Volume 5, Issue 10, October-2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Sentiment Analysis and Opinion Mining For Twitter Data to Understand the Food Price Crisis

D Ramesh

Associate Professor, Department of CSE, JNTUH CEJ,
Telangana State, India

Abstract— The globalization of world economy is increasing the need for basic statistics on supply and demand for agricultural. The rising food prices and food shortage highlights the need for more concrete efforts to harness the potential of the agricultural sector for development since it has remained the dominant sector in most countries economy. The existence of official statistical system especially during crisis which is able to feed in and provide information on regular basis should satisfy the needs for immediate of interventions. The current crisis calls for much closer scrutiny and monitoring of the input to agriculture as well as its production processes and also the problems that cause the food crisis. This in turn will create enormous new challenges of agriculture statistics and related fields. Crisis range in scope and intensity from incidents that directly or indirectly affect a single village to ones that impact the entire country. Rising global food prices affect countries differently depending on whether they are net exporter or importers of food. A complex combination of poor harvest in some major commodity producing countries, increase in the cost of production due to higher fuel and fertilizer prices, diversion of food crops to produce bio fuels and a blockage in global trade is driving food prices up worldwide. The rising food prices and food shortage highlights the need for more concrete efforts to harness the potential of the agricultural sector for development since it has remained the dominant sector in most African economies. Micro blogging today has become a very popular communication tool among Internet users. Millions of users share opinions on different aspects of life every day. Therefore micro blogging web-sites are rich sources of data for opinion mining and sentiment analysis. In this paper, mainly focus on using Twitter, the most popular micro blogging platform, for the task of sentiment analysis and show how to automatically collect a corpus related to Food Crisis for sentiment analysis and opinion mining purposes. And also perform linguistic analysis of the collected corpus related to Food Crisis and explain discovered phenomena. Using the corpus, build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document/.csv file.

Keywords— Agriculture, Internet, Micro blogging, Mining, Twitter, Web site.

I. INTRODUCTION

Today, the World's population is approximately 7.2 billion. With this rapid increase in the population, the problems associated with them are also increasing. Among all the other problems due to population growth, one of the major problems today is Food Price Crisis. Everyone is affected by the growing global food crisis. Those already living on the edge are being pushed over. According to the IMF, food prices have increased on average by more than 40% over the past 12 months. This is driving untold numbers back into poverty and reversing progress towards achieving the Millennium Development Goals. As always the brunt of food insecurity is borne by Africa, Asia and the Caribbean where riots and strikes have occurred and social unrest threatens political stability.

Admittedly a number of factors underlie the present food scarcity. Some of these factors are beyond our control, such as extreme weather phenomena, African countries are affected by serious droughts where as some of the Asian countries where the weather is unpredictable are affected by a series of droughts followed by heavy rains, floods and droughts in some of the areas and changes in consumption patterns with growing affluence in certain parts of the world. However there is an urgent need that address issues such as outdated agricultural practices, inadequate infrastructure and inequitable distribution of land ownership and insecurity of tenure. The list is long, the remedies are known.

The unprecedented level of the price of oil, over a period of almost one year, has dealt a severe blow to the economic growth of many developing countries. Mauritius depends on imported oil for 80 percent of its energy consumption which makes it extremely vulnerable on this count. With informed predictions, warning of a significant rise in temperature, during the 21st century, even the best case scenario is fraught with dire consequences. However, climate change places a particular burden on developing countries which have done the least to cause the problem.

II. INTERNET

The Internet is a global system of inter-connected computer networks that use the Internet protocol suite (TCP/IP) to link several billion devices worldwide. It is a network of networks that consists of millions of private, public, academic, business, and government networks of local to global scope, linked by a broad array of electronic, wireless, and optical networking technologies. The origins of the Internet date back to research commissioned by the United States government in the 1960s to build robust, fault-tolerant communication via computer networks. The primary precursor network, the ARPANET, initially served as a backbone for interconnection of regional academic and military networks in the 1980s. The funding of a new U.S. backbone by the National Science Foundation in the 1980s, as well as private funding for other commercial backbones, led to worldwide participation in the development of new networking technologies, and the merger of many networks.

Most traditional communications media, including telephony and television, are being reshaped or redefined by the Internet, giving birth to new services such as Internet telephony and Internet television. Newspaper, book and other print publishing are adapting to website technology or are reshaped into blogging and web feeds. The entertainment industry, including music, film, and gaming, was initially the fastest growing online segment. The Internet has enabled and accelerated new forms of human interactions through instant messaging, Internet forums and social networking. Online shopping has grown exponentially both for major retailers and small artisans and traders. Business-to-business and financial services on the Internet affect supply chains across entire industries.

Around 40% of the world population has an internet connection today. In 1995, it was less than 1%. The number of internet users has increased tenfold from 1999 to 2013. The first billion was reached in 2005. The second billion in 2010. The third billion in 2014.

A social networking service is a platform to build social networks or social relations among people who share similar interests, activities, backgrounds or real-life connections. A social network service consists of a representation of each user, his or her social links, and a variety of additional services. Social network sites are web-based services that allow individuals to create a public profile, create a list of users with whom to share connections, and view and cross the connections within the system. Most social network services are web-based and provide means for users to interact over the Internet, such as e-mail and instant messaging. Social network sites are varied and they incorporate new information and communication tools such as mobile connectivity, photo/video/sharing and blogging. Online community services are sometimes considered a social network service, though in a broader sense, social network service usually means an individual-centered service whereas online community services are group-centered. Social networking sites allow users to share ideas, pictures, posts, activities, events, and interests with people in their network.

The main types of social networking services are those that contain category places (such as former school year or classmates), means to connect with friends (usually with self-description pages), and a recommendation system linked to trust. Popular methods now combine many of these, with American-based services such as Facebook, Google+, LinkedIn, Instagram, Reddit, Pinterest, Vine, Tumblr, and Twitter widely used worldwide.

Early social networking on the World Wide Web began in the form of generalized online communities such as Theglobe.com (1995), Geocities (1994), Tripod.com (1995). Many of these early communities focused on bringing people together to interact with each other through chat rooms, and encouraged users to share personal information and ideas via personal webpages by providing easy-to-use publishing tools and free or inexpensive webspace. Some communities - such as Classmates.com - took a different approach by simply having people link to each other via email addresses. PlanetAll started in 1996.

