|  |  |
| --- | --- |
| **CH19-TDTU – Mining Massive Data Sets**  **1. Chủ đề:**   * Frequent itemsets and Association rules, * Near Neighbor Search in High Dimensional Data, * Locality Sensitive Hashing (LSH), * Dimensionality reduction, * Recommendation Systems, * Clustering, * Link Analysis, * Large scale supervised machine learning, * Data streams, * Mining the Web for Structured Data, * Web Advertising.   **2. Yêu cầu:**   * Học viên thực hiện đề tài được phân công trong thời gian từ 14/09/2019 đến 27/10/2019 * Học viên nộp bài qua email [hung.buithanhcs@gmail.com](mailto:hung.buithanhcs@gmail.com) trước 12h ngày 26/10/2019 gồm:   + Báo cáo (file \*.docx theo mẫu bằng tiếng Việt hoặc tiếng Anh (tùy chọn)) chú ý tập trung vào các yêu cầu sau:  1. Vấn đề nghiên cứu là gì, đã có ai nghiên cứu chưa, giải pháp đề xuất của mình?  2. Dữ liệu sử dụng?  - Bài toán cần giải quyết sử dụng dữ liệu gì, làm sao để xây dựng được dữ liệu đó?  - Dữ liệu được lấy từ đâu?  3. Giải quyết vấn đề như thế nào?  - Bài toán có cần tiền xử lý không, nếu có hãy liệt kê các bước trong quá trình tiền xử lý đó  - Đã áp dụng các mô hình gì, trình bày cụ thể các bước thực hiện trong mô hình đó  - Các giải thuật sử dụng, các code tham khảo (nếu có)  4. Đánh giá kết quả như thế nào, bằng các độ đo gì, trình bày các kết quả đã đạt được?  + File trình bày (\*.pptx)  + Dữ liệu chương trình  + File chương trình chạy (Tất cả code viết bằng Python file đuôi \*.py)   * Trong trường hợp file lớn quá 10MB, học viên tải lên google drive và gửi đường link vào file text nộp lên E-learning * Tất cả học viên sẽ báo cáo và demo vào Chủ nhật ngày 27/10/2019 bắt đầu từ 13:15, các nhóm khác nghe và đánh giá kết quả của nhóm báo cáo * Khi đi báo cáo, học viên in 2 cuốn báo cáo nộp về Khoa và thầy * Mọi thắc mắc học viên liên hệ với giảng viên: Ts. Bùi Thanh Hùng qua email **hung.buithanhcs@gmail.com** hoặc số điện thoại **0908-542-521** | |
|  |  |
| Đề tài 1 | **Tên đề tài: Phân tích ý kiến người dùng bằng phương pháp học sâu kết hợp CNN-LSTM (Vietnamese Sentiment Analysis using Deep Learning)**  **Lê Anh Khoa – Trương Trọng Vinh**  **Tham khảo**:  1. https://arxiv.org/pdf/1801.09053.pdf  2. https://machinelearningmastery.com/develop-n-gram-multichannel-convolutional-neural-network-sentiment-analysis/  3. https://machinelearningmastery.com/predict-sentiment-movie-reviews-using-deep-learning/  4. https://thesai.org/Downloads/Volume8No6/Paper\_57-Sentiment\_Analysis\_using\_Deep\_Learning.pdf  5. Source for Vietnamese: https://github.com/magizbox/underthesea/wiki/Vietnamese-NLP-Tools  6. Demo: http://undertheseanlp.com  7. https://viblo.asia/p/moi-cac-ban-gop-y-project-sentiment-analysis-su-dung-tf-idf-ap-dung-cho-ngon-ngu-tieng-viet-LzD5d64oZjY  8. Dữ liệu tham khảo: VLSP 2018: https://github.com/undertheseanlp/NLP-Vietnamese-progress |
| Đề tài 2 | **Tên đề tài: Tìm hiểu trình phân loại để phân loại khách truy cập trang web của bạn** (Learn a classifier for categorizing the visitors of your website)  **Đinh Vũ Quốc Trung - Cát nghiêm Hiếu Tuấn**  **Tham khảo**:  1. https://pdfs.semanticscholar.org/c46b/a28b4f468483d388711ab4ed39bc3e0b92f7.pdf  2. https://arxiv.org/pdf/1701.07179.pdf  3. Data: https://gist.github.com/rm-hull/bd60aed44024e9986e3c |
