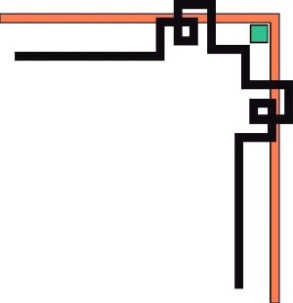
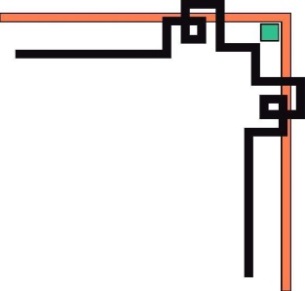
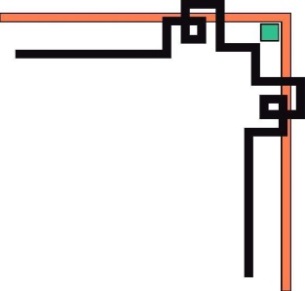
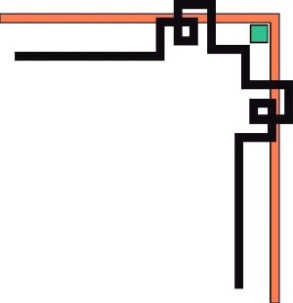
**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**



**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**🕮🕮🕮**

****

**Visualization: Online News Popularity**

**MÔN: BIỂU DIỄN THÔNG TIN**

**GIẢNG VIÊN: LÊ THỊ NHÀN**

* **Nhóm thực hiện:** 
  + 1412558 – Phạm Quốc Toàn
  + 1412646 – Nguyễn Đình Vũ
  + 1412689 – Hoàng Bích Vân
  + 1312618 – Nguyễn Thanh Trà

TP. Hồ Chí Minh, ngày 25 tháng 10 năm 2017

1. **Giới thiệu vấn đề**

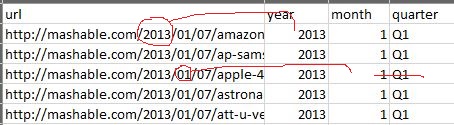
Ngày nay với tốc độ phát triển vượt bậc của công nghệ và internet mà chúng tạo cơ sở cho chúng ta truyền đạt thông tin một cách nhanh chóng, đặc biệt là tin tức trực tuyến. Nhằm trình bày cho tác giả và người đọc có thể biết được đặc điểm của một bài báo tin tức trực tuyến trở nên phổ biến nhờ những yếu tố vào bên trong nó nên chúng tôi sẽ lấy một vài yếu tố cơ bản trong tập yếu tố của nó để phân tích và đưa ra nhận xét sự ảnh hưởng của các yếu tố đó.

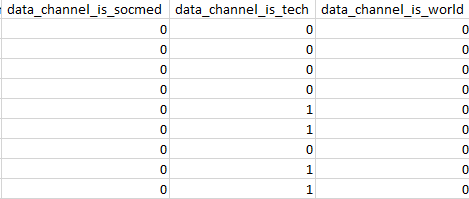
1. **Mô tả bộ dữ liệu**

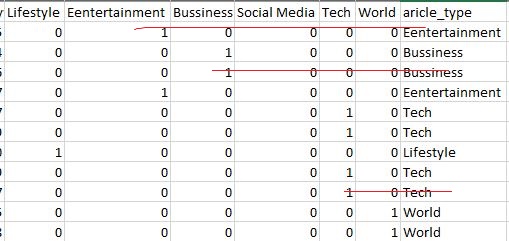
Bộ dữ liệu tóm tắt một số thuộc tính tham gia cấu thành một bài báo và được được xuất bản bởi Mashable, một website tin tức nổi tiếng. Bộ dữ liệu này được thu thập trong hai năm từ 1/2013 đến 12/2014. Mục đích của bộ dữ liệu là cho chúng ta nghiên cứu, phân tích đánh giá các thuộc tính mang tính ảnh hưởng đến sự phổ biến của bài báo.

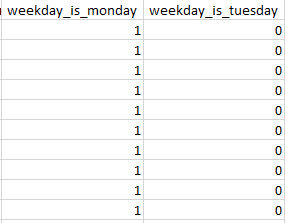
* + - Gồm 39644 dòng (số lượng bài báo), 61 thuộc tính
    - Biến phụ thuộc của bộ dữ liệu: shares
    - Một số bài báo chưa phân loại, có thể đó là Viral
    - Ngưỡng xác định bài cáo có phổ biến hay không là lượng shares >=1400

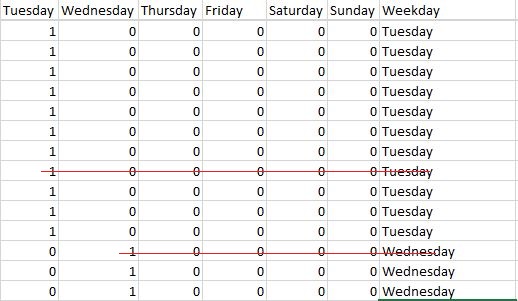
1. **Phân tích vấn đề**
   1. **Xác định nội dung trực quan hóa**
      * Mục tiêu: trả lời được các câu hỏi về các yếu tố ảnh hưởng đến sự phổ biến của bài báo.
      * Chức năng: trình bày và thuyết phục người xem tầm quan trọng của thuộc tính bài báo có ảnh hưởng đến sự phổ biến của bài báo.
      * Kết quả công trình: công cụ biểu diễn trực quan hóa
   2. **Đặt các câu hỏi**
      * 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?
      * C6: Xu hướng quan tâm kênh của người đọc như thế nào theo thời gian (giữa các tháng, quý, năm)?
   3. **Phân tích câu hỏi, tìm phương pháp trả lời bằng trực quan hóa**
      * C1: So sánh số lượng bài viết, lượng shares giữa các kênh
      * C2: Xếp hạng các kênh theo thứ tự về sự phổ phổ biến
      * C3: So sánh số lượng ảnh, video, lượng shares của các kênh
      * C4: Dựa vào số ký tự của tiêu đề, nội dung, lượng share của các kênh
      * C5: Tính tổng lượng share các ngày trong tuần, so sánh
      * C6: Mô phỏng sự thay đổi bảng xếp hạn phổ biến của các channel theo thời gian
   4. **Tiền 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.



* + - Thêm biến article\_type từ biến data\_channel.



* + - Sửa biến weekday.



* + - Thêm biến popular: 1 nếu shares >=1400 ngược lại 0
    - Sữa article\_type = Viral đối với dữ liệu chưa phân loại
  1. **Thiết kế vấn đề**