월간 데이콘 쇼츠 - 뉴스 기사 레이블 복구 해커톤

Public/Private 6th
Gidometa

목차

Pre-Processing

1. URL 제거
2. 해시태그 제거
3. 멘션 제거
4. 이모지 제거
5. 공백 및 특수문자 제거
6. 숫자 제거
7. Html 태그 제거
8. Unicode 제거
9. MinMaxScaler 사용

위 방법 중 일부만 사용

```python
def preprocess_text(text):
    # URL 제거
    text = re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILINE)

    # 해시태그 제거
    text = re.sub(r'#\w+', '', text)

    # 멘션 제거
    text = re.sub(r'@\w+', '', text)

    # 이모지 제거
    text = text.encode('ascii', 'ignore').decode('ascii')

    # 공백 및 특수문자 제거
    text = re.sub(r'\s+', ' ', text).strip()

    # 숫자 제거
    text = re.sub(r'\d+', '', text)

    return text.lower()


df['processed_text'] = df['text'].apply(preprocess_text)
```

Embedding

Hugging Face 의 sentence-transformers 에서 나온 모델들 사용

sentence-transformers/all-distilroberta-v1

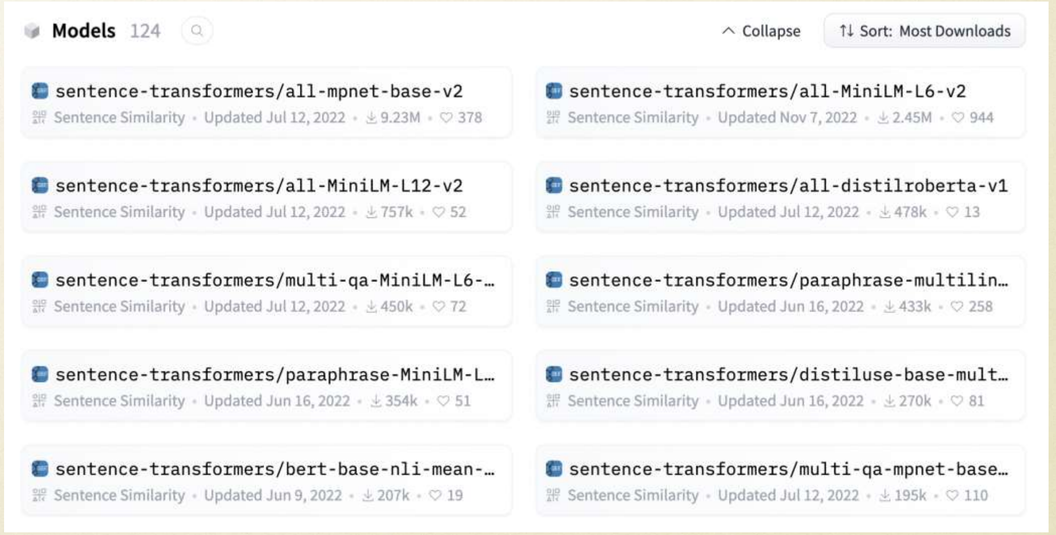
sentence-transformers/all-mpnet-base-v2

sentence-transformers/all-MiniLM-L6-v2

sentence-transformers/all-MiniLM-L12-v2

sentence-transformers/paraphrase-MiniLM-L6-v2

sentence-transformers/paraphrase-MiniLM-L12-v2



https://huggingface.co/sentence-transformers?sort_models=downloads#models

Clustering

Sklearn 에서 제공하는 여러 알고리즘 사용

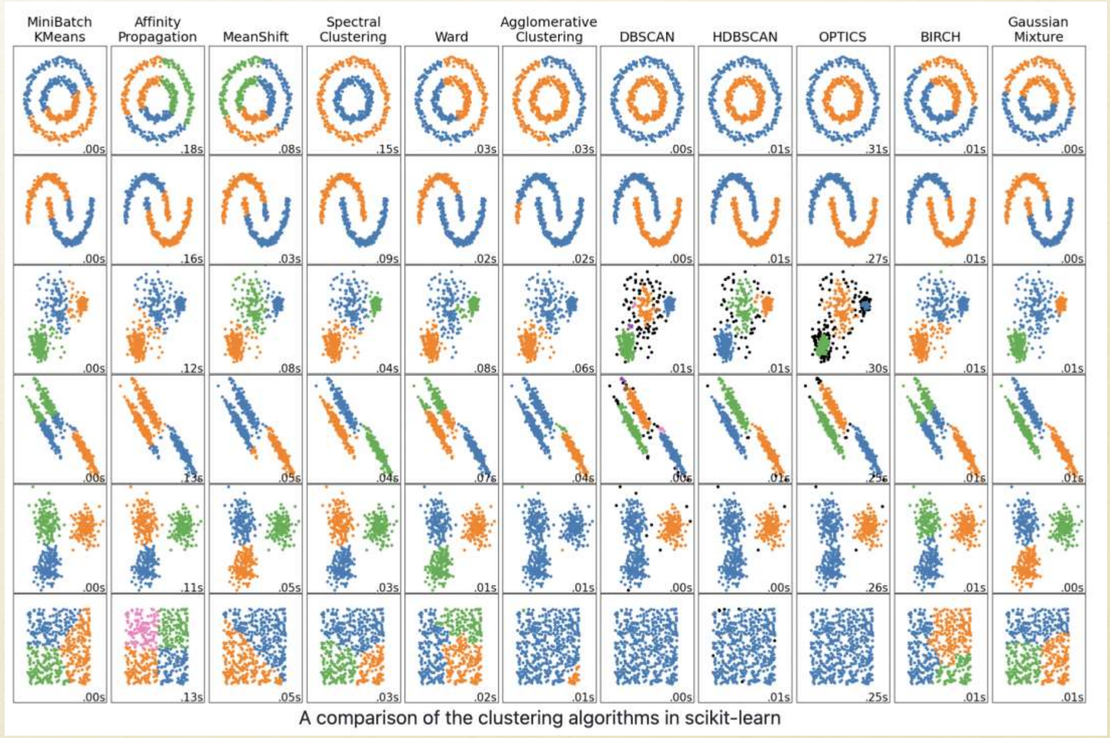
KMeans

Birch

AgglomerativeClustering

BisectingKMeans

GaussianMixture



https://scikit-learn.org/stable/modules/clustering.html

Post-Processing

각 Clustering label 을 일부 출력하여 label mapping

```
 0 ----------------------------------------
Time to Talk Baseball : It's time to talk about the serious risks and potential benefits of building an expensive
Obama Administration Helps Wall Street Criminals Dodge Accountability authors : Obama administration proposal woul
Congress Spikes Handout For Private Equity authors : A few Wall Street firms almost won big.
----------------------------------------
 1 ----------------------------------------
Spanish coach facing action in race row : MADRID (AFP) - Spanish national team coach Luis Aragones faces a formal
GAME DAY PREVIEW Game time: 6:00 PM : CHARLOTTE, North Carolina (Ticker) -- The Detroit Shock face a critical road
Deere's Color Is Green : With big tractors, big sales, and big earnings, Deere's hoeing a profitable row.
----------------------------------------
 2 ----------------------------------------
Bump Stock Maker Resumes Sales One Month After Las Vegas Mass Shooting authors : Move along nothing to see here.
Obama Marks Anniversary Of 9/11 Attacks With Moment Of Silence authors : We stand as strong as ever.