| Đề tài 3 | **Tên đề tài: Phân loại văn bản tiếng Việt sử dụng phương pháp học máy SVM (Vietnamese Document Classify using SVM)**  **Nguyễn Đại Thịnh – Hồng Quang Vinh**  Yêu cầu: Vào đoạn văn bản, xác định thể loại văn bản đó  Cụ thể: Chuẩn bị dữ liệu (10 chủ đề: Chính trị, Đời sống, Kinh doanh, Khoa học, Pháp luật, Sức khỏe, Thế giới, Thể thao, văn hóa, Vi tính)  Dataset tham khảo: https://github.com/duyvuleo/VNTC/tree/master/Data  Training sử dụng Word2Vector (Word Embedding)  Áp dụng giải thuật Deep Learning cho train và predict dữ liệu  **Tham khảo**:  1. http://blog.hocmay.asia/classification/phan-loai-van-ban-tieng-viet-phan-1/  2. Vietnamese Toolkits: http://jvntextpro.sourceforge.net/  3. Demo tham khảo: http://magizbox.com:9386/#!/  4. Source for Vietnamese: https://github.com/magizbox/underthesea/wiki/Vietnamese-NLP-Tools |
| Đề tài 4 | **Tên đề tài: Xác định cộng đồng bằng cách phân tích cấu trúc đồ thị của Wikipedia (D**etect Communities by Analyzing Graph Structure of Wikipedia)  **Trương Trần Vỹ - Huỳnh Minh Phúc**  **Tham khảo**:  1. http://www2009.eprints.org/191/1/p1221.pdf  2. http://snap.stanford.edu/class/cs224w-2010/proj2010/08\_final\_paper.pdf  3. https://github.com/SergiuTripon/msc-thesis-na-epsrc/wiki/Community-Detection-(Wikipedia)  4.http://www.science.smith.edu/classwiki/images/a/a9/Network\_analysis\_collaboration\_wikipedia.pdf  5. Dataset: https://snap.stanford.edu/data/wiki-Talk.html |
| Đề tài 5 | **Tên đề tài: Bảo vệ hệ thống khuyến nghị bằng cách xác định các xếp hạng giả (**Defend a recommender system by identifying fake ratings)  **Nguyễn Duy Hàn Lâm - Lưu Minh Tâm**  **Tham khảo**:  1. http://www.ijarcs.info/index.php/Ijarcs/article/download/4550/4100  2. https://machinelearningcoban.com/2017/05/17/contentbasedrecommendersys/  3. https://machinelearningcoban.com/2017/05/24/collaborativefiltering/  4. https://machinelearningcoban.com/2017/05/31/matrixfactorization/  5. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.329.4891&rep=rep1&type=pdf  6. https://www.math.uci.edu/icamp/courses/math77b/lecture\_12w/pdfs/Chapter%2009%20-%20Attacks%20on%20collaborative%20recommender%20systems.pdf |
| Đề tài 6 | **Tên đề tài: Phân tích miền chủ đề bằng cách sử dụng biểu đồ siêu liên kết thu thập thông tin phổ biến　(A**nalyze topical domain using Common Crawl Hyperlink Graph)  **Diệp Bảo Đạt – Vương Vỹ Thông**  **Tham khảo**:  1. Data: http://webdatacommons.org/hyperlinkgraph/  2.https://pdfs.semanticscholar.org/dcb4/46b2c22fb4bb645e765e73909189d5a6057e.pdf  3. https://www.cis.upenn.edu/~ccb/publications/bitexts-from-common-crawl.pdf  4. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tang\_cikm05.pdf  5. http://www.divms.uiowa.edu/~psriniva/Papers/TKDE06.pdf  6. http://eprints.kingston.ac.uk/32140/1/Patel-A.pdf  7. http://infolab.stanford.edu/~olston/publications/crawling\_survey.pdf  8. Dataset: http://webdatacommons.org/hyperlinkgraph/ |