In the late 1990s, user profiles became a central feature of social networking sites, allowing users to compile lists of "friends" and search for other users with similar interests. New social networking methods were developed by the end of the 1990s, and many sites began to develop more advanced features for users to find and manage friends. This newer generation of social networking sites began to flourish with the emergence of SixDegrees.com in 1997, followed by Makeoutclub in 2000, Hub Culture and Friendster in 2002, and soon became part of the Internet mainstream. Friendster was followed by MySpace and LinkedIn a year later, and eventually Bebo. Friendster became very popular in the Pacific Islands. Orkut became the first popular social networking service in Brazil (although most of its very first users were from the United States) and quickly grew in popularity in India (Madhavan, 2007). Attesting to the rapid increase in social networking sites' popularity, by 2005, it was reported that Myspace was getting more page views than Google. Facebook, launched in 2004, became the largest social networking site in the world in early 2009. Facebook was first introduced (in 2004) as a Harvard social networking site, expanding to other universities and eventually, anyone. Web-based social networking services make it possible to connect people who share interests and activities across political, economic, and geographic borders. Through e-mail and instant messaging, online communities are created where a gift economy and reciprocal altruism are encouraged through cooperation. Information is suited to a gift economy, as information is a nonrival good and can be gifted at practically no cost. Scholars have noted that the term "Social" cannot account for technological features of the social network platforms alone. Hence, the level of network sociability should determine by the actual performances of its users. Companies are using social media as a way to learn about potential employees' personalities and behavior. In numerous situations a candidate who might otherwise have been hired has been rejected due to offensive or otherwise unseemly photos or comments posted to social networks or appearing on a newsfeed.

III. EMERGING TRENDS

While the popularity of social networking consistently rises, new uses for the technology are frequently being observed. At the forefront of emerging trends in social networking sites is the concept of "real-time web" and "location-based." Real-time allows users to contribute contents, which is then broadcast as it is being uploaded. Twitter set the trend for "real-time" services, wherein users can broadcast to the world what they are doing, or what is on their minds within a 140-character limit. Facebook followed suit with their "Live Feed" where users' activities are streamed as soon as it happens. While Twitter focuses on words, Clixtr, another real-time service, focuses on group photo sharing wherein users can update their photo streams with photos while at an event.

One popular use for this new technology is social networking between businesses. Companies have found that social networking sites such as Facebook and Twitter are great ways to build their brand image. According to Jody Nimetz, author of Marketing Jive, there are five major uses for businesses and social media: to create brand awareness, as an online reputation management tool, for recruiting, to learn about new technologies and competitors, and as a lead generation tool to intercept potential prospects. These companies are able to drive traffic to their own online sites while encouraging their consumers and clients to have discussions on how to improve or change products or services. Through data mining, companies are able to improve their sales and profitability. With this data, companies create customer profiles that contain customer demographics and online behavior. A recent strategy has been the purchase and production of "network analysis software". This software is able to sort out through the influx of social networking data for any specific company.

IV. MICRO BLOGGING

Microblogging is a broadcast medium that exists in the form of blogging. A microblog differs from a traditional blog in that its content is typically smaller in both actual and aggregated file size. Microblogs "allow users to exchange small elements of content such as short sentences, individual images, or video links", which may be the major reason for their popularity. These small messages are sometimes called microposts.

Twitter, Facebook, and other microblogging services are also becoming a platform for marketing and public relations, with a sharp growth in the number of social media marketers. The Sysomos study shows that this specific group of marketers on Twitter is much more active than the general user population, with 15% of marketers following over 2,000 people and only .29% of the Twitter public following more than 2,000 people. Microblogging has also emerged as an important source of real-time news updates for recent crisis situations.

Microblogging has noticeably revolutionized the way information is consumed. It has empowered citizens themselves to act as sensors or sources of information that could lead to consequences and influence, or even cause, media coverage. People now share what they observe in their surroundings, information about events, and their opinions about topics from a wide range of fields. Moreover, these services store various metadata from these posts, such as location and time. Aggregated analysis of this data includes different dimensions like space, time, theme, sentiment, network structure etc., and gives researchers an opportunity to understand social perceptions of people in the context of certain events of interest. Microblogging also promotes authorship. On the microblogging platform Tumblr, the reblogging feature links the post back to the original creator..

V. TWITTER AS SOURCE

Social media has become a major platform for information sharing. Due to its openness in sharing data, Twitter is a prime example of social media in which researchers can verify their hypotheses, and practitioners can mine interesting patterns and build realworld applications. This paper takes through the process of harnessing Twitter data to find answers to intriguing questions. In this paper, begin with an introduction to the process of collecting data through Twitter's APIs and proceed to discuss strategies for curating large datasets. And then guide the reader through the process of visualizing Twitter data with realworld examples, present challenges and complexities of building visual analytic tools, and provide strategies to address these issues. And also show by example how some powerful measures can be computed using various Twitter data sources.

Twitter is a massive social networking site tuned towards fast communication. More than 140 million active users publish over 400 million 140- character “Tweets” every day. Twitter’s speed and ease of publication have made it an important communication medium for people from all walks of life. Twitter has played a prominent role in socio-political events, such as the Arab Spring and the Occupy Wall Street movement. Twitter has also been used to post damage reports and disaster preparedness information during large natural disasters, such as the Hurricane Sandy.

Twitter’s popularity as an information source has led to the development of applications and research in various domains. Humanitarian Assistance and Disaster Relief is one domain where information from Twitter is used to provide situational awareness to a crisis situation. Researchers have used Twitter to predict the occurrence of earthquakes and identify relevant users to follow to obtain disaster related information. Studies of Twitter’s use in disasters include regions such as China, and Chile. While a sampled view of Twitter is easily obtained through the APIs, the full view is difficult to obtain. The APIs only grant us access to a 1% sample of the Twitter data, and concerns about the sampling strategy and the quality of Twitter data obtained via the API have been raised recently in. This study indicates that care must be taken while constructing the queries used to collect data from the Streaming API.

Users on Twitter generate over 400 million Tweets everyday. Some of these Tweets are available to researchers and practitioners through public APIs at no cost. APIs to access Twitter data can be classified into two types based on their design and access method:

- REST APIs are based on the REST architecture now popularly used for designing web APIs. These APIs use the pull strategy for data retrieval. To collect information a user must explicitly request it.
- Streaming APIs provides a continuous stream of public information from Twitter. These APIs use the push strategy for data retrieval. Once a request for information is made, the Streaming APIs provide a continuous stream of updates with no further input from the user.

They have different capabilities and limitations with respect to what and how much information can be retrieved. The Streaming API has three types of endpoints:

- Public streams: These are streams containing the public tweets on Twitter.
- User streams: These are single-user streams, with to all the Tweets of a user.
- Site streams: These are multi-user streams and intended for applications which access Tweets from multiple users.

As the Public streams API is the most versatile Streaming API, we will use it in all the examples pertaining to Streaming API.

