Project Resources Documentation

Contents

[Miscellaneous 1](#_Toc359488236)

[Github Instructions: 1](#_Toc359488237)

[Basics: 1](#_Toc359488238)

[Cloning the repository: 3](#_Toc359488239)

[Updating the repository: 3](#_Toc359488240)

[FAQ 4](#_Toc359488241)

[Tweet Classification: 4](#_Toc359488242)

[Medical Related API: 4](#_Toc359488243)

[Medical Hashtags’ Dictionary: 5](#_Toc359488244)

[Twitter Archiving Google Spreadsheet TAGS v5: 5](#_Toc359488245)

[Social Media & Sentiment Analysis(Text Normalization) 5](#_Toc359488246)

[Practices on Health Care 5](#_Toc359488247)

[Objectives and Ideas 5](#_Toc359488248)

# 

Miscellaneous

## Github Instructions:

### Basics:

1. Go to <https://github.com/pratik008/HealthCare_Twitter_Analysis>
2. Click on the "Fork" button, at the top right of the screen. This will create a copy of the project attached to your account. That means that you can make as many changes as you want without any fear to mess up with the official/trusted repository.
3. Go to your github user page (click on your user icon) and select the HealthCare project from the repository list. For instance, my fork lives at: <https://github.com/widged/HealthCare_Twitter_Analysis>
4. Copy the github url to your clipboard. It shows in the text input box at the top of the page.  For mine, it is <https://github.com/widged/HealthCare_Twitter_Analysis.git>. Make sure you have selected the http option and that the url shows as https / read+write access (it should by default if you are logged in into your account).
5. Go to your terminal window. Type

git clone [https://github.com/{your\_github\_id}/HealthCare\_Twitter\_Analysis.git](https://github.com/%7Byour_github_id%7D/HealthCare_Twitter_Analysis.git)

 6. Now, you should have a folder named "HealthCare\_Twitter\_Analysis" on your computer. Go to it, try and execute Tweepy.py (python Tweetpy.py). Install the tweepy library if not available yet (<https://github.com/tweepy/tweepy/blob/master/INSTALL>).    
 7. Make a change to any file. For instance type "touch test.txt"  
 8. Commit your changes, locally. 

git add .

git commit -a -m "comments describing your changes"

 9. If you wish to share with the team, push your changes to github

git push origin master

10. If you think the official version is likely to benefit from your changes, submit a pull request. Don't worry too much about this for now. That can wait. Getting familiar with 1-8 is more important. If you are curious, you can find information there: <https://help.github.com/articles/using-pull-requests>

11. To keep it in sync with the official projects, you have to add the official repository as a remote

git remote add pratik https://github.com/pratik008/HealthCare\_Twitter\_Analysis

Then at any time you will be able to run to pull in the changes not present in your local repository.

git fetch pratik

(it will do its best to "merge" the changes made in Patrik version into your current local version). The different between fetch and pull is explained there: <http://stackoverflow.com/questions/292357/whats-the-difference-between-git-pull-and-git-fetch>  
  
See <https://help.github.com/articles/fork-a-repo> for step by step information of the forking process  
  
That sounds like a lot of information, a lot of new things to absorb. Sure it is. But by experience, it doesn't take long to get a handle on this. The most difficult step is to find the courage to give it a shot.   
  
It really is a worthwhile investment. If you hope to one day get a job as part of a team involved in programming, then that's a must have skill. But even for your personal projects, source control makes a huge difference. Less stress.

All these comments about how to get started with github merged into a single document - <https://github.com/widged/HealthCare_Twitter_Analysis/blob/master/getting_started.md>.

### ****Cloning the repository:****

**Virtual Machine Users**

The repository is already cloned and located in the home directory.

**Windows Users**

If you don't have Git installed, installation instructions can be found here: <http://git-scm.com/download/win>\*\*

You can install a GUI on top of GIt from here: <http://windows.github.com/>

Once the GUI is installed, you can navigate to the repository page, <https://github.com/uwescience/datasci_course_materials>, and click on the 'Clone in Windows' button.

**Linux Users**

If you don't have Git installed, installation instructions can be found here: <http://git-scm.com/download/linux>

If you have Git installed, you need to clone the repository. From a terminal, navigate to where you want the repository to be located and perform a clone.

git clone https://github.com/uwescience/datasci\_course\_materials.git

**Mac Users**

If you don't have Git installed, installation instructions can be found here: <http://git-scm.com/download/mac>

If you have Git installed, you need to clone the repository. From a terminal, navigate to where you want the repository to be located and perform a clone.

git clone https://github.com/uwescience/datasci\_course\_materials.git

### ****Updating the repository:****

**Virtual Machine, Linux, and Mac Users**

Navigate to the repository and perform a git pull

cd datasci\_course\_materials

git pull

**Windows Users**

Perform a pull on the datasci\_course\_materials through the GUI interface. If you set the repository to stay in sync, the gui will give you notifications where there are updates to be pulled.

