Half-Day Tutorial (dg.o 2011 Conference)

Collecting, Analyzing and Visualizing Tweets using Open Source Tools

Seungwon Yang and Andrea L. Kavanaugh
Department of Computer Science, Virginia Tech
{seungwon, kavan}@vt.edu

6/12/2011

This tutorial will teach participants how to collect, analyze and visualize results from twitter data. We will demonstrate several different free, open-source web-based tools that participants can use to collect twitter data (e.g., Archivist, 140kit.com, TwapperKeeper), and show them a few different methods, tools or programs they can use to analyze the data in a given collection. Finally, we will show participants visualization tools and programs they can use to present the analyses, such as tag clouds, graphs and other data clustering techniques. As much as possible this will be a hands-on tutorial, so participants can learn by making their own twitter data collection, analysis and visualization as part of the tutorial.

Table of Contents

Part 1. Web-Based Tools

1. Collecting Tweets

Twitter provides tweets through their REST & Search API and through their Streaming API[1]. The REST & Search API is used to find relevant tweets that are already archived in Twitter's servers. Tweets as old as 7 days usually can be collected from this API. The Streaming API, also called a 'fire hose' API, provides current tweets that are posted in real time. Most tweet collection tools use both of these APIs to archive existing older tweets as well as incoming new tweets.

Terminology

Hashtag
They are a community-driven convention for adding additional context and metadata to your tweets. It is added inline to your post. You create a hashtag simply by prefixing a word with a hash symbol (e.g., #twitter, #japanearthquake, etc.). For more information, please see http://twitter.pbworks.com/w/page/1779812/Hashtags

Retweet (RT)
It is similar to forwarding email to another person. It helps quickly share tweets with all of your followers. The retweeted tweets look like "RT @VerifiedQuotes: I'm not addicted to #twitter. I only tweet…", with 'RT' in front of the text. For more information, please see http://support.twitter.com/entries/77606-what-is-retweet-rt

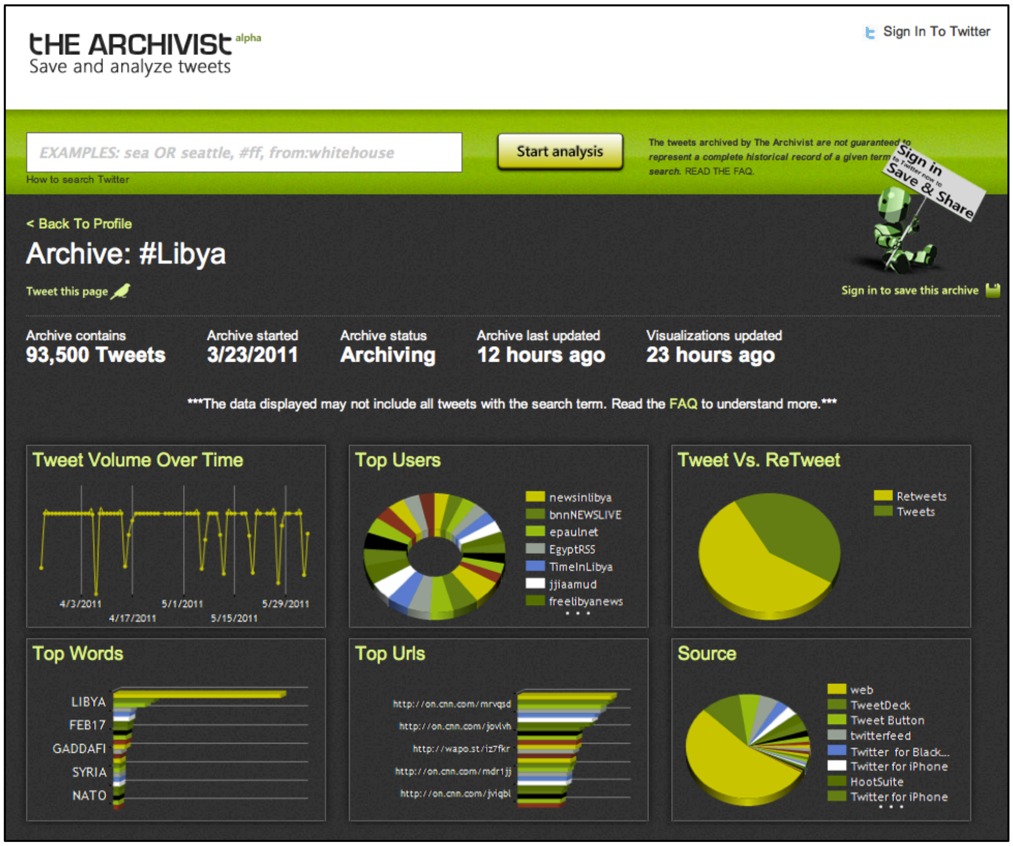
1.1. The Archivist

Web version

The (Online) Archivist[2] developed by Mix Online provides quick and easy creation of tweet visualizations. However, due to Twitter's API Terms and Service, collections created by users reside in the company's servers. In addition, only three collections per account can be created. Export/download of the collected tweets is not allowed. However, it provides six visualizations for the basic analysis of archived tweets:

- Tweet volume over time: the number of tweets posted during a period of time
- Top users: the user IDs of people who tweeted much
- Tweet vs. Retweet: the ratio of original tweets and retweets
- Top words: a list of frequent words in tweets
- Top URLs: a list of frequent URLs found in tweets
- Source: the name of an online service that is used to post tweets

[1] Twitter API FAQ. http://dev.twitter.com/pages/api_faq
[2] http://archivist.visitmix.com/



tHE ARCHIVISt alpha

Save and analyze tweets

Sign In To Twitter

EXAMPLES: sea OR seattle, #ff, from:whitehouse

How to search Twitter

Start analysis

The tweets archived by The Archivist *are not guaranteed to represent a complete historical record of a given term search.* READ THE FAQ.

< Back To Profile

Archive: #Libya

Tweet this page

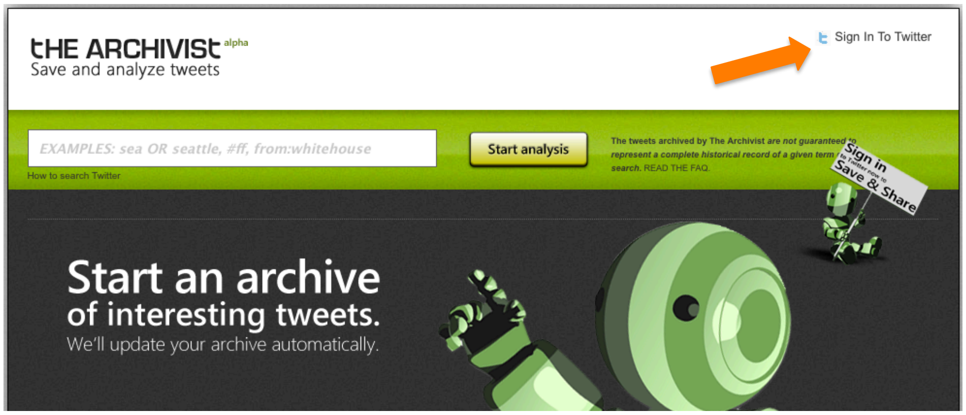
Sign in to save this archive

Archive contains	Archive started	Archive status	Archive last updated	Visualizations updated
93,500 Tweets	3/23/2011	Archiving	12 hours ago	23 hours ago

The data displayed may not include all tweets with the search term. Read the FAQ to understand more.

Instructions to create a new tweet archive and view the visualization of the basic analyses in The Archivist (Web version)

Step 1. You can sign-in using your Twitter account. Click ‘Sign In To Twitter’ link on the top right corner of the main page.



tHE ARCHIVISt alpha

Save and analyze tweets

Sign In To Twitter

EXAMPLES: sea OR seattle, #ff, from:whitehouse

How to search Twitter

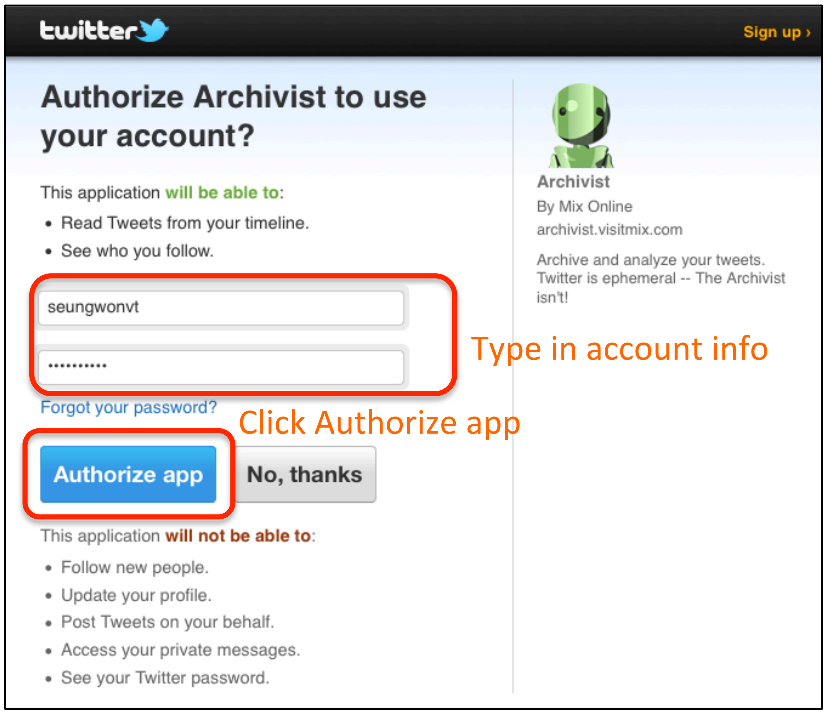
Start analysis

The tweets archived by The Archivist *are not guaranteed to represent a complete historical record of a given term search.* READ THE FAQ.

