ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

ONLINE NEWS POPULARITY

GVLT: Lê Thị Nhàn

NHÓM 10

Mssv	Họ và tên	Email	Vai trò
1412646	Nguyễn Đình Vũ	dinhvuhcmus@gmail.com	Leader
1412558	Phạm Quốc Toàn	bossdiemmaimai@gmail.com	Coder
1412689	Hoàng Bích Vân	Hoangbichvan95@gmail.com	Assitant
1312618	Nguyễn Thanh Trà	tra.it1095@gmail.com	Time keeper

Mục lục

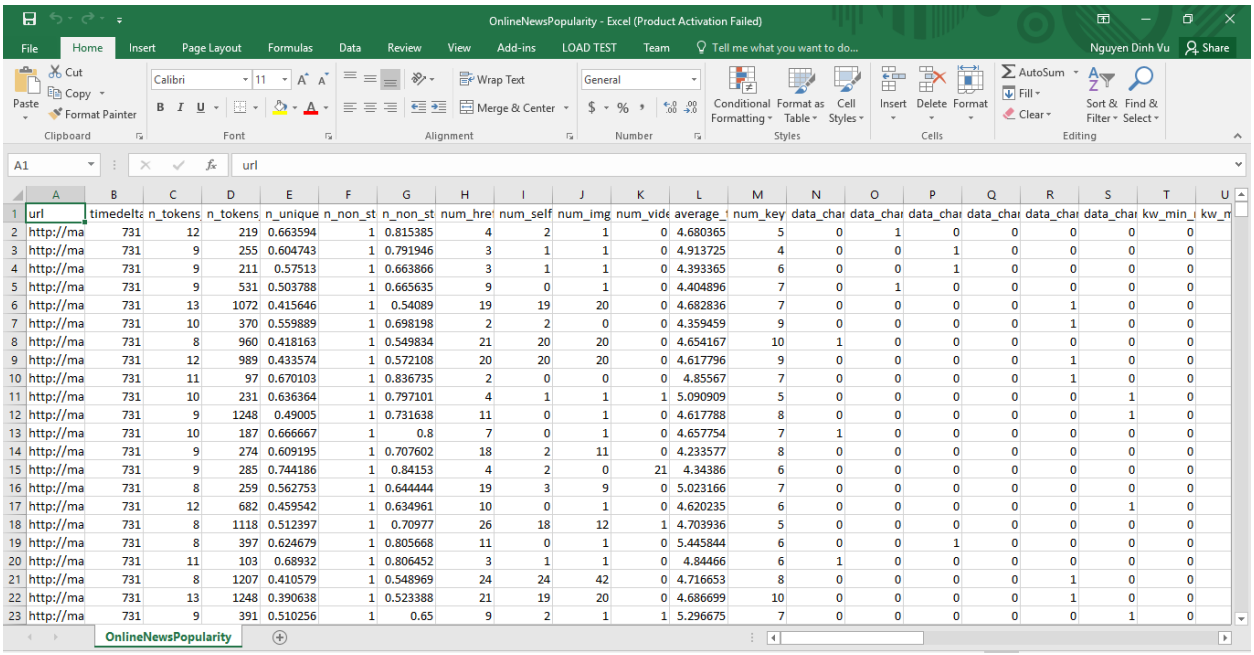
I. Mô tả vấn đề

Nhà xuất bản Mashable hoạt động xuất bản báo online, dựa vào các dữ liệu thu thập được trong 2 năm 2013 và 2014, nhà xuất bản muốn tìm hiểu xem, các yếu tố nào của một bài báo sẽ tạo cho bài báo đó có lượt share nhiều trên mạng xã hội. Từ đó có kế hoạch và chiến lược phát triển hợp lý cho nhà xuất bản.

Mục tiêu: Phân tích đánh giá các thuộc tính ảnh hưởng đến sự phổ biến của bài báo.

