

# 뉴스 기사 레이블 복구 해커톤

2023.09.22 ~ 2023.09.25

호두마르

### **Pre-Processing**



• URL 부분 제거, "short\_description" 문구 제거

Be on TOP: //www.huffingtonpost.com/entry/be-on-top-amazon-best-sel\_b\_12508618.html short\_description

URL 이라고 모두가 불필요한 정보 X, 빨간색 부분 같이 도움이 되는 정보가 있음.

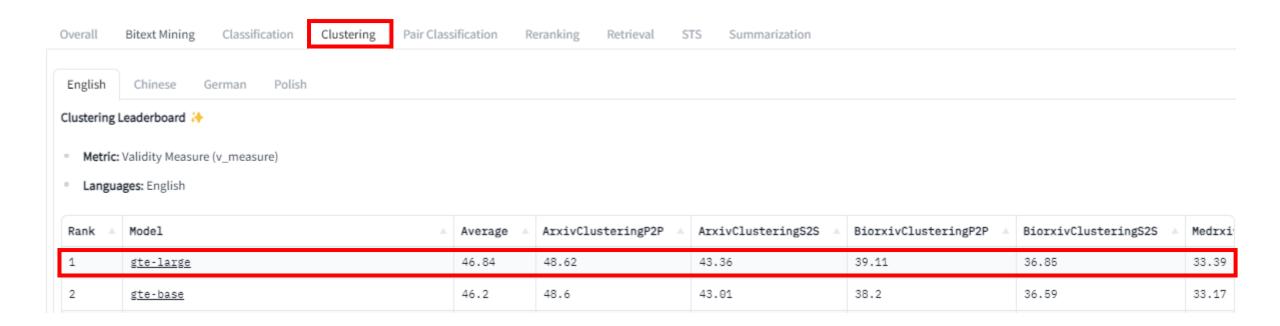
- HTML Tag 제거
- 공백 및 특수문자 제거
- 숫자 제거
- 이모지 제거
- 멘션 제거



#### **Embedding Model**



HuggingFace MTEB Leaderboard Clustering 분야 SOTA model **gte-large** 선택



#### **Post-processing**



Label Mapping

Label의 출력을 보고 mapping 하여 category pseudo labeling

```
# Sentence BERT 임베딩을 사용하여 군집화 수행
kmeans = KMeans(n_clusters=6, random_state=42)

df['kmeans_cluster'] = kmeans.fit_predict(sentence_embeddings)

df[df['kmeans_cluster'] == 3]['text'].head()

v 0.0s

Macromedia contributes to eBay Stores : Macrom...

Qualcomm plans to phone it in on cellular repa...

Thomson to Back Both Blu-ray and HD-DVD : Comp...

Thomson to Back Both Blu-ray and HD-DVD : Comp...

Thomson to Back Both Blu-ray and HD-DVD : Comp...

Sony PSP Draws Crowds and Lines on First Day (...
Name: text, dtype: object
```

```
Mapping
     # 0 busin
     # 1 enter
     # 2 politics
     # 3 sport
     # 4 tech
     # 5 world
     mapping dict = {
        0: 5,
        1: 3,
        2: 2,
         3: 4,
        4: 1,
        5: 0,
```



## Thank you