Start an archive of interesting tweets.

We'll update your archive automatically.

Step 2. Type in your Twitter account info, and click ‘Authorize app’ button.



Authorize Archivist to use your account?

This application **will be able to**:

- Read Tweets from your timeline.
- See who you follow.

seungwonvt

••••••••••

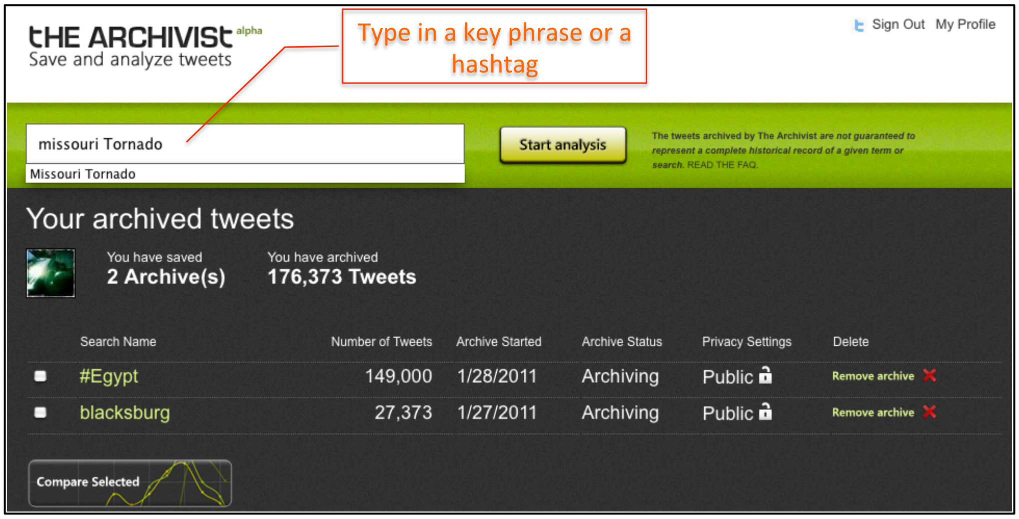
This application **will not be able to**:

- Follow new people.
- Update your profile.
- Post Tweets on your behalf.
- Access your private messages.
- See your Twitter password.

Archivist
By Mix Online
archivist.visitmix.com

Archive and analyze your tweets. Twitter is ephemeral -- The Archivist isn't!

Step 3. Enter a key phrase or a hashtag of your interest. Then click 'Start analysis' button right next to the text box. After some seconds, The Archivist will show six different visualizations, each of which has a link to its full-page view.



missouri Tornado

Missouri Tornado

The tweets archived by The Archivist are not guaranteed to represent a complete historical record of a given term or search. READ THE FAQ.

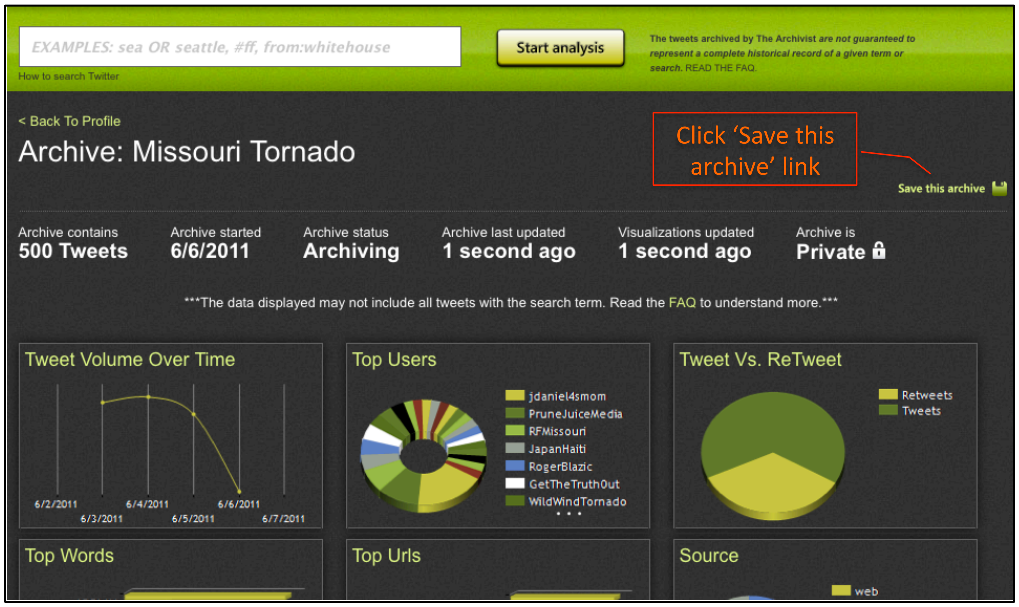
Your archived tweets

You have saved **2 Archive(s)**

You have archived **176,373 Tweets**

	Search Name	Number of Tweets	Archive Started	Archive Status	Privacy Settings	Delete
■	#Egypt	149,000	1/28/2011	Archiving	Public	Remove archive ✕
■	blacksburg	27,373	1/27/2011	Archiving	Public	Remove archive ✕

Step 4. To keep archiving the tweets, you should save the archive in your profile page by clicking 'Save this archive'.



The tweets archived by The Archivist are not guaranteed to represent a complete historical record of a given term or search. READ THE FAQ.

< Back To Profile

Archive: Missouri Tornado

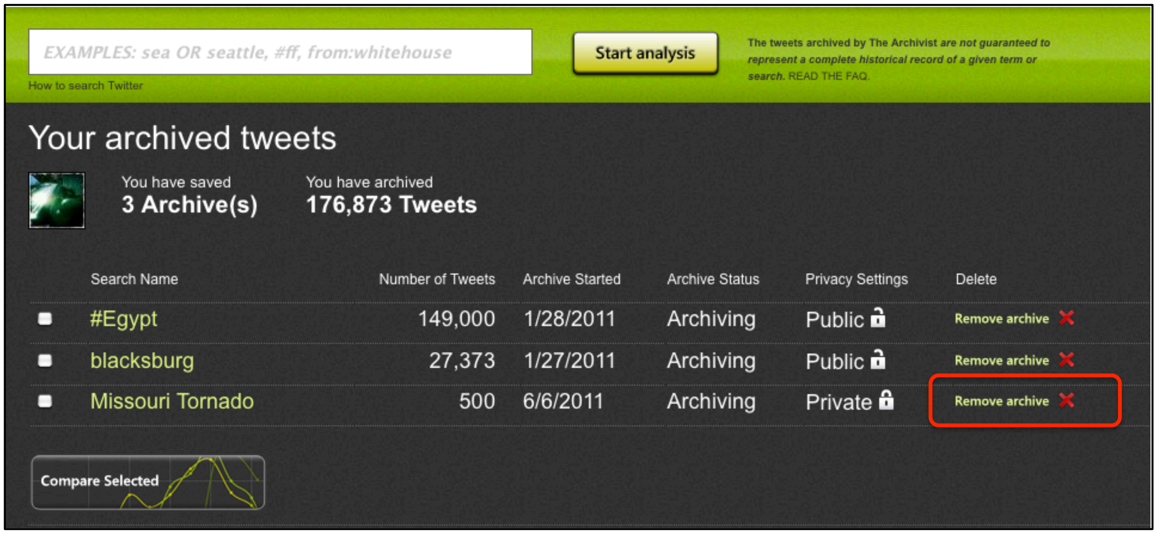
Click 'Save this archive' link

Save this archive

Archive contains 500 Tweets | Archive started 6/6/2011 | Archive status Archiving | Archive last updated 1 second ago | Visualizations updated 1 second ago | Archive is Private

The data displayed may not include all tweets with the search term. Read the FAQ to understand more.