Requests to the APIs contain parameters which can include hashtags, keywords, geographic regions, and Twitter user IDs. Responses from Twitter APIs is in JavaScript Object Notation (JSON) format. JSON is a popular format that is widely used as an object notation on the web. Twitter APIs can be accessed only via authenticated requests. Twitter uses Open Authentication and each request must be signed with valid Twitter user credentials. Access to Twitter APIs is also limited to a specific number of requests within a time window called the rate limit. These limits are applied both at individual user level as well as at the application level. A rate limit window is used to renew the quota of permitted API calls periodically.

A. Introduction to Open Authentication(OAuth):

Open Authentication (OAuth) is an open standard for authentication, adopted by Twitter to provide access to protected information. Passwords are highly vulnerable to theft and OAuth provides a safer alternative to traditional authentication approaches using a three-way handshake. It also improves the confidence of the user in the application as the user’s password for his Twitter account is never shared with third-party applications.

The authentication of API requests on Twitter is carried out using OAuth. Figure 1 summarizes the steps involved in using OAuth to access Twitter API. Twitter APIs can only be accessed by applications.

Below ,detail the steps for making an API call from a Twitter application using OAuth:

- Applications are also known as consumers and all applications are required to register themselves with Twitter. Through this process the application is issued a consumer key and secret which the application must use to authenticate itself to Twitter.
- The application uses the consumer key and secret to create a unique Twitter link to which a user is directed for authentication. The user authorizes the application by authenticating himself to Twitter. Twitter verifies the user’s identity and issues a OAuth verifier also called a PIN.
- The user provides this PIN to the application. The application uses the PIN to request an “Access Token” and “Access Secret” unique to the user.
- Using the “Access Token” and “Access Secret”, the application authenticates the user on Twitter and issues API calls on behalf of the user.
- The “Access Token” and “Access Secret” for a user do not change and can be cached by the application for future requests. Thus, this process only needs to be performed once, and it can be easily accomplished using the method GetUserAccessKeySecret

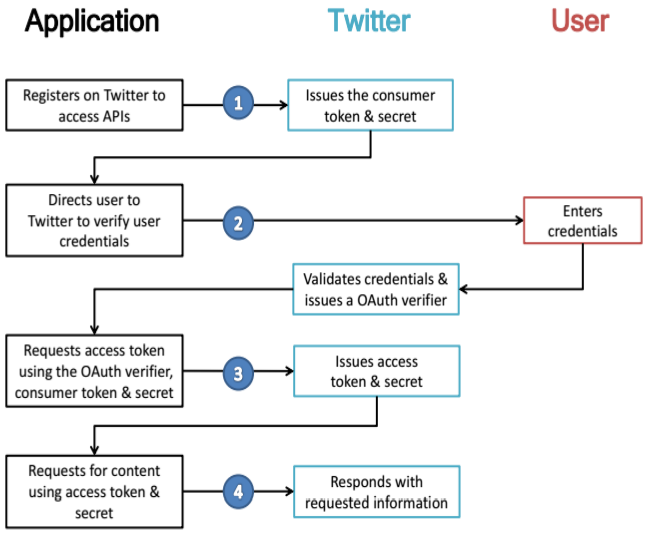

Application
Twitter
User
Registers on Twitter to access APIs
1
Issues the consumer token & secret
Directs user to Twitter to verify user credentials
2
Enters credentials
Validates credentials & issues a OAuth verifier
Requests access token using the OAuth verifier, consumer token & secret
3
Issues access token & secret
Requests for content using access token & secret
4
Responds with requested information

Figure 1: OAuth Overflow

VI. R – PROGRAMMING LANGUAGE

R is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. R is an implementation of the S programming language combined with lexical scoping semantics inspired by Scheme. S was created by John Chambers while at Bell Labs. There are some important differences, but much of the code written for S runs unaltered. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team, of which Chambers is a member. R is named partly after the first names of the first two R authors and partly as a play on the name of S. R is a GNU project. The source code for the R software environment is written primarily in C, Fortran, and R. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. R uses a command line interface; there are also several graphical front-ends for it.

R and its libraries implement a wide variety of statistical and graphical techniques including linear and non-linear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made. For computationally intensive tasks, C, C++, and Fortran code can be linked and called at run time. Advanced users can write C, C++,Java, .NET or Python code to manipulate R objects directly. R is highly extensible through the use of user-submitted packages for specific functions or specific areas of study. Due to its S heritage, R has stronger object-oriented programming facilities than most statistical computing languages. Extending R is also eased by its lexical scoping rules. Another strength of R is static graphics, which can produce publication-quality graphs, including mathematical symbols. Dynamic and interactive graphics are available through additional packages. R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hard copy.

A. Programming Features of R

R is an interpreted language; users typically access it through a command-line interpreter. If a user types 2+2 at the R command prompt and presses enter, the computer replies with 4, as shown below:

```
> 2+2
[1]  4
```

Like other similar languages such as APL and MATLAB, R supports matrix arithmetic. R’s data structures include vectors, matrices, arrays, data frames (similar to tables in a relational database) and lists. R's extensible object system includes objects for (among others): regression models, time-series and geo-spatial coordinates. The scalar data type was never a data structure of R. A scalar is represented as a vector with length one in R.

R supports procedural programming with functions, object oriented programming with generic functions. A generic function acts differently depending on the type of arguments passed to it. In other words, the generic function dispatches the function (method) specific to that type of object. For example, R has a generic print function that can print almost every type of object in R with a simple print (object tname) syntax. Although used mainly by statisticians and other practitioners requiring an environment for statistical computation and software development, R can also operate as a general matrix calculation toolbox with performance benchmarks comparable to GNU Octave or MATLAB. Arrays are stored in column-major order.

B. Packages for R

The capabilities of R are extended through user-created packages, which allow specialized statistical techniques, graphical devices (ggplot2), import/export capabilities, reporting tools (knitr, Sweave), etc. These packages are developed primarily in R, and sometimes in Java, C, C++ and Fortran. A core set of packages is included with the installation of R, with more than 7,000 additional packages (as of August 2015) available at the Comprehensive R Archive Network(CRAN), Biconductor, omegahat, GitHub and other repositories. The "Task Views" page (subject list) on the CRAN website lists a wide range of tasks (in fields such as Finance, Genetics, High Performance Computing, Machine Learning, Medical Imaging, Social Sciences and Spatial Statistics) to which R has been applied and for which packages are available. R has also been identified by the FDA as suitable for interpreting data from clinical research. Other R package resources include Crantastic, a community site for rating and reviewing all CRAN packages, and R-Forge, a central platform for the collaborative development of R packages, R-related software, and projects. R-Forge also hosts many unpublished beta packages, and development versions of CRAN packages. The Bioconductor project provides R packages for the analysis of genomic data, such as Affymetrix and cDNA microarray object-oriented data-handling and analysis tools, and has started to provide tools for analysis of data from next-generation high-throughput sequencing methods.