| Đề tài 7 | **Tên đề tài: Thực nghiệm với hệ thống khuyến nghị lai trong miền ứng dụng cụ thể** (Experiment with hybrid recommenders in specific application domain)  **Nguyễn Văn Khoa – Huỳnh Nhị Đệ**  **Tham khảo**:  1. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.8200&rep=rep1&type=pdf  2. https://machinelearningcoban.com/2017/05/17/contentbasedrecommendersys/  3. https://machinelearningcoban.com/2017/05/24/collaborativefiltering/  4. https://machinelearningcoban.com/2017/05/31/matrixfactorization/  5. https://www.uts.edu.au/sites/default/files/desi-publication-recommender%20system%20application%20developments%20a%20survey-accepted%20menuscript.pdf  6. https://fruct.org/publications/abstract20/files/Dan.pdf  7. Dataset tham khảo: https://github.com/caserec/Datasets-for-Recommneder-Systems |
| Đề tài 8 | **Tên đề tài: Ứng dụng học luật trong phân tích khách hàng**  (Cài đặt được ít nhất 1 giải thuật ASSOCIATION RULES và ứng dụng giải thuật cài đặt vào phân tích khách hàng trên bộ dữ liệu lớn), sử dụng weka để tham khảo kết quả thư viện với giải thuật mình cài đặt)  **Trương Hồng Phát - Diệp Gia Huy**  **Tham khảo:**  1. http://adataanalyst.com/machine-learning/fp-growth-algorithm-python-3/  2. https://acta.mendelu.cz/media/pdf/actaun\_2015063031031.pdf  3. https://ieeexplore.ieee.org/document/4721419  4. Dataset tham khảo: https://www.kaggle.com/c/instacart-market-basket-analysis/data |
| Đề tài 9 | Tên đề tài: Link Prediction based on Graph Neural Networks **Lý Kim Phát - Trương Vũ Hoàng Anh**  **Tham khảo:**  1- https://arxiv.org/abs/1010.0725  2- https://archive.siam.org/meetings/sdm06/workproceed/Link%20Analysis/12.pdf  3- https://arxiv.org/abs/1802.09691  4- Dataset tham khảo: https://www.kaggle.com/c/link-prediction-tu/data |
| Đề tài 10 | **Tên đề tài: Sử dụng Locality Sensitive Hashing giải quyết bài toán Image similarity identification (Action Dataset)**  **Phạm Thị Kim Điệp- Võ Đăng Khoa**  **Tham khảo:**  1- https://en.wikipedia.org/wiki/VisualRank  2- https://github.com/santhoshhari/Locality-Sensitive-Hashing  3- https://journals.sagepub.com/doi/full/10.1177/1550147717694172  4- https://advances.sciencemag.org/content/1/11/e1501057.full?intcmp=trendmd-adv  5- Dataset tham khảo: http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm#action |
| Đề tài 11 | **Tên đề tài: Clustering in Big Data**  **Lưu Thanh Duy – Phan Lý Phương Nghi**  **Tham khảo:**  1- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6832486  2- https://www.mii.lt/paslaugu\_internetas/rodikliai/5veikla/5.3.pdf  3- Dataset tham khảo: https://www.kaggle.com/arjunbhasin2013/ccdata |
| Đề tài 12 | **Tên đề tài: Stock Price Prediction Using Generative Adversarial Network (GAN) (CNN-LSTM)**  **Taymany - Dung Cẩm Quang**  **Tham khảo:**  1- https://towardsdatascience.com/aifortrading-2edd6fac689d  2- https://www.sciencedirect.com/science/article/pii/S1877050919302789  3- Dataset tham khảo: Sử dụng API: Quandl để lấy dữ liệu stock |