**[카카오 아레나 데이터 경진대회 1등 노하우\_정리]**

1. 접속

* **터미널 접속 또는 vscode 접속시**

ID: ubuntu

Password: hunet1234

(**터미널**) 접속시에는

$ ssh ubuntu@14.49.45.241 -p 16022

(**putty**) 접속시에는

위에 IP, port 맞춰서 입력하고 "open" 누른시 다음에 위에 계정이랑 패스워드 입력

* **주피터노트북 접속시**

URL: http://14.49.45.241:8888/tree

password: 세계일등1# (영문자판 활성화 상태에서)

--------------------------------------------------------------------------------------------------------------

1. **상품명**

상품 식별값, 카테고리, 성별, 속성 등 정보 포함 -> 분석하면, 카테고리 분류에 도움

정제 필요 (띄어쓰기,. \_문자, )

\*cf) 상위 입상자 솔루션 : sub-word 단위로 분리하는 BPE(byte pair encoding) 활용

p.21

# **데이터 탐색 요약**

1. 소>세>중>대분류 순서로 상품 카테고리 편중
2. Product와 model의 text에 category 분류 정보가 포함
3. Product -> 비정형 텍스트 형식(문법에 맞지 않음)
4. 상품 image feature, categorical 특징 담고 있음

**2장**

쇼핑몰 상품 카테고리 분류 - 1등 솔루션

transformer기밥 endcoder 활용 for text 처리

기존, lstm -> 집필 transformer

다양한 hidden size로 양상블 -> 최고 성능

5fold 사용

Lstm, transformer : sequence data를 압축된 정보로 인코딩하는 데 사용

값이 -1인 경우, loss function 계산 시, 제외시켜야

**2.1.2 데이터 구성 확인**

데이터 개수, 예측해야하는 카테고리가 어떤 형태로 저장되어 있음?, 예측을 위해 가용한 데이터? 특이사항?

**2.1.3 ml pipeline 구현**

- Data preprocessing

1) df로 변환

Hdf5 포맷?

2) feature engineering

데이터 가공, 추가/삭제

3) 전처리 완료 데이터 csv로 저장

용량 초과로 csv저장이 안되면, hdf5로 별도 저장

P 43

**성능개선 방법**

1. 상품명 전처리 - 특수기호 제거 (2.4.1)
2. 상품명 전처리 – 학습 기반 단어 분절기 사용 (이유: 일반 문장과 다르게, 명사 나열 -> mecab 같은 일반 문장 용도의 한국어 단어 분절기는 성능이 저하될 수 있음 = 상품명으로 학습되는 분절기 사용)
3. 노이즈 심한 column 제외
4. 입력 형태에 대해 추가정보를 segment embedding으로 제공 (상품명 = 단어 나열 -> 띄어쓰기 기준 단어 순서 바뀌어도 의미 크게 변하지 않음, but ‘첫 단어는 브랜드명’과 같이 규칙이 보이기도 함 -> 단어의 위치 정보 반영 + 단어끼리 서로 독립적임 을 모델의 input으로 주기 위함) -> eda필요