C. RStudio

RStudio is a free and open source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. Prepackaged distributions of RStudio Desktop are available for Microsoft Windows, Mac OS X, and Linux. RStudio is written in the C++ programming language and uses the Qt framework for its graphical user interface. Work on RStudio started at around December 2010, and the first public BETA version (v0.92) was officially announced in February 2011.

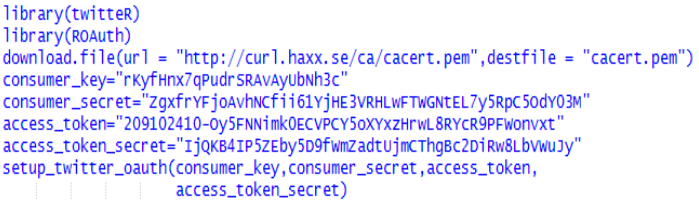
D. Providing Authentication to Twitter app through RStudio

The authentication of API requests on Twitter is carried out using OAuth. Twitter APIs can only be accessed by applications. Below we detail the steps for making an API call from a Twitter application using OAuth through RStudio:

Applications are also known as consumers and all applications are required to register themselves with Twitter. Through this process the application is issued a consumer key and secret which the application must use to authenticate itself to Twitter.

Using the "Access Token" and "Access Secret", the application authenticates the user on Twitter and issues API calls on behalf of the user

RScript for Twitter HandShake :



```
library(twitteR)
library(ROAuth)
download.file(url = "http://curl.haxx.se/ca/cacert.pem",destfile = "cacert.pem")
consumer_key="rKyfHnx7qPudrSRAvAyUbNh3c"
consumer_secret="ZgxfrYFjoAvhNCfii61YjHE3VRHLwFTWGNtEL7y5RpC5OdYO3M"
access_token="209102410-Oy5FNNimkOECVPCY5oXYxzHrwL8RYcR9PFWonvxt"
access_token_secret="IjQKB4IP5ZEby5D9fWmZadtUjmCThgBc2DiRw8LbVWuJy"
setup_twitter_oauth(consumer_key,consumer_secret,access_token,
                    access_token_secret)
```

Library(twitteR)—It provides an Interface to webAPI.

Library(ROAuth) -- Class OAuth wraps and handles OAuth handshakes and signatures for the user within R.

The OAuth class is currently implemented as a reference class. An instance of a generator for this class is provided as a convenience to the user as it is configured to handle most standard cases. To access this generator, use the object OAuthFactory. See theexamples section below for an example of how to instantiate an object of class OAuth.

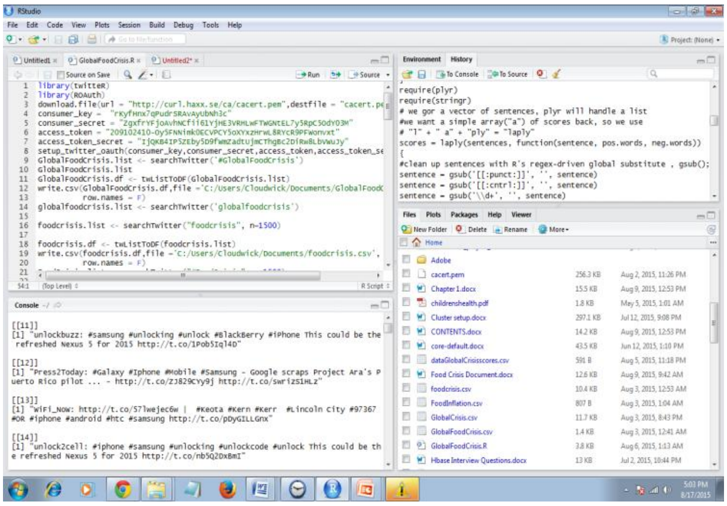
In almost all cases, saving an OAuth object after handshake and loading it into future sessions will allow it to remain authorized without needing any manual intervention that might have been performed initially, such as the PIN step with Twitter authentication. Use the function save to save the credential object to a file and then load in another R session to bring it back in - there should be no reason to undergo another handshake by doing this.

The needs Verifier argument is optional and defaults to TRUE. In almost all cases, the default should be used, the option is primarily provided to enable the examples as the keys provided by the examples are already signed. If you feel that you're in a situation where this should be set to FALSE, it's best to double check this. The sign Method to the handshake method tells the system which OAuth signature hash to use, one of HMAC for HMAC-SHA1 (default), RSA for RSA-SHA1 (not implemented), or text for plaintext. The custom Header argument to OAuth Request can be used to pass additional HTTP header commands to the underlying request.The curl arguments can be used to provide a custom curl header, defaulting to a generic getCurl Handle call.

- consumerKey: The consumer key provided by your application
- consumerSecret: The consumer secret provided by your application

Authentication occurs when the consumer key, consumer secret, access token, access token secret matches the keys and tokens generated during the app creation in twitter. After successful handshake with twitter using twitter app and through Rstudio script. It forms a bridge between the twitter and users to access the data.

User has to provide the hashtags, key words to get the tweets related to the specific tweets. For food crisis tweets hashtags #foodpricecrisis, #crisis, #worldfoodcrisis, #inflation, #droughts #heavyrains #foodinflation #crisis #weathercrisis etc are used. To get the tweets from twitter related to a particular hashtag .



RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
Untitled1 GlobalFoodCrisis.R Untitled2*
Source on Save
Run
Source
54:1 (Top Level) R Script
Console ~/
Environment History
To Console To Source
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home
5:03 PM
8/17/2015

