 원하는 브랜드를 입력하면 해당 브랜드의 모델명, 최소 가격, 최대 가격, 감성 점수를 출력

3.2 주요 동작



	brand	name	min	max	score_min	score_mean	score_max
6	BMW	X4 M	10820.0	10820.0	2.441913	3.714737	4.604698
2	BMW	7시리즈	13700.0	23220.0	2.619893	3.656277	5.215054
5	BMW	X3	6640.0	8420.0	2.486721	3.536335	4.418815
0	BMW	3시리즈	5320.0	6510.0	1.017317	3.450947	5.952371
4	BMW	X2	6190.0	6190.0	1.647904	3.295695	5.168520
8	BMW	Z4	6520.0	9070.0	2.455787	3.295539	4.035342
7	BMW	X5	9790.0	13890.0	1.133526	3.214008	5.199843
1	BMW	5시리즈	6330.0	12220.0	2.009827	3.100927	4.710015

미리 구축한 데이터를 통해 사용자가 원하는 금액이나 브랜드를 입력하면 해당하는 차량에 대한 정보를 표 형태로 보여줍니다.

3.3 회차별 멘토링 결과

회차	내용
1 회차	최초 아이디어는 스마트폰 리뷰 분석 그러나, 삼성-애플-LG 3 개 제조사의 제품이 압도적인 점유율 차지 제품의 다양성 및 실효성에 의문 리뷰 분석할 제품 변경(스마트폰 → 자동차)
2 회차	아이디어 좋음 우리만의 감성사전 구축의 중요성 알게 됨 각 리뷰 별 점수의 가중치? 차종에 따른 리뷰의 수가 다른 경우의 가중치?

4. 기대 효과

4.1 향후 개선 사항

1. 새로운 리뷰글이 올라오면 실시간으로 크롤링하여 데이터베이스에 추가
2. 데이터베이스에 추가된 기사 본문을 자동으로 전처리
3. 이러한 서비스를 다양한 사용자가 이용하는데 편리하도록 챗봇 시스템과 연동

4.2 기대 효과

1. 원하는 금액으로 차량을 검색할 수 있어 브랜드나 모델명과 같은 배경 지식 없이도 차량에 대한 정보를 얻을 수 있음
2. 원하는 브랜드로 차량을 검색할 수 있어 모델명이나 금액대를 모르더라도 각 차량에 대한 정보를 얻을 수 있음
3. 16 개의 사이트 시승후기를 점수화 하여 평균 점수가 높은 차량 순으로 보여줌으로써 원하는 금액내에 속해 있는 차량 중 높은 점수를 받은 차량을 한눈에 알 수 있음

5. 개발 후기

성명	후기
임채명	데이터 전처리에 시간이 많이 필요했습니다. 프로젝트 계획 수립 시 전처리에 더 많은 시간 배분이 필요하다는 것을 배웠습니다. 또한, 프로젝트 계획을 꼼꼼히 검토할수록 불필요한 작업이 없어져 작업이 원활하다는 것을 배웠습니다.
이효정	데이터 전처리를 더 꼼꼼하게 하면 할수록 더 좋은 결과가 나와 시간이 조금 걸리더라도 전처리 과정을 제대로 해야 된다는 것을 배웠습니다. 계획한 일정대로 프로젝트가 진행되지만은 않아서 완료일로부터 여유시간을 두어 진행하는 것이 기한에 맞게 완성할 수 있다는 것을 배웠습니다.
김지현	감성사전을 구축하기 위해 본문 전체에 포함된 모든 단어를 긍정어와 부정어로 직접 라벨링하는 과정에서 반복적인 작업에 지치기도 했지만 다같이 열심히 하는 분위기 속에서 시너지를 얻어 더욱 열심히 할 수 있었습니다.

6. 강사 의견

평 가 요 소	배점	평
아이디어 : 유사한 서비스 존재 유무 및 체계성	/20	
2. 개발 : 실제 구현 정도 및 배포 유무, 코드의 무결성 및 난이도, 현업적용도, 실무기술 반영정도	/30	
3. PJT 수행력 : 일정관리 및 역할분담, 목표 일정 달성도, 팀내 참여도 등	/30	
4. 준비도 : 프리젠테이션 및 프로젝트 준비 정도	/20	
계	/100	강사 의견 필수