Clustering

214683 장인환

목차

1. 모델링 배경 및 목적

뉴스 기사를 자동 분류를 목적으로 모델링 시작

일반적인 분류할 수 있는 데이터셋이 아닌 텍스트를 기반으로 분류 되는 모델을 만들어 보고 싶어 해당 데이터셋을 선택

동계 근로장학중 홈페이지 분류 작업을 맡은 경험을 바탕으로 이번 모델링이 다음에는 학습 모델을 통해 분류 작업에 활용 목적

2. 데이터 설명

2225개의 데이터(뉴스 텍스트) 구성

각 뉴스에는 정수형 레이블이 동시에 존재

정치 = 0
스포츠 = 1
기술 = 2
엔터테인먼트 = 3
비즈니스 = 4

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2225 entries, 0 to 2224
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   Text    2225 non-null   object
 1   Label   2225 non-null   int64 
dtypes: int64(1), object(1)
memory usage: 34.9+ KB
```

3. 모델링 과정

K-means

1. 필요 라이브러리 호출

```python
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
```

2. 데이터 처리

```python
df_clustering = pd.read_csv("C:/Users/user/Desktop/수업관련/3학년 2025/1학기/인공지능
df_clustering.info()

# 텍스트 벡터화
vectorizer = TfidfVectorizer(stop_words='english')
X = vectorizer.fit_transform(df_clustering['Text'])
```

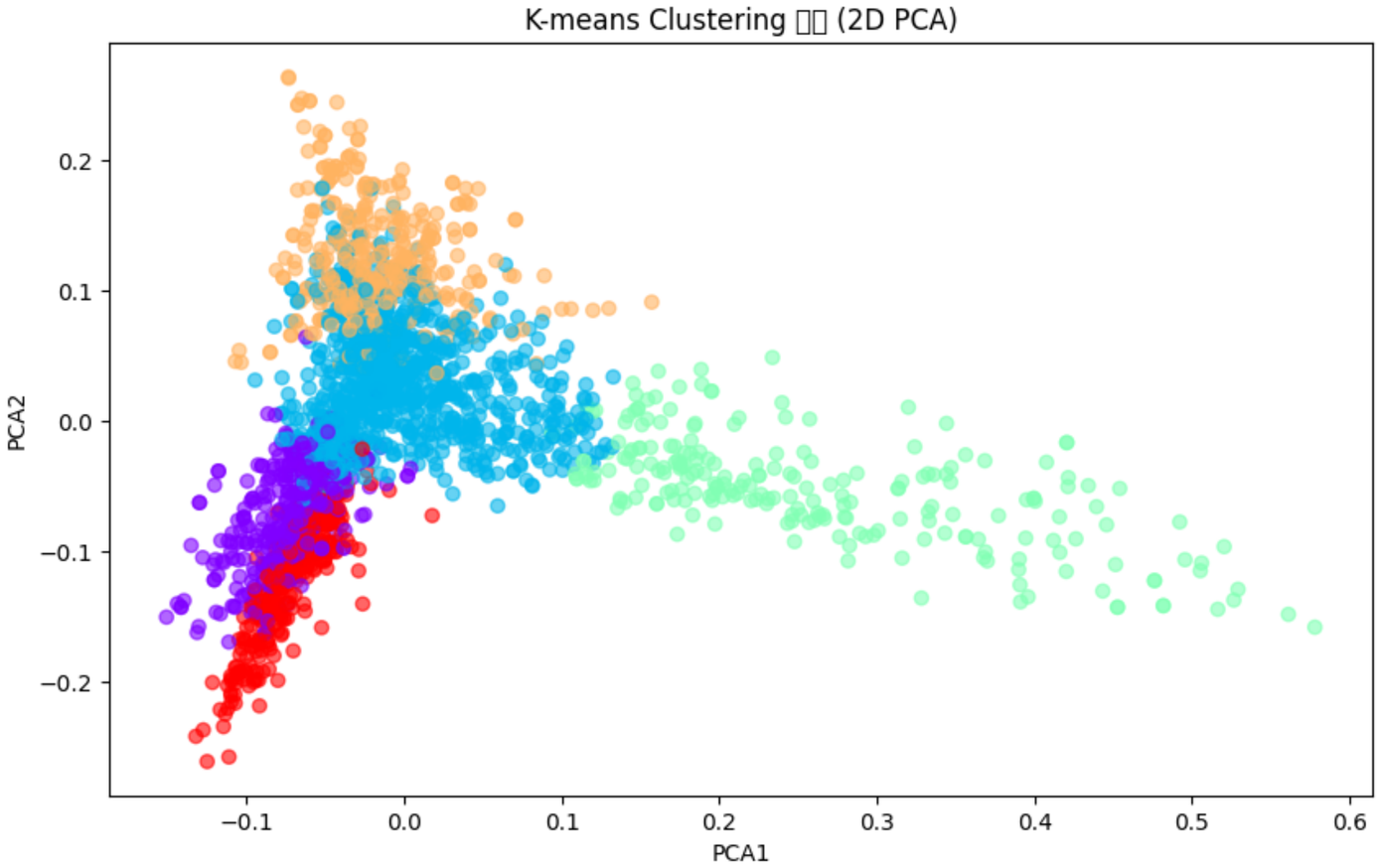
3. K-means 학습

```python
k = 5  # 정치 = 0, 스포츠 = 1, 기술 = 2, 엔터테인먼트 = 3, 비즈니스 = 4
kmeans = KMeans(n_clusters=k, random_state=42)
kmeans.fit(X)
df_clustering['Cluster_KMeans'] = kmeans.labels_