II. Mô tả dữ liệu

1. Bộ dữ liệu nhóm nhận được:



2. Các kiểu dữ liệu
   - Nominal: url
   - Ratio Quantitative:
     1. timedelta
     2. n_tokens_title:
     3. n_tokens_content:
     4. n_unique_tokens:
     5. n_non_stop_words:
     6. n_non_stop_unique_tokens:
     7. num_hrefs: Number of links
     8. num_self_hrefs:
     9. num_imgs:
     10. num_videos:
     11. average_token_length:
     12. num_keywords:
     13. data_channel_is_lifestyle:

14. data_channel_is_entertainment:
15. data_channel_is_bus:
16. data_channel_is_socmed:
17. data_channel_is_tech:
18. data_channel_is_world:
19. kw_min_min:
20. kw_max_min:
21. kw_avg_min:
22. kw_min_max:
23. kw_max_max:
24. kw_avg_max:
25. kw_min_avg:
26. kw_max_avg:
27. kw_avg_avg:
28. self_reference_min_shares:
29. self_reference_max_shares:
30. self_reference_avg_sharess:
31. weekday_is_monday:
32. weekday_is_tuesday:
33. weekday_is_wednesday:
34. weekday_is_thursday:
35. weekday_is_friday:
36. weekday_is_saturday:
37. weekday_is_sunday:
38. is_weekend:
39. LDA_00:
40. LDA_01:
41. LDA_02:
42. LDA_03:
43. LDA_04:
44. global_subjectivity:
45. global_sentiment_polarity:
46. global_rate_positive_words:
47. global_rate_negative_words:
48. rate_positive_words:
49. rate_negative_words:
50. avg_positive_polarity:
51. min_positive_polarity:
52. max_positive_polarity:
53. avg_negative_polarity:

54. min_negative_polarity:
55. max_negative_polarity:
56. title_subjectivity:
57. title_sentiment_polarity:
58. abs_title_subjectivity:
59. abs_title_sentiment_polarity:
60. shares:

3. Ý nghĩa dữ liệu

Attribute Information:

0. url: URL of the article (non-predictive)
1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
2. n_tokens_title: Number of words in the title
3. n_tokens_content: Number of words in the content
4. n_unique_tokens: Rate of unique words in the content
5. n_non_stop_words: Rate of non-stop words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: Number of links
8. num_self_hrefs: Number of links to other articles published by Mashable
9. num_imgs: Number of images
10. num_videos: Number of videos
11. average_token_length: Average length of the words in the content
12. num_keywords: Number of keywords in the metadata
13. data_channel_is_lifestyle: Is data channel 'Lifestyle'?
14. data_channel_is_entertainment: Is data channel 'Entertainment'?
15. data_channel_is_bus: Is data channel 'Business'?
16. data_channel_is_socmed: Is data channel 'Social Media'?
17. data_channel_is_tech: Is data channel 'Tech'?
18. data_channel_is_world: Is data channel 'World'?
19. kw_min_min: Worst keyword (min. shares)
20. kw_max_min: Worst keyword (max. shares)
21. kw_avg_min: Worst keyword (avg. shares)
22. kw_min_max: Best keyword (min. shares)
23. kw_max_max: Best keyword (max. shares)
24. kw_avg_max: Best keyword (avg. shares)
25. kw_min_avg: Avg. keyword (min. shares)
26. kw_max_avg: Avg. keyword (max. shares)
27. kw_avg_avg: Avg. keyword (avg. shares)