Step 5. To remove an archive, simply click the 'Remove archive' link on your profile page.



The tweets archived by The Archivist are not guaranteed to represent a complete historical record of a given term or search. READ THE FAQ.

Your archived tweets

You have saved 3 Archive(s) | You have archived 176,873 Tweets

	Search Name	Number of Tweets	Archive Started	Archive Status	Privacy Settings	Delete
□	#Egypt	149,000	1/28/2011	Archiving	Public	Remove archive ✕
□	blacksburg	27,373	1/27/2011	Archiving	Public	Remove archive ✕
□	Missouri Tornado	500	6/6/2011	Archiving	Private	Remove archive ✕

Note: to have more than 3 archives, you can sign out from Twitter and create an archive. Then, click 'Save this archive' link. It will direct you to the Twitter login page. Once you login, the newly created archive is added to your profile page even if your total number of archive has already exceeded the limit of 3. (May be there is a bug in the software's logic.)

(You need a Twitter account to do this exercise. Please make one if you don't have it yet.)

Exercise

1. Participants get together and have a short brainstorming session.

2. Discuss current issues of interest.
3. Each participant selects an event/issue that he/she would like to collect tweets about.
4. Develop a keyword or a key phrase that represents the event/issue.
5. Use the developed key word or a key phrase as a query to a search engine. See if it returns a good result.
6. Use the same keyword or a key phrase as a search term in Twitter. See if it returns relevant tweets.
7. Modify the keyword or a key phrase.
8. Create an archive using The Archivist Web version using your keyword or key phrase.
9. From six different visualizations, which one(s) do you like the most? Which one will be most useful for your research?

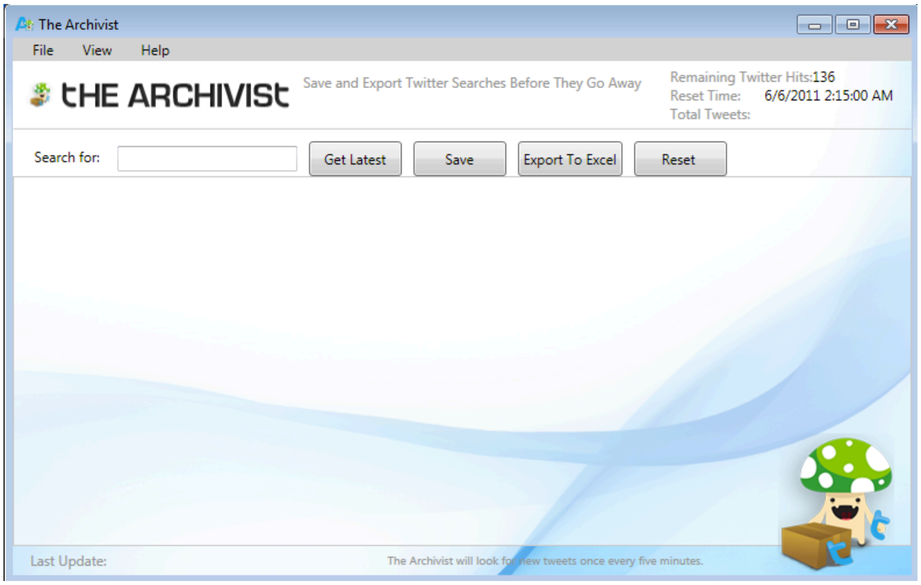
* The same search keyword / key phrase will be used to create archives in other tools in this tutorial session to compare the collection results.

Stand-alone version (Windows PC only)

The Archivist Desktop[3] can run on a user's own **Windows** machine and continuously collect tweets. Tweets can be exported into an XML file or tab-delimited text file for later processing using Excel. Its pie chart visualization shows tweet volumes per Twitter ID. Tweet volume is visualized as a line graph. *The Archivist Desktop* runs continuously in the background and takes up much computing resources so it might result in performance degradation when multiple archives are created in a single machine.

Step 1. Installing the tool
The software can be downloaded from http://visitmix.com/labs/archivist-desktop/
Once installed by saving and double-clicking the setup.exe file, run the tool. You should see its user interface that looks like the image below:

[3] http://visitmix.com/labs/archivist-desktop/



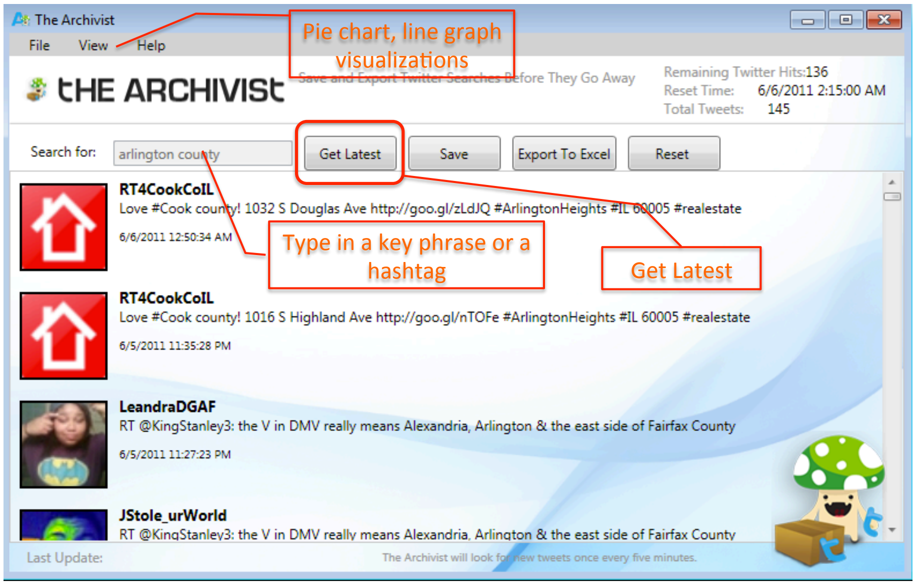
Remaining Twitter Hits:136
Reset Time: 6/6/2011 2:15:00 AM
Total Tweets:

Search for:

Last Update:

The Archivist will look for new tweets once every five minutes.

Step 2. Collecting tweets
Type a key phrase or a hashtag into the box, which is located next to ‘Search for’. Then click ‘Get Latest’ button (the left-most one among control buttons) to begin archiving tweets.


The Archivist
File
View
Help
Pie chart, line graph visualizations
tHE ARCHIVISt
Save and Export Twitter Searches Before They Go Away
Get Latest
Save
Export To Excel
Reset
Type in a key phrase or a hashtag
Get Latest

Remaining Twitter Hits:136
Reset Time: 6/6/2011 2:15:00 AM
Total Tweets: 145

Search for: arlington county

RT4CookCoIL
Love #Cook county! 1032 S Douglas Ave http://goo.gl/zLdJQ #ArlingtonHeights #IL 60005 #realestate
6/6/2011 12:50:34 AM

RT4CookCoIL
Love #Cook county! 1016 S Highland Ave http://goo.gl/nTOFe #ArlingtonHeights #IL 60005 #realestate
6/5/2011 11:35:28 PM

LeandraDGAF
RT @KingStanley3: the V in DMV really means Alexandria, Arlington & the east side of Fairfax County
6/5/2011 11:27:23 PM

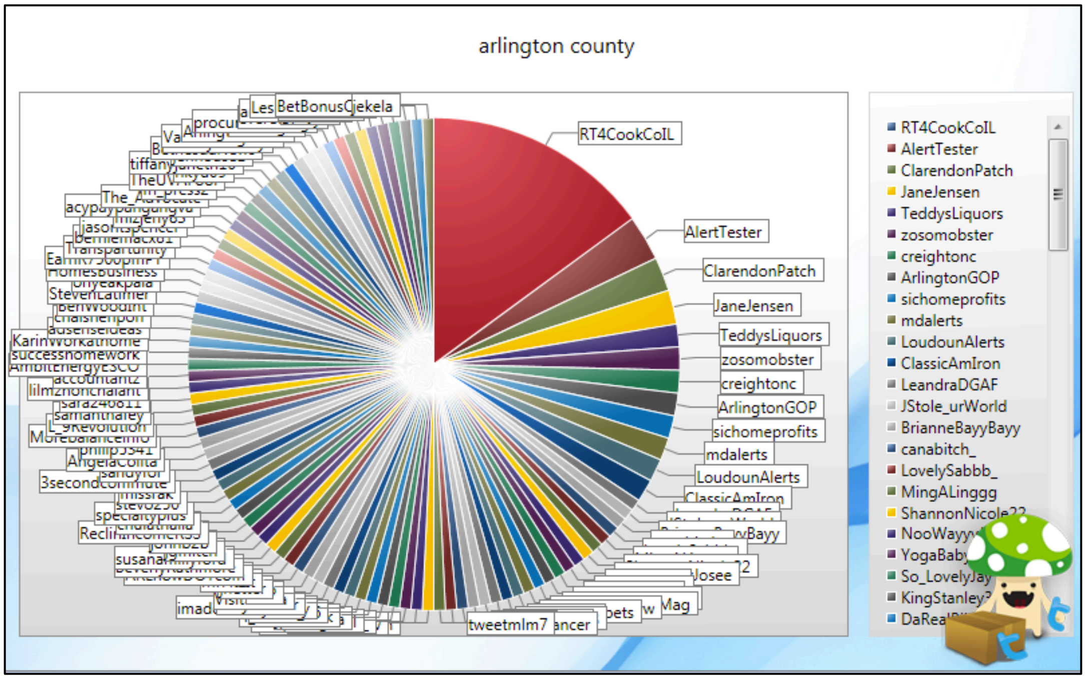
JStole_urWorld
RT @KingStanley3: the V in DMV really means Alexandria, Arlington & the east side of Fairfax County