# 차원 축소
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X.toarray())
```

4. 시각화

```python
# 시각화
plt.figure(figsize=(10, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=kmeans.labels_, cmap='rainbow', alpha=0.6)
plt.title("K-means Clustering 결과 (2D PCA)")
plt.xlabel("PCA1")
plt.ylabel("PCA2")
plt.show()
```

4. 결과
K-means



빨간색 : 클러스터 0번, 정치
파란색 : 클러스터 1번, 스포츠
보라색 : 클러스터 2번, 기술
주황색 : 클러스터 3번, 엔터테인먼트
연두색 : 클러스터 4번, 비즈니스

가로 세로축은 텍스트라는고차원을 축소시켜 유사한 기사들은 가깝도록 만들어줌

3. 모델링 과정

Hierarchical

1. 필요 라이브러리 호출

```
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from scipy.cluster.hierarchy import linkage, dendrogram
from sklearn.metrics.pairwise import cosine_distances
import matplotlib.pyplot as plt
```

2. 데이터 처리

```
df_clustering = pd.read_csv("C:/Users/user/Desktop/수업관련/3학년 2025/1학기/인공
df_clustering.info()

# 벡터 생성
vectorizer = TfidfVectorizer(stop_words='english')
X = vectorizer.fit_transform(df_clustering['Text'])
```

3. K-means 학습

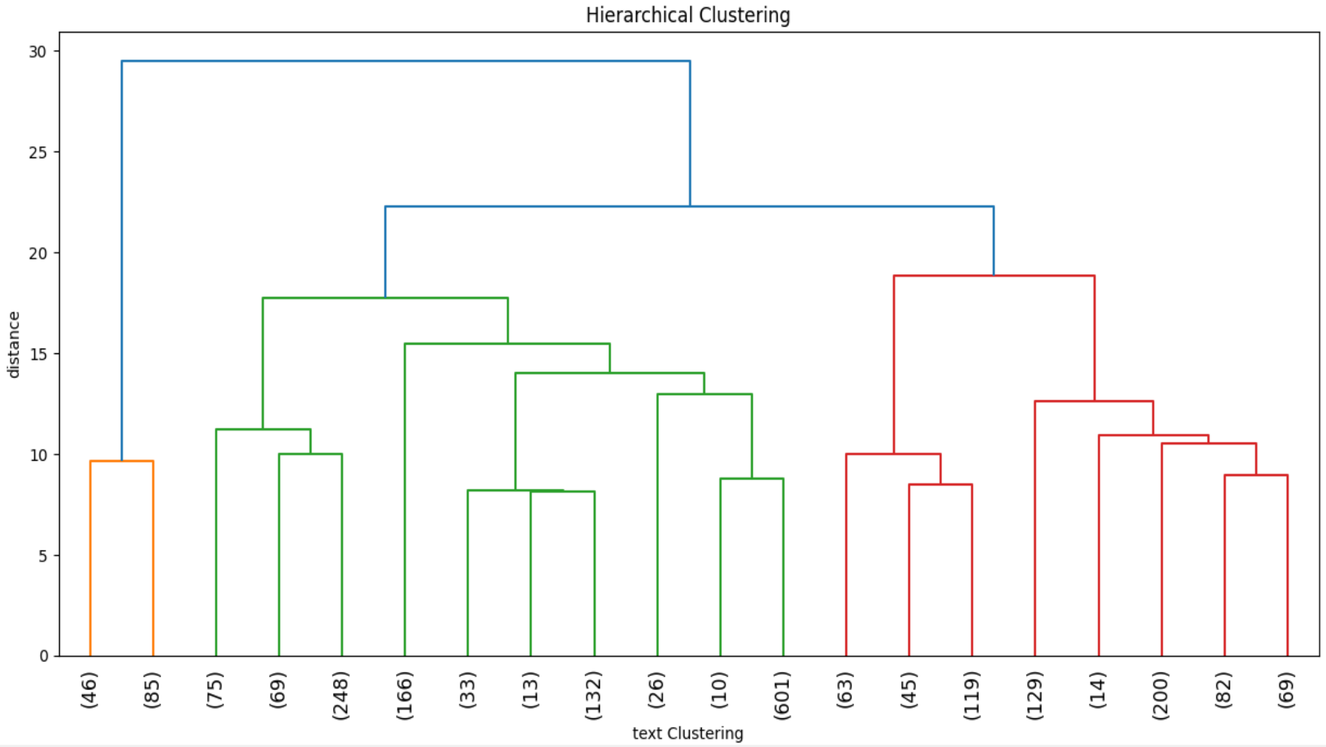
```
# 거리 행렬 계산 (코사인 거리)
distance_matrix = cosine_distances(X)

# 계층적 클러스터링
linkage_matrix = linkage(distance_matrix, method='ward')
```

4. 시각화

```
plt.figure(figsize=(15, 6))
dendrogram(linkage_matrix, truncate_mode='lastp', p=20, leaf_rotation=90., leaf
plt.title("Hierarchical Clustering")
plt.xlabel("text Clustering")
plt.ylabel("distance")
plt.show()
```

4. 결과

Hierarchical



2225개를 모두 표현하기 힘들어 일부분만 출력

가로축의 숫자들은 모두 인덱스 번호, 즉 뉴스 텍스트
세로축은 유사하지 않은 정도로, 낮은 값일 경우 유사하다는 의미

색깔은 그룹으로, 유사 내용을 표현함