Republican Congressman Says Trump Should Apologize To Obama And The UK authors : Best not to hold your breath on t
----------------------------------------
 3 ----------------------------------------
Fischer's Fiancee: Marriage Plans Genuine (AP) : AP - Former chess champion Bobby Fischer's announcement thathe is
Israel Kills 3 Palestinians in Big Gaza Incursion (Reuters) : Reuters - Israeli forces killed three\Palestinians,
The Folly of the Sole Superpower Writ Small authors : Think of this as a little imperial folly up -- and here's th
----------------------------------------
 4 ----------------------------------------
Macromedia contributes to eBay Stores : Macromedia has announced a special version of its Contribute website editi
Qualcomm plans to phone it in on cellular repairs : Over-the-air fixes for cell phones comes to Qualcomm's CDMA.
Thomson to Back Both Blu-ray and HD-DVD : Company, one of the core backers of Blu-ray, will also support its rival
----------------------------------------
 5 ----------------------------------------
Bruce Lee statue for divided city : In Bosnia, where one man #39;s hero is often another man #39;s villain, some c
Only Lovers Left Alive's Tilda Swinton Talks About Almost Quitting Acting and Yasmine Hamdan Performs 'Hal' Live I
Harry #39;s argy-bargy : PRINCE Charles has asked Scotland Yard for an in-depth report on his son Harry #39;s trip
----------------------------------------
```

```python
mapping_dict = {
    0: 0,
    1: 3,
    2: 2,
    3: 5,
    4: 4,
    5: 1
}
```

```python
df['mapping'] = df['agglomerative_cluster2'].apply(lambda x: mapping_dict[x])
sample = pd.read_csv(os.path.join(FILE_PATH ,'sample_submission.csv'))
sample['category'] = df['mapping'].values
sample['category'].head()
sample.to_csv(os.path.join(FILE_PATH, 'st_submit_agglomerative.csv'), index=False)
```

Post-Processing

Ensemble

```python
# 각 파일에서 라벨 추출
all_labels = []
for file_path in file_paths:
    df = pd.read_csv(os.path.join(FOLDER_PATH, file_path))
    labels = df['category'].tolist()
    all_labels.append(labels)

# 각 라벨 리스트에서 가장 빈번한 라벨을 선택 (다수결 투표)
final_labels = []
for i in range(len(all_labels[0])):
    label_votes = [labels[i] for labels in all_labels]
    majority_label = Counter(label_votes).most_common(1)[0][0]
    final_labels.append(majority_label)

# 결과를 CSV 파일로 저장
sample = pd.read_csv(os.path.join(SAVE_PATH ,'sample_submission.csv'))
sample['category'] = final_labels
```

EOF

감사합니다.