ICPSR

SHARING DATA TO ADVANCE SCIENCE

Developing a Social Media Archive at ICPSR

Libby Hemphill

Director, Resource Center for Minority Data

SOMAR Goals

- Evaluation
- Replication
- Novel analysis
- FAIR data principles

SOMAR Audiences

- Sociotechnical researchers studying social media
- SBE researchers using social media data
- Social science methodologists
- Digital archiving and curation researchers
- Researchers in the future

SOMAR Challenges

- Technical infrastructure
- Ethical and legal infrastructure
- Metadata enhancements
- Adoption

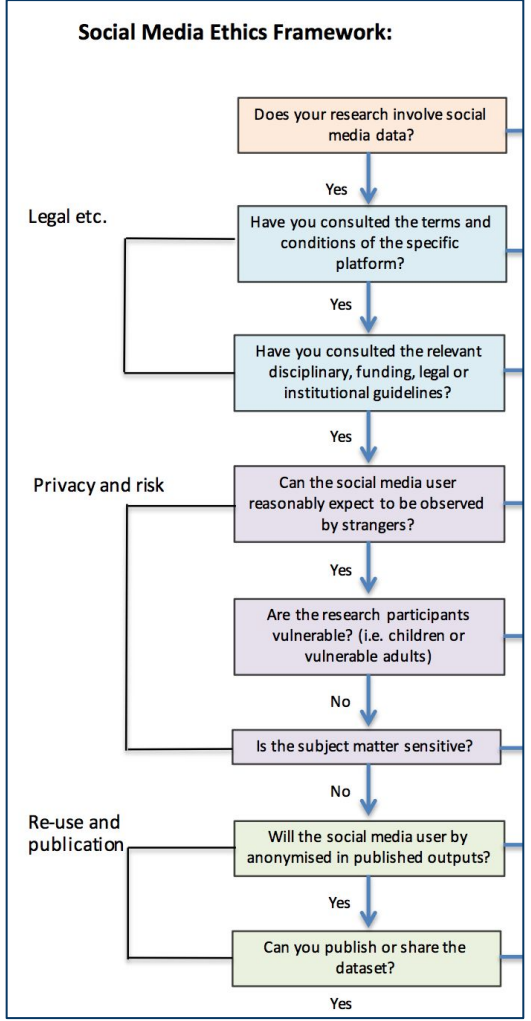
Ethical Considerations

Table 4. "How Would You Feel If a Tweet of Yours Was Used in a Research Study and . . ." ($n = 268$).

