**[ 작업 보고 ]** 20200127 */ 송이레*

## 작업 내용 요약

1. Train label 999인 행 삭제 : part1, part2 담당
2. Rough한 머신러닝 시도
   1. Multinomial model 구현 시도 : 실패, 포기.
   2. Baseline 코드 변경해 random forest model 구현 시도 : 진행 중.

## 작업 내용 설명

1. Train label 999인 행 삭제 : event time 15초 기준으로 잡아놨으므로, 그 이전 행 삭제.
2. Multinomial model 구현 시도 : tensorflow keras 사용해 layer 중첩하려 했으나, 실패.
3. Baseline 코드 변경 : data loader 사용하지 않고, 전처리 + 999인 행 삭제한 파일 기준으로 모델 구현 부분만 따다 쓰려고 진행 중.

## 작업 내용 중 의문

1. Multinomial model 구현 시도 중 궁금했던 부분
   1. Keras 패키지 사용해 layer 구성하고 softmax로 확률 도출하고 싶었다.
   2. 그런데 train\_data set을 만드는 과정에 다음과 같은 문제를 겪어서, 논의하고 싶었다.
      1. 각각의 part에 대해 values로 값을 추출하고, 585\*3273 단위의 한 id 데이터셋이 206개 혹은 207개 들어 있는 numpy array로 reshape하고 싶었다. 일단 이렇게 하면 세션이 터진다.
      2. 세션이 터지지 않았다고 가정하자. 마지막에 dense layer에 1차원 배열이 들어가야 하는데*(ex. 이미지 28\*28 형태로 픽셀 나열했던 것처럼),* 그러면 우리가 가진 데이터는 1초의 상태를 나타내는 데이터 ~ 2초의 상태를 나타내는 데이터 ~ … 이런 식으로 나열될 것이다. 시계열 데이터를 이렇게 재구성해서 학습시켜도 될까?
      3. 혹은, 1차원 배열로 만들지 않고 585\*3273 단위의 id 데이터 셋에 하나의 라벨을 할당한다고 하자. 이건 어떻게 가능할까?
   3. Label을 만드는 과정에 다음과 같은 문제를 겪었다.
      1. 각 part별로 존재하는 label의 수가 다르다. *(ex. Part1에는 103개의 label이 존재한다.)*
      2. 그러면 label을 one\_hot encoding시키더라도 각 part별로 label 배열의 길이가 다르지 않나?
   4. 이러한 생각 끝에, 결과적으로 우리가 설정한 문제, 즉, “한 id당 하나의 label을 설정해 학습”하는 것을 구현하는 것이 어렵다는 생각이 들었다. 이 부분을 논의하고 싶었다.
2. Baseline 코드 변경
   1. baseline에서 random forest 사용하며 진행한 바를 살펴 보면 시계열 데이터의 의미가 없어 보인다.
   2. 한 행당 하나의 label 설정하여 작업해도 될까?

## 논의하고 싶은 사항

1. Data loader v2 사용하지 않아도 될까?
2. 다중 레이블 분류 모델 : 대건오빠가 찾은 것.
3. 체크포인트 사용 방법 숙지.