Healthcare Twitter Analysis

Working Document

George Fisher

August 7, 2014

Abstract

'Healthcare Twitter Analysis' is an Open Source project which intends to investigate ways to improve the quality of medical care with Data Science techniques applied to Twitter. This paper is the record of one participant's progress. The project website is [Mehta and Saama Technologies, 2013]; the GitHub repository for this paper can be found at [Fisher, 2014].

Contents

Part I

Summary

This is just a status update, the project is still in its initial stages.

- **Prior Status**
  It was clear upon joining the project that the data provided would have to be augmented, both by the data from the original tweets and from data elsewhere.

  Prior to building their own NoSQL database, Twitter used MySQL but MongoDB (or maybe CouchDB) are clearly superior to the MySQL implementation.

  I have done some simple Exploratory Data Analyses. I am sorry to say that while these analyses produced interesting pictures, I do not see any obvious connection to improving medical research.

- **Current Status**
  The conversion for full Twitter json in Amazon Web Services (AWS) is underway. With numerous false starts, the process has been running for 5 days and has processed 403 of 896 files.

  Word Frequency analyses have been included:

  - Senator Patty Murray and ex-Senator Scott Brown make the top 25 list of users mentioned in the entire database, so does a UK soccer club.
  - Latent Dirichlet Allocation was run on a fairly large subset and it accurately identified the disease categories.
  - The most-common 1-, 2- and 3-grams associated with each of the project's hashtags were identified.

  I've investigated several other projects with objectives that seem similar to our own to try to learn what they did. Primarily what they seem to do is search for hashtags and phrases that match the conditions they're looking for. Not particularly sophisticated (despite being called Artificial Intelligence) but I'm going to start looking at that, as well as looking at network diagrams.

Part II

Data Acquisition and Management

1 Collect Twitter Data

The first step was to add all the Twitter data to the files provided by the project.

There are 896 csv files provided by Topsy, a Twitter aggregator [Topsy, 2014], containing well over 6 million tweets [Google Drive, 2014]. The files fairly comprehensively cover tweets concerning a wide range of medical conditions, for a six-month period.

However the data included only the text of the tweet, its originating user and a score calculated by Topsy. While the text might be sufficient for a basic textual analysis, the other data provided by Twitter is clearly of value for more extensive analyses, even as simple as filtering by retweet count or plotting geographic incidence.

My GitHub repo for this project [Fisher, 2014] contains a python program that performs two basic tasks:

1. For each tweet in the Topsy data, requests the full json from Twitter
2. For each record it adds
    - All of the data from the Topsy files
    - Location data, including latitude and longitude from Mapquest [Mapquest, 2014]

The additional json fields included are listed in an appendix on page 29.

I transferred all the project files to Amazon Web Service's EC2 service and ran the python program against them all, producing a file that anyone can download for their own use from S3. See the appendix on page 32 for details of the files.

2 CSV vs. JSON

Initially, I focused on creating csv files with this data, and the programs to do so are still on the repo, but after studying Twitter analysis I became convinced that json was more appropriate for two reasons:

1. Every book and paper I have read and every Twitter-analytic program refers to the Twitter data in its json form

2. While MongoDB [MongoDB, 2014] supports csv files, including a utility for csv loading, it is clear that MongoDB's native document structure is that of json and since MongoDB seems like a very useful way to store and access Twitter data, it being the one chosen by most other researchers, storing the data in json format seemed to make the most sense.

3 Text, SQL, NoSQL

On the assumption that this project unearths some really useful analytics that can help medical science, it will need to address the question of the best way to store the data. We're using text files at the moment which are easy to use but they get clumsy quickly.

Through 2010 Twitter used MySQL for its data storage. ([highscalability.com, 2011, Wired, 2014, Quora, 2012]). Subsequent volume growth and the need to serve data from many locations worldwide prompted Twitter to build its own NoSQL database [Twitter, 2014a, Computerworld, 2010].

Initially I thought we could consider MySQL since Twitter itself had used it but the way it was used was to store the json in a single long text field and to use UDFs to parse it. This is clearly inferior to using MongoDB which does essentially the same thing but is specifically built for indexed json queries.

MongoDB, like all NoSQL datastores, requires map/reduce to perform join operations. To use json, therefore, we must load into MongoDB records that have all the data we need for each record. To the extent that we find useful data in addition to the Twitter data, it will have to be incorporated in the json, one way or another.

I have MongoDB on my machine and access to it on AWS; it's on my to-do list to learn MongoDB well enough to have an informed opinion about it and maybe actually do something useful with it.

4 Supplemental Data

Finding additional data for the tweets to allow more extensive analyses is clearly a very important area of research but I am aware of only one effort in the group to do this.

4.1 Ontologies

Tim Cook [Cook, 2014] has begun work on adding ontology data from BioPortal [BioPortal, 2014].

5 Online Twitter Access

5.1 Online with Python

The main section of the repo contains `Instructions for python.pdf` which provides instructions for setting up Python, IPython and installing the prerequisites for online Twitter access.

The `code` folder of the repo contains an IPython notebook `Online Twitter Basics.ipynb` that walks through the process of making online-queries of Twitter and doing simple analyses of the responses. From the notebook you can combine the static project data with real-time queries.

5.2 Online with R

The main section of the repo contains `Instructions for r.pdf` which will get you set up for online Twitter access from RStudio, which is where I did most of the analyses in this document, some of it using the static project data, some it doing real-time Twitter queries.

Part III

Exploratory Data Analysis

6 Sentiment

6.1 Summary of Sentiment Analyses

While the pictures are very pleasing, it is not at all clear to me how rudimentary sentiment analysis will provide any value to medical researchers; presumably there are deeper, more sophisticated techniques that provide more useful insights.

6.2 Analysis of Breen, AFINN, Score

There are two sentiment measuring systems which popped up in my initial studies of the subject: Jeffrey Breen's [Breen, 2011b, Breen, 2011a] and AFINN [Nielsen, 2011]. In addition, the Topsy data includes a measure called score [Topsy, 2010].

I wondered how the two sentiment measures compared to each other and whether sentiment and score had any relation. Loading the data on Cancer, Cardiovascular and Digestive into R, I had a look:

	breen	afinn	score
min	-6.00	-10.00	6.02
mean	0.00	0.50	8.36
median	0.00	0.00	7.58
stdev	1.23	2.10	1.66
skew	0.00	0.65	1.05
npskew	0.00	0.24	0.47
kurtosis	0.85	2.76	-0.15
max	6.00	16.00	14.62

Table 1: **Statistical Comparison of Sentiments and Score**

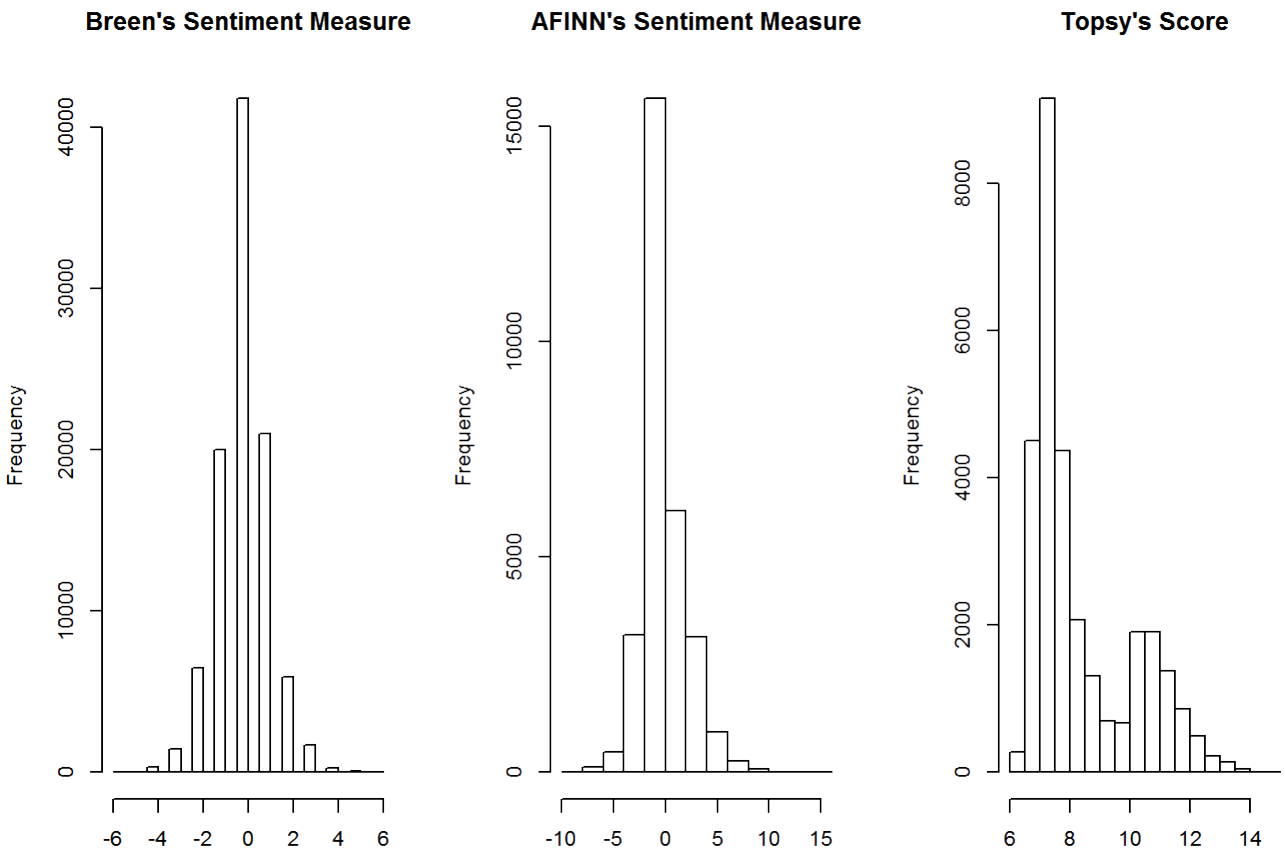


Figure 1: **Distribution of Sentiment Measures and Score** Breen and AFINN are more similar to each other than to score: both have a mean of nearly zero and both are symmetrical around it; but AFINN has a much greater variance and non-normal tail behavior. Score has more of a log or Poisson shape to its distribution, which is bimodal, and is clearly different from the two sentiment scores.

