Báo cáo bài tập lớn

IT5427E - BIG DATA INTEGRATION AND PROCESSING

Nhóm:

- Hoàng Văn Công
- Vũ Đình Nguyên

Tổng quan

- Tên chương trình:

  Phân tích dữ liệu các bài báo online

- Mục đích:
  - Tìm hiểu các thông tin từ dữ liệu thu thập được từ các trang báo online.
  - Đưa ra được một số nhận định từ dữ liệu.

- Phân công:
  - Nguyên:
    - Thu thập dữ liệu
    - Tiền xử lý dữ liệu
    - Cài đặt phân cụm
  - Công:
    - Tiền xử lý dữ liệu
    - Phân tích dữ liệu

Content

1. Thu thập dữ liệu
2. Tiền xử lý dữ liệu
3. Phân tích dữ liệu
4. Cài đặt phân cụm

1. Thu thập dữ liệu

1. Thu thập dữ liệu

- Nguồn báo:
  - 24h.
  - VietnamNet.
  - VNExpress.
- Công cụ: Python với Scrapy
- Dữ liệu:
  - 24h.csv
  - Vietnamnet.csv
  - VNExpress.csv
- Kích thước dữ liệu: 100M

1. Thu thập dữ liệu

❖ Cấu trúc dữ liệu:

Name	Type	Desc
headline	string	Tiêu đề (title) của bài báo
description	string	Mô tả (subtitle) của bài báo, 1 câu ngắn gọn để mô tả bài báo.
date_published	string	Thời gian đăng bài viết dạng: dd/mm/yyyy
content	string	Nội dung bài báo
category	string	Category của bài báo

2. Tiền xử lý dữ liệu

2. Tiền xử lý dữ liệu

a. Tổng hợp các nguồn dữ liệu

- ✓ Merge 3 data frame vào thành 1

```
# init source
data_24h_df = data_24h_df.withColumn('source', F.lit("24h"))
data_VNE_df = data_VNE_df.withColumn('source', F.lit("VNExpress"))
data_VNN_df = data_VNN_df.withColumn('source', F.lit("VietNamNet"))

# Merge
data_df = data_VNE_df.union(data_VNN_df).union(data_24h_df)
```

2. Tiền xử lý dữ liệu

b. Thay đổi time format

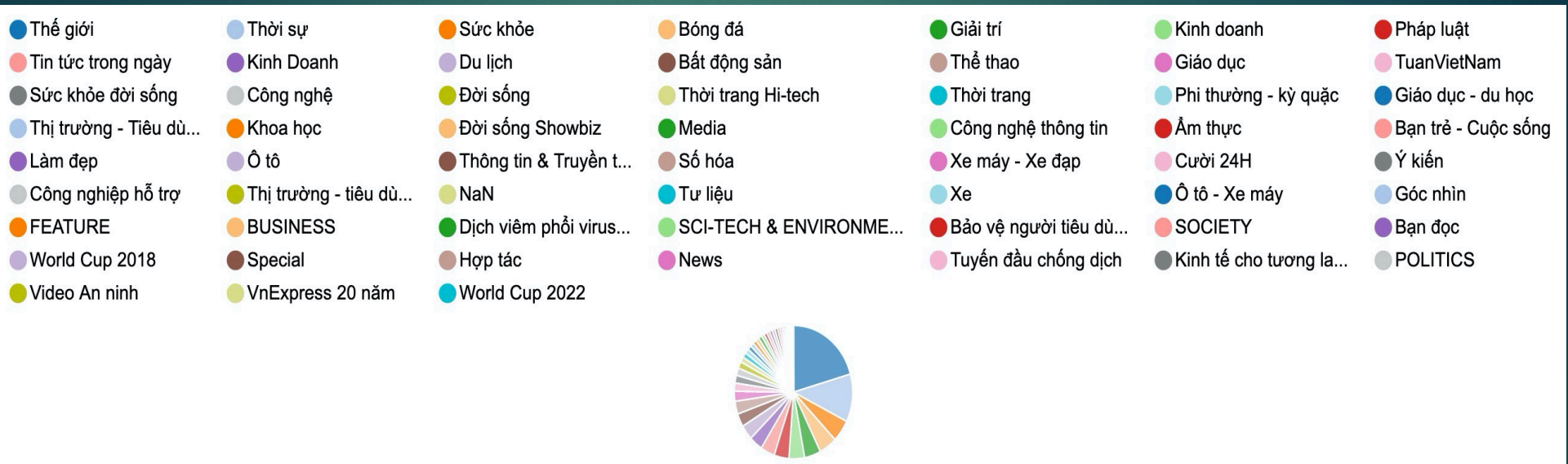
- ✓ Chuyển từ dd/mm/yyyy sang yyyy/mm và yyyy/mm/dd

```
%spark.pyspark
data_df = data_df.withColumn("pub_date",
                  reformatDate_UDF(data_df.date_published))

data_df = data_df.withColumn("pub_month",
                  reformatMonth_UDF(data_df.date_published))

z.show(data_df.select('pub_date','pub_month'))
```

