ROBOT MEDIA LABORATORY

뉴스 기사 레이블 복구 해커톤

2023.09.22 ~ 2023.09.25

호두마르

ROBOT MEDIA LABORATORY

- 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

Overall | Bitext Mining | Classification | Clustering | Pair Classification | Reranking | Retrieval | STS | Summarization

English | Chinese | German | Polish

Clustering Leaderboard ✨

- **Metric:** Validity Measure (v_measure)
- **Languages:** English

Rank	Model	Average	ArxivClusteringP2P	ArxivClusteringS2S	BiorxivClusteringP2P	BiorxivClusteringS2S	Medrxi
1	gte-large	46.84	48.62	43.36	39.11	36.85	33.39
2	gte-base	46.2	48.6	43.01	38.2	36.59	33.17

ROBOT MEDIA LABORATORY

- 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()
```

✓ 0.0s

```
3     Macromedia contributes to eBay Stores : Macrom...
4     Qualcomm plans to phone it in on cellular repa...
5     Thomson to Back Both Blu-ray and HD-DVD : Comp...
23    FTC Files First Lawsuit Against Spyware Concer...
31    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,
}
```

✓ 0.0s

ROBOT MEDIA LABORATORY

Thank you