Last Update:

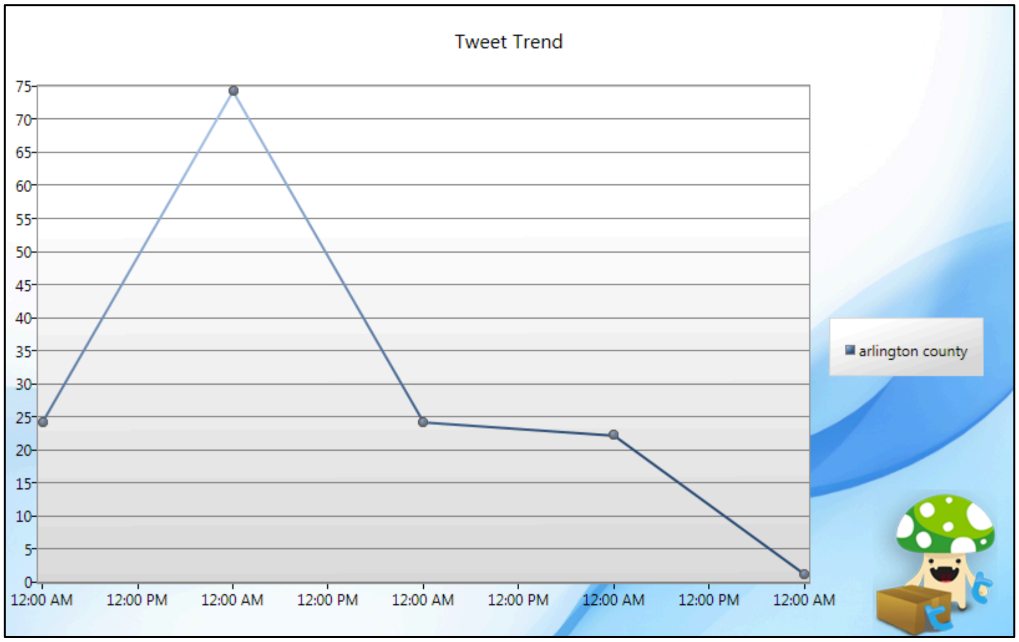
The Archivist will look for new tweets once every five minutes.

Step 3. Two visualizations.

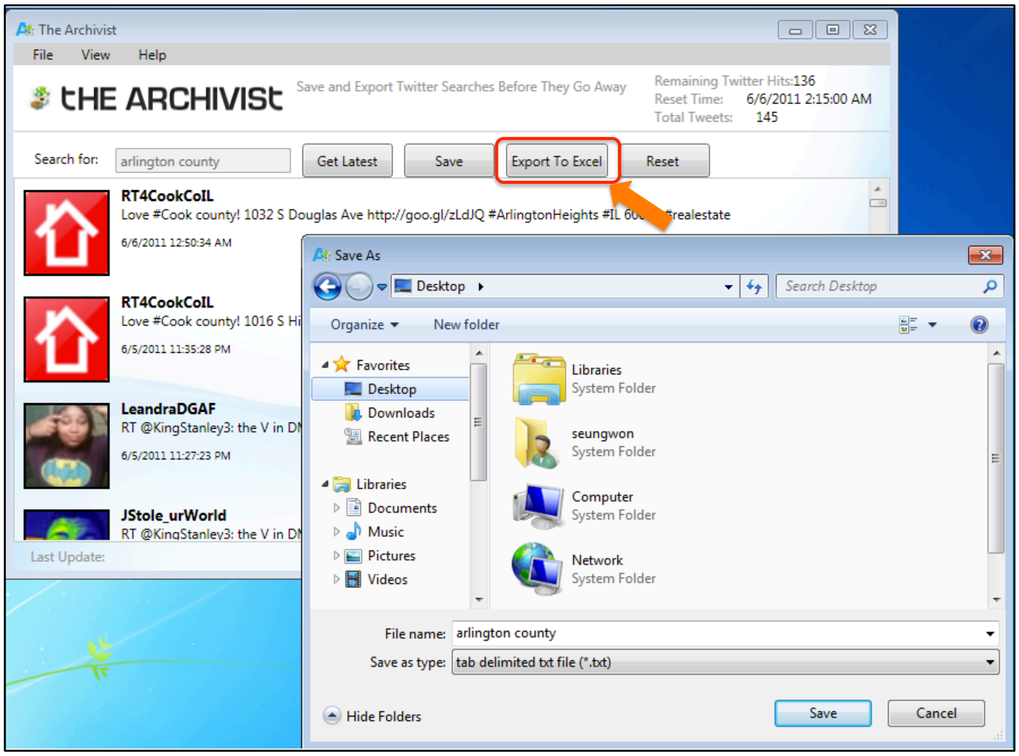
Pie chart visualization of tweets per tweet IDs:
Go to ‘View’ on the task bar and select ‘View Pie’. The pie chart shows the proportion of tweets from each person.



Tweet volume visualization overtime:
Go to ‘View’ on the task bar and select ‘View Chart’. The graph shows the number of tweets over a period of time.



Step 4. Exporting tweets.
There are two options for exporting collected tweets. First option is to export as a tab-delimited text file, which can be imported into Excel for analysis. For this, click ‘Export to Excel’ button on the software interface as shown in Figure 16. It will open a window, where you can select the folder to store the exported file.



Remaining Twitter Hits:136
Reset Time: 6/6/2011 2:15:00 AM
Total Tweets: 145

Search for: arlington county

RT4CookCoIL
Love #Cook county! 1032 S Douglas Ave http://goo.gl/zLdJQ #ArlingtonHeights #IL 600 #realestate
6/6/2011 12:50:34 AM

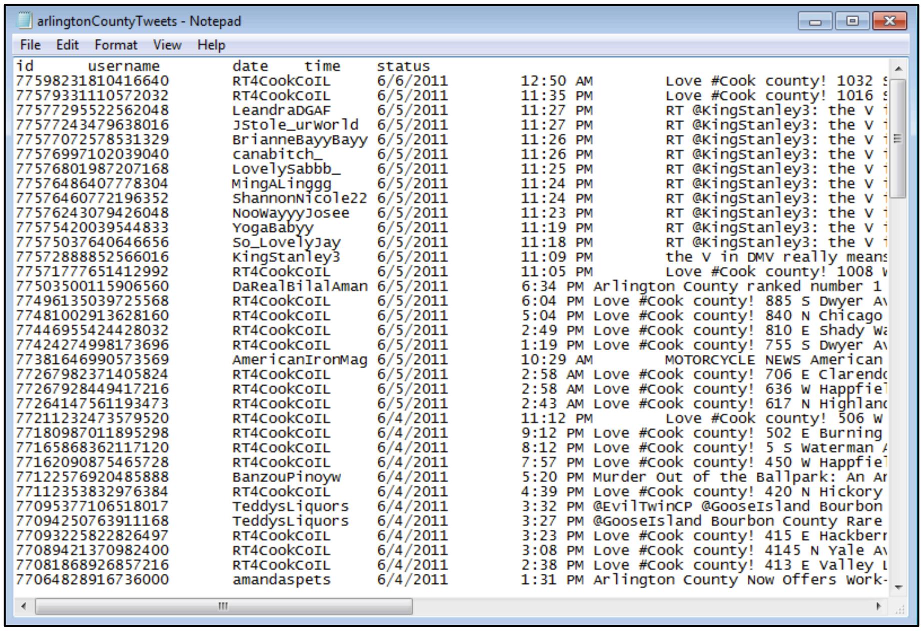
RT4CookCoIL
Love #Cook county! 1016 S Hi
6/5/2011 11:35:28 PM

LeandraDGAF
RT @KingStanley3: the V in DM
6/5/2011 11:27:23 PM

JStole_urWorld
RT @KingStanley3: the V in DM

File name: arlington county
Save as type: tab delimited txt file (*.txt)