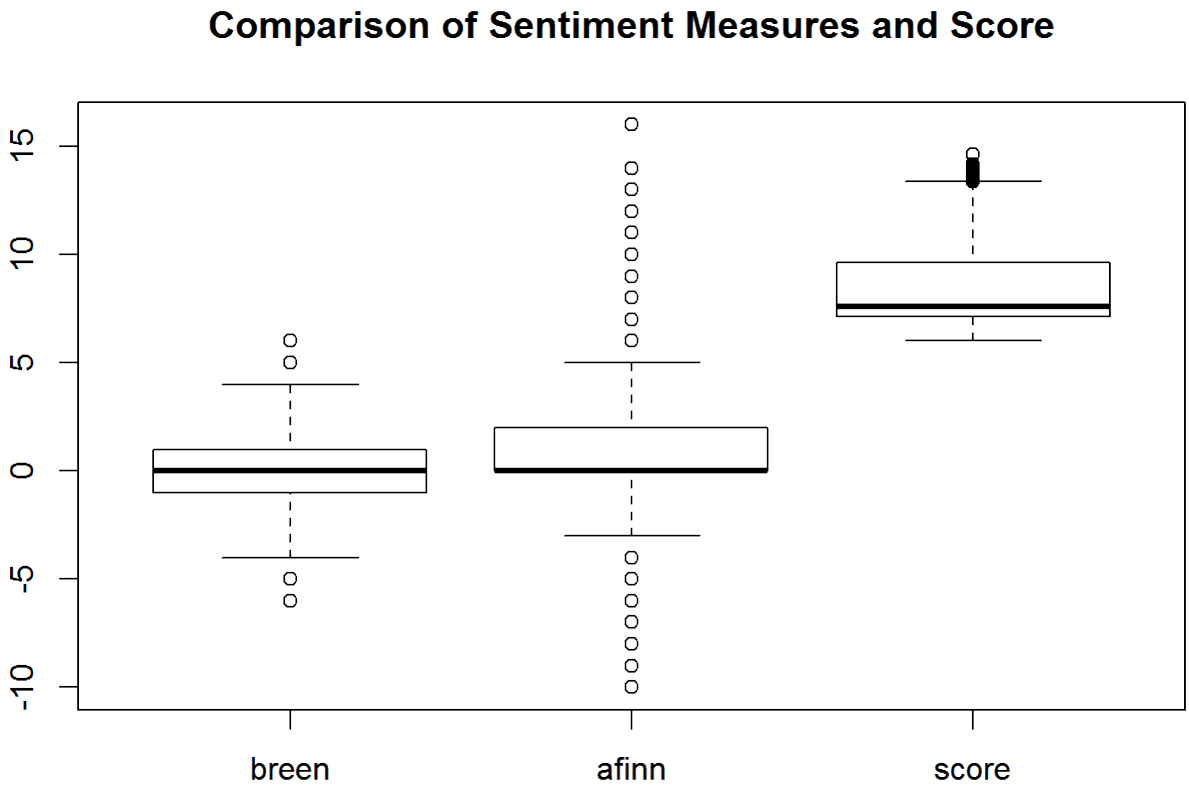


Figure 2: **Distribution of Sentiment Measures and Score** A box plot shows even more starkly the difference in the distributions of these three measures.

First It would seem that score is not created from or predicted by either sentiment measure.

Second The question arises as to which sentiment measure is preferable, if indeed either is adequate: AFINN has a much greater dispersion of its measures, which perhaps is to be expected when dealing with life-destroying diseases; on the other hand, Breen produces a more-nearly-normal distribution and by some accident of Providence, most naturally-occurring phenomena are normally distributed, perhaps including peoples' feelings.

6.3 Digging Deeper into Sentiment Measures

Gaston Sanchez wrote a series in 2012 about Twitter analysis [Sanchez, 2012]. His work provides an interesting overview of general summary analyses that people do on Twitter data and I have reproduced some of his work here, using R and the Breen sentiment scoring system [Breen, 2011b], with data from this project in four (randomly-chosen) categories :

1. Blood Disorders
2. Cancer
3. Cardiovascular Diseases
4. Digestive Disorders

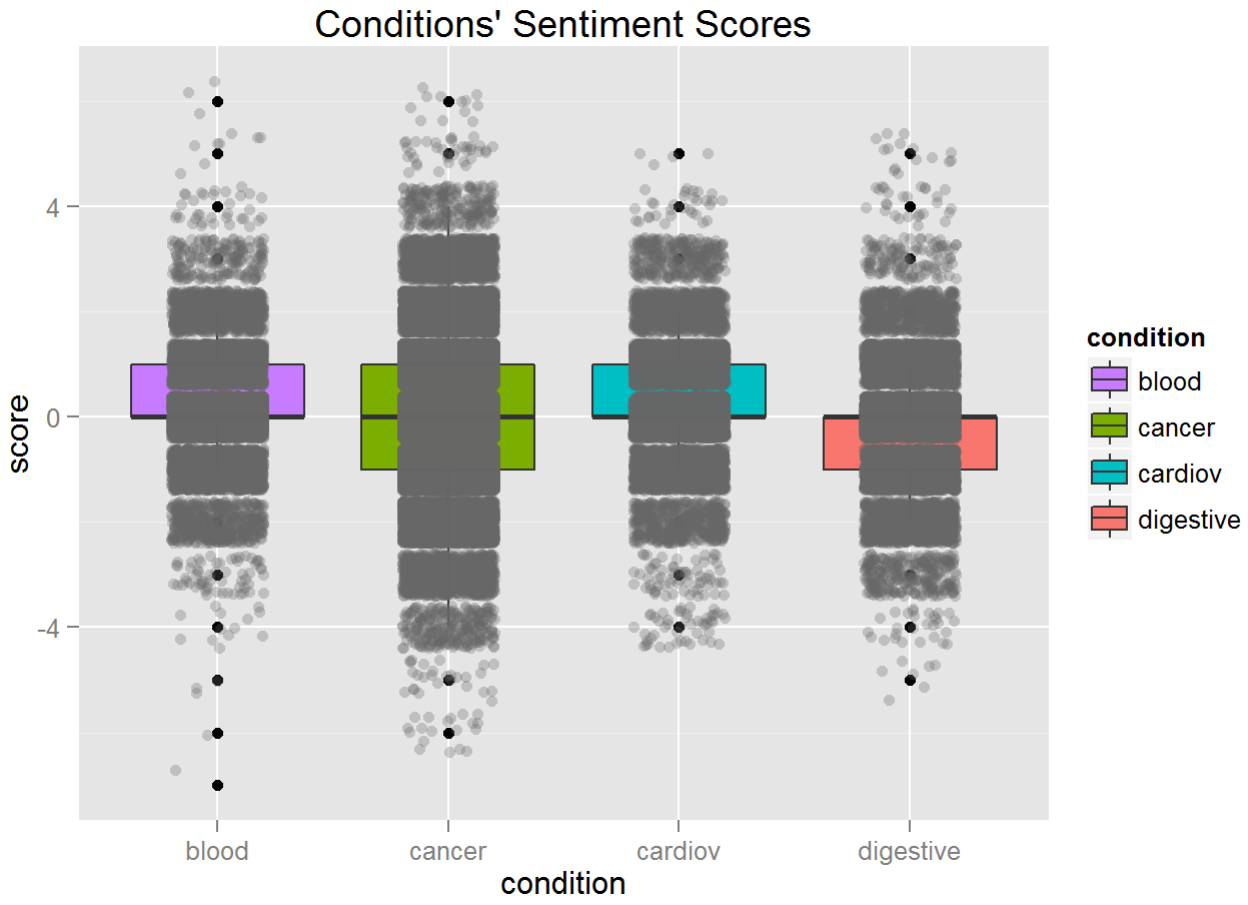


Figure 3: **Distribution of sentiment: Boxplots** The dark gray dots represent the individual data points, roughly 14,000 per condition. The boxes in color represent the inter-quartile distribution of the sentiment for each condition, with bold dots above and below representing outliers beyond the inter-quartile ranges.