### FAQ

1. I added my name to my local file Project\_members, I then did git push origin master, but my changes don't seem to appear in the shared file.

Steps:

Your changes will appear in your fork. To have them appear in the official repo, two options. 

1. Make a pull request. See <https://help.github.com/articles/using-pull-requests>
2. Be granted write access to the official project. You will have to contact @Pratik for this. Then you can do

git remote add pratik https://github.com/pratik008/HealthCare\_Twitter\_Analysis

git push pratik master

A priori, pull requests are the best way to go for teams made of persons from diverse organisations with variable levels of skills. It provides an extra security net to all parties.   
  
An inconvenient, though is that they need to be approved... and we all come from different timezones. Another one is that pull requests are pretty much useless on binary content such as excel files. The file would need to be turned into a text one to allow for successive merges... or moved to a google spreadsheet.   
  
Or perhaps there is no need for an excel file. If people create a github account and fork, we de facto have a list of all contributors. It can be seen at: [https://github.com/pratik008/HealthCare\_Twitter\_Analysis/network/members](https://github.com/pratik008/HealthCare_Twitter_Analysis/network). Clicking on a user name will get you to their profile.

## **Tweet Classification:**

[SNOMED](http://en.wikipedia.org/wiki/SNOMED_CT) : **Med**icine **C**linical **T**erms

## **Medical Related API:**

<http://www.nlm.nih.gov/api/>

<http://blog.programmableweb.com/2012/05/16/72-medical-apis-avvo-national-library-of-medicine-and-nhs/>

looking into relevant APIs for "over-the-counter" drug names:

found RxMix / RxNav (<http://rxnav.nlm.nih.gov/APIsOverview.html>)

and DailyMed (<http://dailymed.nlm.nih.gov/dailymed/>), to be useful. DailyMed contains information of all drug labels, including over-the-counter (OTC) drugs submitted into the FDA.

## **List of MD Anderson Cancer Center Physicians and Leaders**

## [**Twitter Archiving Google Spreadsheet TAGS v5**](http://mashe.hawksey.info/2013/02/twitter-archive-tagsv5/)**:**

<http://mashe.hawksey.info/2013/02/twitter-archive-tagsv5/>

## **Social Media & Sentiment Analysis(Text Normalization):**

<http://www.slideshare.net/thewilde/social-media-sentiment-analysis-splunk-conf2012>

## **Resources and Tools**

## Data Collection :

* Some tools that may be useful for analyzing/visualizing results <http://data.gov.au/resources/>
* In case we need to analyze past twits (but only for 2006-2009) <http://www.infochimps.com/collections/twitter-census>
* Disease Hashtags

<http://www.symplur.com/healthcare-hashtags/diseases/>

* Found these from the Forum, looks useful <http://hedonometer.org/index.html> (I see people were happier during recession, interesting - Pratik)
* Medicare Charge data by billing code submitted by Patrick Salazar <http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/index.html>
* Twitter analysis using Association analysis <http://wwwconference.org/www2011/proceeding/companion/p111.pdf>
* Comparative Study of Clustering Techniques for Short Text Documents <http://bpavlyshenko.blogspot.in/>
* Identifying Symptom Groups from Emergency Department Presenting Complaint Free Text using SNOMED CT<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243271/>
* Automatic Meaning Discovery Using Google <http://homepages.cwi.nl/~paulv/papers/amdug.pdf>

## **Research Papers**

-Bilge, U.; Bozkurt, S.; Yolcular, B. O. & Ozel, D. **Can social web help to detect influenza related illnesses in Turkey**? Stud Health Technol Inform, Department of Biostatistics and Medical Informatics, Faculty of Medicine, Akdeniz University, Antalya, Turkey.[ubilge@akdeniz.edu.tr](mailto:ubilge@akdeniz.edu.tr), 2012, 174, 100-104

**Abstract**: In this study, a content analysis of Twitter is carried out to determine the frequency of tweets mentioning influenza like illnesses (swine flu, cold etc.) and results have been compared with news articles, Google search trends and national health statistics in Turkey. Between 1st January 2009 and 31st December 2010, over 4,165 influenza-related Turkish posts on Twitter and 10,000 news articles in three leading Turkish newspaper websites containing influenza related keywords have been analyzed. No strong correlation has been found between influenza related online data and the real world records. Although there is no significant relationship, this study shows that there is a huge amount of data can be harvested from the social web applications such as Twitter.