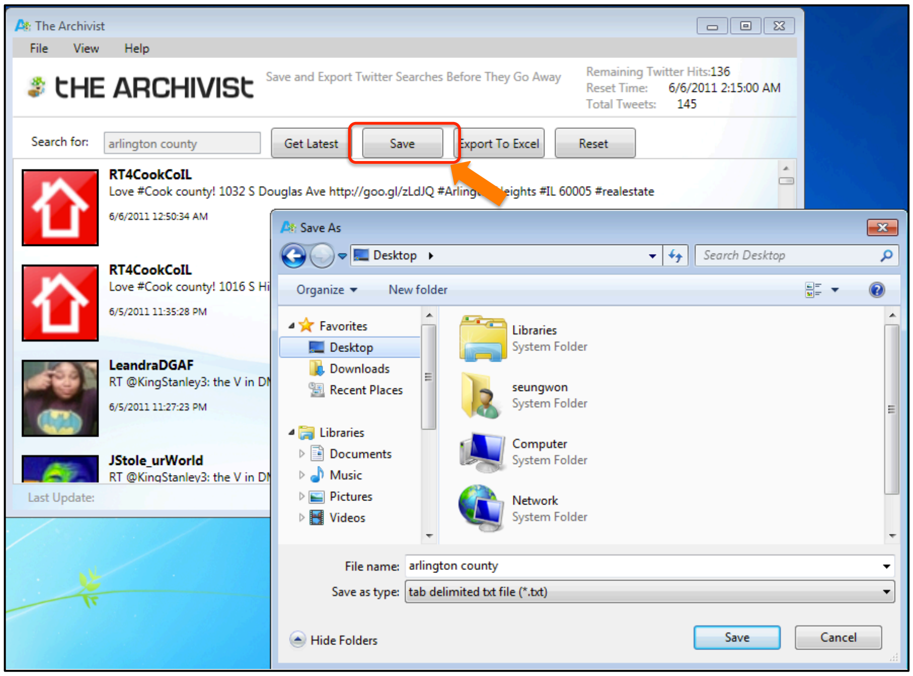
(Tab-delimited text file example)



```
id      username        date    time    status
77598231810416640       RT4CookCoIL     6/6/2011        12:50 AM        Love #Cook county! 1032
77579331110572032       RT4CookCoIL     6/5/2011        11:35 PM        Love #Cook county! 1016
77577295522562048       LeandraDGAF     6/5/2011        11:27 PM        RT @KingStanley3: the V
77577243479638016       JStole_urWorld  6/5/2011        11:27 PM        RT @KingStanley3: the V
77577072578531329       BrianneBayyBayy 6/5/2011        11:26 PM        RT @KingStanley3: the V
77576997102039040       canabitch_      6/5/2011        11:26 PM        RT @KingStanley3: the V
77576801987207168       LovelySabbb_    6/5/2011        11:25 PM        RT @KingStanley3: the V
77576486407778304       MingALinggg     6/5/2011        11:24 PM        RT @KingStanley3: the V
77576460772196352       ShannonNicole22 6/5/2011        11:24 PM        RT @KingStanley3: the V
77576243079426048       NooWayyyJosee   6/5/2011        11:23 PM        RT @KingStanley3: the V
77575420039544833       YogaBabyy       6/5/2011        11:19 PM        RT @KingStanley3: the V
77575037640646656       So_LovelyJay    6/5/2011        11:18 PM        RT @KingStanley3: the V
77572888852566016       KingStanley3    6/5/2011        11:09 PM        the V in DMV really mean
77571777651412992       RT4CookCoIL     6/5/2011        11:05 PM        Love #Cook county! 1008
77503500115906560       DaRealBilalAman 6/5/2011        6:34 PM Arlington County ranked number 1
77496135039725568       RT4CookCoIL     6/5/2011        6:04 PM Love #Cook county! 885 S Dwyer A
77481002913628160       RT4CookCoIL     6/5/2011        5:04 PM Love #Cook county! 840 N Chicago
77446955424428032       RT4CookCoIL     6/5/2011        2:49 PM Love #Cook county! 810 E Shady W
77424274998173696       RT4CookCoIL     6/5/2011        1:19 PM Love #Cook county! 755 S Dwyer A
77381646990573569       AmericanIronMag 6/5/2011        10:29 AM        MOTORCYCLE NEWS American
77267982371405824       RT4CookCoIL     6/5/2011        2:58 AM Love #Cook county! 706 E Clarend
77267928449417216       RT4CookCoIL     6/5/2011        2:58 AM Love #Cook county! 636 W Happfie
77264147561193473       RT4CookCoIL     6/5/2011        2:43 AM Love #Cook county! 617 N Highlan
77211232473579520       RT4CookCoIL     6/4/2011        11:12 PM        Love #Cook county! 506 W
77180987011895298       RT4CookCoIL     6/4/2011        9:12 PM Love #Cook county! 502 E Burning
77165868362117120       RT4CookCoIL     6/4/2011        8:12 PM Love #Cook county! 5 S Waterman
77162090875465728       RT4CookCoIL     6/4/2011        7:57 PM Love #Cook county! 450 W Happfie
77122576920485888       BanzouPinoyw    6/4/2011        5:20 PM Murder Out of the Ballpark: An A
77112353832976384       RT4CookCoIL     6/4/2011        4:39 PM Love #Cook county! 420 N Hickory
77095377106518017       TeddysLiquors   6/4/2011        3:32 PM @EvilTwinCP @GooseIsland Bourbon
77094250763911168       TeddysLiquors   6/4/2011        3:27 PM @GooseIsland Bourbon County Rare
77093225822826497       RT4CookCoIL     6/4/2011        3:23 PM Love #Cook county! 415 E Hackber
77089421370982400       RT4CookCoIL     6/4/2011        3:08 PM Love #Cook county! 4145 N Yale A
77081868926857216       RT4CookCoIL     6/4/2011        2:38 PM Love #Cook county! 413 E Valley
77064828916736000       amandaspets     6/4/2011        1:31 PM Arlington County Now Offers Work-
```

Note: After exporting tweets as a tab-delimited text file, please change the filename to avoid the file content becoming an xml file.

The other option is to export as an XML file that can be parsed with scripts written in Python or PHP. For this, click the ‘Save’ button.



Remaining Twitter Hits:136
Reset Time: 6/6/2011 2:15:00 AM
Total Tweets: 145

Search for: arlington county

RT4CookCoIL
Love #Cook county! 1032 S Douglas Ave http://goo.gl/zLdJQ #Arlingt...eights #IL 60005 #realestate
6/6/2011 12:50:34 AM

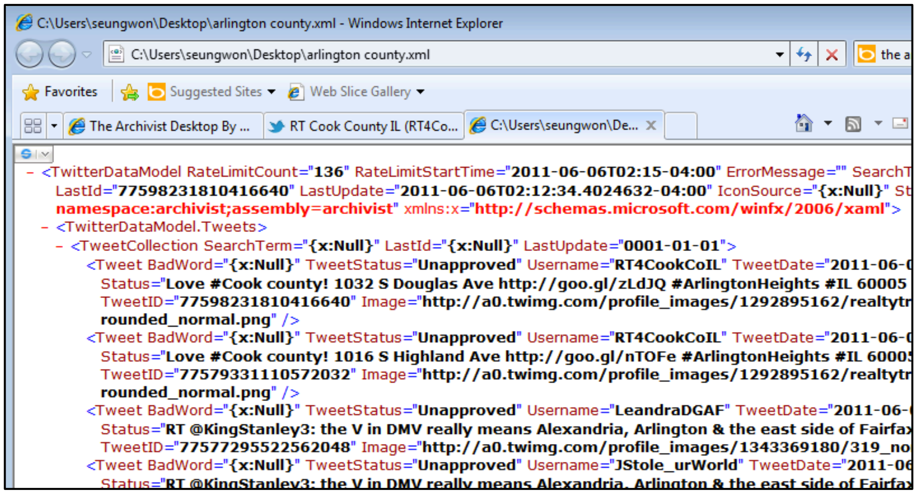
RT4CookCoIL
Love #Cook county! 1016 S Hi
6/5/2011 11:35:28 PM

LeandraDGAF
RT @KingStanley3: the V in DI
6/5/2011 11:27:23 PM

JStole_urWorld
RT @KingStanley3: the V in DI

File name: arlington county
Save as type: tab delimited txt file (*.txt)

