캠퍼님들 안녕하세요!
이번 WEEK4 자연어처리, 금요일의 NER 과제를 맡은 문영기 조교라고 합니다.
이번 과제는 여러분들께서 자주 접하시게 될 캐글 형태의 대회를 미리 체험해보고, 연습하자는 취지에서 기획되었습니다. 열심히 참여하셔서 리더보드 상위권을 찍으시고 (1등이면 더 좋겠죠!), 관련한 후기들을 기록으로 남기시면 좋은 포트폴리오가 되지 않을까 생각합니다!

과제 task 설명

과제는 큰 틀에서는 pretrained language model을 finetuning하여 개체명 인식(Named Entity Recognition)을 수행하는 것에 주안점을 두고 있습니다. 학습 데이터로는 모두의말뭉치에서 제공하는 개체명인식말뭉치를 사용하고, 캐글 리더보드의 테스트용으로는 Upstage에서 제공하는 개체명인식 말뭉치를 사용하고 있습니다. 모두의 말뭉치는 코랩 폴더 내에 별도의 전처리 코드를 첨부하였고, 코드의 결과물로 train.tsv, dev.tsv를 얻으실 수 있습니다. Upstage 말뭉치는 캐글 대회 페이지에서 받으실 수 있는 test.txt 파일로서, 대회 진행을 용이하게 하기 위해 제가 약간의 전처리를 진행하였습니다. test.txt 파일을 확인하실 때에 이상한 점이 있더라도 문제가 있는 것은 아니니, 너그러이 넘어가주시면 감사하겠습니다 :)

대회를 시작하기에 앞서 개체명인식에 대해 말씀드리고자 합니다. 개체명 인식이란 문장 내에서 인물, 시간, 장소 등을 인식하는 task입니다. 여기서 인물, 시간, 장소 등을 tag라고 부르는데, 대회 데이터에서 사용하는 tag는 PS, LC, OG, DT, TI, QT를 사용하고 있습니다. 각각 PS는 사람(Person), LC는 장소(Location), OG는 단체(Organization), DT는 날짜(DATE), TI는 시간(TIME), QT는 계측단위(Quantity)를 뜻합니다. 이러한 TAG의 기준은 TTA 표준 개체명 태그셋을 기준으로 결정되었으며, 이러한 기준은 Upstage에서 제작한 데이터에도 반영되어 있습니다. 개체명 인식을 문장으로 예시를 들어보자면 다음과 같습니다. “<대한민국:OG> <3대:QT> 온천으로 <충주 수안보:LC>, <온양 온천:LC>, <울진 백암온천:LC>이 있고, 그 외에도 힐링을 위한 온천여행 패키지 상품은 다양하답니다.”

아마 캠퍼님들께서 데이터를 열어보셨을 때 tag에서 생소한 부분이 보일겁니다. PS, LC 와 같은 tag가 아니라 PS-B, PS-I,O 등의 태그가 보이실텐데요. 이러한 방식은 BIO 표기법이라고 합니다. 위의 예시에 있는 “충주 수안보, 온양 온천, 울진 백암온천”를 보시면 모두 다 LC로 표기하게 되면 각 개체들을 구분할 수가 없게 됩니다. 그래서 충주(LC-B) 수안보(LC-I), 온양(LC-B), 온천(LC-I)로 표기하여 각 단어들을 구분할 수 있게 합니다. O는 개체명이 아닌 것들에 붙는 tag입니다. Tag에 대한 세부적인 내용들은 위에서 언급한 “ TTA 표준 개체명 태그셋” 문서에서 확인하실 수 있습니다. 해당 문서는 구글에서 검색하시면 바로 확인하실 수 있습니다.

과제 수행 방법

앞 설명이 너무 길었네요. 바로 본론으로 들어가겠습니다. 여러분들께서 참조 하실 링크는 아래 2가지 입니다.

쉽게 수행하실 수 있도록 설명들은 적어보았는데, **과제 진행에서 막히는 부분이 있으시다면 슬랙 #질의응답 채널에 올려주시면 최대한 빠르게 답변 달아드리겠습니다.**