They all have their median nearly at zero with a very wide dispersion in both the positive and negative direction. Blood and Cardiovascular disorders seem to be somewhat skewed toward positive overall sentiment while Digestive disorders are skewed toward the negative ranges.

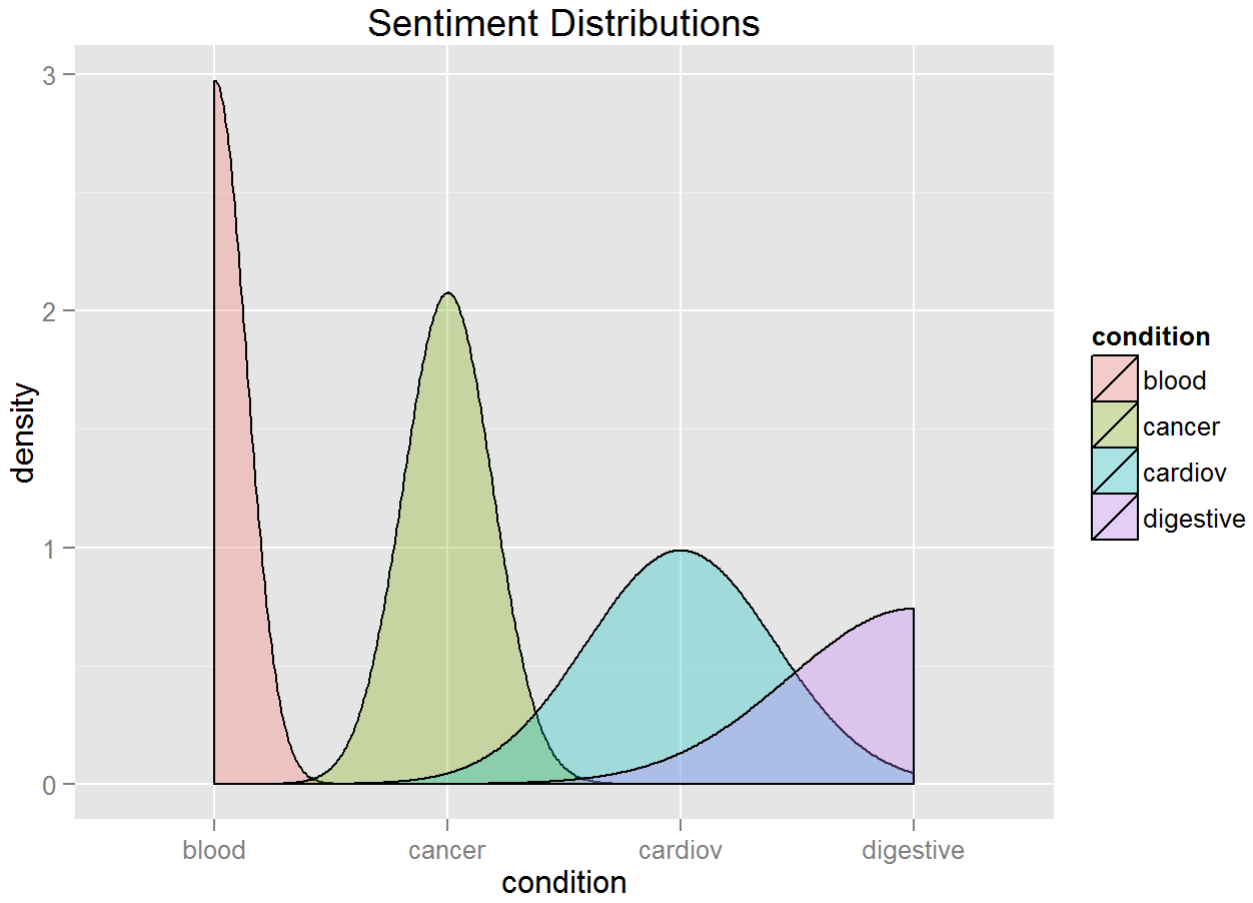


Figure 4: **Distribution of sentiment: Histograms** Another way to look at the distribution of sentiment is to show a smoothed histogram. For each condition, the vertical white line over the label is plotted over the average for that category and the plot shows the distribution around the mean although the left-tail of Blood and the right tail of Digestive are not plotted due to size constraints but they are roughly symmetric. In the study of sentiment measures in section 6.2 beginning on page 8, it was shown that the Breen sentiment measure is symmetric in general and the measures for these specific conditions reflect that.

Blood is in a tight range around its mean, while Digestive has the greatest dispersion.

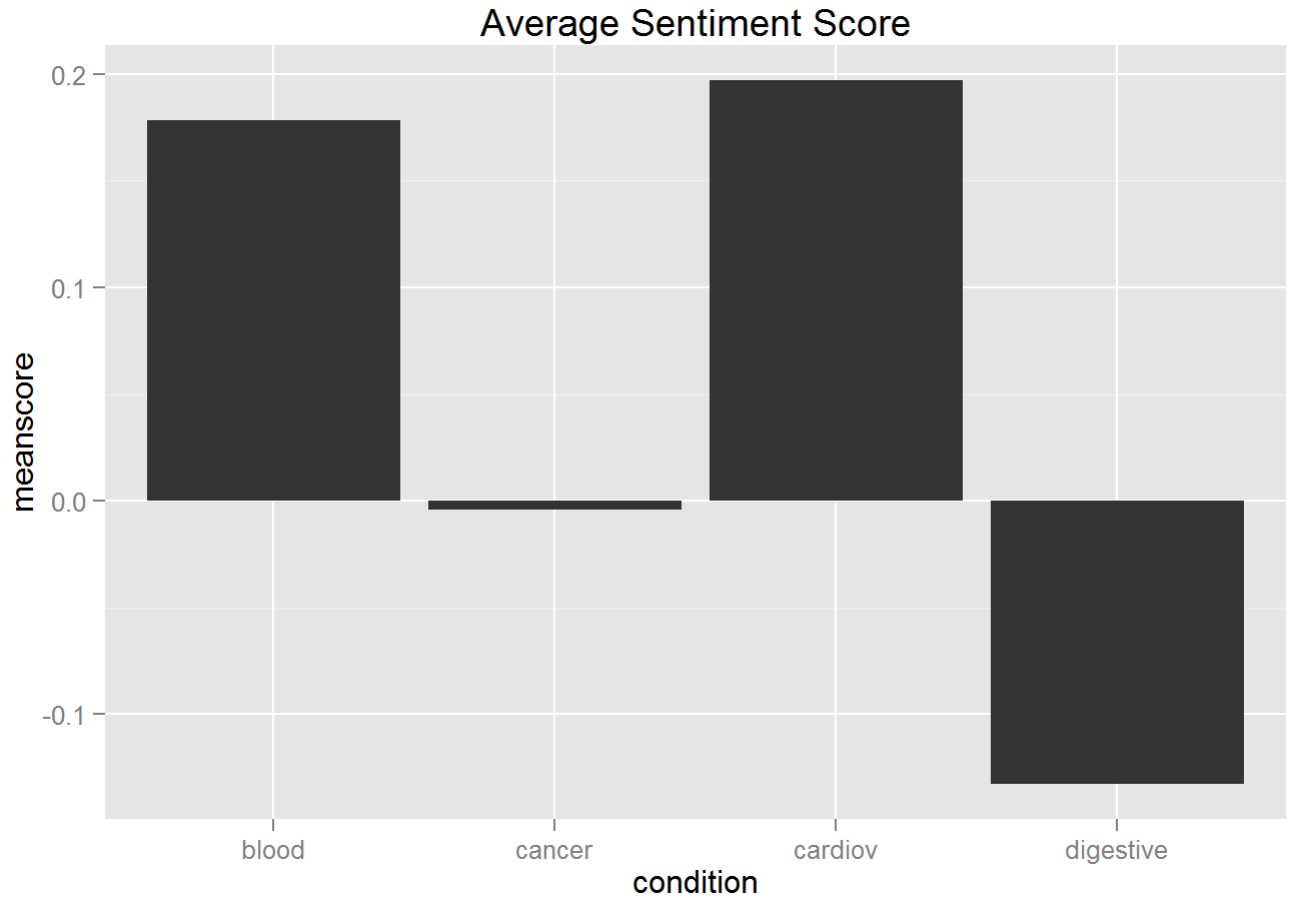


Figure 5: **Average Scores** The averages show us very starkly what we saw in the distributions: Digestive disorders seem to have by far the most negative effect on their suffers and/or those who tweet about them.

