**Đề thi:**

# MACHINE LEARNING WITH PYTHON

**Thời hạn nộp bài:**

\*\*\* HV tạo 1 thư mục **LDS6\_k267\_HoVaTen\_Cuoi\_ky** trong thư mục LDS6\_k264\_HoVaTen trên Google Drive đã share, lưu tất cả bài làm vào để chấm điểm \*\*\*

\*\*\* HV sẽ bị trừ điểm nếu bài làm giống nhau \*\*\*

\*\*\* HV phải gửi mail đính kèm link của thư mục **LDS6\_k264\_HoVaTen\_Cuoi\_ky** đúng hạn nộp bài, sau hạn nộp bài nếu HV không gửi thì sẽ không được chấm điểm \*\*\* **Chú ý, với mỗi câu :**

* HV cần kiểm tra xem dữ liệu đã sạch, chuẩn và dùng được hay chưa, nếu chưa thì cần tiền xử lý dữ liệu trước khi làm bài.
* Lần lượt thực hiện các bước làm bài theo quy trình đã được hướng dẫn làm demo/bài tập trong lớp.
* Mỗi câu là một file viết trên jupyter notebook, các yêu cầu nhận xét kết quả trong từng câu được viết trong cell dưới định dạng Markdown.
* Mỗi câu đều phải đưa ra nhận xét, giải pháp cho các lựa chọn.
* Câu nào có phần trực quan hóa kết quả thì vừa phải trực quan vừa phải giải thích.

## 1. Clustering

* Cho dữ liệu **bbc** (giải nén file bbc-fulltext). Đọc tất cả dữ liệu vào dataframe **news** theo định dạng sau:

|  |  |  |
| --- | --- | --- |
| id (số thự tự tự động) | **content** (nội dung của tin tức trong từng tập tin) | class (một trong 5 loại : business, entertainment, politics, sport, tech) – là tên của folder chứa tin tức |
| 0 | … | … |
| … | … | … |

* Yêu cầu : Chuẩn hóa dữ liệu (nếu cần) và **chọn một thuật toán thích hợp** để thực hiện việc **phân cụm dữ liệu** dựa trên cột **content** của dataframe **news** trên.
  1. Áp dụng thuật toán thích hợp. Dựa trên cơ sở nào để cho rằng thuật toán này thích hợp?
  2. Tìm kết quả => có bao nhiêu cụm => mẫu (tin tức) nào thuộc cụm nào?
  3. Nhận xét trên từng cụm: mỗi cụm có những keywords nào ? Vẽ WordClouds cho từng cụm.
  4. Kiểm chứng lại với **class** đã có

## 2.