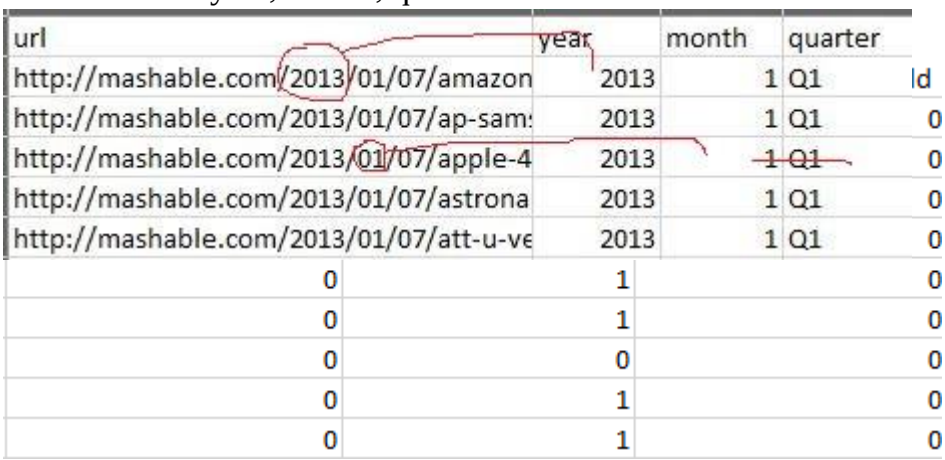
28. self_reference_min_shares: Min. shares of referenced articles in Mashable
29. self_reference_max_shares: Max. shares of referenced articles in Mashable
30. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
31. weekday_is_monday: Was the article published on a Monday?
32. weekday_is_tuesday: Was the article published on a Tuesday?
33. weekday_is_wednesday: Was the article published on a Wednesday?
34. weekday_is_thursday: Was the article published on a Thursday?
35. weekday_is_friday: Was the article published on a Friday?
36. weekday_is_saturday: Was the article published on a Saturday?
37. weekday_is_sunday: Was the article published on a Sunday?
38. is_weekend: Was the article published on the weekend?
39. LDA_00: Closeness to LDA topic 0
40. LDA_01: Closeness to LDA topic 1
41. LDA_02: Closeness to LDA topic 2
42. LDA_03: Closeness to LDA topic 3
43. LDA_04: Closeness to LDA topic 4
44. global_subjectivity: Text subjectivity
45. global_sentiment_polarity: Text sentiment polarity
46. global_rate_positive_words: Rate of positive words in the content
47. global_rate_negative_words: Rate of negative words in the content
48. rate_positive_words: Rate of positive words among non-neutral tokens
49. rate_negative_words: Rate of negative words among non-neutral tokens
50. avg_positive_polarity: Avg. polarity of positive words
51. min_positive_polarity: Min. polarity of positive words
52. max_positive_polarity: Max. polarity of positive words
53. avg_negative_polarity: Avg. polarity of negative words
54. min_negative_polarity: Min. polarity of negative words
55. max_negative_polarity: Max. polarity of negative words
56. title_subjectivity: Title subjectivity
57. title_sentiment_polarity: Title polarity
58. abs_title_subjectivity: Absolute subjectivity level
59. abs_title_sentiment_polarity: Absolute polarity level
60. shares: Number of shares (target)

III. Phân tích vấn đề

1. Xác định nội dung trực quan hóa

- Mục tiêu: trả lời được các câu hỏi về các yếu tố ảnh hưởng đến sự phổ biến của bài báo.
- Chức năng: trình bày và thuyết phục người xem tầm quan trọng của thuộc tính bài báo có ảnh hưởng đến sự phổ biến của bài báo.
- Kết quả công trình: công cụ biểu diễn trực quan hóa

2. Khảo sát và phân tích bộ dữ liệu
   - Gồm 39644 dòng, 61 thuộc tính
   - Biến phụ thuộc của bộ dữ liệu: shares
   - Một số bài báo chưa phân loại, có thể đó là Viral
   - Ngưỡng xác định bài cáo có phổ biến hay không là shares >=1400
3. Đặt câu hỏi để phân tích
   - C1: Có phải một kênh phổ biến vì nó có nhiều bài viết?
   - C2: Kênh nào phổ biến nhất và ít phổ biến nhất?
   - C3: Nhiều ảnh, video có ảnh hưởng đến lượng đọc, chia sẻ bài báo không?
   - C4: Người đọc thích bài báo có tiêu đề, nội dung ngắn/ hay dài. Nên đặt tiêu đề bài báo bao nhiêu ký tự là phù hợp?
   - C5: Đăng bài thời điểm nào trong tuần thu hút người xem nhất?
4. Xử lý dữ liệu
   - Biến dùng trực tiếp: shares, n_tokens_title, num_imgs, num_videos, n_tokens_content
   - Thêm biến year, month, quarter tách từ url.



url	year	month	quarter	
http://mashable.com/2013/01/07/amazon	2013	1	Q1	ld
http://mashable.com/2013/01/07/ap-sam:	2013	1	Q1	0
http://mashable.com/2013/01/07/apple-4	2013	1	Q1	0
http://mashable.com/2013/01/07/astrona	2013	1	Q1	0
http://mashable.com/2013/01/07/att-u-ve	2013	1	Q1	0
0	1			0
0	1			0
0	0			0
0	1			0
0	1			0

   - Thêm biến article_type từ biến data_channel.