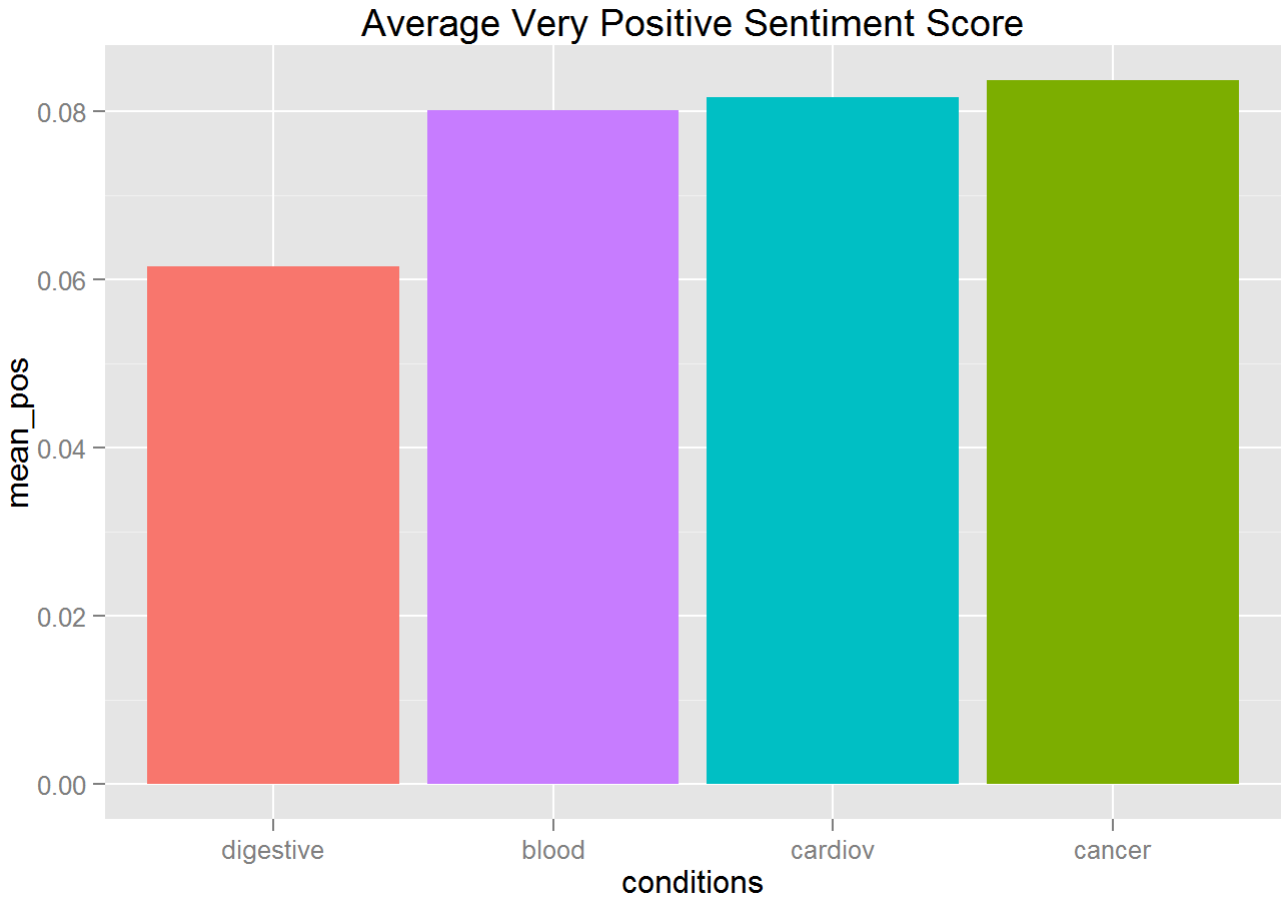


Figure 6: **Average Positive Scores** Looking at the mean scores for only those with a positive sentiment provides more reinforcement for what we have already seen: digestive disorders have a negative psychological effect to the extent of having the lowest mean positive scores.

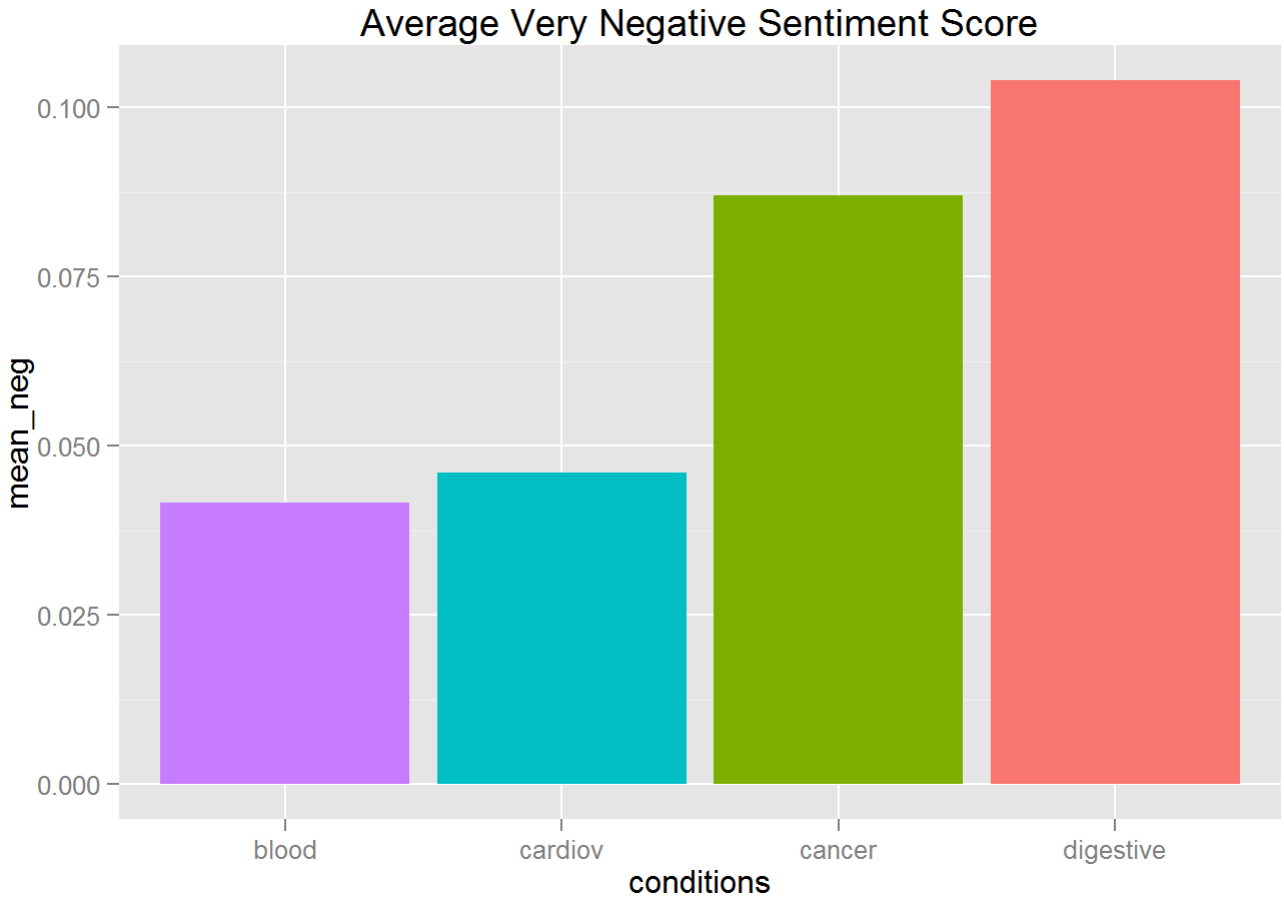


Figure 7: **Average Negative Scores** Looking at the mean scores for only those with a negative sentiment tells the same story: none are good, but of these four, tweets about Digestive Disorders show the greatest tendency toward negativity.

7 Analysis of Text

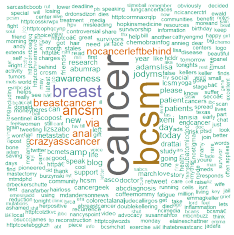
7.1 Summary

Similar to my observation about Sentiment Analysis, I find the pictures for the simple, common textual analyses interesting but I do not see a connection to helping medical research. If value is to be added in this area it will have to come either from the inclusion of additional tags found outside the tweets themselves or else from more sophisticated techniques; perhaps both.

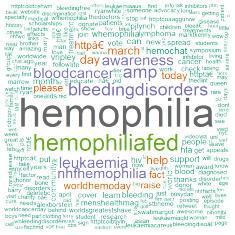
7.2 Word Clouds

Word Clouds are a very popular EDA technique for text and again with help from Gaston Sanchez [Sanchez, 2012] I have produced a sampling with R and datasets created using the technique described in section 1 starting on page 5.

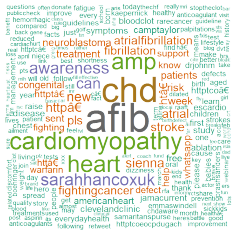
The corpus was restricted to the first 10,000 tweets in the database for each condition and then further reduced to include only those that had been retweeted more than three times; without these filters the pictures were an incomprehensible mess.



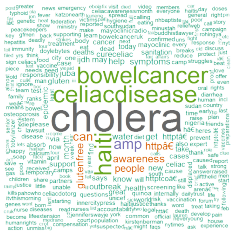
(a) Cancer



(b) Blood Diseases



(c) Cardiovascular



(d) Digestive

Figure 8: Simple Word Clouds for Four Medical Conditions

(a) Comparative (b) Commonality

Figure 9: **Comparison Word Clouds** show the words specific to the individual conditions. **Commonality Word Clouds** show the words that tweets about the four conditions have in common.