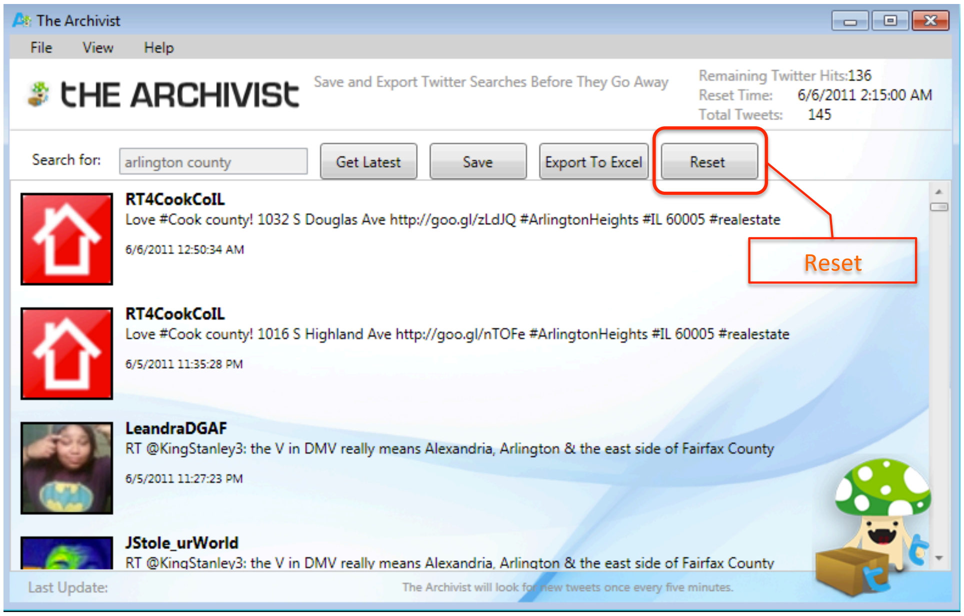
(XML export file example)


C:\Users\seungwon\Desktop\arlington county.xml - Windows Internet Explorer
C:\Users\seungwon\Desktop\arlington county.xml
Favorites
Suggested Sites
Web Slice Gallery
The Archivist Desktop By ...
RT Cook County IL (RT4Co...
C:\Users\seungwon\De...

```
- <TwitterDataModel RateLimitCount="136" RateLimitStartTime="2011-06-06T02:15-04:00" ErrorMessage="" SearchT
    LastId="77598231810416640" LastUpdate="2011-06-06T02:12:34.4024632-04:00" IconSource="{x:Null}" St
    namespace:archivist;assembly=archivist" xmlns:x="http://schemas.microsoft.com/winfx/2006/xaml">
  - <TwitterDataModel.Tweets>
    - <TweetCollection SearchTerm="{x:Null}" LastId="{x:Null}" LastUpdate="0001-01-01">
        <Tweet BadWord="{x:Null}" TweetStatus="Unapproved" Username="RT4CookCoIL" TweetDate="2011-06-0
          Status="Love #Cook county! 1032 S Douglas Ave http://goo.gl/zLdJQ #ArlingtonHeights #IL 60005
          TweetID="77598231810416640" Image="http://a0.twimg.com/profile_images/1292895162/realtytr
          rounded_normal.png" />
        <Tweet BadWord="{x:Null}" TweetStatus="Unapproved" Username="RT4CookCoIL" TweetDate="2011-06-0
          Status="Love #Cook county! 1016 S Highland Ave http://goo.gl/nTOFe #ArlingtonHeights #IL 6000
          TweetID="77579331110572032" Image="http://a0.twimg.com/profile_images/1292895162/realtytr
          rounded_normal.png" />
        <Tweet BadWord="{x:Null}" TweetStatus="Unapproved" Username="LeandraDGAF" TweetDate="2011-06-
          Status="RT @KingStanley3: the V in DMV really means Alexandria, Arlington & the east side of Fairfax
          TweetID="77577295522562048" Image="http://a0.twimg.com/profile_images/1343369180/319_no
        <Tweet BadWord="{x:Null}" TweetStatus="Unapproved" Username="JStole_urWorld" TweetDate="2011-06
          Status="RT @KingStanley3: the V in DMV really means Alexandria, Arlington & the east side of Fairfax
```

Step 5. Stop collecting tweets.

After running the tool for a period of time to archive tweets, you may want to stop archiving. Clicking the 'Reset' button or closing the tool's interface by clicking 'X' on the top right corner will stop archiving the tweets. Please don't forget to save your tweets.



Remaining Twitter Hits:136
Reset Time: 6/6/2011 2:15:00 AM
Total Tweets: 145

Search for: arlington county

RT4CookCoIL
Love #Cook county! 1032 S Douglas Ave http://goo.gl/zLdJQ #ArlingtonHeights #IL 60005 #realestate
6/6/2011 12:50:34 AM

RT4CookCoIL
Love #Cook county! 1016 S Highland Ave http://goo.gl/nTOFe #ArlingtonHeights #IL 60005 #realestate
6/5/2011 11:35:28 PM

LeandraDGAF
RT @KingStanley3: the V in DMV really means Alexandria, Arlington & the east side of Fairfax County
6/5/2011 11:27:23 PM

JStole_urWorld
RT @KingStanley3: the V in DMV really means Alexandria, Arlington & the east side of Fairfax County

Last Update:
The Archivist will look for new tweets once every five minutes.

Note: Since *The Archivist Desktop* instances are running in the computer's main memory and not storing the collected tweets into a database. Therefore, all the tweets will be lost if the tool stops running (e.g., due to automatic rebooting of the computer, or by the user's decision to stop archiving tweets.). If the computer turns into a sleep mode, the collection process stops, and starts again once the computer returns to a normal mode.

1.2. yourTwapperKeeper

To avoid violation of Twitter's API Terms of Service, *TwapperKeeper*[4], the public web service for archiving tweets, had discontinued its export and download features, which are essential for researchers to analyze tweet content.

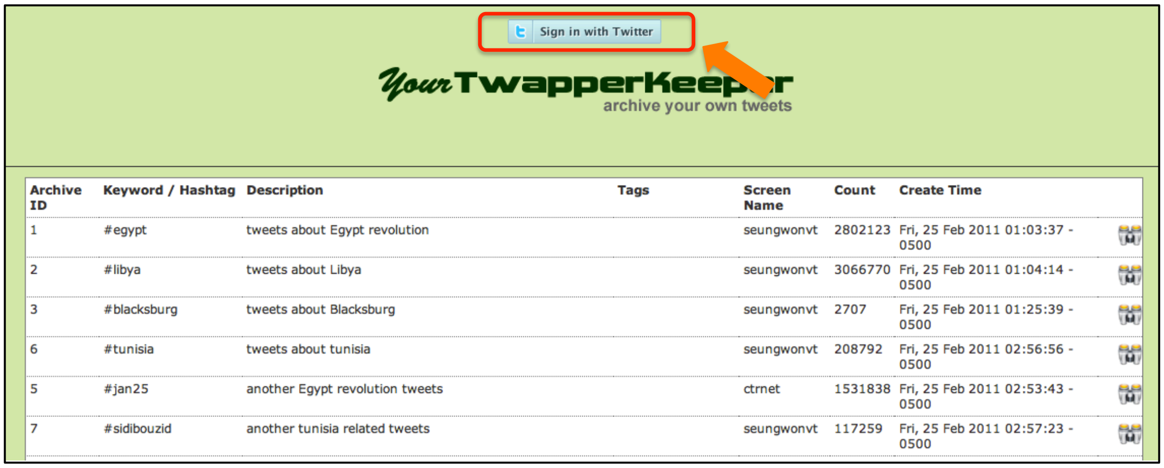
YourTwapperKeeper[5], the open source version of *TwapperKeeper*, can run on a user's own machine to archive tweets with certain hashtags (e.g., #libya) and key words (e.g., japan earthquake). Users can export and download archived tweets into various formats such as an Excel file, RSS, JSON and HTML. Because *YourTwapperKeeper* resides on a non-public machine, exporting tweets is not a violation of any terms of service.

