

Đề thi:

BIG DATA IN MACHINE LEARNING

Hạn chót nộp bài: 23h30 – Chủ Nhật ngày 27/03/2022

****** HV tạo 1 thư mục **LDS9_K273_HoVaTen_Cuoi_ky** trong thư mục **LDS9_K273_ONLINE_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 **LDS9_K273_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.
- Cần hiển thị thông tin chung của dữ liệu để có cái nhìn ban đầu về dữ liệu.
- Trong dữ liệu có thể có rất nhiều thông tin (feature/column), cần xác định xem thông tin nào thật sự cần thiết dùng trong thuật toán thì đưa vào, không cần thiết thì không đưa.
- 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 trực quan hóa kết quả thì vừa phải trực quan vừa phải giải thích.
- *Cần phải in các dòng hiển thị kết quả sau từng bước để GV đọc và chấm điểm*, GV chỉ « run » lại bài làm của HV khi thấy bài làm có vấn đề vì thời gian để thực thi cho một bài khá dài.

Câu 1: Classification - Fake and real news dataset (1.0 điểm)

Use **fake-and-real-news-dataset** to build a model to determine if an article is **fake news or not**.

Read more information here:

<https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset>

Câu 2: Classification – Coronavirus tweets NLP (1.0 điểm)

Use data in **covid_text_classification** folder to build a model to determine if an **OriginalTweet** is Extremely Negative/ Negative/ Neutral/ Positive/ Extremely Positive.

Read more information here:

<https://www.kaggle.com/datatattle/covid-19-nlp-text-classification>

Câu 3: Regression - Combined Cycle Power Plant (1.0 điểm)

Use **CCPP** dataset to build the model to predict the **net hourly electrical energy output (EP)** of the plant based on Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V)

Read more information here:

<http://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant>

Câu 4: Clustering – Movies Clustering (1.0 điểm)

Build a clustering model to [cluster the movies](#) in **tmdb_5000_movies.csv** dataset. Explain the main characteristics of each cluster. Use Word Cloud to visualize each cluster.

(Hint: Use some features such as 'title', 'tagline', 'overview', 'popularity'...)

Câu 5: Recommendation - Amazon - Office Products (1.0 điểm)

[If you have more than two laptops/computers or one 32GB-laptop/computer:](#)

- Use the information "reviewerID", "asin" (ProductID), and "overall" (users' ratings for each product) in dataset **ratings_Office_Products.csv** to build a model to [predict overalls for products](#) that have not been selected by users. Then [make recommendations](#) to some users: AZZ56WF4X19G2, A335QXPTV1RIV1, ATIMW8SYGAASW

[If you only have one laptop/computer:](#)

- Use the information "reviewerID", "asin" (ProductID), and "overall" (users' ratings for each product) in dataset **Office_Products_5.json** (unzip file reviews_Office_Products_5.json.gz) to build a model to [predict overalls for products](#) that have not been selected by users. Then [make recommendations](#) to some users: A3CJ7MHAS9IMAM, A3LGT6UZL99IW1, A21KNRUAA5RK5E

Read more information here:

<http://jmcauley.ucsd.edu/data/amazon/>

Câu 6: Association Rules – BAKERY (1.0 điểm)

Use dataset **75000** (select one file in this folder that is suitable for you) to build the model to [identify sets of items](#) that are frequently bought together (please use Flavor and Food name (in **goods.csv**) instead of Id).

Read more information here:

Dataset: <http://users.csc.calpoly.edu/~dekhtyar/466-Spring2018/labs/lab01.html>

--- Good luck 🍀 ---