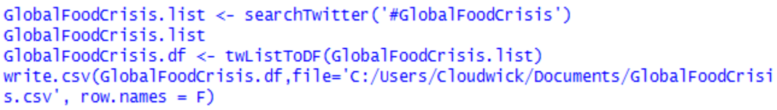
```
library(twitteR)
library(ROAuth)
download.file(url = "http://curl.haxx.se/ca/cacert.pem",destfile = "cacert.pe
consumer_key = "rKyfHnx7qPudrSRAvAyUbNh3c"
consumer_secret = "ZgxfrYFjoAvhNCfii61YjHE3VRHLwFTWGNtEL7y5RpC5OdYO3M"
access_token = "209102410-Oy5FNNimkOECVPCY5oXYxzHrwL8RYcR9PFWonvxt"
access_token_secret = "IjQKB4IP5ZEby5D9fWmZadtUjmCThgBc2DiRw8LbVWuJy"
setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_token_se
GlobalFoodCrisis.list <- searchTwitter('#GlobalFoodCrisis')
GlobalFoodCrisis.list
GlobalFoodCrisis.df <- twListToDF(GlobalFoodCrisis.list)
write.csv(GlobalFoodCrisis.df,file ='C:/Users/Cloudwick/Documents/GlobalFoodC
          row.names = F)
globalfoodcrisis.list <- searchTwitter('globalfoodcrisis')

foodcrisis.list <- searchTwitter("foodcrisis", n=1500)

foodcrisis.df <- twListToDF(foodcrisis.list)
write.csv(foodcrisis.df,file ='C:/Users/Cloudwick/Documents/foodcrisis.csv',
          row.names = F)
```

```
[[11]]
[1] "unlockbuzz: #samsung #unlocking #unlock #BlackBerry #iPhone This could be the
 refreshed Nexus 5 for 2015 http://t.co/1Pob5Iql4D"

[[12]]
[1] "Press2Today: #Galaxy #Iphone #Mobile #Samsung - Google scraps Project Ara's P
uerto Rico pilot ... - http://t.co/ZJ829CYy9j http://t.co/SwrIZS1HLz"

[[13]]
[1] "WiFi_Now: http://t.co/S7lwejec6w |  #Keota #Kern #Kerr  #Lincoln City #97367
#OR #iphone #android #htc #samsung http://t.co/pDyGILLGnx"

[[14]]
[1] "unlock2cell: #iphone #samsung #unlocking #unlockcode #unlock This could be th
e refreshed Nexus 5 for 2015 http://t.co/nb5Q2DxBmI"
```

```
require(plyr)
require(stringr)
# we gor a vector of sentences, plyr will handle a list
#we want a simple array("a") of scores back, so we use
# "l" + " a" + "ply" = "laply"
scores = laply(sentences, function(sentence, pos.words, neg.words))
{
#clean up sentences with R's regex-driven global substitute , gsub();
sentence = gsub('[[:punct:]]', '', sentence)
sentence = gsub('[[:cntrl:]]', '', sentence)
sentence = gsub('\\d+', '', sentence)
```

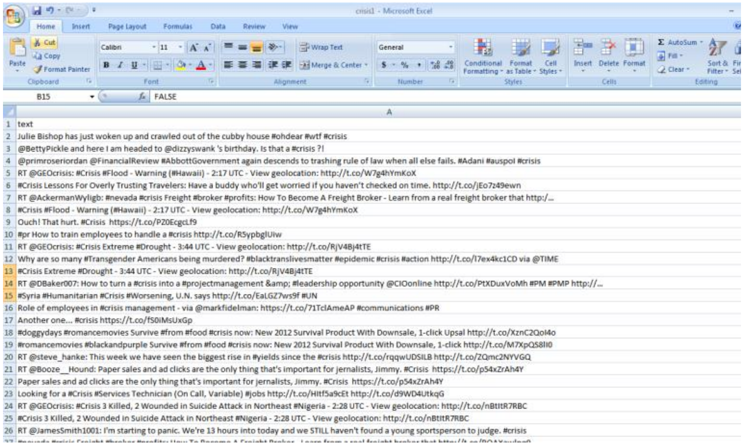
Adobe		
cacert.pem	256.3 KB	Aug 2, 2015, 11:26 PM
Chapter 1.docx	15.5 KB	Aug 9, 2015, 12:53 PM
childrenshealth.pdf	1.8 KB	May 5, 2015, 1:01 AM
Cluster setup.docx	297.1 KB	Jul 12, 2015, 9:08 PM
CONTENTS.docx	14.2 KB	Aug 9, 2015, 12:53 PM
core-default.docx	43.5 KB	Jun 12, 2015, 1:10 PM
dataGlobalCrisisscores.csv	591 B	Aug 5, 2015, 11:18 PM
Food Crisis Document.docx	12.6 KB	Aug 9, 2015, 9:42 AM
foodcrisis.csv	10.4 KB	Aug 3, 2015, 12:53 AM
FoodInflation.csv	807 B	Aug 3, 2015, 1:04 AM
GlobalCrisis.csv	11.7 KB	Aug 3, 2015, 8:43 PM
GlobalFoodCrisis.csv	1.4 KB	Aug 3, 2015, 12:41 AM
GlobalFoodCrisis.R	3.8 KB	Aug 6, 2015, 1:13 AM
Hbase Interview Questions.docx	13 KB	Jul 2, 2015, 10:44 PM

The minimum tweet count is 20 and maximum tweets limit can be upto 1500, but maximum limit is not achieved.



```
GlobalFoodCrisis.list <- searchTwitter('#GlobalFoodCrisis')
GlobalFoodCrisis.list
GlobalFoodCrisis.df <- twListToDF(GlobalFoodCrisis.list)
write.csv(GlobalFoodCrisis.df,file='C:/Users/Cloudwick/Documents/GlobalFoodCrisi
s.csv', row.names = F)
```

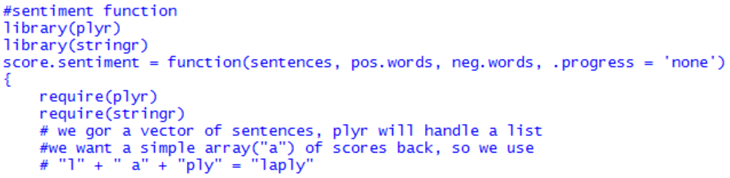
foodcrisis.list<-searchTwitter("foodcrisis",n=1500) foodcrisis.df <- twListToDF(foodcrisis.list)
write.csv(foodcrisis.df,file='C:/Users/Cloudwick/Documents/foodcrisis.csv', row.names = F)