2. Tiền xử lý dữ liệu

c. Xử lý category

- ✓ Chuyển từ 59 categories về 9 categories



2. Tiền xử lý dữ liệu

c. Xử lý category

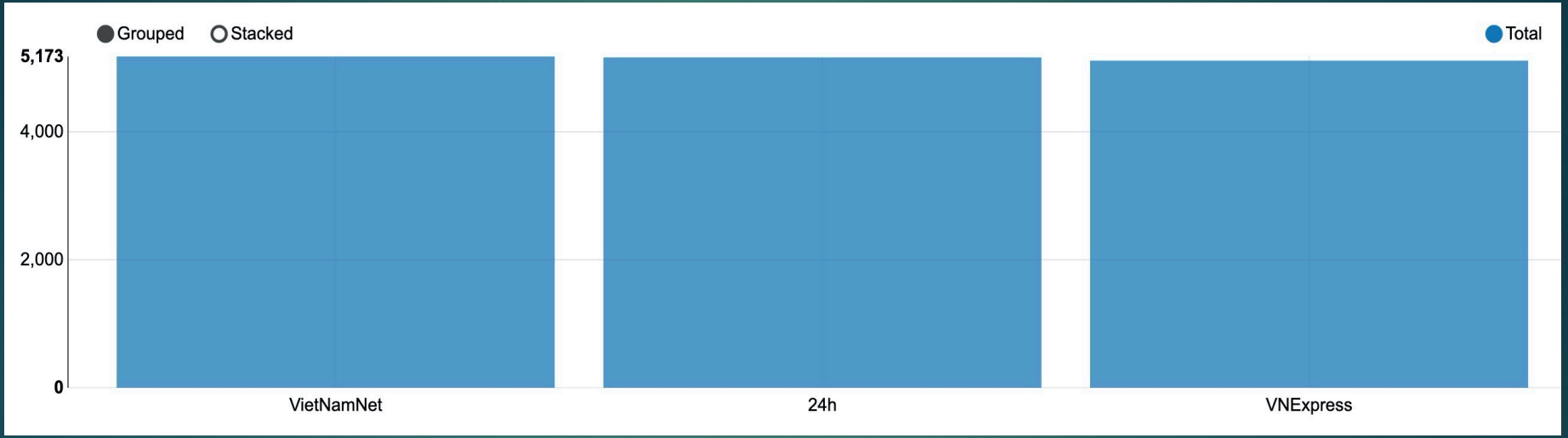
- ✓ Chuyển từ 59 categories về 9 categories

Name	Sub-Categories
Tin tức	Thời sự, Tin tức trong ngày, TuanVietNam, Thế giới, Special, POLITICS, SOCIETY, FEATURE, Video An ninh
Sức khoẻ- Đời sống	Sức khỏe đời sống, Sức khỏe, Đời sống, Bạn trẻ - Cuộc sống, Tuyến đầu chống dịch, Dịch viêm phổi virus corona, Ẩm thực, Du lịch
Giải trí	Giải trí , Đời sống, Showbiz, Cười 24H, Media
Kinh doanh	Kinh doanh, Kinh Doanh, Thị trường - Tiêu dùng, Thị trường - tiêu dùng, BUSINESS, Bảo vệ người tiêu dùng, Hợp tác, Bất động sản, Kinh tế cho tương lai, Pháp luật
Thể thao	Thể thao, World Cup 2022, Bóng đá, World Cup 2018
Giáo dục	Giáo dục, Giáo dục - du học
Công nghệ	Công nghệ, Khoa học, Công nghệ thông tin, SCI-TECH & ENVIRONMENT, Thông tin & Truyền thông, Số hóa
Thời trang	Thời trang, Làm đẹp, Thời trang Hi-tech
Xe	Xe, Ô tô - Xe máy, Ô tô, Xe máy - Xe đạp
Others	Góc nhìn, Ý kiến, Tư liệu, Phi thường - kỳ quặc, Công nghiệp hỗ trợ, Bạn đọc, VnExpress 20 năm