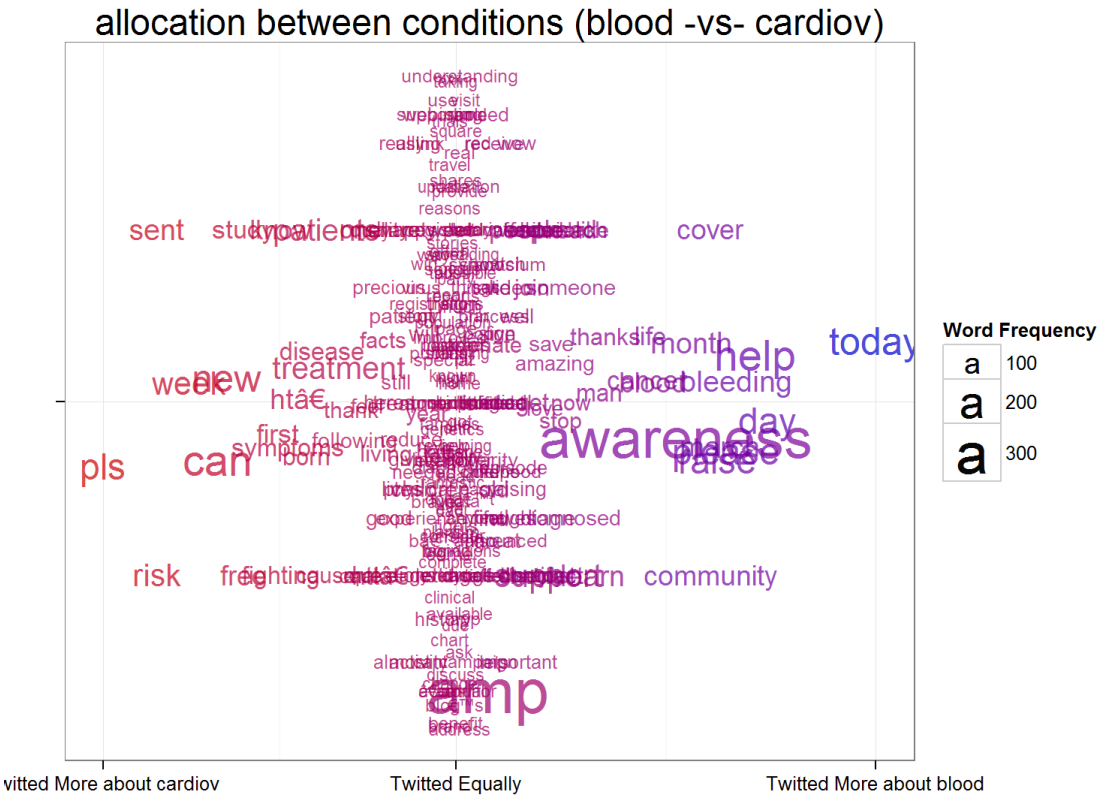


Figure 10: **Conway Comparative Word Cloud of Two Medical Conditions** Comparative word clouds compare all categories together. Conway word clouds show how two categories allocate words between them.

7.3 Word Frequency

The text field of a tweet has four kinds of 'tokens':

- Hashtags, beginning with '#', indicating a topic
- User Mentions, beginning with '@', indicating a message to/about about a particular user
- URLs, links to other pages or media
- Words, including some emoticons

I have parsed every text field in the database into these four token types, removing stop-words and nuisance strings such as 'rt' in the case of words, looking at the various frequencies of tokens:

7.3.1 Overall Frequencies

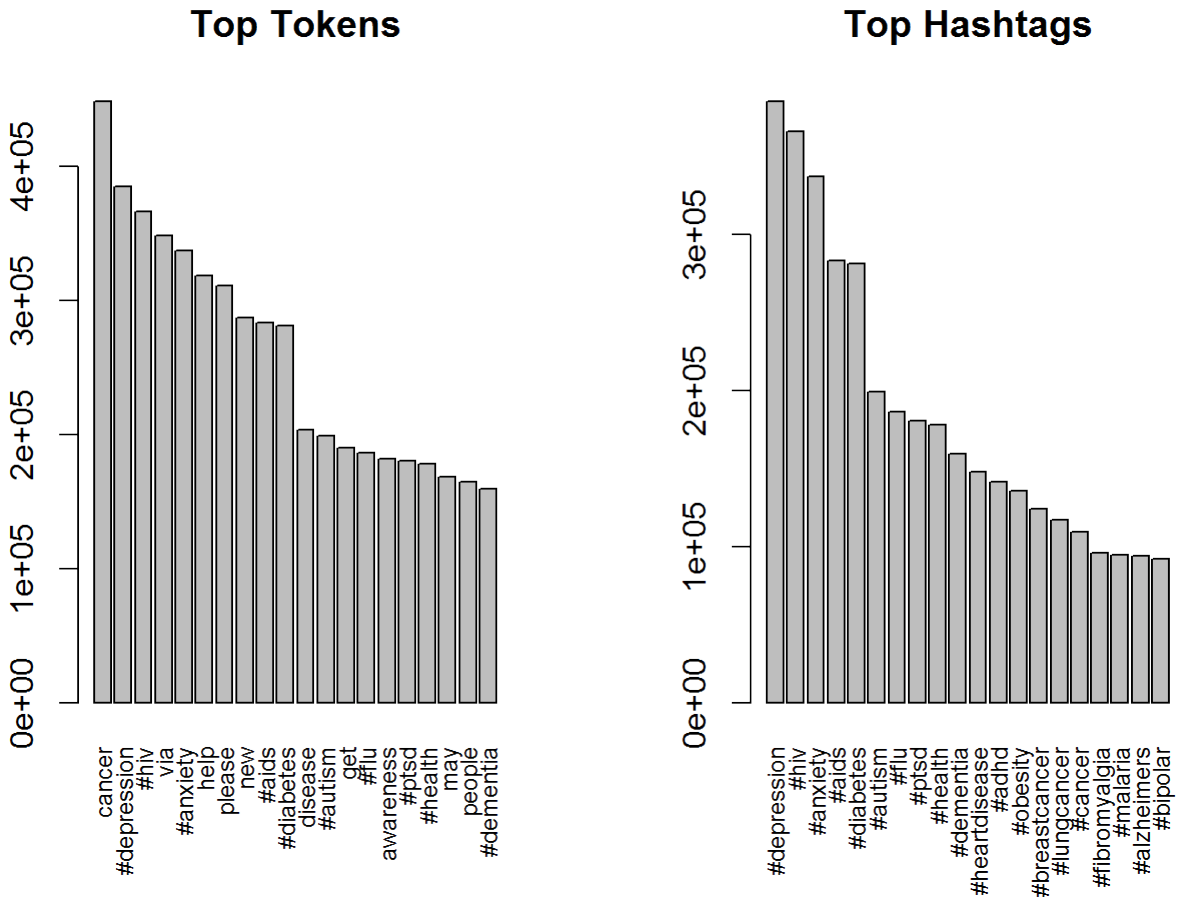


Figure 11: **Most Common Tokens Overall and Most Common Hashtags** The following hashtags are among the top hashtags mentioned in the entire dataset but are not on the list provided by the project:

- #health
- #cancer
- #mentalhealth
- #fibro
- #love
- #pain
- #awareness
- #veterans
- #asd
- #spoonie
- #disability
- #sex
- #weightloss
- #stress
- #glutenfree
- #advice
- #dating