B15 FALSE

	A
1	text
2	Julie Bishop has just woken up and crawled out of the cubby house #ohdear #wtf #crisis
3	@BettyPickle and here I am headed to @dizzyswank 's birthday. Is that a #crisis ?!
4	@primroseriordan @FinancialReview #AbbottGovernment again descends to trashing rule of law when all else fails. #Adani #auspol #crisis
5	RT @GEOcrisis: #Crisis #Flood - Warning (#Hawaii) - 2:17 UTC - View geolocation: http://t.co/W7g4hYmKoX
6	#Crisis Lessons For Overly Trusting Travelers: Have a buddy who'll get worried if you haven't checked on time. http://t.co/jEo7z49ewn
7	RT @AckermanWyligb: #nevada #crisis Freight #broker #profits: How To Become A Freight Broker - Learn from a real freight broker that http:/...
8	#Crisis #Flood - Warning (#Hawaii) - 2:17 UTC - View geolocation: http://t.co/W7g4hYmKoX
9	Ouch! That hurt. #Crisis https://t.co/PZ0EcgcLf9
10	#pr How to train employees to handle a #crisis http://t.co/R5ypbglUiw
11	RT @GEOcrisis: #Crisis Extreme #Drought - 3:44 UTC - View geolocation: http://t.co/RjV4Bj4tTE
12	Why are so many #Transgender Americans being murdered? #blacktranslivesmatter #epidemic #crisis #action http://t.co/l7ex4kc1CD via @TIME
13	#Crisis Extreme #Drought - 3:44 UTC - View geolocation: http://t.co/RjV4Bj4tTE
14	RT @DBaker007: How to turn a #crisis into a #projectmanagement &amp; #leadership opportunity @CIOonline http://t.co/PtXDuxVoMh #PM #PMP http://...
15	#Syria #Humanitarian #Crisis #Worsening, U.N. says http://t.co/EaLGZ7ws9f #UN
16	Role of employees in #crisis management - via @markfidelman: https://t.co/71TclAmeAP #communications #PR
17	Another one... #crisis https://t.co/fS0iMsUxGp
18	#doggydays #romancemovies Survive #from #food #crisis now: New 2012 Survival Product With Downsale, 1-click Upsal http://t.co/XznC2QoI4o
19	#romancemovies #blackandpurple Survive #from #food #crisis now: New 2012 Survival Product With Downsale, 1-click http://t.co/M7XpQS8lI0
20	RT @steve_hanke: This week we have seen the biggest rise in #yields since the #crisis http://t.co/rqqwUDSILB http://t.co/ZQmc2NYVGQ
21	RT @Booze__Hound: Paper sales and ad clicks are the only thing that's important for jernalists, Jimmy. #Crisis https://t.co/p54xZrAh4Y
22	Paper sales and ad clicks are the only thing that's important for jernalists, Jimmy. #Crisis https://t.co/p54xZrAh4Y
23	Looking for a #Crisis #Services Technician (On Call, Variable) #jobs http://t.co/Hltf5a9cEt http://t.co/d9WD4UtkqG
24	RT @GEOcrisis: #Crisis 3 Killed, 2 Wounded in Suicide Attack in Northeast #Nigeria - 2:28 UTC - View geolocation: http://t.co/nBtItR7RBC
25	#Crisis 3 Killed, 2 Wounded in Suicide Attack in Northeast #Nigeria - 2:28 UTC - View geolocation: http://t.co/nBtItR7RBC
26	RT @JamesSmith1001: I'm starting to panic. We're 13 hours into today and we STILL haven't found a young sportsperson to judge. #crisis

The tweets to be stored in R are stored primarly stored in the list datastructure. This list of tweets are further converted to dataframe using twListToDF method, which converts a list of files to a dataframe, the tweets in dataframe are further stored as csv file for analysis.



```
#sentiment function
library(plyr)
library(stringr)
score.sentiment = function(sentences, pos.words, neg.words, .progress = 'none')
{
    require(plyr)
    require(stringr)
    # we gor a vector of sentences, plyr will handle a list
    #we want a simple array("a") of scores back, so we use
    # "l" + " a" + "ply" = "laply"
```

E. Library (plyr): Tools for Splitting, Applying and Combining Data

plyr is an R package that makes it simple to split data apart, do stuff to it, and mash it back together. This is a common data-manipulation step. Importantly, plyr makes it easy to control the input and output data format from a syntactically consistent set of functions.

A set of tools that solves a common set of problems: you need to break a big problem down into manageable pieces, operate on each piece and then put all the pieces back together. For example, you might want to fit a model to each spatial location or time point in your study, summarise data by panels or collapse high-dimensional arrays to simpler summary statistics. The development of 'plyr' has been generously supported by 'Becton Dickinson'.

F. Library (stringr): Simple, Consistent Wrappers for Common String Operations

A consistent, simple and easy to use set of wrappers around the fantastic 'stringi' package. All function and argument names (and positions) are consistent, all functions deal with NAs and zero length vectors in the same way, and the output from one function is easy to feed into the input of another.

G. Basic String Operations

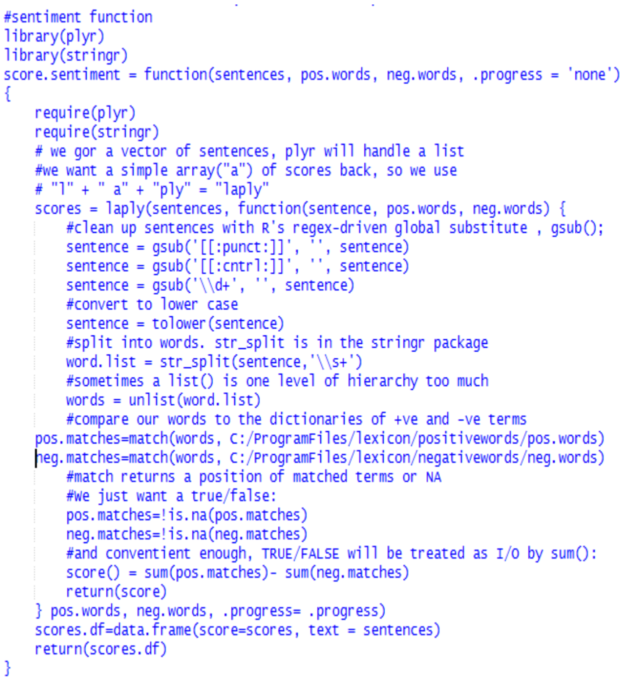
There are three string functions that are closely related to their base R equivalents, but with a few enhancements:

- str_c()is equivalent to paste(), but it uses the empty string (“ ”) as the default separator and silently removes NULL inputs.
- str_lenght()is equivalent to nchar() , but it preserves NA’s (rather than giving them length 2) and converts factors to characters (not integers).
- str_sub() is equivalent to substr() but it returns a zero length vector if any of its inputs are zero length, and otherwise expands each argument to match the longest. It also accepts negative positions, which are calculated from the left of the last character. The end position defaults to -1, which corresponds to the last character.
- str_str<-is equivalent to sub_str<-, but like str_sub it understands negative indices, and replacement strings not do need to be the same length as the string they are replacing.

Three functions add new functionality:

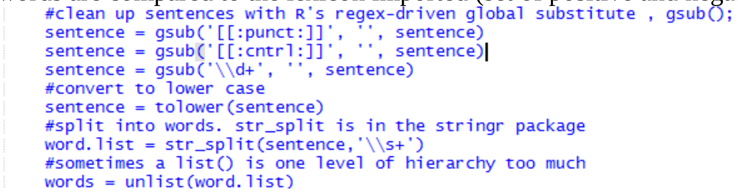
- str_dup() to duplicate the characters within a string.
- str_trim() to remove leading and trailing whitespace.
- str_pad() to pad a string with extra whitespace on the left, right, or both sides.

laply(x.list,function) - lapply(x) returns a list of the same length of x. The sentences, the positive and negative words are compared. lapply returns a list of the same length as X, each element of which is the result of applying FUN to the corresponding element of X.



