PROJECT MÔN HỌC MÁY

Giảng viên: **Tiến sĩ Bùi Thanh Hùng**
Trưởng Lab Khoa học Phân tích dữ liệu và Trí tuệ nhân tạo
Giám đốc chương trình Trí tuệ nhân tạo và Hệ thống thông tin
Đại học Thủ Dầu Một
hung.buithanhcs@gmail.com
Website: **https://sites.google.com/site/hungthanhbui1980**
Nộp bài: Google Classroom (code+dữ liệu+report),
Deadline: 24h trước Session 15 hai ngày
Chấm bài: **Session 15**

Project

Thực hiện các yêu cầu sau và viết báo cáo theo mẫu gửi kèm:

1. Phân tích yêu cầu bài toán: Phân tích được yêu cầu của bài toán là gì (1 điểm)
2. Phương pháp giải quyết: Trình bày được các phương pháp giải quyết bài toán, Giải thích lý do tại sao chọn phương pháp này, Vẽ được sơ đồ tổng quát giải quyết bài toán (2 điểm)
3. Hiện thực - Viết code theo phương pháp giải quyết ở trên: Trình bày được cụ thể giải thuật sử dụng để giải quyết bài toán (Lưu đồ giải thuật), các tham số sử dụng, các thư viện sử dụng, code của bài toán … (4.5 điểm)
4. Đánh giá kết quả đạt được: So sánh với ít nhất 1 phương pháp khác, Vẽ được biểu đồ so sánh giữa các phương pháp theo các độ đo ví dụ như: Accuracy, MSE, RMSE, MAP, …. (hãy lựa chọn ít nhất 2 độ đo trong các độ đo phổ biến để đánh giá bài toán trên) (1.5 điểm)
5. Hướng phát triển trong tương lai: Đưa ra được hướng phát triển trong tương lai và giải thích lý do tại sao lại đưa ra hướng phát triển đó (0.25 điểm).
6. Báo cáo theo mẫu (0.25 điểm)
7. Điểm làm việc nhóm (0.5 điểm)

DANH SÁCH ĐỀ TÀI

Đề tài 01
Nhận dạng tên riêng tiếng Việt

Tham khảo:
1. http://www.aclweb.org/anthology/W17-2630
2. http://www.aclweb.org/anthology/W15-3907
3. https://pdfs.semanticscholar.org/719f/f2b6c3d1ac54a0a44dc3570e0d1d795f3a89.pdf
4. github tham khảo: https://github.com/phuonglh/vn.vitk
5. https://arxiv.org/pdf/1705.04044.pdf
6. https://cs.nyu.edu/~thien/pubs/thesis-thien.pdf
7. http://lamda.nju.edu.cn/nguyenct/files/papers/ncamtu-09-paper_ner.pdf
8. Source for Vietnamese: https://github.com/magizbox/underthesea/wiki/Vietnamese-NLP-Tools
9. Demo http://undertheseanlp.com

Đề tài 02
Trích xuất từ khóa tiếng Việt
(Vietnamese Key Pharse Extraction)

Tham khảo:
1. https://arxiv.org/pdf/1704.06879.pdf
2. http://acl2014.org/acl2014/P14-1/pdf/P14-1119.pdf
3. https://aclweb.org/anthology/D16-1080
4. https://pdfs.semanticscholar.org/bdbf/25f3dcf63d38cdb527a9ffca269fa0b8046b.pdf
5. Source for Vietnamese: https://github.com/magizbox/underthesea/wiki/Vietnamese-NLP-Tools

Đề tài 03
Tóm tắt văn bản tiếng Việt tự động
(Text Summarization for Vietnamese Language)

Tham khảo:
1. http://www.aclweb.org/anthology/N16-1007
2. https://www.sciencedirect.com/science/article/pii/S2314728817300582
3. https://www.cs.cornell.edu/~oirsoy/files/emnlp14drnt.pdf
4. https://aclanthology.info/pdf/D/D15/D15-1168.pdf
5. Source for Vietnamese: https://github.com/magizbox/underthesea/wiki/Vietnamese-NLP-Tools

Đề tài 04

Phân lớp văn bản tiếng Việt

Phân lớp văn bản tiếng Việt bằng phương pháp học sâu LSTM (hoặc CNN).

- Hãy tự crawl dữ liệu từ trang web vnexpress.net hay vietnamenet theo các chủ đề: Có thể tham khảo dữ liệu sau để huấn luyện:
  https://github.com/duyvuleo/VNTC/tree/master/Data
- Sử dụng phương pháp học sâu để phân lớp văn bản trên

Đề tài 05

Vietnamese Relation Extraction

Dữ liệu và yêu cầu đề bài có ở đường link sau.

https://vlsp.org.vn/vlsp2020/eval/re

Đề tài 06

Hệ thống gợi xem phim

Hãy xây dựng hệ thống gợi xem phim (Recommander System) bằng phương pháp kết hợp giữa Content –Base và Collaborative Filtering

Tham khảo: Machine Learning cơ bản: Bài 23 + 24

Yêu cầu:

- Thu thập dữ liệu xem phim tự động từ trang web xem phim bất kỳ. Có thể tham khảo bộ dữ liệu IMDB và cào dữ liệu cũng như xây dựng dữ liệu tương đương với bộ dữ liệu này.
- Xây dựng hệ thống gợi ý xem phim

Đề tài 07

Phân tích ý kiến người dùng bằng phương pháp học sâu

Dữ liệu và yêu cầu đề bài có ở đường link sau.

https://www.aivivn.com/contests/1