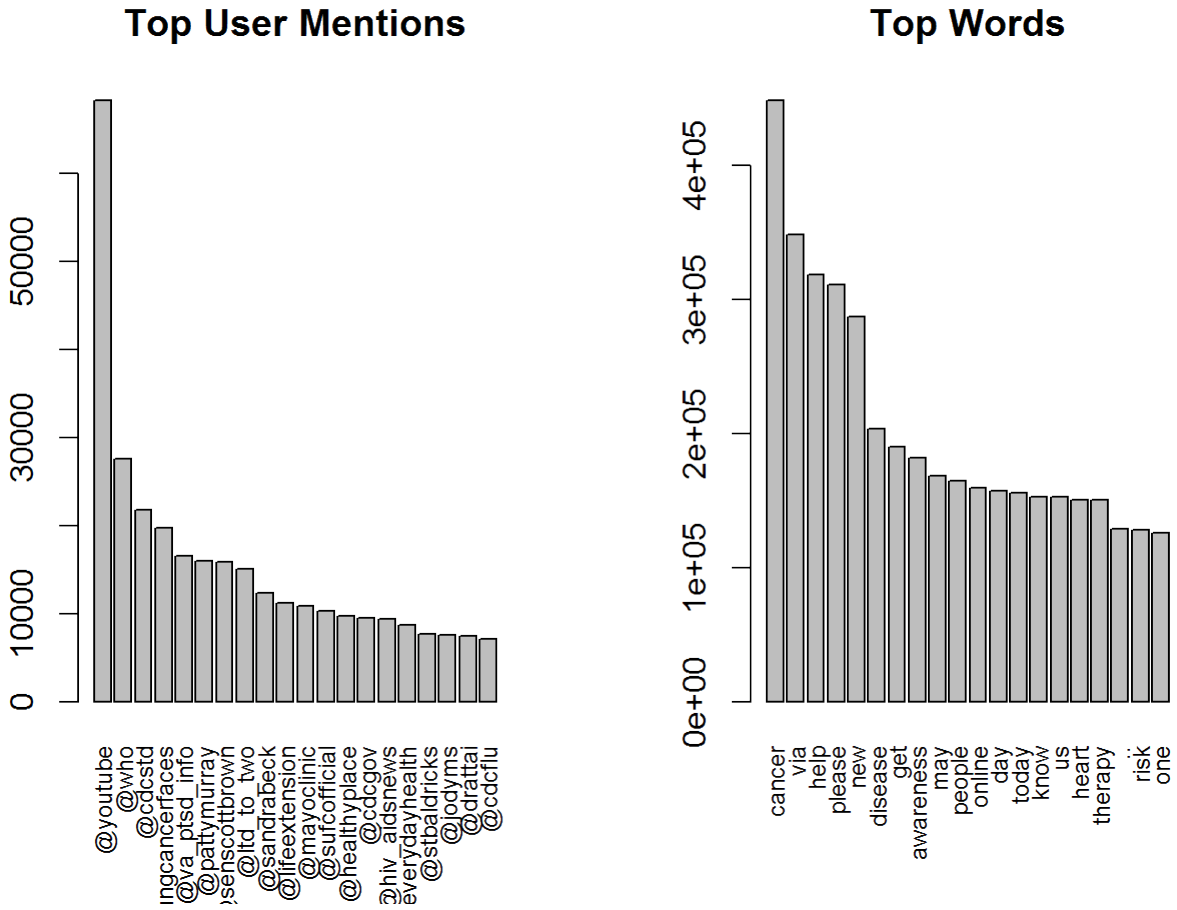


Figure 12: **Most Common Users Mentioned and Most Common Words Used**

Screen Name	Twitter Description
@youtube	Tweets on news, music and trends from all your favorite channels.
@who	Official Twitter account of the World Health Organization
@cdcstd	Helping people to be safer and healthier by the prevention of STDs
@lungcancerfaces	Faces of Lung Cancer
@va_ptsd_info	National Center for PTSD
@pattymurray	Senator Patty Murray
@senscottbrown	Scott P. Brown
@ltd_to_two	Multiple Sclerosis (PRMS), Fibro may have limited me but it can't destroy me.
@sandrabeck	Sandra Beck #TalkRadio Host #divorce #death #illness #recovery #faith #spiritu
@lifeextension	The latest research on health, wellness, nutrition, & aging.
@mayoclinic	The Mayo Clinic
@sufcofficial	Official Twitter site of Scunthorpe United football club.
@healthyplace	Trusted information on psychological disorders and treatments,
@cdcgov	Centers for Disease Control & Prevention
@hiv_aidsnews	News and developments in the global fight against HIV and AIDS.
@everydayhealth	Powerful weight-loss tools, expert advice & health news and information.
@stbaldricks	Charity funding the world's most promising research to #ConquerKidsCancer.
@jodyms	Writer, blogger. Optimist. Cancer Advocate.
@drattai	Breast Surgeon, President-Elect of @ASBrS
@cdcflu	Flu-related updates from the Centers for Disease Control & Prevention.
@icombat_stress	Motivational Mentor. Hope. Help. Healing. You CAN Turn Your Life Around.
@pozmagazine	The premier HIV/AIDS advocacy
@mndassoc	The Motor Neurone Disease Association.
@clevelandclinic	The Cleveland Clinic
@alldiabetesnews	The Most Comprehensive Diabetes News Aggregator on the Web.

Table 2: **Top Users Mentioned** One current and one ex Senator make the top users mentioned? A soccer club? . . . must have been gathered during the World Cup.

7.3.2 Co-Occurrence With Top Hashtags

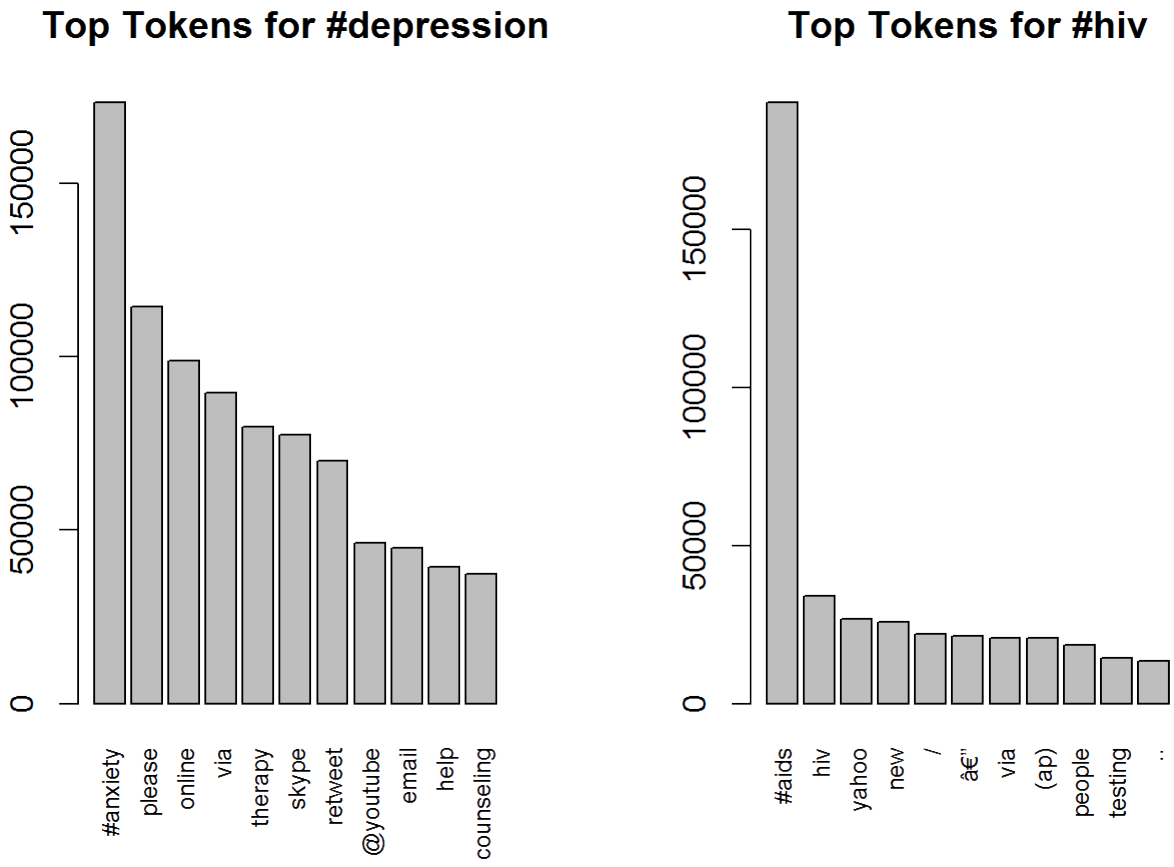


Figure 13: **Tokens Most-Commonly Co-Occurring With Two Top Hashtags**

7.4 Latent Dirichlet Allocation

I loaded 40,000 tweets of the Blood category into a matrix in R and asked it to tell me the topics; it did a pretty good job: it said there were three:

- sepsis
- myeloma
- hemophilia

True, but unedifying. Perhaps there's a better use for this tool.

7.5 1-, 2-, and 3-Grams for each Hashtag

The following are samples of three csv files in the repo that contain the most-common 1-, 2- and 3-grams associated with each of the hashtags for this project:

hashtag	1-gram	count
rettsyndrome	help	724
influenza	flu	5826
caudaequina	syndrome	20
schizofrenie	van	7
bedwetting	child	95
epilepsy	help	3913
dysautonomia	sharing	915
ppd	postpartum	1089
eds	awareness	1402
sarcoidosis	via	406
trichotillomania	hair	362
afib	atrial	1036
gallbladder	pain	599
testicularcancer	awareness	861
hernia	surgery	123

hashtag	2-gram	count
rettsyndrome	awareness for	250
influenza	out stories	990
caudaequina	please watch	7
bedwetting	your child	59
epilepsy	check out	878
dysautonomia	for sharing	843
ppd	postpartum depression	508
eds	ehlers-danlos syndrome	562
sarcoidosis	news daily	334
trichotillomania	check out	106
afib	atrial fibrillation	931
gallbladder	can cause	274
testicularcancer	to check	225
hernia	detailed general	30

hashtag	3-gram	count
rettsyndrome	$ awareness for	220
influenza	is out stories	990
caudaequina	please watch share	7
schizofrenie	dialoog finse blijkt	3
bedwetting	fitted mattress cover	23
epilepsy	thanks for the	649
dysautonomia	thanks for sharing	760
ppd	should feel ashamed	165
eds	info 085251378519 atau	352
sarcoidosis	news daily review	330
trichotillomania	support this eye-opening	90
afib	with atrial fibrillation	156
gallbladder	can cause severe	273
testicularcancer	about going through	156
hernia	general surgery videos	30
incontinence	disposable pads shaped	211