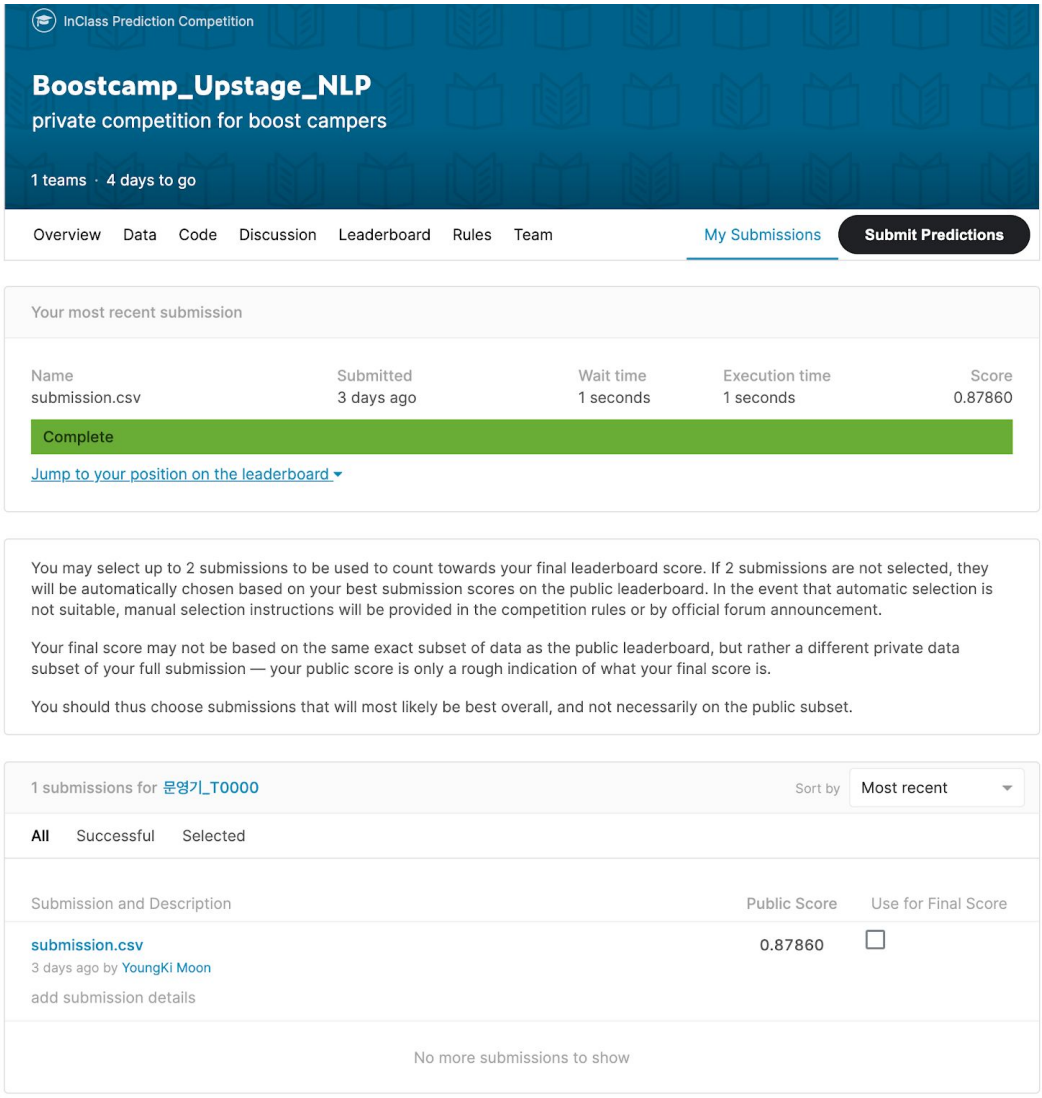
캐글 대회 참여 링크 : https://www.kaggle.com/t/00100e33445f4032bebb53c44fd0c4e5
코랩 폴더 링크 :
https://drive.google.com/drive/folders/1140kPuewjZ3rasT4DvkHaLm0ZqFwpV59?usp=sharing

대회 관련 안내사항

1. 본 과제는 참여하는 모든 캠퍼님들이 모두의말뭉치 - 개체명인식 말뭉치에 대한 사용 권한을 승인 받았다는 전제하에서 진행됩니다. **사용 권한으로 인해 생긴 문제는 주최, 운영 측에서 책임지지 않습니다.**
2. 위의 캐글 대회 참여링크를 눌러 대회에 참여할 수 있습니다.
3. 개인 참여만 허용하며, 팀 참여는 허용하지 않습니다. 제출시에 팀 이름은 슬랙과 같이 “이름_캠퍼ID”로 구성해주세요. 팀 이름은 대회에서만 사용하는 이름입니다. **캐글 가입시 아이디는 편하신대로 해주세요!**
4. 대회 참여 후 하루에 20회까지 제출이 가능합니다.
5. 대회는 UTC 기준 2월 21일 일요일 오전 9시, 한국시간 기준 2월 21일 일요일 오후 6시까지입니다.
6. Train, Dev 데이터는 모두의 말뭉치를 전처리하여 얻어집니다. 코랩 폴더에 동봉된 preprocessing.ipynb를 참조해주세요. Test 데이터는 Upstage에서 제작하여 제공해주신 데이터이고 캐글 대회 data 섹션에서 다운로드 받으실 수 있습니다.
7. 모델 학습은 코랩 폴더에 동봉된 run.ipynb를 참조하셔서 진행해주세요.
8. 캐글 대회에서의 제출은 “submit predictions” 버튼을 통해 제출할 수 있습니다.
9. 초기에 공개되는 리더보드는 “Public Leaderboard”로서 전체 데이터의 85%로만 채점된 결과를 보여줍니다. 대회 종료 후 전체 데이터를 이용하여 채점함 “Private Leaderboard”가 공개됩니다. 대회를 진행하실 때에 일반화 성능도 놓치지말고 챙겨주세요.
10. 과제 제출 확인은 캐글 제출 완료 후 제출 화면 캡쳐본을 edwith.org의 과제 게시글에 댓글로 달아주세요. 잘 모르시겠으면 제가 댓글로 남겨놓은 제출 예시를 참조해주세요! **팀 이름이 “본인이름_캠퍼ID”로 되어있는지 꼭 확인해주세요!!!** edwith에 제출을 한 번 하고 난 이후에는 자유롭게 일요일까지 대회에 참가해 주시면 됩니다.

과제 결과 제출 예시

n submissions for “TEAM_NAME” 에서 TEAM_NAME 꼭 본인의 “이름_캠퍼ID”인지 확인해주세요!



캠퍼님들께서 앞으로 인공지능 분야에서 일하시면서 최소 한 번 이상은 캐글 포맷의 대회에 참여하실 일이 생길것이라 생각되어 해당 과제를 준비하게 되었습니다. 이번 과제를 통해 여러분들의 실력이 다른 캠퍼님들과 비교했을 때 어느정도인지도 확인해보시고, 캐글 포맷 대회에 대한 경험도 얻어가셨으면 좋겠습니다. 감사합니다!