3. Phân tích dữ liệu

3. Phân tích dữ liệu

a. Thống kê bài báo theo nguồn

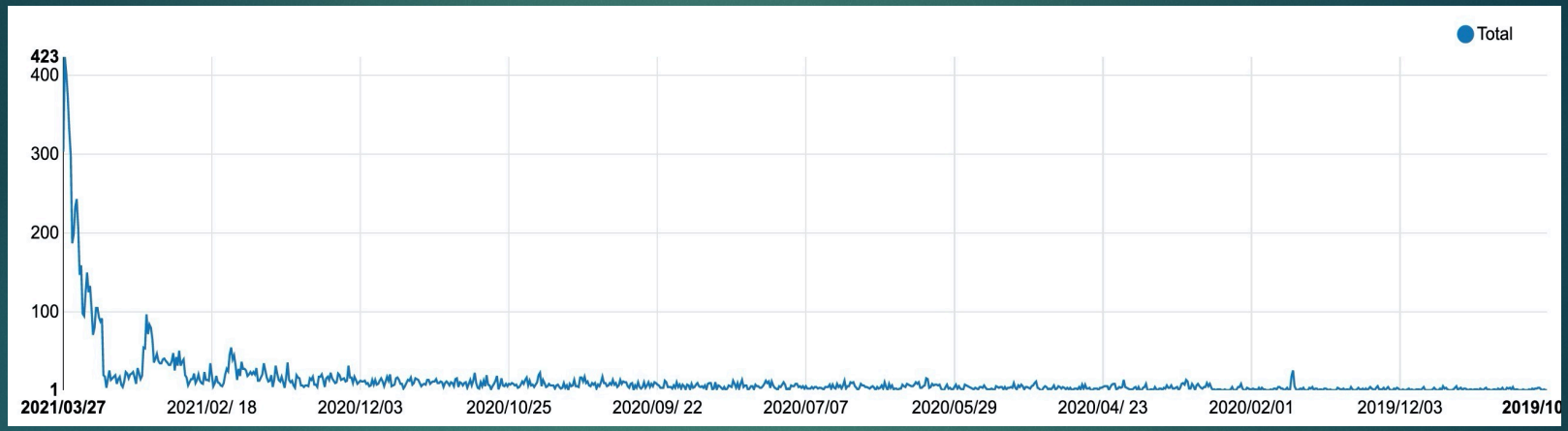


- Số lượng bài viết gần tương đương nhau ~ 5.100 bài/ source

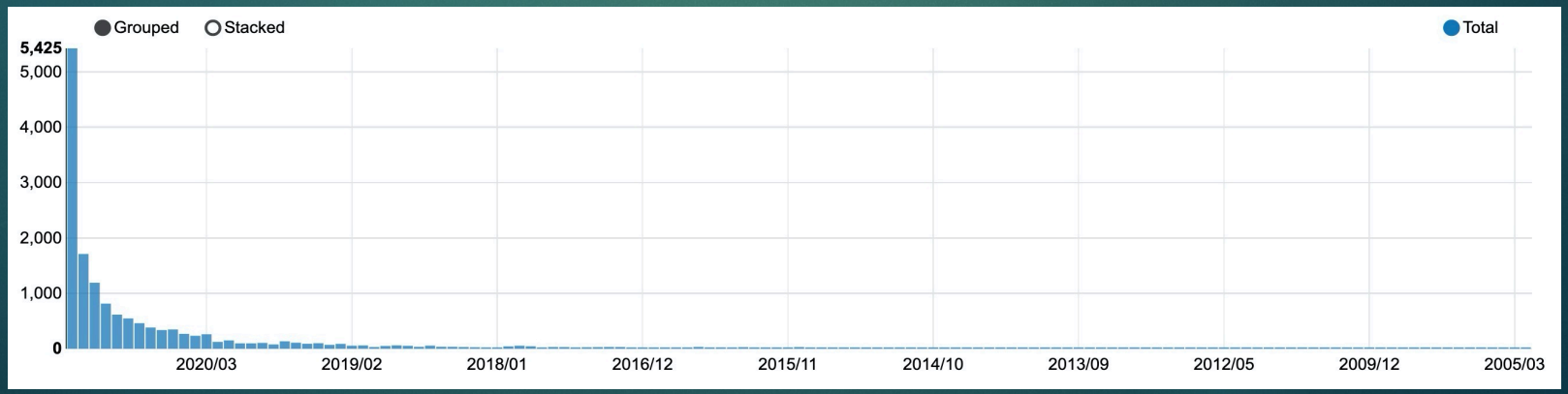
3. Phân tích dữ liệu

b. Thống kê bài báo theo ngày / tháng

- Theo ngày



