資料探勘HW4

AI三 B1228005 胡樂麒

1. 資料分析

本作業使用 Kaggle：News Aggregator Dataset

- 共 500 筆新聞標題（TITLE）

- 新聞類別：CATEGORY（如 business、science、technology 等）

```
Documents loaded: 500
Cluster 0 contains 20 samples
Cluster 1 contains 24 samples
Cluster 2 contains 52 samples
Cluster 3 contains 195 samples
Cluster 4 contains 22 samples
Cluster 5 contains 25 samples
Cluster 6 contains 34 samples
Cluster 7 contains 74 samples
Cluster 8 contains 27 samples
Cluster 9 contains 27 samples
Cluster 0 contains 53 samples
 Most important terms
 1) ebay (score: 0.1456)
 2) icahn (score: 0.1452)
 3) north (score: 0.1315)
 4) metro (score: 0.1289)
 5) killed (score: 0.1228)
Cluster 1 contains 246 samples
 Most important terms
 1) the (score: 0.0414)
 2) to (score: 0.0372)
 3) mcdonald (score: 0.0354)
 4) market (score: 0.0346)
 5) sales (score: 0.0312)
Cluster 2 contains 45 samples
 Most important terms
 1) transit (score: 0.2244)
 2) public (score: 0.2218)
 3) americans (score: 0.1310)
 4) record (score: 0.1209)
 5) in (score: 0.1161)
Cluster 3 contains 48 samples
 Most important terms
 1) stocks (score: 0.1946)
 2) on (score: 0.1507)
 3) asia (score: 0.1254)
 4) lower (score: 0.1167)
 5) news (score: 0.1099)
Cluster 4 contains 22 samples
 Most important terms
 1) jetblue (score: 0.3169)
 2) american (score: 0.2910)
 3) airlines (score: 0.2792)
 4) agreement (score: 0.2247)
 5) end (score: 0.2005)
Cluster 5 contains 86 samples
 Most important terms
 1) gox (score: 0.1856)
 2) mt (score: 0.1856)
 3) bitcoin (score: 0.0942)
 4) bankruptcy (score: 0.0883)
 5) hackers (score: 0.0830)
```

---

2. 文字前處理

使用 TfidfVectorizer(max_df=0.4) 進行向量化，包含：

- 小寫化(文字轉換成詞彙前，先把所有字詞統一轉成小寫)

- 停用詞處理(移除無意義標點符號與無效 token)
- TF-IDF 權重(TF:「某個詞在某篇文件中出現多不多」 IDF:「這個詞是否只在少數文章中出現」)
- 最大詞頻限制 max_df=0.4(某個詞在超過 40% 的文章中都出現，這個詞會被直接去掉)
- 使用 unigram (1-gram)(把每個單字視為一個 token)

Online clustering 使用的是 HashingVectorizer。

---

3. Clustering Algorithms 三種分群方法

3.1 KMeans

KMeans 設KMeans(n_clusters=6)

3.2 EAC (Ensemble Agglomerative Clustering)

使用：

- n_clusterings=20,多次 KMeans 20 次
- 共現矩陣 (Co-association matrix)

- linkage(average) 群聚層次法
- final_k=6,最後分成 6 群

3.3 Online MiniBatchKMeans（Online Learning）

Online clustering 設計：

- HashingVectorizer → 使用 hash 函數把字詞映射到固定維度空間。
- MiniBatchKMeans(n_clusters=10)
- 每批 batch_size=10
- partial_fit（分成多次少量進行訓練）

---

4. 評估指標

使用兩種評估方法：

(1) Intrinsic — Silhouette Score

- 衡量分群的內部品質
- 越高越好（-1 ~ 1）

(2) Extrinsic — BCubed Precision / Recall

- 使用 “ CATEGORY ” 當真實類別
- Precision 高 → 每群不混雜
- Recall 高 → 同類不被切太散

---

5. 執行結果

5.1 分群數量

方法	群數
KMeans	6
EAC	6
Online MiniBatchKMeans	10

5.2 評估指標分數

KMeans（k=6）

- Silhouette：0.6024

- BCubed Precision：0.9960
- BCubed Recall：0.3028

→ 分群乾淨，但把同類新聞切得較散（Recall 較低）

```
=== Number of Clusters (KMeans) ===
KMeans clusters: 6

=== KMeans Evaluation ===
Silhouette Score: 0.602417217934363
BCubed Precision: 0.996016260162603
BCubed Recall: 0.3027735470941877

Top 6 topic clusters: [1, 5, 0, 3, 2, 4]
```

EAC（final_k=6）

- Silhouette：0.4211
- BCubed Precision：0.9960
- BCubed Recall：0.3157

→ Precision 高且 Recall 相比Kmeans更好一點

```
=== Number of Clusters (EAC) ===
EAC clusters: 6

=== EAC Evaluation ===
Silhouette: 0.42110100470547207
BCubed Precision: 0.9960157480314965
BCubed Recall: 0.315663326653309
```