Lifestyle	Eentertainment	Bussiness	Social Media	Tech	World	aricle_type
0	1	0	0	0	0	Eentertainment
0	0	1	0	0	0	Bussiness
0	0	1	0	0	0	Bussiness
0	1	0	0	0	0	Eentertainment
0	0	0	0	1	0	Tech
0	0	0	0	1	0	Tech
1	0	0	0	0	0	Lifestyle
0	0	0	0	1	0	Tech
0	0	0	0	1	0	Tech
0	0	0	0	0	1	World
0	0	0	0	0	1	World

- Sửa biến weekday.

weekday_is_monday	weekday_is_tuesday
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0

Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Weekday
1	0	0	0	0	0	Tuesday
1	0	0	0	0	0	Tuesday
1	0	0	0	0	0	Tuesday
1	0	0	0	0	0	Tuesday
1	0	0	0	0	0	Tuesday
1	0	0	0	0	0	Tuesday
1	0	0	0	0	0	Tuesday
~~1~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~Tuesday~~
1	0	0	0	0	0	Tuesday
1	0	0	0	0	0	Tuesday
1	0	0	0	0	0	Tuesday
0	~~1~~	~~0~~	~~0~~	~~0~~	~~0~~	~~Wednesday~~
0	1	0	0	0	0	Wednesday
0	1	0	0	0	0	Wednesday