-Borondo, J.; Morales, A. J.; Losada, J. C. & Benito, R. M. Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish Presidential election as a case study. Chaos, Grupo de Sistemas Complejos and Departamento de Física y Mecánica, Universidad Politécnica de Madrid, ETSI Agrónomos, 28040 Madrid, Spain., 2012, 22, 023138

**Abstract**: Transmitting messages in the most efficient way as possible has always been one of politicians' main concerns during electoral processes. Due to the rapidly growing number of users, online social networks have become ideal platforms for politicians to interact with their potential voters. Exploiting the available potential of these tools to maximize their influence over voters is one of politicians' actual challenges. To step in this direction, we have analyzed the user activity in the online social network Twitter, during the 2011 Spanish Presidential electoral process, and found that such activity is correlated with the election results. We introduce a new measure to study political sentiment in Twitter, which we call the relative support. We have also characterized user behavior by analyzing the structural and dynamical patterns of the complex networks emergent from the mention and retweet networks. Our results suggest that the collective attention is driven by a very small fraction of users. Furthermore, we have analyzed the interactions taking place among politicians, observing a lack of debate. Finally, we develop a network growth model to reproduce the interactions taking place among politicians.

-Chew, C. & Eysenbach, G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One, Centre for Global eHealth Innovation, University Health Network, Toronto, Canada., 2010, 5, e14118

**Abstract**: Surveys are popular methods to measure public perceptions in emergencies but can be costly and time consuming. We suggest and evaluate a complementary "infoveillance" approach using Twitter during the 2009 H1N1 pandemic. Our study aimed to: 1) monitor the use of the terms "H1N1" versus "swine flu" over time; 2) conduct a content analysis of "tweets"; and 3) validate Twitter as a real-time content, sentiment, and public attention trend-tracking tool.Between May 1 and December 31, 2009, we archived over 2 million Twitter posts containing keywords "swine flu," "swineflu," and/or "H1N1." using Infovigil, an infoveillance system. Tweets using "H1N1" increased from 8.8% to 40.5% (R(2) = .788; p.001), indicating a gradual adoption of World Health Organization-recommended terminology. 5,395 tweets were randomly selected from 9 days, 4 weeks apart and coded using a tri-axial coding scheme. To track tweet content and to test the feasibility of automated coding, we created database queries for keywords and correlated these results with manual coding. Content analysis indicated resource-related posts were most commonly shared (52.6. 4.5% of cases were identified as misinformation. News websites were the most popular sources (23.2, while government and health agencies were linked only 1.5% of the time. 7/10 automated queries correlated with manual coding. Several Twitter activity peaks coincided with major news stories. Our results correlated well with H1N1 incidence data.This study illustrates the potential of using social media to conduct "infodemiology" studies for public health. 2009 H1N1-related tweets were primarily used to disseminate information from credible sources, but were also a source of opinions and experiences. Tweets can be used for real-time content analysis and knowledge translation research, allowing health authorities to respond to public concerns.

-King, D.; Ramirez-Cano, D.; Greaves, F.; Vlaev, I.; Beales, S. & Darzi, A. Twitter and the health reforms in the English National Health Service. Health Policy, Centre for Health Policy, Imperial College London, 10th Floor QEQM, St Mary's Hospital, London W2 1NY, United Kingdom. Electronic address: dominic.king@imperial.ac.uk., 2013, 110, 291-297

**Abstract**: Social media (for example Facebook and YouTube) uses online and mobile technologies to allow individuals to participate in, comment on and create user-generated content. Twitter is a widely used social media platform that lets users post short publicly available text-based messages called tweets that other users can respond to. Alongside traditional media outlets, Twitter has been a focus for discussions about the controversial and radical reforms to the National Health Service (NHS) in England that were recently passed into law by the current coalition Government. Looking at over 120,000 tweets made about the health reforms, we have investigated whether any insights can be obtained about the role of Twitter in informing, debating and influencing opinion in a specific area of health policy. In particular we have looked at how the sentiment of tweets changed with the passage of the Health and Social Care Bill through Parliament, and how this compared to conventional opinion polls taken over the same time period. We examine which users appeared to have the most influence in the 'Twittersphere' and suggest how a widely used metric of academic impact - the H-index - could be applied to measure context-dependent influence on Twitter.