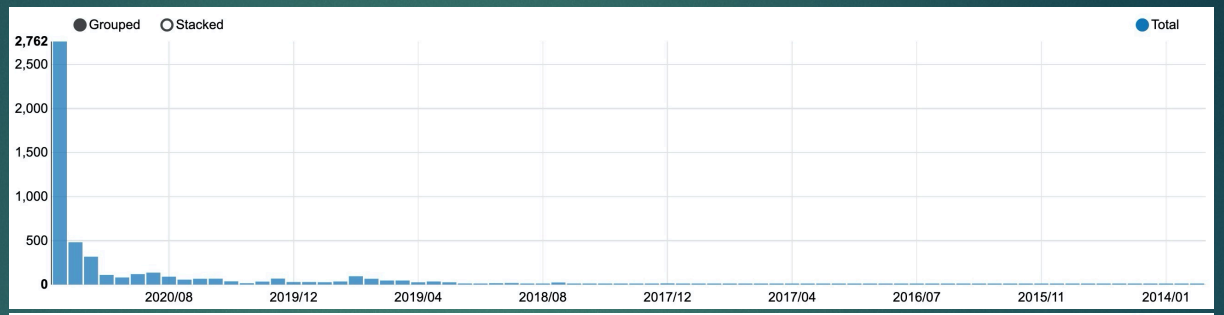
- Theo tháng



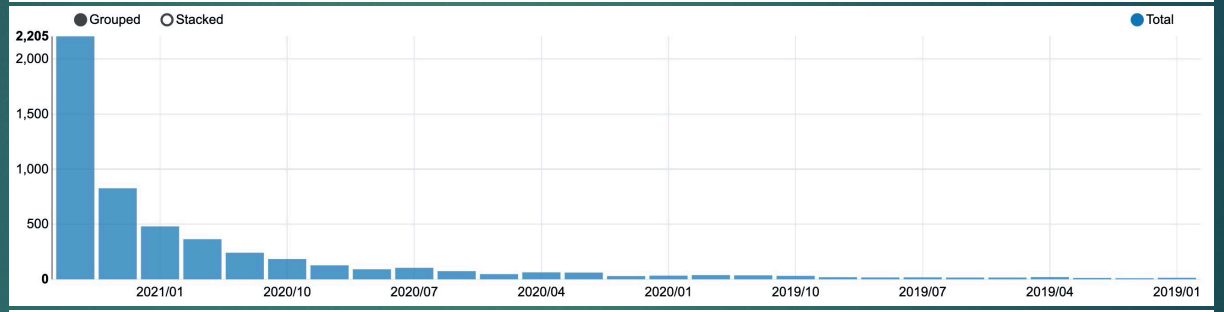
3. Phân tích dữ liệu

b. Thống kê bài báo tháng theo source

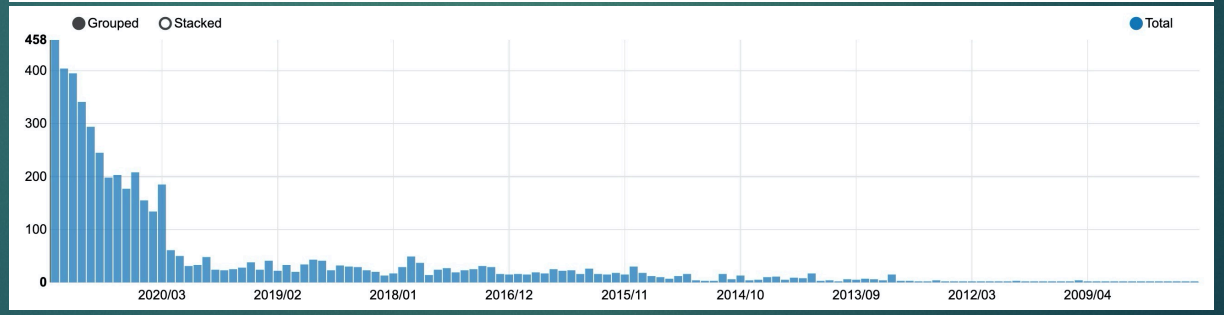
- 24h:



- Vietnam Net:



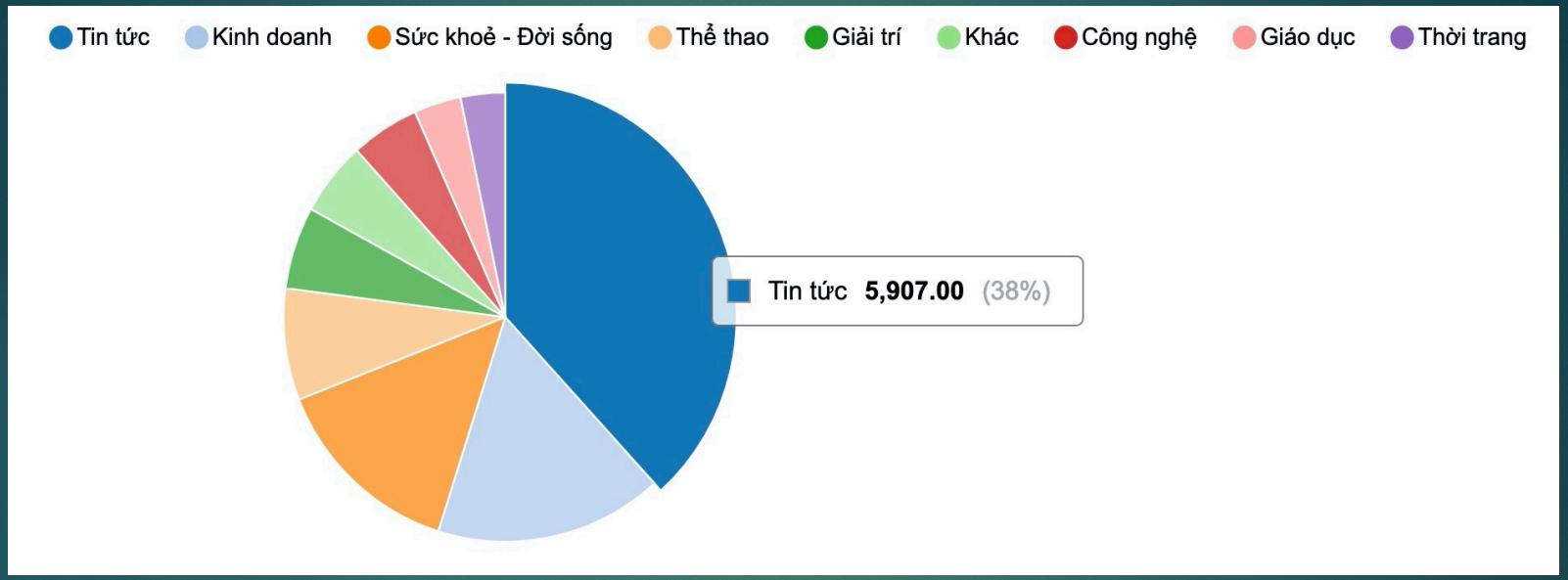
- VNExpress:



- Data phân bố không đồng đều:
  - Hiệu chiểu view của các query
  - Nhớ rằng data có thể không thể hiện đúng về số lượng

3. Phân tích dữ liệu

c. Thống kê theo category

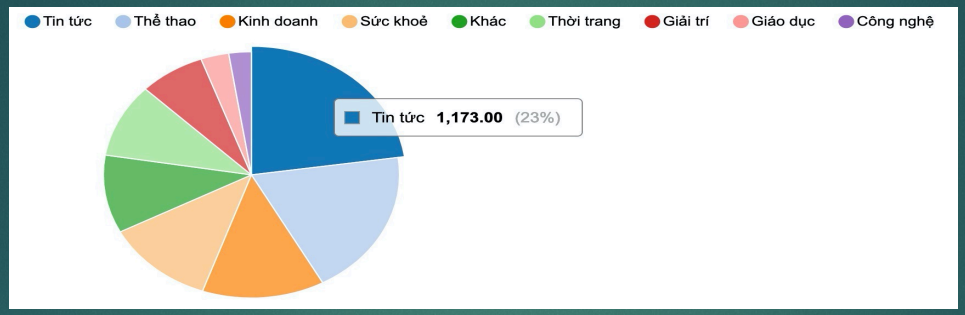


- Số lượng bài viết lớn nhất thuộc thể loại Tin tức với 38%
- Tiếp theo sau là 2 mảng: Kinh doanh (17%) và Sức khoẻ - Đời sống (14%)

3. Phân tích dữ liệu

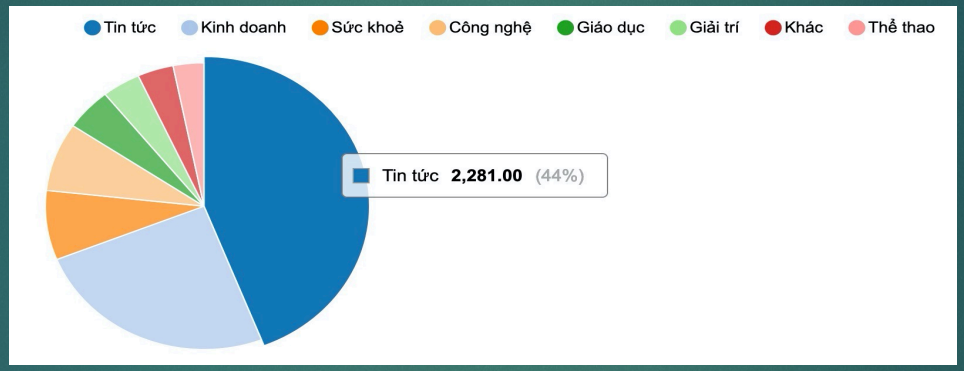
d. Thống kê category theo nguồn

- 24h:



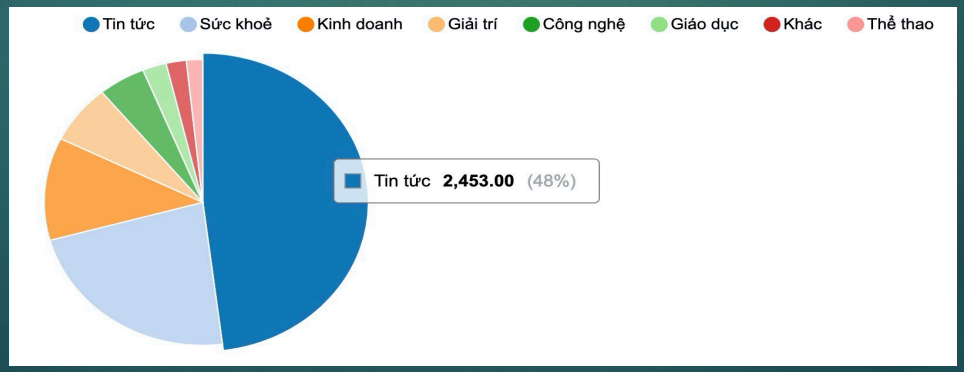
- Tin tức: 23%
- Thể thao: 19%
- Kinh doanh: 13%
- Sức khoẻ: 12%

- Vietnam Net:



- Tin tức: 44%
- Kinh doanh: 25%
- Sức khoẻ: 8%

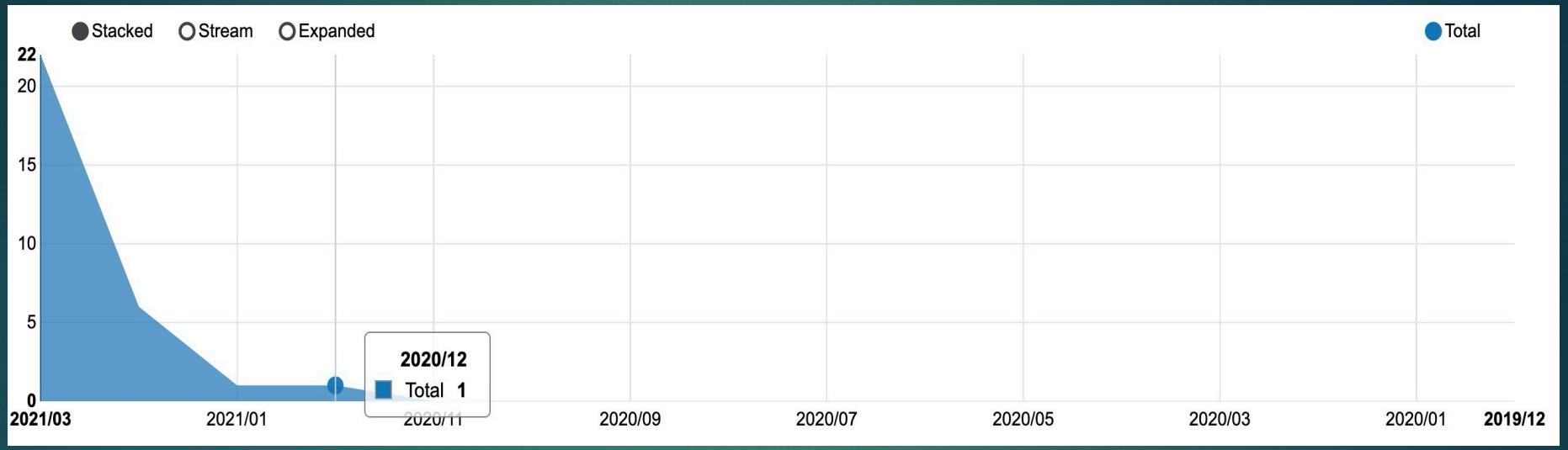
- VNExpress:



- Tin tức: 48%
- Sức khoẻ: 23%
- Kinh doanh: 12%

3. Phân tích dữ liệu

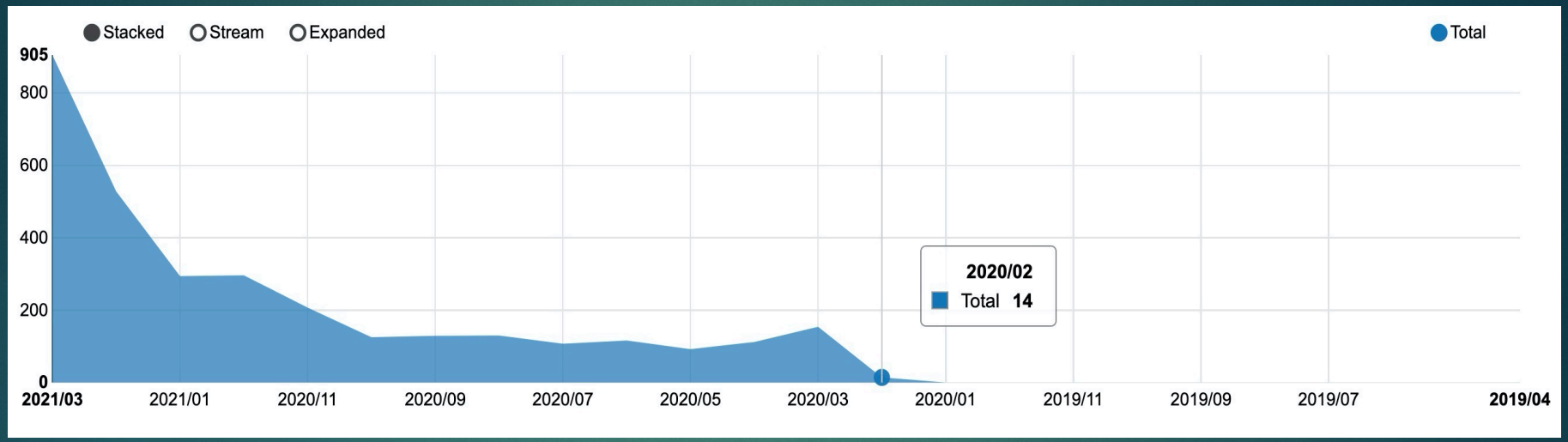
e. Thống kê số lượng bài báo đề cập đến Bitcoin