- Thêm biến popular: 1 nếu shares >=1400 ngược lại 0

- Sữa article_type = Viral đối với dữ liệu chưa phân loại

IV. Thiết kế vấn đề

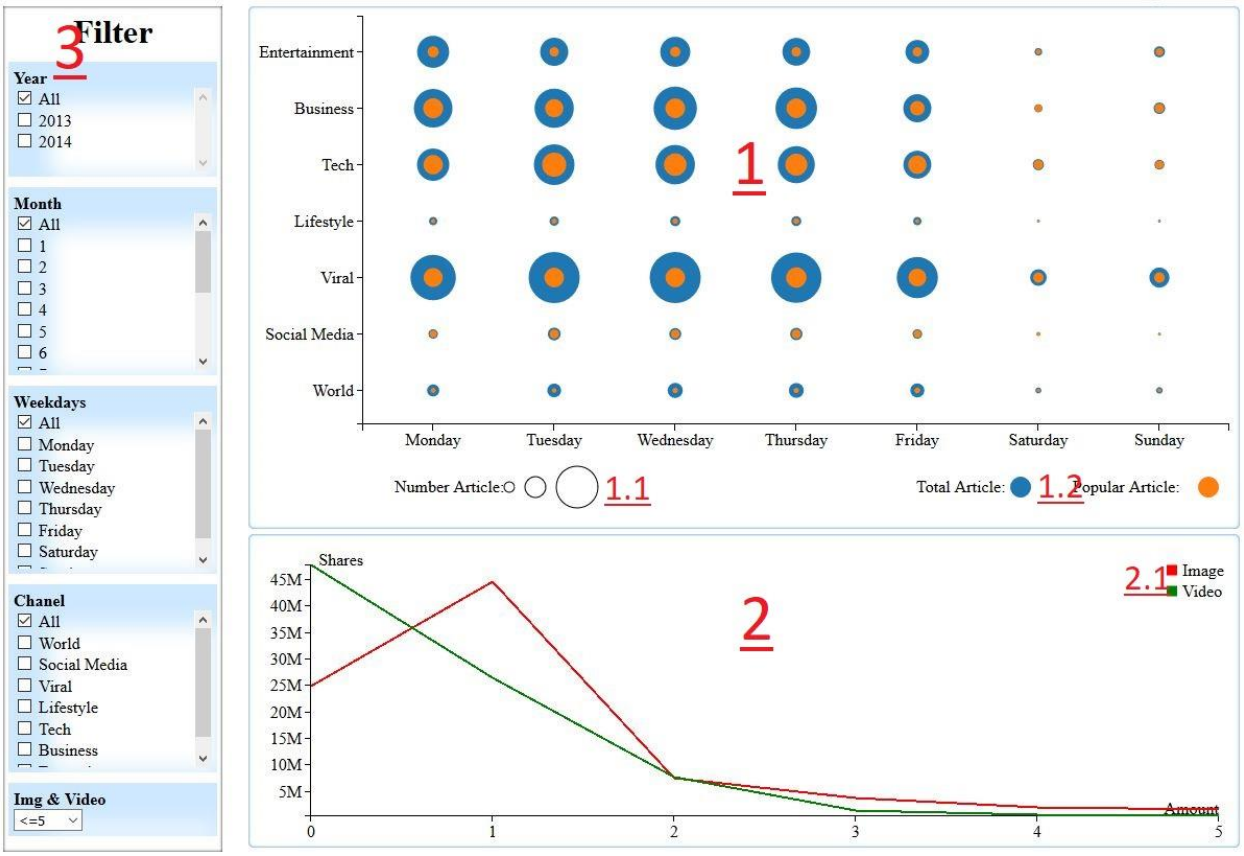
1. Nhận thấy rằng, các câu hỏi C1, C2, C5 đều có chung đặc điểm là sử dụng thuộc tính channel và weekday, ta có thể kết hợp chúng lại để biểu diễn chung với nhau như một ma trận, các phần tử tại vị trí của ma trận sẻ biểu diễn các dữ liệu độ Shares của bài báo. Thực hiện được như vậy bởi channel và weekday là kiểu Interval Quantity nhưng miền giá trị nó thấp và giá trị củng biểu diễn được dưới dạng Nominal. Khi thực hiện như vậy theo 2 trục Ox, Oy thì phần tử biểu diễn là rơi vào góc phần tư thứ nhất của biểu đồ, do đó ta dường như vẽ ra một biểu đồ Scatterplot. Với mỗi điểm Scatterplot sẽ bao gồm tổng số bài báo và số bài phổ biến.
    - Channel và weekday được biểu diễn bằng Ox, Oy
    - Xác định vị trí của một bài báo thuộc channel và weekday dùng mark là Point
    - Lượng bài báo (giá trị Point biểu diễn) – biến viusal là kiểu Volumn (là kích thước đường kính của Point, độ to của hình tròn)
    - Tổng số bài báo và số bài báo phổ biến được phân biệt bởi màu sắc – biến visual là kiểu Color (Point có type là Color)
2. Đối với câu 2: vì mục tiêu quan sát là số lượng ảnh, clip mà phạm vi nó rất rộng, co giá trị từ 0 cho đến 130 nên khó có thể phân tích. Do đó giải pháp đưa ra là vẽ biểu đồ đường.
    - Ox biểu diến số lượng ảnh, clip, Oy là lượng Shares

- Line biểu diễn sự ảnh hưởng của số lượng ảnh, clip tới số lượng chia sẻ (nhìn theo hướng từ trái qua phải)

3. Đối với câu 4: số lượng kí tự trong tiêu đề và nội dung có phạm vi cực lớn, thay đổi nhanh chóng, khó tính tuần suất nên không biết cách xử lý.

V. Cài đặt

- Cài đặt mã nguồn:
  - D3 để vẽ SVG
  - Jquery để sử dụng bộ chọn linh hoạt
  - D3.tip hỗ trợ hiển thị html
  - File css mà mã nguồn tự viết là mystyle.css và myscript.js.
- Các hàm của D3 sử dụng trong đồ án
  - d3.csv: đọc dữ liệu từ file
  - d3.nest: gom nhóm dữ liệu
  - d3.append : thêm đối tượng mới vào vùng SVG
  - d3.selectAll: khai báo hoặc chọn các đối tượng theo bộ chọn
  - d3.scale: hỗ trợ chia bin và tính toán vị trí của các biến dữ liệu
  - …
- Cài đặt các hàm chính: gồm 4 hàm chính
  - Type_WeekDay_Scaterplot: cài đặt cho thiết kế vấn đề số 1.
  - Img_Video_Line: cài đặt cho thiết kế vấn đề số 2
  - Filters: hàm lọc dữ liệu khi có tương tác với người
  - Visualize: hàm chủ chốt, nằm mọi chỗ mỗi khi có sự kiện xảy ra, mục đích gọi các hàm khác khi có sự kiện lọc dữ liệu, nhận bộ dữ liệu đã lọc sau đó gọi các hàm để visualize vẽ ảnh ra trang html.
- Hình ảnh công trình



