Các bạn chọn 1 trong 2 task sau:

1. **Natural Language Processing task:**

Dưới đây là tổng hợp của 8 bộ dataset dành cho bài toán text classification. Mỗi bộ khi các bạn download về đều có 1 file mô tả ở bên trong. Tổng quát về các bộ dataset như sau:

Dataset	Classes	Train samples	Test samples
AG's News	4	120 000	7 600
Sogou News	5	450 000	60 000
DBPedia	14	560 000	70 000
Yelp Review Polarity	2	560 000	38 000
Yelp Review Full	5	650 000	50 000
Yahoo! Answers	10	1 400 000	60 000
Amazon Review Full	5	3 000 000	650 000
Amazon Review Polarity	2	3 600 000	400 000

Các bạn hãy chọn 1 trong số các bộ dữ liệu này

2. **Computer Vision task**

Các bạn hãy chọn 1 trong số 2 bộ dữ liệu sau:
CIFAR 10: https://www.cs.toronto.edu/~kriz/cifar.html
Fashion MNIST: https://github.com/zalandoresearch/fashion-mnist

<u>Yêu cầu:</u>

1. Dựa vào bộ dữ liệu đã chọn, hãy xây dựng 1 mô hình classification tương ứng. Các bạn có thể dùng sklearn, hoặc các thư viện Machine Learning khác, ví dụ XGBoost.

2. (Optional) Hãy thử trực quan hóa bộ dữ liệu mà các bạn đã chọn trong không gian 2/3 chiều mà không sử dụng label của dữ liệu. Các bạn có thể làm theo các bước sau:

- Đọc dữ liệu
- Biến đổi dữ liệu thành vector (word2vec, tfidf, làm phẳng ảnh, ….)
- Giảm chiều dữ liệu với PCA hoặc t-SNE
- Trực quan hóa