Online MiniBatchKMeans (k=10)

- Silhouette：0.0093
- BCubed Precision：0.9960
- BCubed Recall：0.5133

→ 因為使用 HashingVectorizer，Silhouette 很低
→ 但把同類新聞分得更細，因此 Recall 最高（0.513）

```
=== Number of Clusters (Online Clustering) ===
Online clusters: 10

=== Online Clustering Evaluation ===
Silhouette Score: 0.009333913250835588
BCubed Precision: 0.9960114942528794
BCubed Recall: 0.5132665330661345
```

---

6. Top 6 Topics（作業 b）

6.1 KMeans — Top Six Topics

Cluster	Keywords
1	the, to, mcdonald, market, sales, bull, in, bitcoin
5	gox, mt, bitcoin, bankruptcy, hackers, files, exchange, ceo
0	ebay, icahn, north, metro, killed, worker, train, by
3	stocks, on, asia, lower, news, data, from, us
2	transit, public, americans, record, in, numbers, ridership, highest
4	jetblue, american, airlines, agreement, end, interline, ticketing

```
Top 6 topic clusters: [1, 5, 0, 3, 2, 4]

===== Six Main Topics =====

Topic Cluster 1:
Keywords: the, to, mcdonald, market, sales, bull, in, bitcoin

Topic Cluster 5:
Keywords: gox, mt, bitcoin, bankruptcy, hackers, files, exchange, ceo

Topic Cluster 0:
Keywords: ebay, icahn, north, metro, killed, worker, train, by

Topic Cluster 3:
Keywords: stocks, on, asia, lower, news, data, from, us

Topic Cluster 2:
Keywords: transit, public, americans, record, in, numbers, ridership, highest

Topic Cluster 4:
Keywords: jetblue, american, airlines, agreement, end, interline, and, ticketing
```

6.2 EAC — Top Six Topics

Cluster	Keywords
5	the, to, mcdonald, market, sales, american, bitcoin, jetblue
0	gox, mt, bankruptcy, bitcoin, files, hackers, for, us
2	stocks, on, data, lower, china, asia, weak, after
1	public, transit, americans, record, in, numbers, ridership, highest
4	ebay, icahn, nominees, board, rejects, carl, shareholders, asks
3	north, metro, killed, worker, train, by, struck, harlem

```
===== EAC — Six Main Topics =====
Top 6 EAC clusters: [5, 0, 2, 1, 4, 3]

EAC Cluster 5:
Keywords: the, to, mcdonald, market, sales, american, bitcoin, jetblue

EAC Cluster 0:
Keywords: gox, mt, bankruptcy, bitcoin, files, hackers, for, us

EAC Cluster 2:
Keywords: stocks, on, data, lower, china, asia, weak, after

EAC Cluster 1:
Keywords: public, transit, americans, record, in, numbers, ridership, highest

EAC Cluster 4:
Keywords: ebay, icahn, nominees, board, rejects, carl, shareholders, asks

EAC Cluster 3:
Keywords: north, metro, killed, worker, train, by, struck, harlem
```

6.3 Online Clustering — Top Six Topics（依照群數前六名）

Cluster	Keywords
0	on, in, stocks, public, transit, the, market, record
5	to, ecb, mcdonald, euro, sales, ebay, nominees, for
8	mt, gox, hackers, ceo, bitcoin, blog, claim, bitcoins
4	bankruptcy, files, mt, gox, for, us, protection, exchange
7	failing, eu, banks, aims, tackling, deal, for, focus
3	unwarranted, creates, says, pressure, noyer, euro, strong, economic

---

7. 三種分群方法比較

KMeans

- 三種方法Precision幾乎一樣，Precision 皆為0.996
- Silhouette 最高(群與群之間距離明顯、邊界清楚)
- 但 Recall 最低 → 代表語意相近的標題容易被切散

➡ 適合：快速、固定群數的情況

EAC

用多個子模型一起表決，把多次聚類中一致出現的關係保留，不一致的排除，讓最終群組更加準確。

- Silhouette 次高
- BCubed Recall 高於 KMeans
- 反映 ensemble 的分群較穩定的特色

➡ 適合：融合多個模型、提升穩定性

Online MiniBatchKMeans

- k=10 群
- 使用 HashingVectorizer 導致 Silhouette 極低(群與群之間界線模糊、不易形成明確主題)
- 但 Recall 最高 (0.513)(語意相近的文章最不容易被拆散)

➡ 適合：大量資料、即時學習

---

8. 結論

1. 三種方法的主題一致性高：股票、美股、比特幣、航空、地鐵意外等主題都有被識別。

2. KMeans 結果最穩定、類群最乾淨。

3. EAC 有提升 Recall 的趨向，整體分群更平衡。

4. Online clustering 切得更細，雖然 Silhouette 較低，但仍能有效捕捉新聞主題。