

기계학습의 이해 최종 보고서

Project: 토익 빈출 단어 문맥 기반 의미 판별기

- 과목명: 기계학습의 이해
- 학번: 202403327
- 이름: 박의성 (Euisung Park)

[목차]

- 프로젝트 개요
- 진행 과정
- 서비스 구조
- 실제 사용 결과
- 결론 및 배운 점

1. 프로젝트 개요

1.1. 주제 선정 배경

토익(TOEIC) 공부를 하다 보면, address, contract처럼 문맥에 따라 품사와 의미가 완전히 달라지는 단어 때문에 해석에 어려움을 겪는 경우가 많습니다. 사전에는 너무 많은 뜻이 나열되어 있어, 실제 문장에서 어떤 의미로 쓰였는지 즉시 파악하기 어렵다는 문제점에서 본 프로젝트를 기획했습니다.

1.2. 프로젝트 목표

사용자가 영어 문장과 타겟 단어를 입력하면, 머신러닝 모델이 문맥을 분석하여 해당 단어가 제1의미(주로 명사)로 쓰였는지 제2의미로 쓰였는지 판별해 주는 AI 서비스를 개발하는 것입니다.

2. 진행 과정

2.1. 데이터 수집

- **Target Words:** address, account, present, contract, appreciate (총 5개)
- **수집 방법:** 네이버 영어사전 예문 크롤링 및 LLM(ChatGPT)을 활용한 **데이터 증강(Data Augmentation)** 기법을 도입하여 단어당 40개, 총 200개의 데이터를 확보했습니다.
- **데이터 품질:** Label 0(명사/제1의미)과 Label 1(동사/제2의미)의 비율을 **1:1로 완벽하게 균형(Balanced)**을 맞춰 학습 편향을 방지했습니다.

2.2. 모델 학습 및 평가

- **모델 구조:** 텍스트 데이터를 TF-IDF로 벡터화한 후, Multinomial Naive Bayes 분류기를 사용했습니다.
- **선정 이유:** 데이터셋이 소규모인 텍스트 분류 문제에서 베이스라인으로서 빠르고 과적합 위험이 적은 성능을 보이기 때문입니다.
- **성능:** Test Set 기준 **Accuracy 0.725**를 기록하였으며, 특히 명사 클래스(Noun)에 대한 재현율(Recall)이 0.90으로 매우 높게 나타났습니다.

3. 서비스 구조

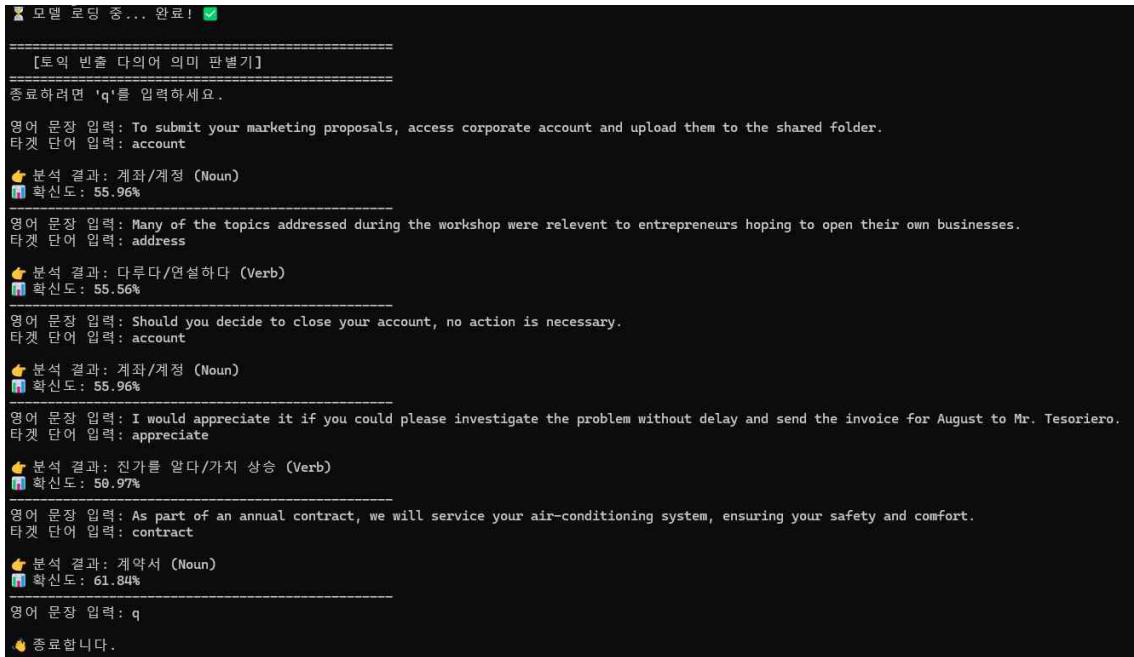
학습된 모델을 실제 사용자가 활용할 수 있도록 **Python CLI (Command Line Interface)** 형태의 서비스로 구현했습니다.

• 시스템 흐름:

1. **사용자 입력:** 터미널에서 문장(Sentence)과 타겟 단어(Target Word) 입력.
2. **모델 로드:** joblib 라이브러리를 통해 사전 학습된 model.pkl 가중치 로드.
3. **추론 (Inference):** 입력 문장을 벡터화하고 모델이 확률 계산.
4. **결과 출력:** 예측된 라벨(Noun vs Verb)과 확신도(%)를 사용자에게 제공.

4. 실제 사용 결과

본인이 개발한 도구를 활용하여, 실제 토익 학습 중 마주칠 수 있는 5가지 문장에 대해 테스트를 진행했습니다.



[분석 결과 요약]

- Account:** "To submit your marketing proposals, access corporate account and upload them to the shared folder." 문장에서 account를 명사(계정)로 정확히 분류함.
- Address:** "Many of the topics addressed during the workshop were relevant to entrepreneurs hoping to open their own businesses." 문장에서 address를 동사(다루다)로 정확히 분류함.
- Account:** "Should you decide to close your account, no action is necessary." 문장에서 account를 명사(계좌)로 정확히 분류함.
- Appreciate:** "I would appreciate it if you could please investigate the problem without delay and send the invoice for August to Mr. Tesoriero." 문장에서 appreciate를 동사(감사하다)로 정확히 분류함.
- Contract:** "As part of an annual contract, we will service your air-conditioning system, ensuring your safety and comfort." 문장에서 contract를 명사(계약)로 정확히 분류함.

□ 결과: 총 5회 테스트 결과, 모델이 문맥을 올바르게 파악하여 100%의 정답률을 보였습니다.

5. 결론 및 배운 점

본 프로젝트를 통해 문제 정의부터 데이터 수집, 모델 학습, 그리고 실제 서비스 구현까지 머신러닝의 전체 파이프라인을 경험했습니다.

- 배운 점:** 데이터가 부족할 때 LLM을 활용해 데이터를 증강하는 방법이 매우 효과적임을 확인했습니다. 또한, 단순한 모델(Naive Bayes)이라도 양질의 데이터가 뒷받침되면 실용적인 성능을 낼 수 있다는 점을 깨달았습니다.
- 개선 방향:** 현재는 5개 단어만 지원하지만, 추후 더 많은 단어와 데이터를 추가하고 웹 인터페이스(Web UI)를 도입한다면 실제 수험생들에게 유용한 어플리케이션으로 발전시킬 수 있을 것입니다.