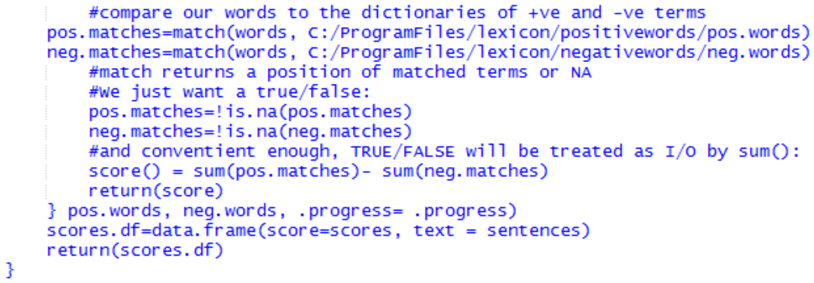
```
#sentiment function
library(plyr)
library(stringr)
score.sentiment = function(sentences, pos.words, neg.words, .progress = 'none')
{
    require(plyr)
    require(stringr)
    # we gor a vector of sentences, plyr will handle a list
    #we want a simple array("a") of scores back, so we use
    # "l" + " a" + "ply" = "laply"
    scores = laply(sentences, function(sentence, pos.words, neg.words) {
        #clean up sentences with R's regex-driven global substitute , gsub();
        sentence = gsub('[[:punct:]]', '', sentence)
        sentence = gsub('[[:cntrl:]]', '', sentence)
        sentence = gsub('\\d+', '', sentence)
        #convert to lower case
        sentence = tolower(sentence)
        #split into words. str_split is in the stringr package
        word.list = str_split(sentence,'\\s+')
        #sometimes a list() is one level of hierarchy too much
        words = unlist(word.list)
        #compare our words to the dictionaries of +ve and -ve terms
    pos.matches=match(words, C:/ProgramFiles/lexicon/positivewords/pos.words)
    neg.matches=match(words, C:/ProgramFiles/lexicon/negativewords/neg.words)
        #match returns a position of matched terms or NA
        #We just want a true/false:
        pos.matches=!is.na(pos.matches)
        neg.matches=!is.na(neg.matches)
        #and conventient enough, TRUE/FALSE will be treated as I/O by sum():
        score() = sum(pos.matches)- sum(neg.matches)
        return(score)
    } pos.words, neg.words, .progress= .progress)
    scores.df=data.frame(score=scores, text = sentences)
    return(scores.df)
}
```

Punctuations and control statements are replaced with spaces from the sentence using global substitute. The sentence is converted to lowercase since analysis is carried out only if the sentence is in lowercase. The sentence is further split into the list of words and these words are compared to the lexicon imported (set of positive and negative words).



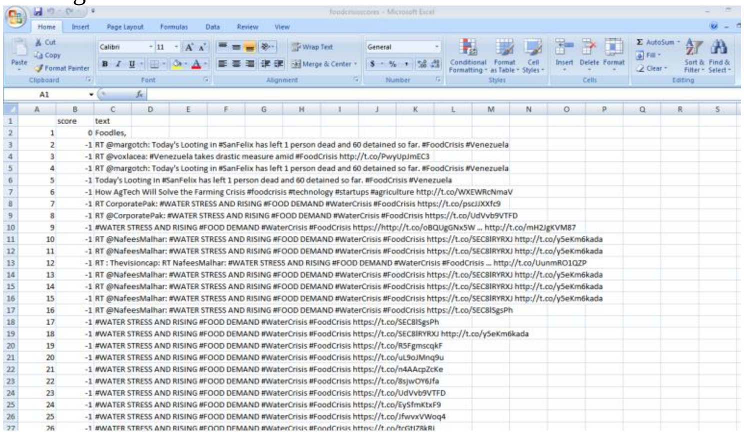
```
#clean up sentences with R's regex-driven global substitute , gsub();
sentence = gsub('[[:punct:]]', '', sentence)
sentence = gsub('[[:cntrl:]]', '', sentence)
sentence = gsub('\\d+', '', sentence)
#convert to lower case
sentence = tolower(sentence)
#split into words. str_split is in the stringr package
word.list = str_split(sentence,'\\s+')
#sometimes a list() is one level of hierarchy too much
words = unlist(word.list)
```

The list of words are compared to the set of lexicon (positive and negative words). The final result is the position of matched terms. We compared whether the word is a positive or a negative word. The final score is the difference between pos.matches and neg.matches.



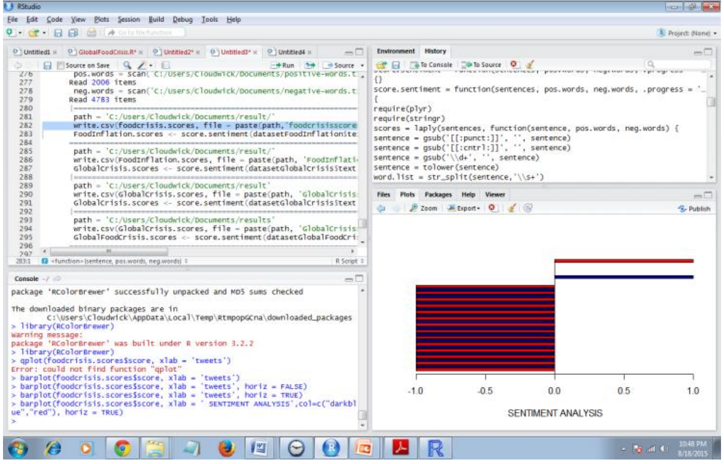
```
        #compare our words to the dictionaries of +ve and -ve terms
    pos.matches=match(words, C:/ProgramFiles/lexicon/positivewords/pos.words)
    neg.matches=match(words, C:/ProgramFiles/lexicon/negativewords/neg.words)
        #match returns a position of matched terms or NA
        #We just want a true/false:
        pos.matches=!is.na(pos.matches)
        neg.matches=!is.na(neg.matches)
        #and conventient enough, TRUE/FALSE will be treated as I/O by sum():
        score() = sum(pos.matches)- sum(neg.matches)
        return(score)
    } pos.words, neg.words, .progress= .progress)
    scores.df=data.frame(score=scores, text = sentences)
    return(scores.df)
}
```

Score 0 indicates the tweets is neutral.
Score 1 indicates the tweet is positive.
Score -1 indicates the tweet is negative.