- Giải thích chức năng công cụ
  - Vùng 1 là nơi biểu diễn scatterplot. Đặc tả biến visual nằm ở mục IV
  - Vùng 1.1 – chú giải: đặc tả cho biết kích thước bán kính của hình tròn biểu diễn cho số lượn bài báo
  - Vùng 1.2 – chú giải: là vị trí tương tác khi rê chuột vào 2 hình tròn màu xanh hoặc cam. Chức năng khi rê vào màu xanh thì hình tròn màu cam trên scatterplot sẽ biến mất và ngược lại. Sự tương tác này giúp người nhìn dễ quan sát từng hình ảnh hơn khi nhìn vào biểu đồ.
  - Vùng 2 là nơi biểu diễn biểu đồ line. Đặc tả ở mục IV.
  - Vùng 2.1 – chú giải: vừa là chú giải chỉ màu của line, vừa có sự tương tác rê chuột sẽ ẩn line có màu khác nhằm hiển thị rõ line hơn khi chúng chồng chéo.
  - Vùng 3: Tương tác cho phép người dùng lọc theo các điều kiện thiết thiết lập sẵn.

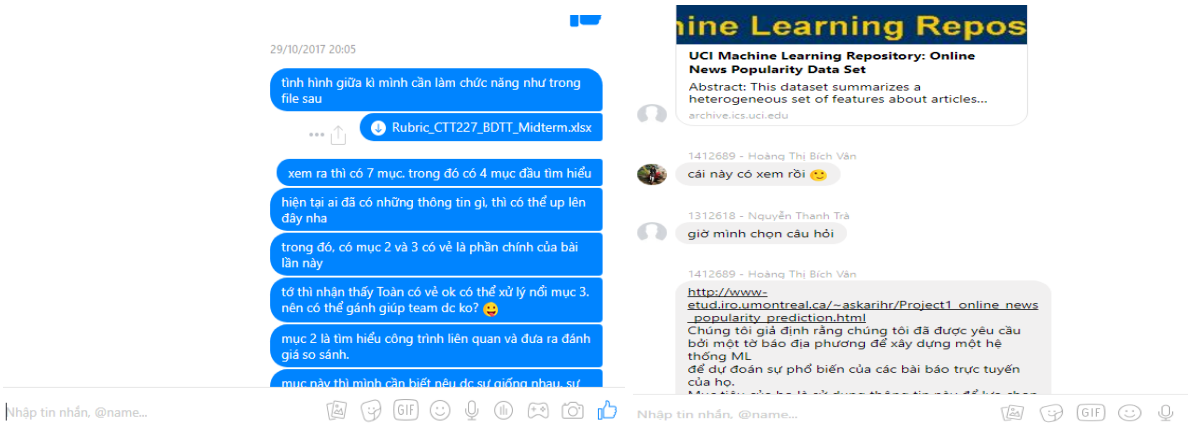
VI. Biên bản họp nhóm

- Họp nhóm lần 1.

