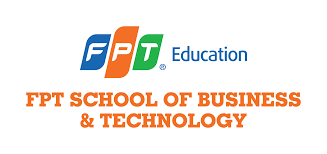
VIỆN QUẢN TRỊ & CÔNG NGHỆ FSB



BÀI TỔNG KẾT MÔN HỌC – CLUSTERING DATA

Giảng viên hướng dẫn: Quản Thành Thơ

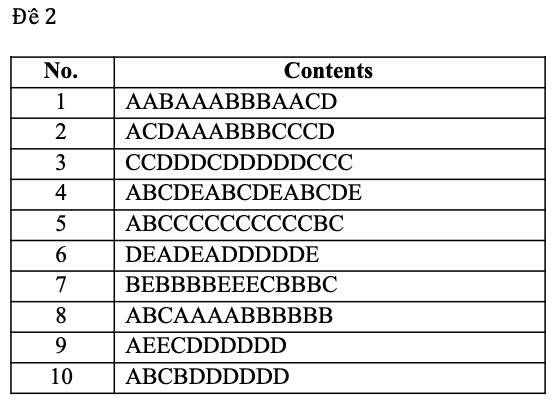
Tác giả: Nguyễn Hoài Phương

Lớp: MSE09

# **1 – Giới thiệu bài toán**

Phân cụm tập dữ liệu cho trước bằng thuật toán k-means và HAC.

**2 – Tập dữ liệu**

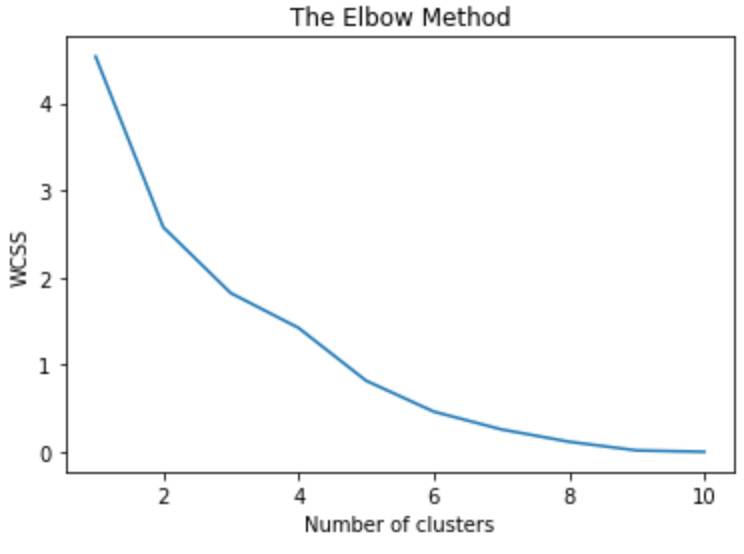


# 3 – Các bước thực hiện

* Đầu tiên, chúng ta phải vectorize tập dữ liệu đã cho thành các vector. Tfidf được sử dụng để vectorize tập dữ liệu trên
* Tiếp theo, sử dụng các thuật toán phân cụm như k-means và HAC (với nhiều phương thức gom cụm khác nhau) để phân cụm cho tập dữ liệu đã cho
* Đánh giá kết quả

# 4 – Sử dụng k-means

## 4.1 – Lựa chọn số cụm tối ưu



Từ đồ thị elbow ở trên, chúng ta có thể nhận ra n\_clusters = 5 hoặc n\_clusters = 6 là lựa chọn tốt nhất.

## 4.2 – Số lượng cluster bằng 5

Cluster 1:

No 3 : CCDDDCDDDDDCCC

No 10 : ABCBDDDDDD

Cluster 2:

No 1 : AABAAABBBAACD

No 2 : ACDAAABBBCCCD

No 8 : ABCAAAABBBBBB

Cluster 3:

No 4 : ABCDEABCDEABCDE

No 7 : BEBBBBEEECBBBC

Cluster 4:

No 5 : ABCCCCCCCCCCBC

Cluster 5:

No 6 : DEADEADDDDDE

No 9 : AEECDDDDDD

silhouette\_score : 0.29770943824670415

## **4.2 – Số lượng cluster bằng 6**

Cluster 1:

No 6 : DEADEADDDDDE

No 9 : AEECDDDDDD

Cluster 2:

No 1 : AABAAABBBAACD

No 2 : ACDAAABBBCCCD

No 8 : ABCAAAABBBBBB

Cluster 3:

No 5 : ABCCCCCCCCCCBC

Cluster 4:

No 3 : CCDDDCDDDDDCCC

No 10 : ABCBDDDDDD

Cluster 5:

No 7 : BEBBBBEEECBBBC

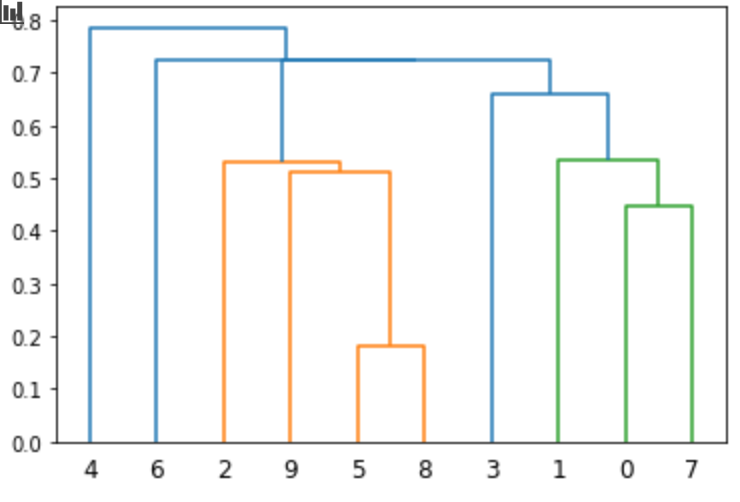
Cluster 6:

No 4 : ABCDEABCDEABCDE

silhouette\_score : 0.24713976988796443

# **5 – Sử dụng HAC**

## **5.1 – Gom cụm bằng phương pháp Single Link**

****

Từ biểu đồ dendrogram, n\_clusters = 9 là lựa chọn tối ưu.

Cluster 1:

No 6 : DEADEADDDDDE

No 9 : AEECDDDDDD

Cluster 2:

No 8 : ABCAAAABBBBBB

Cluster 3:

No 10 : ABCBDDDDDD

Cluster 4:

No 5 : ABCCCCCCCCCCBC

Cluster 5:

No 4 : ABCDEABCDEABCDE

Cluster 6:

No 7 : BEBBBBEEECBBBC

Cluster 7:

No 3 : CCDDDCDDDDDCCC

Cluster 8:

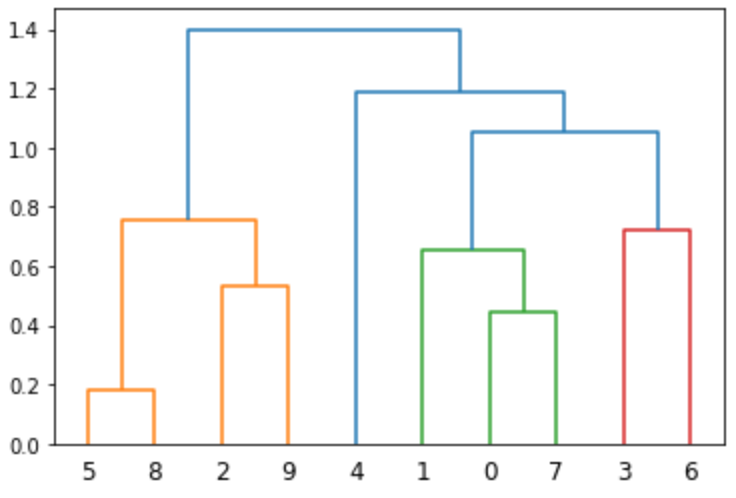
No 2 : ACDAAABBBCCCD

Cluster 9:

No 1 : AABAAABBBAACD

silhouette\_score : 0.13514177939197886

## **5.2 – Gom cụm bằng phương pháp Complete Link**

****

## Từ đồ thị dendrogram, n\_clusters = 4 sẽ là lựa chọn tốt nhất.

Cluster 1:

No 3 : CCDDDCDDDDDCCC

No 6 : DEADEADDDDDE

No 9 : AEECDDDDDD

No 10 : ABCBDDDDDD

Cluster 2:

No 4 : ABCDEABCDEABCDE

No 7 : BEBBBBEEECBBBC

Cluster 3:

No 5 : ABCCCCCCCCCCBC

Cluster 4:

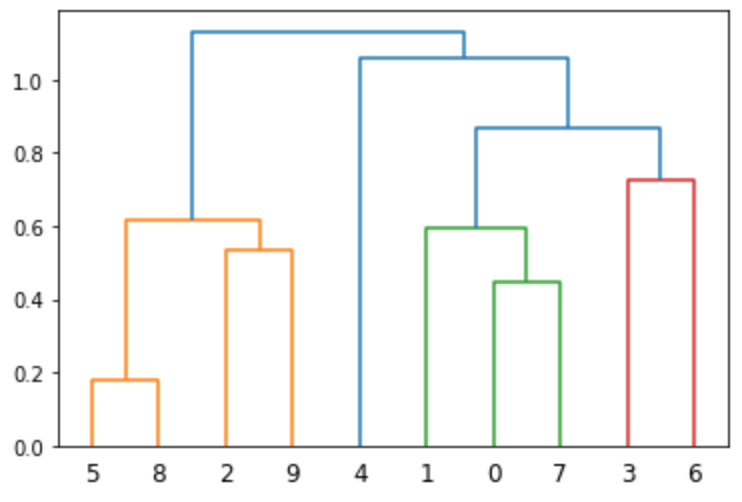
No 1 : AABAAABBBAACD

No 2 : ACDAAABBBCCCD

No 8 : ABCAAAABBBBBB

silhouette\_score : 0.3250339344884349

## **5.3 – Gom cụm bằng phương pháp Average**



## Từ đồ thị dendrogram, n\_clusters = 9 là tối ưu.

Cluster 1:

No 6 : DEADEADDDDDE

No 9 : AEECDDDDDD

Cluster 2:

No 10 : ABCBDDDDDD

Cluster 3:

No 8 : ABCAAAABBBBBB

Cluster 4:

No 7 : BEBBBBEEECBBBC

Cluster 5:

No 4 : ABCDEABCDEABCDE

Cluster 6:

No 5 : ABCCCCCCCCCCBC

Cluster 7:

No 2 : ACDAAABBBCCCD

Cluster 8:

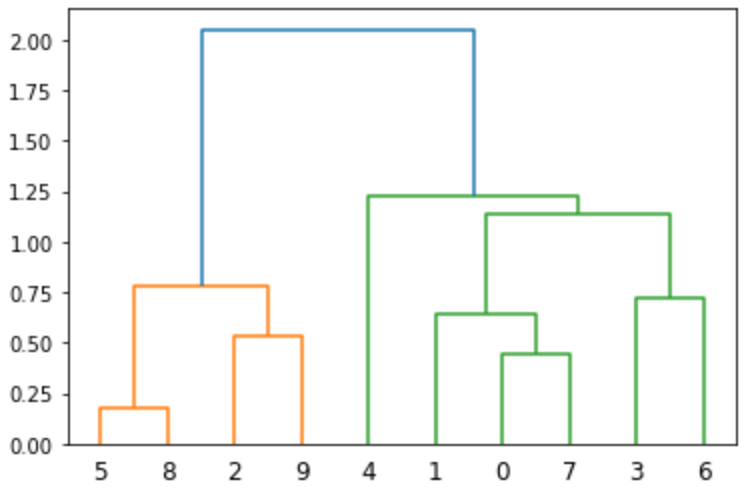
No 3 : CCDDDCDDDDDCCC

Cluster 9:

No 1 : AABAAABBBAACD

silhouette\_score : 0.13514177939197886

## **5.4 – Gom cụm bằng phương pháp Ward**



## Từ đồ thị dendrogram, n\_clusters = 2 là tối ưu.

Cluster 1:

No 1 : AABAAABBBAACD

No 2 : ACDAAABBBCCCD

No 4 : ABCDEABCDEABCDE

No 5 : ABCCCCCCCCCCBC

No 7 : BEBBBBEEECBBBC

No 8 : ABCAAAABBBBBB

Cluster 2:

No 3 : CCDDDCDDDDDCCC

No 6 : DEADEADDDDDE

No 9 : AEECDDDDDD

No 10 : ABCBDDDDDD

silhouette\_score : 0.3487772838410113

# 6 – Kết luận

Với phương pháp gom cụm, chúng ta chỉ có thể so sánh các kết quả có cùng số cụm, xem thử kết quả nào tốt hơn kết quả nào dựa trên silhouette chứ không thể so sánh các kết quả có số cụm khác nhau.

Nếu tập dư liệu là “2D” hoặc “3D”, chúng ta có thể vẽ ra phân bố của các điểm trong không gian 2 – 3 chiều và dễ dàng kiểm chứng kết quả phân cụm có hợp lý không. Tuy nhiên, khi dữ liệu là n chiều với n > 3, thật sự khó kiểm chứng sự hợp lý của bài toán phân cụm.