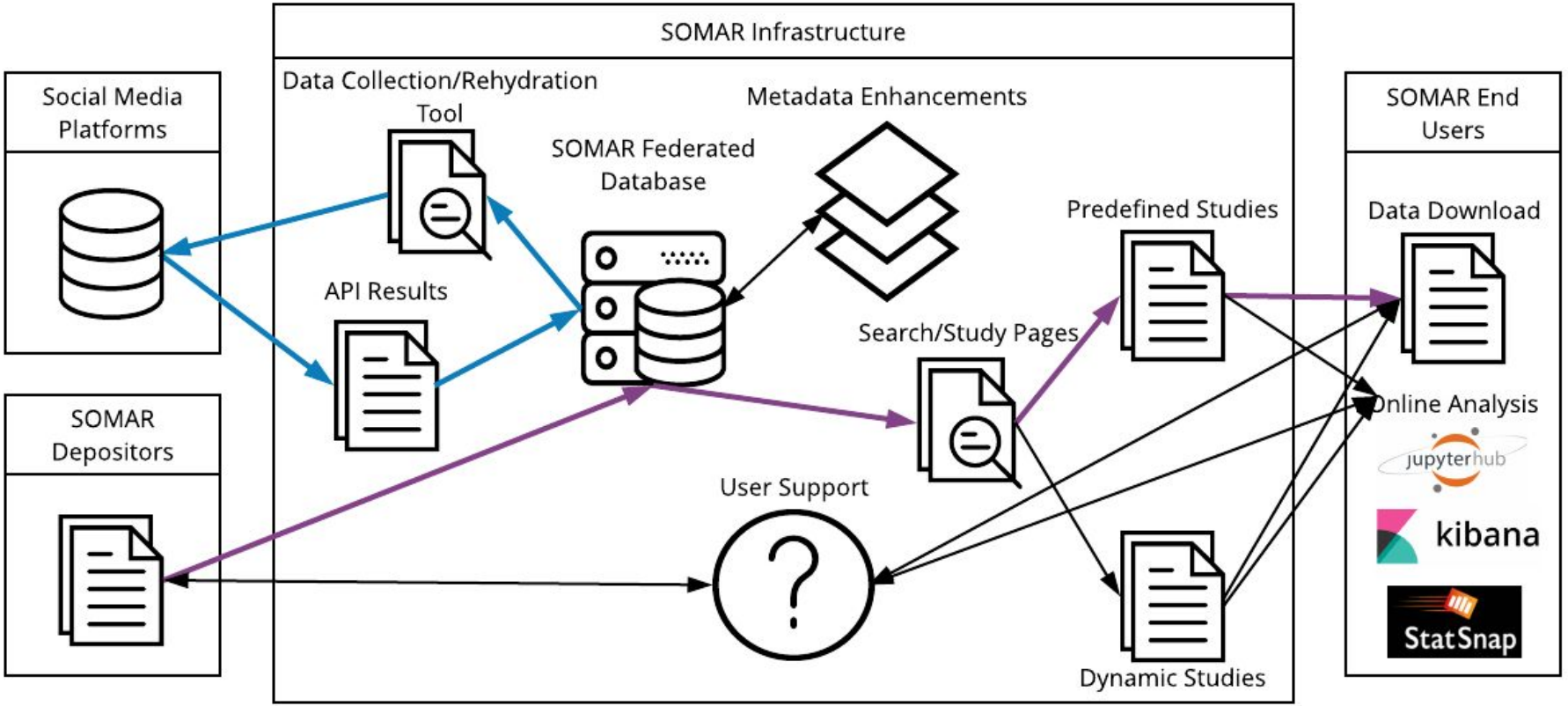
	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
. . . you were not informed at all?	35.1%	31.7%	16.4%	13.4%	3.4%
. . . you were informed about the use after the fact?	21.3%	29.1%	20.5%	22.0%	7.1%
. . . it was analyzed along with millions of other tweets?	2.6%	18.7%	25.5%	30.0%	23.2%
. . . it was analyzed along with only a few dozen tweets?	16.5%	30.3%	24.0%	20.2%	9.0%
. . . it was from your "protected" account?	54.9%	20.5%	13.8%	6.0%	4.9%
. . . it was a public tweet you had later deleted?	31.3%	32.5%	20.5%	10.4%	5.2%
. . . no human researchers read it, but it was analyzed by a computer program?	2.6%	14.3%	30.5%	32.3%	20.3%
. . . the human researchers read your tweet to analyze it?	9.7%	27.6%	25.0%	25.4%	12.3%
. . . the researchers also analyzed your public profile information, such as location and username?	32.2%	23.2%	21.0%	13.9%	9.7%
. . . the researchers did not have any of your additional profile information?	4.9%	15.4%	25.1%	34.1%	20.6%
. . . your tweet was quoted in a published research paper, attributed to your Twitter handle?	34.3%	21.6%	21.6%	13.1%	9.3%
. . . your tweet was quoted in a published research paper, attributed anonymously?	9.0%	16.8%	26.5%	28.4%	19.4%

Note. The shading was used to provide a visual cue about higher percentages.

