**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

**BÀI TẬP THỰC HÀNH**

**MÔN HỌC**

**MÁY HỌC TRONG THỊ GIÁC MÁY TÍNH**

**CLUSTERING**

**GV: TS. LÊ ĐÌNH DUY**

**SVTH: Trần Quốc Long**

**TP. HCM, ngày 19 tháng 10 năm 2017**

***NHÓM NGHIÊN CỨU BIG DATA & DEEP LEARNING, UIT***

**Mục tiêu:**

1. Bài toán clustering.

Bài toán clustering là bài toán mà ở đó, từ một tập dữ liệu đầu vào, ta tìm được cách phân chia dữ liệu thành các nhóm dữ liệu giống nhau (liên quan với nhau) về một mặt nào đó.

Ở đây, mức độ liên quan được đánh giá trên các độ đo tương đồng trong không gian vector, như Eucledian, Cosine.

Mỗi phần tử trong tập dữ liệu được biểu diễn dưới dạng một vector n chiều. Từ đó, ta có thể ước lượng được độ liên quan giữa các vector.

1. Cách đánh giá (độ đo) của một thuật toán clustering.
2. Các hàm thư viện của scikit-learn cho bài toán clustering

cho các thuật toán như Kmeans clustering, Spectral clustering, DBSCAN, Agglomerative clustering.

1. Áp dụng một số thuật toán clustering trên các tập dữ liệu:
2. So sánh và phân tích điểm mạnh/yếu của các thuật toán clustering.

References:

* 1. <https://machinelearningcoban.com/2017/01/01/kmeans/>