PROJECT REPORT

MID-PROJECT

팀 개요

팀명	파워레인조	프로젝트 기간	5일(22.04.19 ~ 22.04.25)

역할 분담

번호	이름	분담 내용	역할
1	이시현	데이터 전처리, t-sne 시각화, 웹페이지 제작	팀장
2	김윤민	모델 선택 및 최적화	팀원
3	이민형	데이터 전처리, 모델링	팀원
4	정승욱	EDA, 모델링, 발표 자료 제작	팀원
5	최보근	데이터 전처리, 시각화	팀원
6			

프로젝트 계획(팀별)

프로젝트 주제와 목적

[주제] : 뉴스 제목에 따른 토픽 분류 모델 개발 및 서비스화

[목적] :
뉴스 제목을 기준으로 뉴스를 7개의 토픽으로 분류하는 모델을 개발합니다.

토픽은 IT과학, 경제, 사회, 생활문화, 세계, 스포츠, 정치로 분류되며, 뉴스 토픽 분류 홈페이지를 제작하여
토픽 분류 서비스 제공하는 것을 목표로 합니다.

활용 데이터 및 출처

Dacon의 '뉴스 토픽 분류 AI 경진대회' - YNAT(주제 분류를 위한 연합 뉴스 헤드라인) 데이터 세트 활용

[Supervised ML NLP, Topic Classification] by Klue Data

예상 결과물

[stream lit] 으로 제작한 웹 페이지를 통해 뉴스 링크를 받으면 토픽을 분류해주는 서비스 제공

프로젝트 결과(팀별)

프로젝트 진행 내용

- EDA를 통해 전처리 아이디어 공유

[Phase 01]

- 데이터 전처리 - 토큰화, 품사 태깅, 텍스트 전처리 (한자 대체, 불용어 제거, 공백 제거 등)
- 단어사전 생성(tfidf matrix) 후 다양한 모델링 적용 및 모델 선택 (linear SVC)
- 단어 사전 기반 데이터 시각화(t-SNE), LDA 차원축소 후 텍스트 재 전처리 필요성 공유

[Phase 02]

- 추가 불용어 제거 작업
- linearSVC 모델링 K-fold, Grid Search로 하이퍼파라미터 튜닝 후 최종 모델 선정

[최종 시각화] Plotly, Plot 패키지 사용해 토픽별 단어 시각화

[서비스 구현] 모델 적용해 Stream lit으로 서비스 구현

[전후 분석]

t-SNE로 분류한 모델 토픽별로 시각화 했으나 sparse하게 잘 나오지 않음, PCA후 축별 단어 추출 필요성 제기

뉴스 본문 내용도 분석한다면 발전 가능성 공유

결론(향후 방안 및 활용성)

[결론]
- LinearSVC를 하이퍼파라미터 튜닝한 결과, 정확도 0.8466로 가장 높은 성능을 보임
- 웹 페이지로 뉴스의 토픽 분류 서비스 제공
- 향후 뉴스 헤드라인 뿐만 아니라, 뉴스 본문 내용까지 포함시켜 더 분명한 토픽 분류로 발전시킬 것

[프로젝트 기대효과, 활용방안]
- 기자가 뉴스 작성 후 토픽을 지정하지 않고 홈페이지에서 자동으로 분류해 게시
- 홈페이지의 토픽 분류 변경할 때마다 일일이 수작업으로 바꿔주지 않아도 됨

프로젝트 회고(개인별)

기술력 성장에 대한 회고

협업능력 및 커뮤니케이션 성장에 대한 회고