Fiesler and Proferes 2018



Townsend and Wallace 2016


SOMAR Infrastructure
Social Media Platforms
Data Collection/Rehydration Tool
Metadata Enhancements
SOMAR Federated Database
API Results
Predefined Studies
Search/Study Pages
SOMAR End Users
Data Download
Online Analysis
jupyterhub
kibana
StatSnap
SOMAR Depositors
User Support
Dynamic Studies

cbuntain / TutorialSocialMediaCrisis

Watch 3 | Star 6 | Fork 4

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights

Branch: master | TutorialSocialMediaCrisis / notebooks / T04-08 - Twitter Analytics.ipynb | Find file | Copy path

cbuntain Updated for Python 3 and new Reddit API — 4657d8c on May 24, 2017

1 contributor

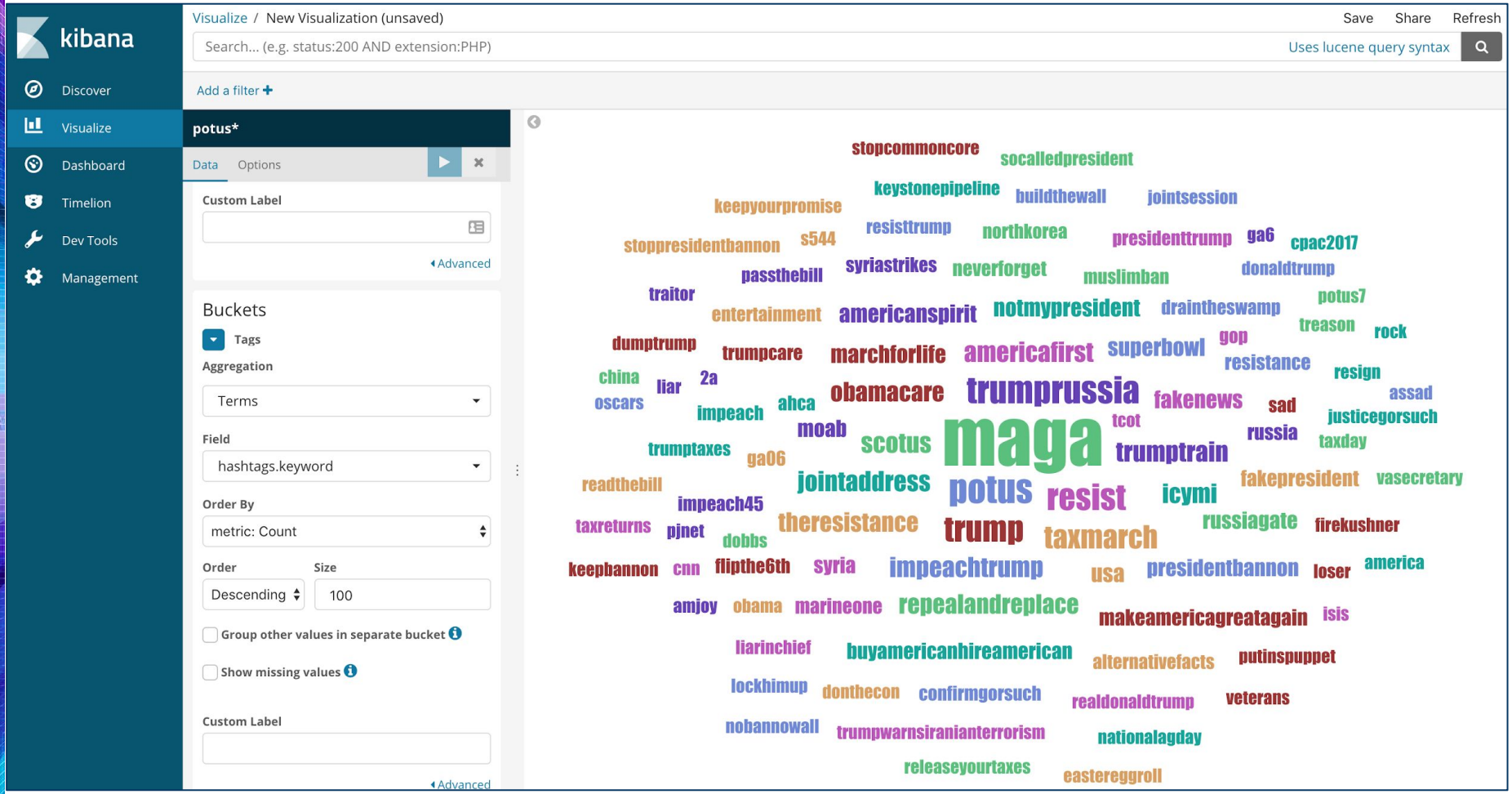
1762 lines (1761 sloc) | 59.4 KB | Raw | Blame | History