Instructions to develop an archive and export its content

Creating an archive

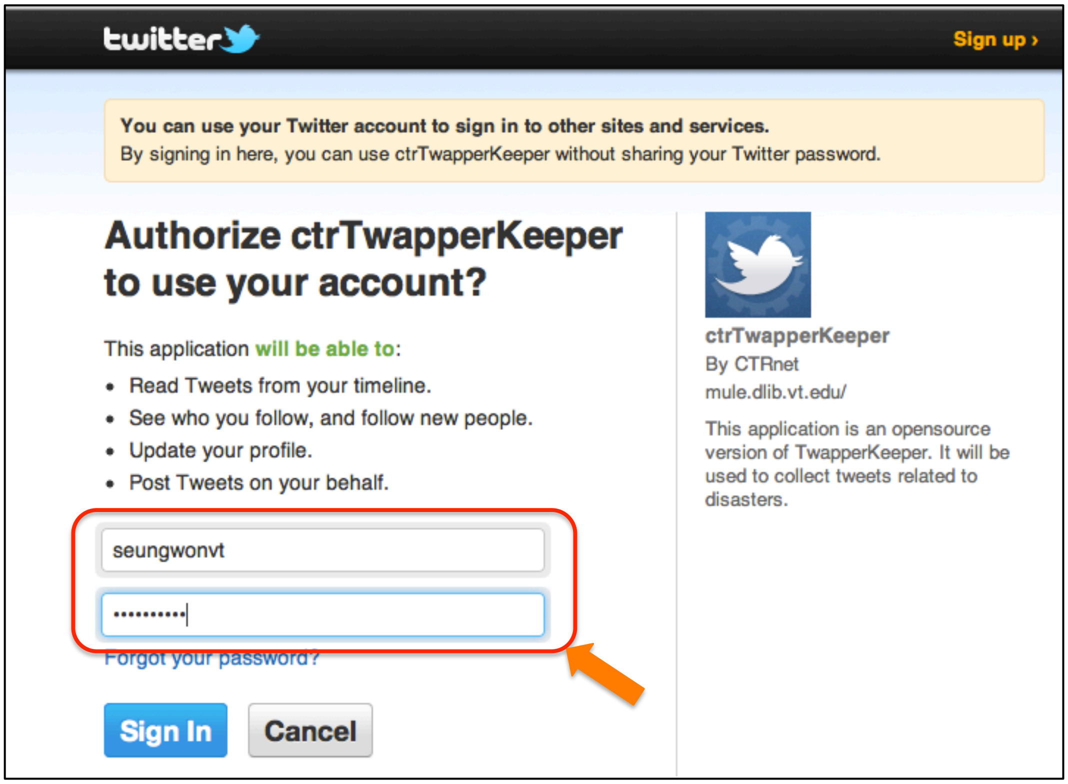
Step 1. Click the blue button on top of the user interface.

[4] http://twapperkeeper.com/index.php
[5] http://your.twapperkeeper.com/



Archive ID	Keyword / Hashtag	Description	Tags	Screen Name	Count	Create Time
1	#egypt	tweets about Egypt revolution		seungwonvt	2802123	Fri, 25 Feb 2011 01:03:37 -0500
2	#libya	tweets about Libya		seungwonvt	3066770	Fri, 25 Feb 2011 01:04:14 -0500
3	#blacksburg	tweets about Blacksburg		seungwonvt	2707	Fri, 25 Feb 2011 01:25:39 -0500
6	#tunisia	tweets about tunisia		seungwonvt	208792	Fri, 25 Feb 2011 02:56:56 -0500
5	#jan25	another Egypt revolution tweets		ctrnet	1531838	Fri, 25 Feb 2011 02:53:43 -0500
7	#sidibouzid	another tunisia related tweets		seungwonvt	117259	Fri, 25 Feb 2011 02:57:23 -0500

Step 2. You let *yourTwapperKeeper* access your Twitter account info. Then the control is passed back to the application.



You can use your Twitter account to sign in to other sites and services.
By signing in here, you can use ctrTwapperKeeper without sharing your Twitter password.

Authorize ctrTwapperKeeper to use your account?

This application will be able to:

- Read Tweets from your timeline.
- See who you follow, and follow new people.
- Update your profile.
- Post Tweets on your behalf.

seungwonvt

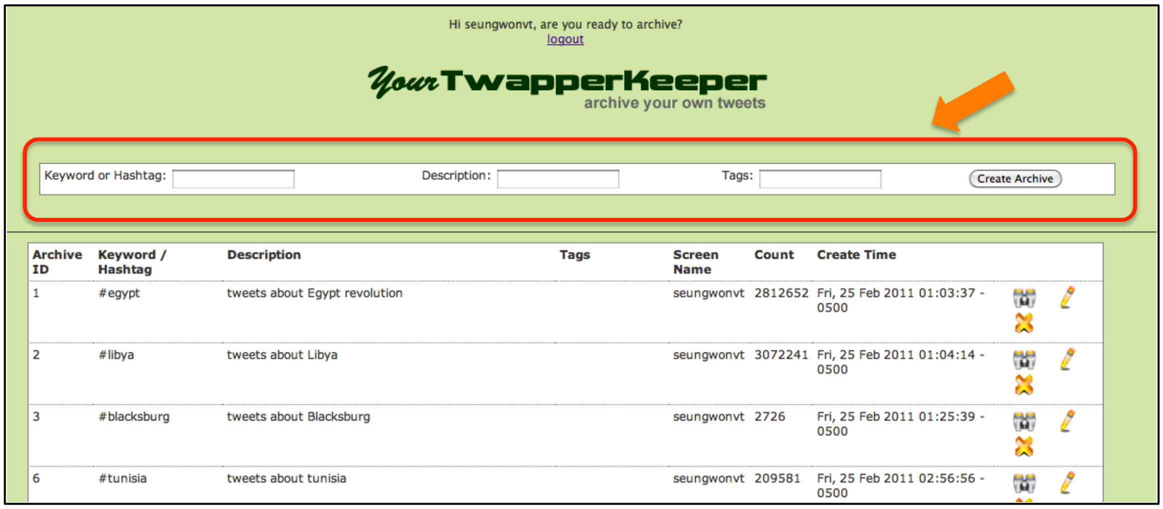
••••••••••

Forgot your password?

ctrTwapperKeeper
By CTRnet
mule.dlib.vt.edu/

This application is an opensource version of TwapperKeeper. It will be used to collect tweets related to disasters.

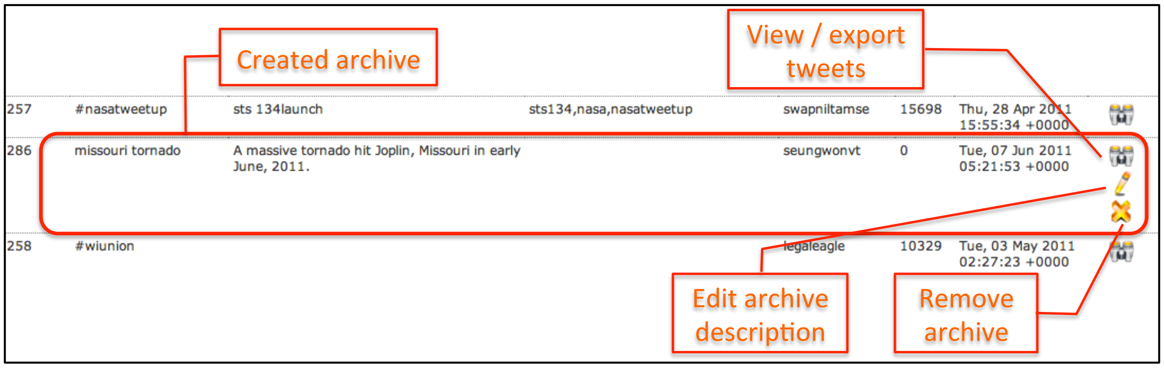
Step 3. Type in a key phrase/hashtag, description of your archive and optional tags. Then, click ‘Create Archive’ button.



Keyword or Hashtag: Description: Tags:

Archive ID	Keyword / Hashtag	Description	Tags	Screen Name	Count	Create Time
1	#egypt	tweets about Egypt revolution		seungwonvt	2812652	Fri, 25 Feb 2011 01:03:37 -0500
2	#libya	tweets about Libya		seungwonvt	3072241	Fri, 25 Feb 2011 01:04:14 -0500
3	#blacksburg	tweets about Blacksburg		seungwonvt	2726	Fri, 25 Feb 2011 01:25:39 -0500
6	#tunisia	tweets about tunisia		seungwonvt	209581	Fri, 25 Feb 2011 02:56:56 -0500

Step 4. Once the archive is successfully created, it is displayed on the list of archives as shown below. Only the creator of an archive can remove the archive or edit its description. Users have access to all archives. They can view and export tweets from archives that were created by anyone.



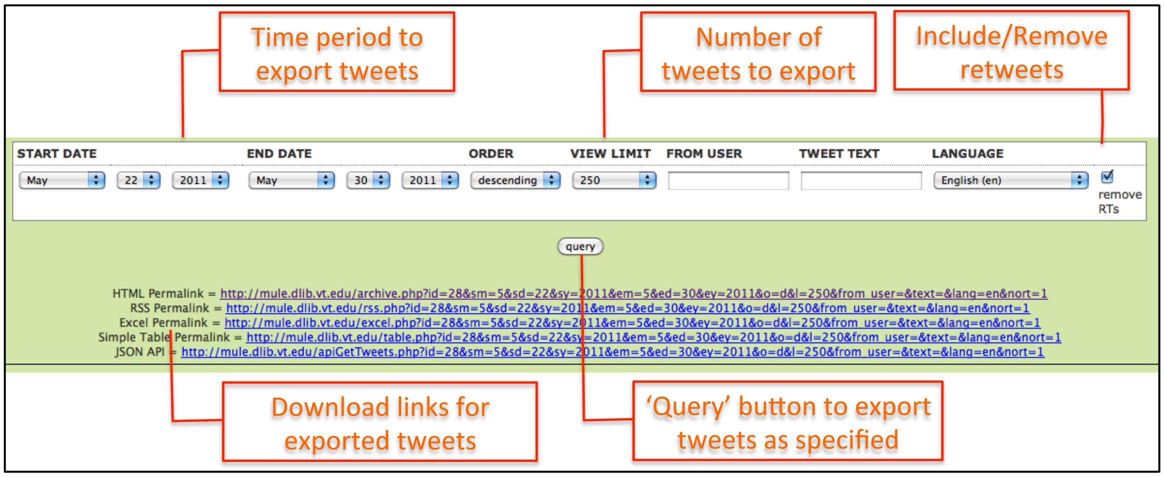
257	#nasatweetup	sts 134launch	sts134,nasa,nasatweetup	swapniltamse	15698	Thu, 28 Apr 2011 15:55:34 +0000
286	missouri tornado	A massive tornado hit Joplin, Missouri in early June, 2011.		seungwonvt	0	Tue, 07 Jun 2011 05:21:53 +0000
258	#wiunion			legaleagle	10329	Tue, 03 May 2011 02:27:23 +0000

Clicking a binocular icon on the right side of an archive opens a tab on a browser, where the raw tweets can be exported in various formats such as HTML, RSS, Excel, Simple Table, or JSON API.