  Họp trực tiếp. https://drive.google.com/open?id=0B4wt_bgbZtoUMmw0LWtqYk1yN2M

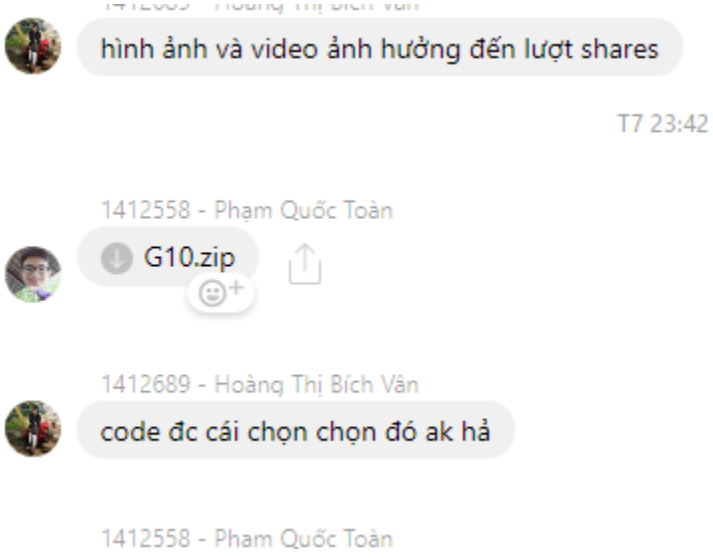
➢ Họp nhóm lần 2: 29/10/2017

- Họp qua Facebook.
- Thành phần tham dự: đầy đủ 4 thành viên.
- Nội dung: Bàn về đồ án, quá trình giữa kì
- Hình ảnh:



➢ Họp nhóm lần 3: 7/1/2018

- Họp qua Facebook.
- Thành phần tham dự: đầy đủ 4 thành viên.
- Nội dung: Bàn về đồ án, quá trình làm đồ án, hỗ trợ nhau làm việc theo sự phân công và hoàn thành đồ án
- Hình ảnh:



➢ Họp nhóm lần 5: 13/01/2018.

- Họp trực tiếp
- Nội dung: Tổng kết đồ án. Làm poster, Báo cáo, và review lại chương trình.

VII. Phân công công việc

https://docs.google.com/spreadsheets/d/1LvsVFFc8D2tbp3jWD6W66UIvAcClUOTHfy0yz-ig-Bo/edit?usp=sharing

	B	C	D	E	F	G	H
	2	Toàn, Vũ, Vân, Trà	- Họp nhóm, tổng hợp nội dung tìm hiểu	6	14/10/2017	14/10/2017	
	3	Toàn, Vân	- Viết báo cáo	6	14/10/2017	15/10/2017	
	4	Vân, Trà	- So sánh cách công trình	7	16/10/2017	19/10/2017	
	5	Toàn, Vũ, Vân, Trà	- Đặt câu hỏi, mỗi người đặt ít nhất 2 câu về bộ dữ liệu - Phân tích hướng giải quyết các câu hỏi	7	16/10/2017	19/10/2017	
	6	Toàn	- Tiền xử lý dữ liệu cho phù hợp để trả lời câu hỏi	7	20/10/2017	21/12/2017	
	7	Vân	- Tổng hợp tài liệu, viết báo cáo	7	22/10/2017	22/10/2017	
	8	Vân	- Thiết kế biến visual, vẽ mô phỏng trực quan giải quyết câu hỏi 1	8	23/10/2017	31/10/2017	
	9	Trà	- Thiết kế biến visual, vẽ mô phỏng trực quan giải quyết câu hỏi 2	8	23/10/2017	31/10/2017	
	10	Vũ	- Thiết kế biến visual, vẽ mô phỏng trực quan giải quyết câu hỏi 3,4	8	23/10/2017	31/10/2017	
	11	Toàn	- Thiết kế biến visual, vẽ mô phỏng trực quan giải quyết câu hỏi 5,6	8	23/10/2017	31/10/2017	
	12	Vũ, Vân	- Tạo ppt thuyết trình giữa kỳ	9	01/11/2017	02/11/2017	
	13	Toàn, Vũ, Vân, Trà	- Họp nhóm, tổng hợp và review các trực quang mô phỏng	9	03/11/2017	05/11/2017	
	14	Cả nhóm (Phụ trách chính: Toàn)	Code, xây dựng chương trình	10, 11, 12	04/11/2017	07/01/2018	
	15	Vân phụ trách chính	- Làm Poster	13	07/01/2018	14/01/2018	
	16	Vũ phụ trách chính	- Làm Report.	13	07/01/2018	14/01/2018	