- Thời điểm tháng 12.2020 là thời điểm Bitcoin vượt ngưỡng 20.000$

3. Phân tích dữ liệu

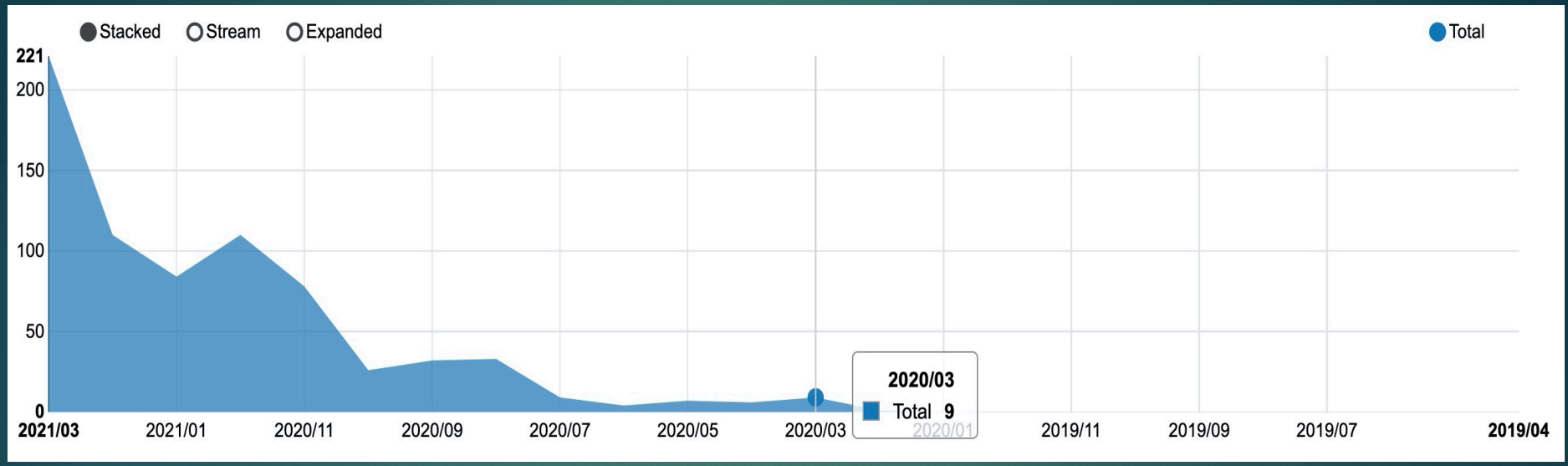
f. Thống kê số lượng bài báo đề cập đến Covid



- Thời điểm tháng 2.2020 là thời điểm Việt Nam xuất hiện bệnh nhân Covid đầu tiên

3. Phân tích dữ liệu

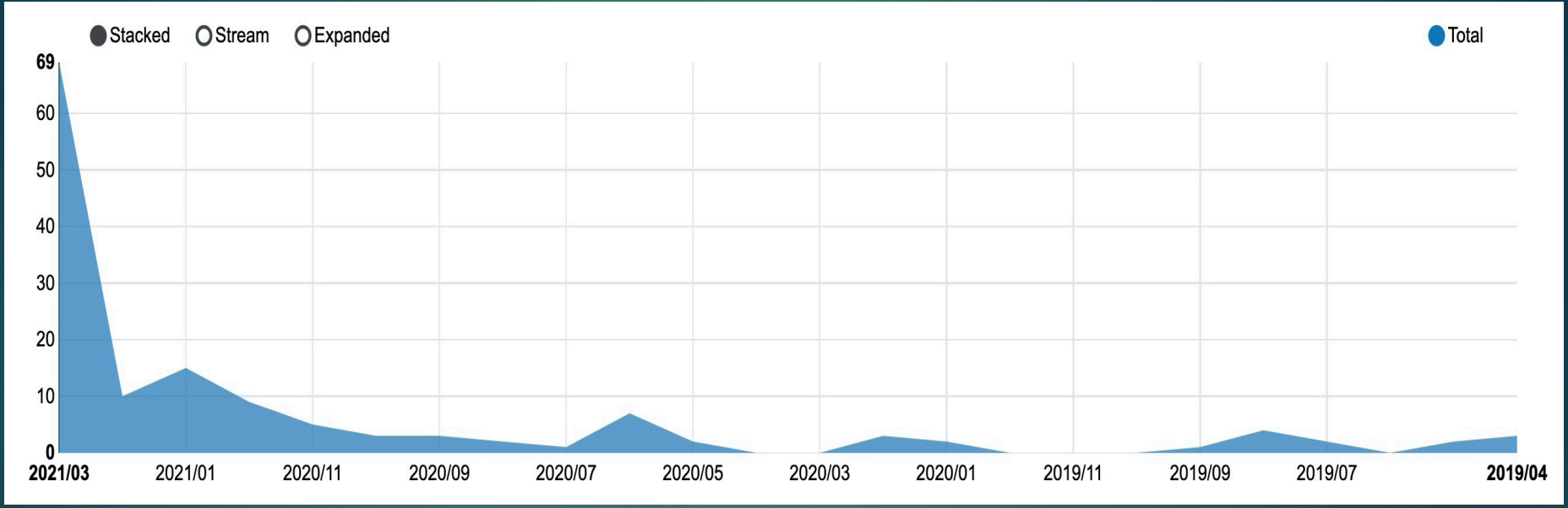
g. Thống kê số lượng bài báo đề cập đến Vaccine Covid



- Việc đề cập đến vaccine covid cũng được quan tâm ngay tại thời điểm xuất hiện dịch

