BÁO CÁO GIỮA KỲ KHÓA LUẬN TỐT NGHIỆP
HK II - NĂM HỌC: 2017-2018

SV: (Ký và ghi rõ họ tên)

TÊN ĐỀ TÀI: PHÂN TÍCH VÀ KHAI THÁC DỮ LIỆU WEB

Tuần	Nội dung	Mức độ hoàn thành	Nguyên nhân chưa hoàn thành	Kế hoạch tiếp theo	Ghi chú
1-2 (03/12/2017 đến 17/12/2017)	- Tìm hiểu và giới thiệu nội dung đề tài. - Khảo sát làm rõ về đề tài. - Ước lượng về tài chính. - Xác định và phân rã bài toán. - Tìm và đưa ra các giải pháp. - Tiếp nhận yêu cầu của giảng viên. - Xác định mục tiêu cần đạt được của đề tài. - Lập kế hoạch cụ thể cho từng giai đoạn và từng tuần: ước lượng thời gian thực hiện, công việc cụ thể,... - Vẽ biểu đồ Gantt trên word.	**100%**			- Học kỳ I (2017-2018)
3 (18/12/2017 đến 24/12/2017)	- Tìm hiểu các kiến thức nền về khai phá và phân tích dữ liệu, các công nghệ sử dụng, kiến thức cơ bản về đại số tuyến tính trong phân tích dữ liệu. - Tìm hiểu về các dự án thực tế và hướng đi của đề tài trong tương lai.	**0%**	- Tạm dừng đồ án để ôn thi cuối kỳ I 2017-2018	- Chuyển nội dung tuần thứ 3 thực hiện sang tuần tiếp theo	- Thi cuối kỳ I (2017 - 2018)
4 (18/12/2017 đến 24/12/2017)	- Thực hiện nội dung tuần thứ 3. - Thay đổi tiến độ công việc cho phù hợp. - Thay đổi biểu đồ Gantt từ word sang visio theo yêu cầu của giảng viên.	**100%**		- Đẩy nhanh một số công việc để phù hợp với kế hoạch.	- Thời gian làm vào buổi tối vì thực tập ban ngày. (2h mỗi ngày trừ thứ 5 và chủ nhật đi học)
5 - 6 (25/12/2017 đến 07/01/2018)	- Tìm hiểu API của mạng xã hội facebook và twiter - Chạy thử nghiệm API. - Tìm hiểu về cách lưu trữ dữ liệu. - Lựa chọn lưu trữ và API của mạng xã hội.	**100%**			- Chọn mạng xã hội facebook và lưu trữ bằng file excel.

7-8-9 (08/01/2018 đến 28/01/2018)	- Xây dựng chương trình để lấy dữ liệu và lưu trữ tự động. - Thực hiện lấy dữ liệu về và lưu trữ. - Tìm hiểu về các thư viện của python: numpy, pandas, matploid,..	**50%**	- Chưa tìm hiểu kỹ về chính sách sử dụng dữ liệu của facebook. - Chưa lựa chọn được trang có dữ liệu phù hợp. - Một số lỗi về định dạng và kết quả trả về. - Đường truyền mạng bị gián đoạn trong quá trình lấy dữ liệu	- Tìm hiểu về chính sách sử dụng dữ liệu của facebook. - Hỏi ý kiến giảng viên về việc lựa chọn cho phù hợp.	- Giải pháp lựa chọn là trang tin tức BBCNews. Hoặc các trag tin tức tương tự nêu không đáp ứng được về mặt dữ liệu.
10 (29/01/2018 đến 04/02/2018)	- Tiếp tục thực hiện việc lấy dữ liệu: + Khắc phục lỗi định dạng. + Khắc phục lỗi ngắt kết nối và thêm khả năng tự lấy dữ liệu khi có kết nối mạng. + Khắc phục lỗi khi đọc file quá lớn.	**100%**			- Đổi dạng lưu trữ từ excel sang json.
11 (05/02/2018 đến 11/02/2018)	- Xử lý dữ liệu thô: + Lọc lấy những dữ liệu cần thiết và loại bỏ dữ liệu thừa. + Xây dựng chương trình để xử lý dữ liệu. - Tìm hiểu về thuật toán phân tích nội dung. - Tìm hiểu về thuật toán phân cụm.	**80%**	- Cần nhiều thời gian để tìm hiểu và chạy thử thực tế.	- Thay đổi phương án sử dụng số liệu để phù hợp với thời gian làm đồ án	
12 - 13 - 14 (12/02/2018 đến 04/03/2018)	- Xử lý nhiễu dữ liệu: + Loại bỏ các ký tự đặc biệt. + Thêm dữ liệu cho các trường trống. + Chia dữ liệu làm nhiều trường. - Xây dựng chương trình xử lý.	**100%**			- Nghỉ tết nguyên đán (2 tuần). Thời gian trong kế hoạch chưa tính đến nên bị trễ so với kế hoạch.
15 (05/03/2018 đến 18/03/2018)	- Khắc phục các lỗi khi xử lý nhiễu. - Sửa lỗi dữ liệu do chương trình chạy sai. - Thực hiện lấy dữ liệu một số trang để đánh giá dữ liệu có bị lỗi hay không.	**80%**	- Chương trình lấy dữ liệu ban đầu chưa mã hoá dữ liệu nên khi lưu trữ sẽ bị lỗi. (Chỉ xảy ra với các ký tự unicode)	- Giải pháp khắc phục là chạy lại chương trình sau khi cải tiến hoặc chỉ sử dụng dữ liệu không phải ký tự unicode	
16 (19/03/2018 đến 25/03/2018)	- Xây dựng bài toán dự đoán. - Tìm hiểu về mô hình Marakor chain. - Xây dựng và vẽ biểu đồ. - Tìm hiểu về biểu đồ Pareto.	**70%**	- Biểu đồ chưa biểu diễn được theo mong muốn.	- Tìm hiểu cách thiết kế biểu đồ cho phù hợp với bài toán.	- Sử dụng jupyter notebook để biểu diễn tốt hơn.
17 (26/03/2018 đến 01/04/2018)	- Viết báo cáo giữa kỳ. - Tổng hợp nhật ký trên giấy, kế hoạch, các lỗi lên trello. - Lấy ý kiến và đánh giá của giáo viên để chuẩn bị cho báo cáo. - Tiếp tục công việc tuần 16.	**70%**			
ĐÁNH GIÁ TỔNG KẾT GIAI ĐOẠN GIỮA KỲ		**70%**			