자연어 기반 인공지능 산업분류 자동화 경진대회

대회소개

▮ 통계데이터 인공지능 활용대회란?

■ 통계 데이터의 새로운 활용 분야를 발굴하고 데이터의 잠재적 가치를 제고하고자, 자연어 기반의 인공지능 분류 학습에 필요한 양질의 통계데이터를 개방하여, 이를 통해 인공지능 학습 및 활용 저변을 확대하고자 개최하는 대회

지원 대상

■ 통계데이터와 인공지능(AI) 및 기계학습에 관심 있는 대한민국 국적의 누구나 지원 가능 ※개인 또는 팀(3명 이하)의 내국인, 외국인의 경우 개인 참여는 어려우며 한국인이 포함된 팀의 팀원으로 참가 가능

▍공모 주제

- "자연어 기반 인공지능 산업분류 자동화"
- 자연어 기반의 통계데이터를 인공지능으로 자동 분류하는 기계학습 모델 발굴로 통계 데이터 활용 저변 확대

주관 및 주최 : 한국통계진흥원

Contents

01. 경진대회 소개

02. 아키텍쳐 개요

03. 결과 정리

04. Appendix

경진대회 소개

대회소개

▮ 통계데이터 인공지능 활용대회란?

■ 통계 데이터의 새로운 활용 분야를 발굴하고 데이터의 잠재적 가치를 제고하고자, 자연어 기반의 인공지능 분류 학습에 필요한 양질의 통계데이터를 개방하여. 이를 통해 인공지능 학습 및 활용 저변을 확대하고자 개최하는 대회

지원 대상

- 통계데이터와 인공지능(AI) 및 기계학습에 관심 있는 대한민국 국적의 누구나 지원 가능
- ※개인 또는 팀(3명 이하)의 내국인, 외국인의 경우 개인 참여는 어려우며 한국인이 포함된 팀의 팀원으로 참가 가능

공모 주제

- "자연어 기반 인공지능 산업분류 자동화"
- 자연어 기반의 통계데이터를 인공지능으로 자동 분류하는 기계학습 모델 발굴로 통계 데이터 활용 저변 확대

진행일정

| 참가신청 | 자료 분석 기간 | 결과물 제출 | 심사 및 결과,시상 |
|--------------------|---------------------|--------|--------------------|
| 3.7.(월) ~ 3.21.(월) | 3.14.(월) ~ 4.13.(수) | | 4.14.(목) ~ 5.4.(수) |

심사방법

- 정확도 평가 및 전문가 평가
- (1차 평가) 예측값의 정확도 기준 시상 인원의 2~3배수 선출
- (2차 평가) 1차 심사 대상자의 코드설명자료로 코드작성의 적정성 평가

01

경진대회 소개 – 주요 특징

- ✓ 3개의 feature 이용해 3가지 target을 각각 분류하는 문제
- 각 targe은 의존성이 존재 (target1 ⊃ target2 ⊃ target3)

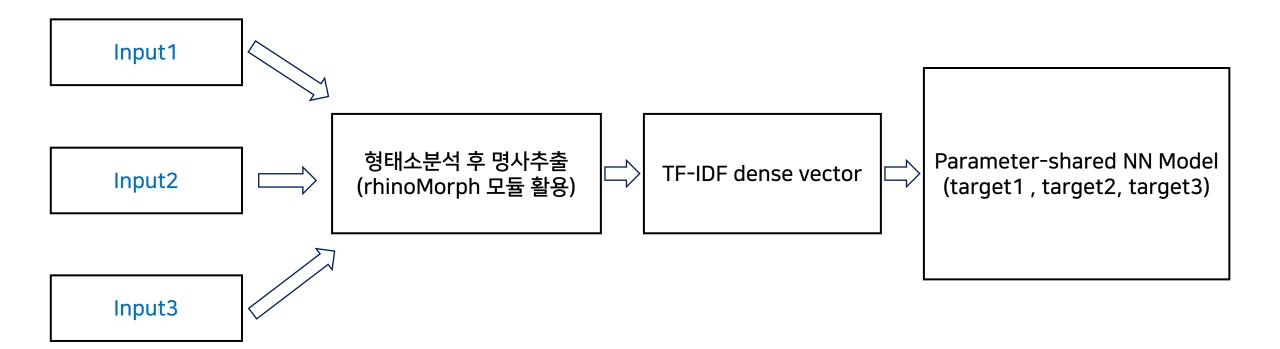


아키텍쳐 개요1 – Stage Learning

각 target에 대한 모델을 따로 학습시키고, 의존성이 있는 target의 모델은 의존된 target을 학습시킨 모델의 최종 output을 input에 결합하여 학습 NN Model (target1) Input1 softmax output **NN Model** 형태소분석 후 명사추출 TF-IDF dense vector Input2 (target2) (rhinoMorph 모듈 활용) softmax output Input3 **NN Model** (target3)

아키텍쳐 개요2 - Multi Task Learning

✓ 파라미터를 공유하는 모델을 생성하여 학습 (Tensorflow Recommendation System 내 예시 코드를 수정하여 활용)



03 결과 정리

★순위현황 *(22.06.08. 기준)

| 현재 순위 | 응시 번호 | Accuracy | F1-score |
|----------|----------|----------|----------|
| 1등 | AI0149 | 91.33 | 81.57 |
| 2등 | AI0135 | 91.26 | 82.31 |
| 3등 | AI0288 | 91.25 | 81.31 |

^{*} 최종 심사결과와 상이할 수 있음.



| 순위 | 응시 번호 | Accuracy | F1-score |
|-----|----------|----------|----------|
| 59등 | AI0184 | 88.38 | 76.67 |

^{*} 순위현황, 나의 점수: 휴일은 집계되지 않음.

- ✓ Multi Task Learning 방식으로 Accracy 88.38 달성
- ✓ 3가지 targe에 대해 하나의 모델로만 학습하여 상위권 score와 필적할 만한 성능을 보임
- ✓ 적은 파라미터로 높은 performance의 모델을 설계

감사합니다