-Greaves, F.; Ramirez-Cano, D.; Millett, C.; Darzi, A. & Donaldson, L. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. BMJ Qual Saf, Department of Primary Care and Public Health, Imperial College London, UK.[felix.greaves08@imperial.ac.uk](mailto:felix.greaves08@imperial.ac.uk), 2013, 22, 251-255

**Abstract**: Recent years have seen increasing interest in patient-centred care and calls to focus on improving the patient experience. At the same time, a growing number of patients are using the internet to describe their experiences of healthcare. We believe the increasing availability of patients' accounts of their care on blogs, social networks, Twitter and hospital review sites presents an intriguing opportunity to advance the patient-centred care agenda and provide novel quality of care data. We describe this concept as a 'cloud of patient experience'. In this commentary, we outline the ways in which the collection and aggregation of patients' descriptions of their experiences on the internet could be used to detect poor clinical care. Over time, such an approach could also identify excellence and allow it to be built on. We suggest using the techniques of natural language processing and sentiment analysis to transform unstructured descriptions of patient experience on the internet into usable measures of healthcare performance. We consider the various sources of information that could be used, the limitations of the approach and discuss whether these new techniques could detect poor performance before conventional measures of healthcare quality.

-R Lyles, C.; López, A.; Pasick, R. & Sarkar, U. "5 Mins of uncomfyness is better than dealing with cancer 4 a lifetime": an exploratory qualitative analysis of cervical and breast cancer screening dialogue on twitter.\_ J Cancer Educ\_, Division of General Internal Medicine, Department of Medicine, University of California San Francisco, San Francisco, CA, USA, LylesC@medsfgh.ucsf.edu., 2013, 28, 127-133

**Abstract**: Twitter.com is a "micro-blogging" website. Although Twitter use is growing rapidly, little is known about health behavior discussions on this site, even though a majority of messages are publicly available. We retrieved publicly available Twitter messages during a 5-week period in early 2012, searching separately for the terms "Pap smear" and "mammogram." We used content analysis to code each 140-character message, generating a separate coding framework for each cancer screening term and calculating the frequencies of comments. Using the brief account description, we also coded the author as individual, organization, or news media outlet. There were 203 Pap smear and 271 mammogram messages coded, over three fourths of which were from individual accounts. Overall, 22 % of Pap smear messages and 25 % of mammogram messages discussed personal experiences, including attending appointments, negative sentiment about the procedure, and results. Other messages from both individuals and organizations (8 % Pap smear, 18 % mammogram) promoted screening. About one quarter of the messages expressed personal experiences with cancer screening. This demonstrates that Twitter can be a rich source of information and could be used to design new health-related interventions.

-Signorini, A.; Segre, A. M. & Polgreen, P. M. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One, Department of Computer Science, University of Iowa, Iowa City, Iowa, United States of America., 2011, 6, e19467

**Abstract**: Twitter is a free social networking and micro-blogging service that enables its millions of users to send and read each other's "tweets," or short, 140-character messages. The service has more than 190 million registered users and processes about 55 million tweets per day. Useful information about news and geopolitical events lies embedded in the Twitter stream, which embodies, in the aggregate, Twitter users' perspectives and reactions to current events. By virtue of sheer volume, content embedded in the Twitter stream may be useful for tracking or even forecasting behavior if it can be extracted in an efficient manner. In this study, we examine the use of information embedded in the Twitter stream to (1) track rapidly-evolving public sentiment with respect to H1N1 or swine flu, and (2) track and measure actual disease activity. We also show that Twitter can be used as a measure of public interest or concern about health-related events. Our results show that estimates of influenza-like illness derived from Twitter chatter accurately track reported disease levels.

* Integrating Social Media into Emergency-Preparedness Efforts Raina M. Merchant, M.D., Stacy Elmer, M.A., and Nicole Lurie, M.D., M.S.P.H. N Engl J Med 2011; 365:289-291July 28, 2011DOI: 10.1056/NEJMp1103591 <http://www.nejm.org/doi/full/10.1056/NEJMp1103591>

Last edited by hongsonnghiem, a month ago

**Hospitals on Twitter**

<http://ebennett.org/hsnl/hospitals-on-twitter/>

This is a list of MD Anderson Cancer Center Physicians and Leaders on twitter:

<https://twitter.com/PhysRelations/md-anderson-faculty>

Here is a link to the top hospitals on twitter by hospitals and tweets.

<http://ebennett.org/hsnl/hospitals-on-twitter/>

Cancer patients on Twitter: a novel patient community on social media.