8 Time Series Analyses

8.1 Historical Time Series

8.2 Real-Time Time Series

9 Network Analyses

Appendices

A Other Medicine-Related Twitter Projects

- How Twitter Is Studied in the Medical Professions:
  A Classification of Twitter Papers Indexed in PubMed
  [Williams et al., 2013a]
- What do people study when they study Twitter?
  [Williams et al., 2013b]
- Pandemics in the Age of Twitter:
  Content Analysis of Tweets during the 2009 H1N1 Outbreak
  [Chew and Eysenbach, 2010]
- The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic
  [Signorini et al., 2011]
- The potential of social networks for early warning and outbreak detection systems:
  The swine flu Twitter study
  [Kostkova et al., 2010]
- Using Twitter and other social media platforms to provide situational awareness during an incident
  [Tobias, 2011]
- The other Twitter revolution:
  How social media are helping to monitor the NHS reforms
  [McKee et al., 2011]
- A visual backchannel for large-scale events
  [Dork et al., 2010]
- Dissemination of health information through social networks:
  Twitter and antibiotics
  [DScanfeld et al., 2010]
- Twitter as a communication tool for orthopedic surgery.
  [Franko, 2011]
- Machine intelligence for health information:
  Capturing concepts and trends in social media via query expansion
  [Su et al., 2011]

- Social Internet sites as a source of public health information
  [Vance et al., 2009]
- Hospitals are finding ways to use the social media revolution to raise money, engage patients and connect with their communities
  [Galloro, 2011]
- Twitter mining for fine-grained syndromic surveillance
  [syn, 2014]
- Now Trending #health In My Community
  [Department of Health and Human Services, 2012, US Dept. pf Health & Human Services, 2012]
- Physicians On Twitter
  [Sabine Tejpar et al., 2011]
- Agencies Use Social Media to Track Food-born Illnesses
  [BM, 2014]
- Social media in vascular surgery
  [Indes et al., 2013]
- Decoding Twitter:
  Surveillance and trends for cardiac arrest and resuscitation communication
  [Bosley et al., 2012]
- Twitter as a tool for ophthalmologists
  [Micieli and Micieli, 2012]
- Dissemination of health information through social networks
  Twitter and antibiotics
  [Scanfeld et al., 2010]
- All Atwitter About Radiation Oncology:
  A Content Analysis of Radiation Oncology-related Traffic on Twitter
  [Jhawar et al., 2012]

B fields_added_to_twitter_json.txt

This file shows examples of the json fields added to the full Twitter data by the python program `get_twitter_json.py`, which can be found in the `code` folder on GitHub [Fisher, 2014]. The program's docstring contains examples of using the file this program produces in Python and R as well as loading it into a MongoDB collection. A sample file produced can be found in the `files` folder of the GitHub repo.

The official guide to Twitter's json structures is here: [Twitter, 2014b]; the guide to Mapquest programming is here: [Mapquest, 2014].

The reason for adding the Unix timestamps is for efficient searching in MongoDB; I expect these dates will be part of an eventual index structure and dates in text format are useless for this.

Note: a spot check of the geo tagging says it's pretty good. However, while 'Toronto Canada' gives what you'd expect, 'Toronto' by itself serves up a little town in Ohio; also, while a human might make something of 'Vancouver & Caracas', my program cannot. Caveat Emptor.

```
get_twitter_json.py augments the Twitter json with the following fields
========================================================================

"timestamp": 1389010334.0              # unix timestamp for Twitter's 'created_at' field

"user": {

    "location": "Perth, Western Australia",                    # given

    "location_geoinfo": {                                      # derived
            "sideOfStreet": "N",
            "linkId": "282859405",
            "mapUrl": "http://www.mapquestapi.com/staticmap/v4/getmap?key=Fmjtd|luur2008n9,
            "displayLatLng": {
                "lat": -31.952712,
                "lng": 115.86048
            },
            "adminArea6Type": "Neighborhood",
            "adminArea3Type": "State",
            "dragPoint": false,
            "geocodeQualityCode": "A5XAX",
            "adminArea1": "AU",
            "geocodeQuality": "CITY",
            "adminArea3": "Western Australia",
            "adminArea4": "",
            "adminArea5": "Perth",
            "adminArea6": "",
            "postalCode": "",
            "type": "s",
            "adminArea1Type": "Country",
            "adminArea5Type": "City",
            "latLng": {
                "lat": -31.952712,
                "lng": 115.86048
            },
            "adminArea4Type": "County",
            "unknownInput": "",
            "street": ""
```

```
        },

"topsy": {

    "firstpost_date": "01/06/14",
    "timestamp": 1388984400.0,          # unix timestamp for "firstpost_date" field

    "url": "http://twitter.com/primary_immune/status/420090415086198784",
    "score": 7.2846317,
    "trackback_author_nick": "primary_immune",
    "trackback_author_url": "http://twitter.com/primary_immune",
    "trackback_permalink": "http://twitter.com/Primary_Immune/status/420090415086198784",


    "file_counter": 2,                  # info about get_twitter_json.py process
    "short_file_name": "Jan to May\\Blood\\Tweets_BloodCancer.csv"
}
```

C Details of the S3 Twitter json Data File

1. The original dataset consisted of 896 csv files with 6,543,272 lines.
2. The raw json file is ??? Gb
   https://foo.bar
3. The json file zipped is ???? Gb
   https://foo.zip
4. There are ??? individual json entities
   (a) ??? lines (???%) were omitted because Twitter did not return any data for their id
   (b) ??? lines (???%) have non-empty locations
   (c) ??? of lines with locations (???%) have geo information

The data is processed in batches of 100 tweets, a limit imposed by Twitter for automated requests, and a text file is appended and saved after each batch is processed.

The process of running the program is fairly fast: on a Dell Windows 7 laptop it processes about 1,700 tweets per minute; however, the elapsed time is much longer because Twitter imposes a limit of around 14,000 tweets per 15-minute interval, so the program goes to sleep every 13,500.

On Amazon's EC2, the program ran to completion in ??? hours. The program terminated several times because of network errors and I lost several hours each time before I noticed and restarted the process. At the end of the process I will have to knit the several files into one and I will take that opportunity to improve the geo tagging since, by watching the extended log, I noticed ways to to get better accuracy.

D Amazon Web Services EC2 & S3

AWS EC2 and S3 have rather obscure documentation and operate in basic command-line mode, however once you've mastered them they are quite useful since you can get essentially as much computing power, storage and Internet access as you could possibly need on demand.

EC2 is the name for the service that provides either Unix or Windows servers on demand. S3 is the name of bit-bucket data storage.

On top of the base operating system you have to build your own programming environment. I used IPython, see pages 34 and 35.

In addition to being quite useful, it is also inexpensive: even with numerous false starts my total bill for this project was only $?????.

Amazon Web Services for background Python

I assume you have an AWS account and an access Key pair for SSH access. On Windows I used Putty as my SSH terminal and WinSCP for FTP; on my iPhone I used Server Auditor.

Setup:

1. I started an EC2 Ubuntu Server 14.04 LTS (HVM), SSD Volume Type - ami-e7b8c0d7 on an x86_64 t2.micro configuration. The SSH logon for such an image is ubuntu@Public IP.
2. I struggled for an entire day trying to figure out how to access files on S3 from EC2; I gave up and FTP'd the entire bunch from Google Drive to the image I had just started. There is a nice tool at http://timkay.com/aws/ that is helpful but not for 897 files in recursive folder structures.
3. I also had to download
    1. `get_twitter_json.py` and edit it a little for Ubuntu file formats
    2. `mapquest_key.txt`
    3. `twitter_credentials.py`
    4. `twitter_functions.py`
    5. `filename_list.csv` had to be re-created for the Ubuntu file names and locations

Install Python:

1. I downloaded the Anaconda distro:
   `wget http://09c8d0b2229f813c1b93-c95ac804525aac4b6dba79b00b39d1d3.r79.cf1.rackcdn.com/Anaconda-2.0.1-Linux-x86_64.sh`
2. ... and installed it
   `bash Anaconda-2.0.1-Linux-x86_64.sh`
   Note: 'q' gets you out of the license agreement
3. Reloaded the .bashrc ...
   `source .bashrc`
4. ... and issued the following commands:
   `sudo -i`
   `apt-get update`
   `apt-get install python-pip`
   `pip install oauth2`
   `apt-get install ipython`

Then I started up the python program in the background

```
nohup python get_twitter_json.py "filename_list.csv" 1 0 &
```

... and exited the shell
`exit`

As the program churned through the files I was able to sign on and monitor progress via the `nohup.out` file. I could also watch system statistics through the AWS Management Console and on the iPhone AWS app. I probably could have used `boto` but I didn't try it.