	A	B	C
1		score	text
2	1	0	Foodles,
3	2	-1	RT @margotch: Today's Looting in #SanFelix has left 1 person dead and 60 detained so far. #FoodCrisis #Venezuela
4	3	-1	RT @voxlacea: #Venezuela takes drastic measure amid #FoodCrisis http://t.co/PwyUpJmEC3
5	4	-1	RT @margotch: Today's Looting in #SanFelix has left 1 person dead and 60 detained so far. #FoodCrisis #Venezuela
6	5	-1	Today's Looting in #SanFelix has left 1 person dead and 60 detained so far. #FoodCrisis #Venezuela
7	6	-1	How AgTech Will Solve the Farming Crisis #foodcrisis #technology #startups #agriculture http://t.co/WXEWRcNmaV
8	7	-1	RT CorporatePak: #WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/pscJJXXfc9
9	8	-1	RT @CorporatePak: #WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/UdVvb9VTFD
10	9	-1	#WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://http://t.co/oBQUgGNx5W … http://t.co/mH2JgKVM87
11	10	-1	RT @NafeesMalhar: #WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/SEC8lRYRXJ http://t.co/y5eKm6kada
12	11	-1	RT @NafeesMalhar: #WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/SEC8lRYRXJ http://t.co/y5eKm6kada
13	12	-1	RT : Thevisioncap: RT NafeesMalhar: #WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis … http://t.co/UunmRO1QZP
14	13	-1	RT @NafeesMalhar: #WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/SEC8lRYRXJ http://t.co/y5eKm6kada
15	14	-1	RT @NafeesMalhar: #WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/SEC8lRYRXJ http://t.co/y5eKm6kada
16	15	-1	RT @NafeesMalhar: #WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/SEC8lRYRXJ http://t.co/y5eKm6kada
17	16	-1	RT @NafeesMalhar: #WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/SEC8lSgsPh
18	17	-1	#WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/SEC8lSgsPh
19	18	-1	#WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/SEC8lRYRXJ http://t.co/y5eKm6kada
20	19	-1	#WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/R5FgmscqkF
21	20	-1	#WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/uL9oJMnq9u
22	21	-1	#WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/n4AAcpZcKe
23	22	-1	#WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/8sjwOY6Jfa
24	23	-1	#WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/UdVvb9VTFD
25	24	-1	#WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/EySfmKtxF9
26	25	-1	#WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/JfwvxVWoq4
27	26	-1	#WATER STRESS AND RISING #FOOD DEMAND #WaterCrisis #FoodCrisis https://t.co/tcGtlZ8kRI

Output: (Rstudio Output)
Output graphs are generated using the scores of the tweets.



RStudio
File Edit Code View Plots Session Build Debug Tools Help
Untitled1 GlobalFoodCrisis.R* Untitled2* Untitled3* Untitled4
Source on Save Run Source
Environment History To Console To Source
Files Plots Packages Help Viewer Zoom Export Publish
Project: (None)
-1.0
-0.5
0.0
0.5
1.0
SENTIMENT ANALYSIS
10:48 PM
8/16/2015

```
 pos.words = scan('C:/Users/Cloudwick/Documents/positive-words.t
Read 2006 items
 neg.words = scan('C:/Users/Cloudwick/Documents/negative-words.t
Read 4783 items

 path = 'C:/Users/Cloudwick/Documents/result/'
 write.csv(foodcrisis.scores, file = paste(path,'foodcrisisscore
 FoodInflation.scores <- score.sentiment(datasetFoodInflation$te

 path = 'C:/Users/Cloudwick/Documents/result/'
 write.csv(FoodInflation.scores, file = paste(path, 'FoodInflati
 Globalcrisis.scores <- score.sentiment(datasetGlobalcrisis$text

 path = 'C:/Users/Cloudwick/Documents/result'
 write.csv(GlobalCrisis.scores, file = paste(path, 'GlobalCrisis
 GlobalCrisis.scores <- score.sentiment(datasetGlobalCrisis$text

 path = 'C:/Users/Cloudwick/Documents/results'
 write.csv(GlobalCrisis.scores, file = paste(path, 'GlobalCrisis
 GlobalFoodCrisis.scores <- score.sentiment(datasetGlobalFoodCri
```

```
283:1  <function>(sentence, pos.words, neg.words)  R Script
```

```
{}
score.sentiment = function(sentences, pos.words, neg.words, .progress = '
{
require(plyr)
require(stringr)
scores = laply(sentences, function(sentence, pos.words, neg.words) {
sentence = gsub('[[:punct:]]', '', sentence)
sentence = gsub('[[:cntrl:]]', '', sentence)
sentence = gsub('\\d+', '', sentence)
sentence = tolower(sentence)
word.list = str_split(sentence,'\\s+')
```

```
package 'RColorBrewer' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Cloudwick\AppData\Local\Temp\RtmpopGCna\downloaded_packages
> library(RColorBrewer)
Warning message:
package 'RColorBrewer' was built under R version 3.2.2
> library(RColorBrewer)
> qplot(foodcrisis.scores$score, xlab = 'tweets')
Error: could not find function "qplot"
> barplot(foodcrisis.scores$score, xlab = 'tweets')
> barplot(foodcrisis.scores$score, xlab = 'tweets', horiz = FALSE)
> barplot(foodcrisis.scores$score, xlab = 'tweets', horiz = TRUE)
> barplot(foodcrisis.scores$score, xlab = ' SENTIMENT ANALYSIS',col=c("darkbl
ue","red"), horiz = TRUE)
>
```

VII. CONCLUSIONS

Driven by real-world applications and key industrial stakeholders and initialized by national funding agencies, managing and mining Big Data have shown to be a challenging yet very compelling task. While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations. Such combined characteristics suggest that Big Data require a “big mind” to consolidate data for maximum values.

To explore Big Data, we have analyzed several challenges at the data, model, and system levels. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values.

In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data.

At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors. We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time.

The new applications are generating vast amount of data in structured and unstructured form. Big data is able to process and store that data and probably in more amounts in near future. Hopefully, Hadoop will get better. New technologies and tools that have ability to record, monitor measure and combine all kinds of data around us, are going to be introduced soon. We will need new technologies and tools for anonym zing data, analysis, tracking and auditing information, sharing and managing, our own personal data in future. So many aspects of life health, education, telecommunication, marketing, sports and business etc that manages big data world need to be polished in future.

REFERENCES

[1] Hadoop The Definitive Guide, Storage and Analysis at Internet Scale, O’Reilly, 4thEdition,Tom White.

[2] Alekh Agarwal and Pushpak Bhattacharyya, Sentiment Analysis: “A New Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified” International Conference on Natural Language Processing (ICON 05), IIT Kanpur, India, December, 2005.

[3] Programming Hive, O’Reilly, Jason Rutherglen, DeanWampler & Edward Capriolo.

[4] Liu, Bing , "Sentiment Analysis and Subjectivity", In Indurkhya, N.; Damerau, F. J., 2010.

[5] "A Twitter and web sentiment analysis tool". werfamous.com.

[6] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[7] Pang, Bo; Lee, Lillian, "4.1.2 Subjectivity Detection and Opinion Identification". Opinion Mining and Sentiment Analysis. 2008.

[8] Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining" . Proceedings of LREC. pp. 2200–2204.