3. Phân tích dữ liệu

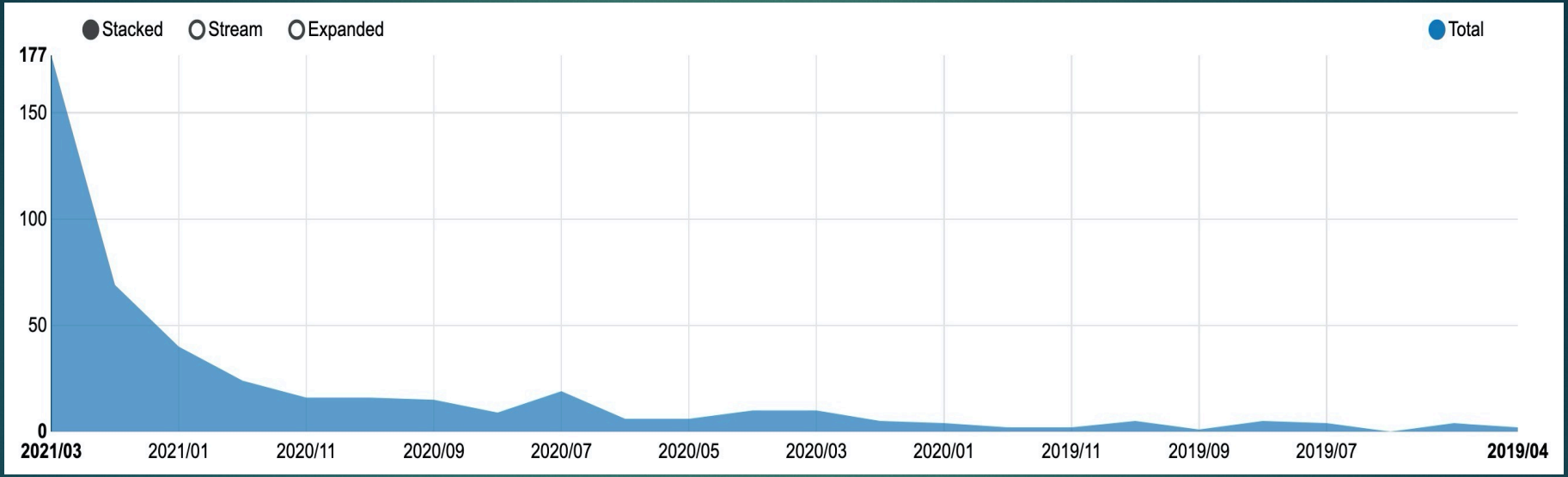
h. Thống kê số lượng bài báo đề cập đến những vấn đề không tốt (c****, g****, h****)



- Số lượng vấn đề không tốt có vẻ tăng

3. Phân tích dữ liệu

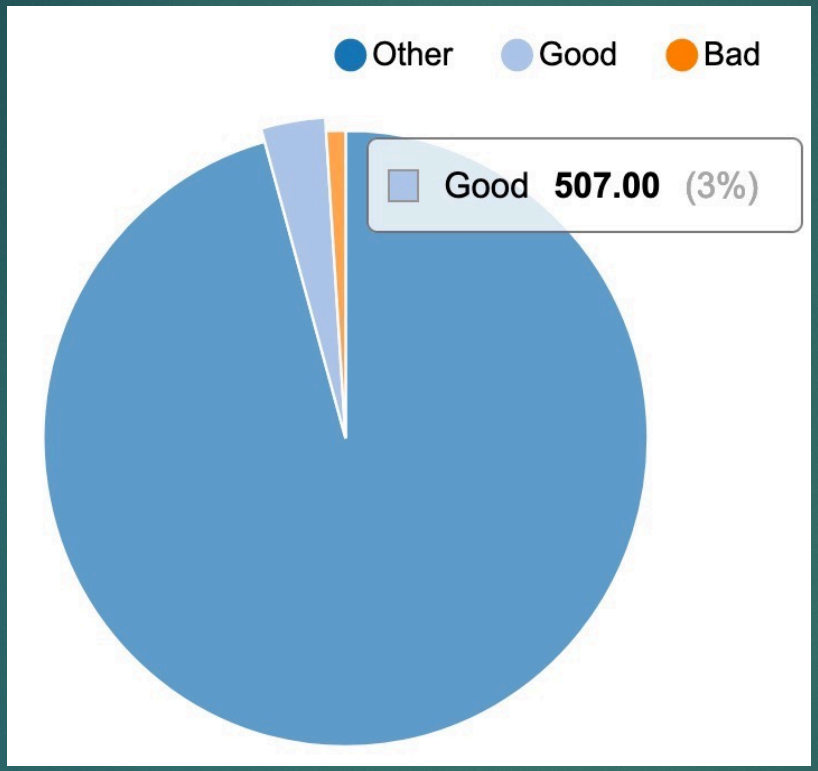
i. Thống kê số lượng bài báo đề cập đến những vấn đề tốt (tình yêu, niềm vui, hạnh phúc,...)



- Số lượng vấn đề tốt cũng có vẻ tăng

3. Phân tích dữ liệu

k. Tỷ lệ tin tốt và không tốt



- Nhìn chung số lượng việc tốt vẫn nhiều hơn việc không tốt

4. Cài đặt phân cụm

4. Cài đặt phân cụm

a. Cài đặt môi trường

b. Cài đặt master node

c. Cài đặt worker node

Q&A?

Thank you!