Create S3 zip file

The first step was to compress it:

```
# see http://pymotw.com/2/zipfile/

infilename  = 'bigtweet_file001.json'
outfilename = 'bigtweet_file001.zip'

import zipfile
try:
    import zlib
    compression = zipfile.ZIP_DEFLATED
except:
    compression = zipfile.ZIP_STORED

zf = zipfile.ZipFile(outfilename, mode='w')
try:
    zf.write(infilename, compress_type=compression)
finally:
    zf.close()
```

... and then to move it to S3

Install utilities from `http://timkay.com/aws/`

```
 * sudo -i
 * apt-get install curl
 * curl https://raw.githubusercontent.com/timkay/aws/master/aws -o aws
 * vi ~/.awssecret # AWS credentials Ctrl+o :w <enter> Ctrl-o :q <enter>
 * perl aws --install
 * chmod +x aws
 * cd /home/ubuntu
```

Then you can enter `s3put <S3 bucket name> <local file to be transferred into S3>`

References

[syn, 2014] (2014). Twitter mining for fine-grained syndromic surveillance. Artifical Intelligence in Medicine 61 (2014) 153-163.

[BioPortal, 2014] BioPortal (2014). Bioportal, the worlds most comprehensive repository of biomedical ontologies. http://bioportal.bioontology.org/.

[BM, 2014] BM, K. (2014). Agencies use social media to track foodborne illness. JAMA, 312(2):117–118.

[Bosley et al., 2012] Bosley, J. C., Zhao, N. W., Hill, S., Shofer, F. S., Asch, D. A., Becker, L. B., and Merchant, R. M. (2012). Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. http://www.resuscitationjournal.com/article/S0300-9572(12)00871-4/abstract.

[Breen, 2011a] Breen, J. (2011a). Github twitter-sentiment-analysis-tutorial-201107. https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107. Code provided in conjunction with tutorial slides.

[Breen, 2011b] Breen, J. (2011b). slides from my r tutorial on twitter text mining #rstats. http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/.

[Chew and Eysenbach, 2010] Chew, C. and Eysenbach, G. (2010). Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0014118.

[Computerworld, 2010] Computerworld (2010). Twitter growth prompts switch from mysql to 'nosql' database. http://www.computerworld.com/s/article/9161078/Twitter_growth_prompts_switch_from_MySQL_to_NoSQL_database.

[Cook, 2014] Cook, T. (2014). Tim Cook healthcare twitter analysis repo. https://github.com/twcook/TweetMapping.

[Department of Health and Human Services, 2012] Department of Health and Human Services (2012). Now Trending: #Health in My Community. http://nowtrending.hhs.gov/. This contest challenged entrants to create a web-based application that searched open source Twitter data for health topics and delivered analyses of that data for both a specified geographic area and the national level.

[Dork et al., 2010] Dork, M., Gruen, D., Williamson, C., and Carpendale, S. (2010). A visual backchannel for large-scale events. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5613451.

[DScanfeld et al., 2010] DScanfeld, Scanfeld, V., and Larson, E. (2010). Dissemination of health information through social networks: Twitter and antibiotics. http://www.ajicjournal.org/article/S0196-6553(10)00034-9/abstract.

[Fisher, 2014] Fisher, G. (2014). George Fisher healthcare twitter analysis github repo. https://github.com/grfiv/healthcare_twitter_analysis.

[Franko, 2011] Franko, O. (2011). Twitter as a communication tool for orthopedic surgery. http://www.ncbi.nlm.nih.gov/pubmed/22050252?dopt=Abstract.

[Galloro, 2011] Galloro, V. (2011). Hospitals are finding ways to use the social media revolution to raise money, engage patients and connect with their communities. http://www.ncbi.nlm.nih.gov/pubmed/21513035?dopt=Abstract.

[Google Drive, 2014] Google Drive (2014). Healthcare twitter analysis data files. https://drive.google.com/folderview?id=0B2io9_E3COquYWdlWjdzU3ozbzg&usp=sharing.

[highscalability.com, 2011] highscalability.com (2011). How twitter stores 250 million tweets a day using mysql. http://highscalability.com/blog/2011/12/19/how-twitter-stores-250-million-tweets-a-day-using-mysql.html.

[Indes et al., 2013] Indes, J. E., Gates, L., Mitchell, E. L., and Muhs, B. E. (2013). Social media in vascular surgery. http://www.jvascsurg.org/article/S0741-5214(12)02104-0/abstract.

[Jhawar et al., 2012] Jhawar, S., Sethi, R., Yuhas, C., and Schiff, P. (2012). All atwitter about radiation oncology: A content analysis of radiation oncology-related traffic on twitter. http://www.redjournal.org/article/S0360-3016(12)02712-5/abstract.

[Kostkova et al., 2010] Kostkova, P., de Quincey, E., and Jawaheer, G. (2010). The potential of social networks for early warning nad outbreak detection systems: the swine flu twitter study. http://ijidonline.com/article/S1201-9712(10)00507-2/abstract.

[Mapquest, 2014] Mapquest (2014). Mapquest developer api. http://developer.mapquest.com/.

[McKee et al., 2011] McKee, M., Cole, K., Hurst, L., Aldridge, R., and Horton, R. (2011). The other twitter revolution: how social media are helping to monitor the nhs reforms. http://www.ncbi.nlm.nih.gov/pubmed/21325389?dopt=Abstract.

[Mehta and Saama Technologies, 2013] Mehta, P. and Saama Technologies (2013). Healthcare twitter analysis website. https://www.coursolve.org/need/184.

[Micieli and Micieli, 2012] Micieli, R. and Micieli, J. A. (2012). Twitter as a tool for ophthalmologists. http://www.canadianjournalofophthalmology.ca/article/S0008-4182(12)00294-3/abstract.

[MongoDB, 2014] MongoDB (2014). Mongodb website. http://www.mongodb.org/.

[Nielsen, 2011] Nielsen, F. Å. (2011). Afinn. http://www2.imm.dtu.dk/pubdb/views/bibtex.php?id=6010. Informatics and Mathematical Modelling, Technical University of Denmark.

[Quora, 2012] Quora (2012). Twitter: Which database system(s) does twitter use? https://www.quora.com/Twitter-1/Which-database-system-s-does-Twitter-use.

[Sabine Tejpar et al., 2011] Sabine Tejpar, MD, P., Wendy De Roock, MD, P., and Derek Jonker, M. (2011). Physicians on twitter. JAMA, 305(6):566–568.

[Sanchez, 2012] Sanchez, G. (2012). Mining twitter. https://github.com/gastonstat/Mining_Twitter. An interesting collection of R programs to do analyses of Twitter data. His use of ggplot was out of date but aside from fixing that, the R code worked pretty well in the context of the data for this project.

[Scanfeld et al., 2010] Scanfeld, D., Scanfeld, V., and Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. http://www.ajicjournal.org/article/S0196-6553(10)00034-9/abstract.

[Signorini et al., 2011] Signorini, A., Segre, A. M., and Polgreen, P. M. (2011). The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0019467.

[Su et al., 2011] Su, X., Suominen, H., and Hanlen, L. (2011). Machine intelligence for health information: capturing concepts and trends in social media via query expansion. http://www.ncbi.nlm.nih.gov/pubmed/21893923?dopt=Abstract.

[Tobias, 2011] Tobias, E. (2011). Using twitter and other social media platforms to provide situational awareness during an incident. http://www.ncbi.nlm.nih.gov/pubmed/22130339?dopt=Abstract.

[Topsy, 2010] Topsy (2010). Cool tool: Topsy finds most influential tweeters on any topic. http://gigaom.com/2010/07/15/cool-tool-topsy-finds-most-influential-tweeters-on-any-topic/.

[Topsy, 2014] Topsy (2014). Topsy website. http://topsy.com/.

[Twitter, 2014a] Twitter (2014a). Manhattan, our real-time, multi-tenant distributed database for twitter scale. https://blog.twitter.com/2014/manhattan-our-real-time-multi-tenant-distributed-database-for-twitter-scale.

[Twitter, 2014b] Twitter (2014b). Twitter json documentations. https://dev.twitter.com/docs.

[US Dept. pf Health & Human Services, 2012] US Dept. pf Health & Human Services (2012). A partnership between the public and the government to solve important challenges. https://challenge.gov/?q=334-now-trending-health-in-my-community.

[Vance et al., 2009] Vance, K., Howe, W., and Dellavalle, R. (2009). Social internet sites as a source of public health information. http://www.ncbi.nlm.nih.gov/pubmed/19254656?dopt=Abstract.

[Williams et al., 2013a] Williams, S. A., Terras, M., and Warwick, C. (2013a). How twitter is studied in the medical professions: A classification of twitter papers indexed in pubmed. http://www.medicine20.com/2013/2/e2/. Medicine 2.0: Social Media, Mobile Apps, and Internet/Web 2.0 in Health, Medicine and Biomedical Research.

[Williams et al., 2013b] Williams, S. A., Terras, M. M., and Warwick, C. (2013b). What do people study when they study twitter? classifying twitter related academic papers. https://www.emeraldinsight.com/journals.htm?articleid=17088387.

[Wired, 2014] Wired (2014). This is what you build to juggle 6,000 tweets a second. http://www.wired.com/2014/04/twitter-manhattan/.