<http://www.ncbi.nlm.nih.gov/pubmed/23270426>

**Python Resources :**

Python - Natural Language Toolkit

<http://nltk.org/>

<http://nltk.googlecode.com/svn/trunk/doc/howto/collocations.html>

Python - Machine Learning Toolkit

<http://scikit-learn.org/stable/>

**Visualization :**

Real-Time Heatmap

[blog.comsysto.com/2012/07/10/real-time-twitter-heat-map-with-mongodb/](http://blog.comsysto.com/2012/07/10/real-time-twitter-heat-map-with-mongodb/)

Some tools that may be useful for analyzing/visualizing results

<http://data.gov.au/resources/>

**R Graphical Library :**

<http://gallery.r-enthusiasts.com/>

**General Analysis :**

Twitter analysis using Association analysis <http://wwwconference.org/www2011/proceeding/companion/p111.pdf>

Comparative Study of Clustering Techniques for Short Text Documents <http://bpavlyshenko.blogspot.in/>

In case we need to analyze past tweets <http://www.infochimps.com/collections/twitter-census>

Found these from the Forum, looks useful <http://hedonometer.org/index.html>

Identifying Symptom Groups from Emergency Department Presenting Complaint Free Text using

SNOMED CT <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243271/>

Automatic Meaning Discovery Using Google <http://homepages.cwi.nl/~paulv/papers/amdug.pdf>

**Government Data :**

Medicare Billing/Charge data

<http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/index.html>

Medicare Charge data by billing code submitted by Patrick Salazar <http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/index.html>

Practices on Health Care

## Objectives and Ideas

**Goal 1 : Identify Medical and Non-Medical Tweets**

**1) Algorithm for classifying Medical Tweets - Iga Korneta**

pick some obvious words ("doctor", "hospital", "nurse", "wound", "sick"...) as start, then check what are the most popular nontrivial words that co-occur with those, and then iterate until the tweets start being "clearly" nonmedical. (we could then do regression/clustering to find out how much each of these terms predicts whether a tweet is medical or not in a training set, and then in a test set.)

**2) Twitter Spam detection - Hugo Toledo**

Create list of publicly available/usable spam keyword corpus

Review their efficacy in our effort

Test against a twitter stream and/or previously collected sample

Rank effectiveness of each list

Collectively discuss next step on this work stream

**Goal 2 : Trends of Diseases in US**

**3) Algorithm for Classifying Medical Disease - Finding trends**

Ideas very similar to topic 2. But here we take some tweets related to common diseases like (“Cancer”, “Asthma”, “Diabetes”,”COPD” etc.)

We then find related diseases and also the trends of these diseases over US.

Finally compare the findings to government data available.

**4) Disease Hashtag Heat Map - Patrick Salazar**

Correlate one or two common diseases and display the data in a visual form using a real-time heat map.

<http://blog.comsysto.com/2012/07/10/real-time-twitter-heat-map-with-mongodb/>

**Goal 3 : Sentiment of patients on twitter**

**5) Patient voice around treatments and medication - Sentiment analysis.**

Discover tweets about patients

Find the themes discussed and the sentiment around these themes (Example : co-pay at hospital too high, extremely frustrated - negative sentiment about bill).

Generalize findings to understand which states / institutes are happier and better at which services?

**Goal 4 : Disaster Recovery**

**6) Quick Information extraction and summarization.**

Collect tweets related to a disaster site.

Summarize vast quantity of data

Extract key information that could help in recovery

Identify and isolate SOS messages

**Other Ideas**

**7) Underlying network - Martina Pugliese**

Twitter network is, apart from being directed, also exponential in the distribution of degrees. This means that you usually have a lot of highly connected nodes and a few hubs, and this follows an exp. law. I think that the netowork should be taken into account, i.e., that information on who says what, if the tweet has a @ or a RT, is important. For example, if I am disappointed with the treatment I received in a hospital, I may generate a cascade of tweets from other people who agree with me.

**8) Twitter Analysis - A Graph Problem**

If we follow important medical tweets and their re-tweets, we might be able to create a graph of nodes in the medical tweets world. High level – Solve by thinking twitter analysis as a graph problem

one idea -<http://ebiquity.umbc.edu/blogger/2007/04/19/twitter-social-network-analysis/>

node - hashtags (or medical words)

edges - number of co-occurrence

color - classification ( 1. medical or not 2. sentiment 3. emotion)

size of node - volume of tweets of a hashtag (or medical word)

**Other Tasks :**

Medical API’s research

Top Hospital Hashtags / Usernames on Twitter

Top Medical Websites - Important Twitter Hashtags collection.

Relevant white-papers and algorithms.

Analyze twitter attributes