In [ ]:
```
%matplotlib inline

import time
import calendar
import codecs
import datetime
import json
import sys
import gzip
import string
import glob
import requests
import os

import numpy as np
```

Twitter Crisis Anlytics

The following notebook walks us through a number of capabilities or common pieces of functionality one may want when analyzing Twitter following a crisis. We will start by defining information for a set of events for which we have data.

In [ ]:
```
crisisInfo = {
    "boston": {
        "name": "Boston Marathon Bombing",
        "time": 1366051740, # Timestamp in seconds since 1/1/1970, UTC
                            # 15 April 2013, 14:49 EDT -> 18:49 UTC
```


kibana
Discover
Visualize
Dashboard
Timelion
Dev Tools
Management
Visualize / New Visualization (unsaved)
Save
Share
Refresh
Search... (e.g. status:200 AND extension:PHP)
Uses lucene query syntax
Add a filter +
potus*
Data
Options
Custom Label
Advanced
Buckets
Tags
Aggregation
Terms
Field
hashtags.keyword
Order By
metric: Count
Order
Descending
Size
100
Group other values in separate bucket
Show missing values
Custom Label
Advanced
stopcommoncore
socalledpresident
keystonepipeline
buildthewall
jointsession
keepyourpromise
resisttrump
northkorea
presidenttrump
ga6
cpac2017
stoppresidentbannon
s544
syriastrikes
neverforget
muslimban
donaldtrump
passthebill
traitor
potus7
entertainment
americanspirit
notmypresident
draintheswamp
treason
rock
dumptrump
trumpcare
marchforlife
americafirst
superbowl
gop
resistance
china
liar
2a
resign
assad
oscars
impeach
ahca
obamacare
trumprussia
fakenews
sad
justicegorsuch
moab
tcot
russia
trumptaxes
ga06
scotus
maga
trumptrain
taxday
readthebill
jointaddress
potus
resist
icymi
fakepresident
vasecretary
impeach45
taxreturns
pjnet
dobbs
theresistance
trump
taxmarch
russiagate
firekushner
keepbannon
cnn
flipthe6th
syria
impeachtrump
usa
presidentbannon
loser
america
amjoy
obama
marineone
repealandreplace
makeamericagreatagain
isis
liarinchief
buyamericanhireamerican
alternativefacts
putinspuppet
lockhimup
donthecon
confirmgorsuch
realdonaldtrump
veterans
nobannowall
trumpwarnsiranianterrorism
nationalagday
releaseyourtaxes
easteregroll

SOMAR Open Questions

- What about content that’s integral but not native to the social media post (e.g., links, images, videos)?
- What are the right metadata enhancements?
- How should SOMAR fit/model data management practices?
- How should we connect to existing collections and tools?
- How should we sustain the enterprise?

SOMAR Next Steps

- SFM test bed on ICPSR infrastructure
- Tweet Sets test bed on ICPSR infrastructure
- Reddit harvester for SFM
- Automate rehydration loop
- Kibana exploration
- Jupyter connector to ElasticSearch
- URL expansion and web archiving

More Resources

- ICPSR Data Management and Curation
- FAIR principles for scientific data management
- Documenting the Now
- Social Feed Manager
- Tweet Sets
- GESIS Twitter Dataset example
- GETAR