Exporting an archive as an Excel file

Step 5. Select settings for tweet export. Users can select tweets within a certain time period, the maximum number of tweets to export, tweets from a certain user, tweets that has a certain text, and whether to include retweets or not.

After selecting and typing in all the settings, click 'Query' button to generate download links for various export formats.



Time period to export tweets

Number of tweets to export

Include/Remove retweets

START DATE			END DATE			ORDER	VIEW LIMIT	FROM USER	TWEET TEXT	LANGUAGE	
May	22	2011	May	30	2011	descending	250			English (en)	☑ remove RTs

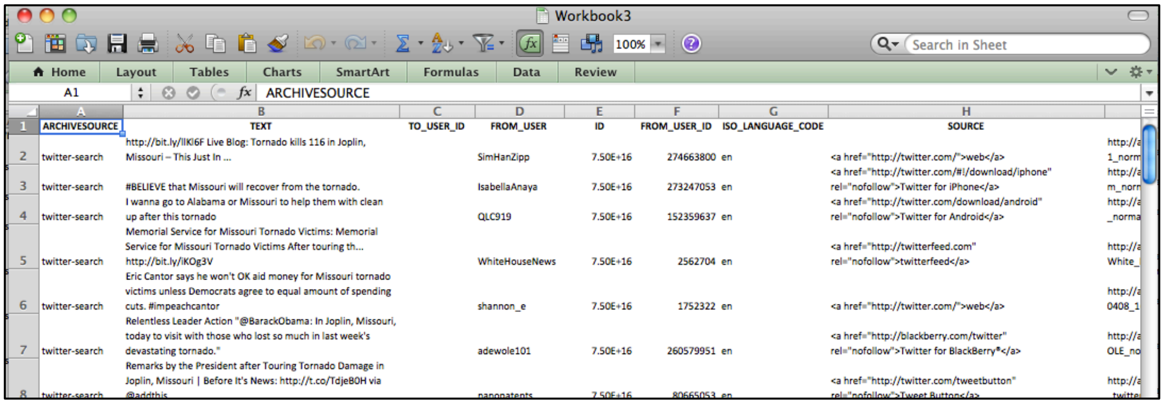
query

HTML Permalink = http://mule.dlib.vt.edu/archive.php?id=28&sm=5&sd=22&sy=2011&em=5&ed=30&ey=2011&o=d&l=250&from_user=&text=&lang=en&nort=1
RSS Permalink = http://mule.dlib.vt.edu/rss.php?id=28&sm=5&sd=22&sy=2011&em=5&ed=30&ey=2011&o=d&l=250&from_user=&text=&lang=en&nort=1
Excel Permalink = http://mule.dlib.vt.edu/excel.php?id=28&sm=5&sd=22&sy=2011&em=5&ed=30&ey=2011&o=d&l=250&from_user=&text=&lang=en&nort=1
Simple Table Permalink = http://mule.dlib.vt.edu/table.php?id=28&sm=5&sd=22&sy=2011&em=5&ed=30&ey=2011&o=d&l=250&from_user=&text=&lang=en&nort=1
JSON API = http://mule.dlib.vt.edu/apiGetTweets.php?id=28&sm=5&sd=22&sy=2011&em=5&ed=30&ey=2011&o=d&l=250&from_user=&text=&lang=en&nort=1

Download links for exported tweets

'Query' button to export tweets as specified

Step 6. Exported tweets in an Excel file.
Columns include the tweet text, to-user ID, from-user, tweet ID, language setting, geo location coordinates, tweet time, etc.



A1 fx ARCHIVESOURCE

	A	B	C	D	E	F	G	H	
1	ARCHIVESOURCE	TEXT	TO_USER_ID	FROM_USER	ID	FROM_USER_ID	ISO_LANGUAGE_CODE	SOURCE	
2	twitter-search	http://bit.ly/lIKl6F Live Blog: Tornado kills 116 in Joplin, Missouri – This Just In ...		SimHanZipp	7.50E+16	274663800	en	<a href="http://twitter.com/">web</a>	http://a 1_norm
3	twitter-search	#BELIEVE that Missouri will recover from the tornado.		IsabellaAnaya	7.50E+16	273247053	en	<a href="http://twitter.com/#!/download/iphone" rel="nofollow">Twitter for iPhone</a>	http://a m_norn
4	twitter-search	I wanna go to Alabama or Missouri to help them with clean up after this tornado		QLC919	7.50E+16	152359637	en	<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>	http://a _norma
5	twitter-search	Memorial Service for Missouri Tornado Victims: Memorial Service for Missouri Tornado Victims After touring th... http://bit.ly/iKOg3V		WhiteHouseNews	7.50E+16	2562704	en	<a href="http://twitterfeed.com" rel="nofollow">twitterfeed</a>	http://a White_
6	twitter-search	Eric Cantor says he won't OK aid money for Missouri tornado victims unless Democrats agree to equal amount of spending cuts. #impeachcantor		shannon_e	7.50E+16	1752322	en	<a href="http://twitter.com/">web</a>	http://a 0408_1
7	twitter-search	Relentless Leader Action "@BarackObama: In Joplin, Missouri, today to visit with those who lost so much in last week's devastating tornado."		adewole101	7.50E+16	260579951	en	<a href="http://blackberry.com/twitter" rel="nofollow">Twitter for BlackBerry®</a>	http://a OLE_no
8	twitter-search	Remarks by the President after Touring Tornado Damage in Joplin, Missouri | Before It's News: http://t.co/TdjeB0H via @addthis		nanopatents	7.50E+16	80665053	en	<a href="http://twitter.com/tweetbutton" rel="nofollow">Tweet Button</a>	http://a _twitte

Exercise

1. Create an archive using yourTwapperKeeper tool at http://virginia.cc.vt.edu/
2. Use your keyword or key phrase as the search term. You may add a hash '#' to make your keyword a hashtag (Search Twitter or The Archivist Web version with your hashtag and see if it returns a reasonable result).
3. Refresh the browser to update the number of tweets archived.
4. Open a tweet export page by clicking the binocular icon in one of the archives.
5. Adjust the export setting and export the tweets following the instructions above.
6. Download the excel file and examine its content.

1.3. 140kit

140kit[6] is online services to collect, analyze, and visualize tweets. Collection of tweets can continue for a maximum of seven days, and then potentially be extended for further archiving. Once the collection is completed, the tool provides basic analyses and graph visualizations of tweets. CSV files of the basic analyses can be downloaded from the site.

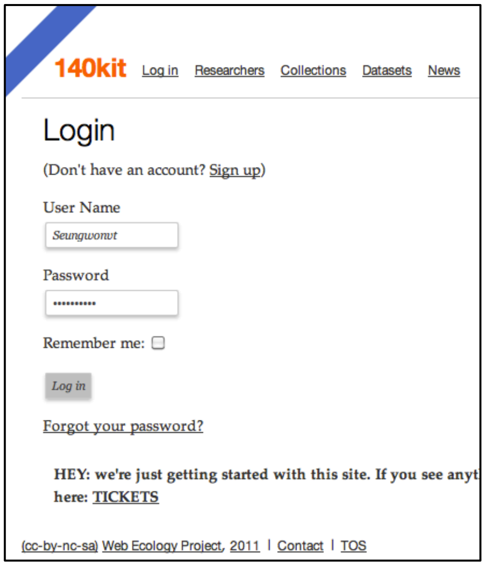
Users can search existing collections and visualize the tweets by using the features provided. 140kit used to allow exporting of the raw tweet archives, but this feature is no longer available due to the updated Twitter API Terms of Service at http://dev.twitter.com/pages/api_terms

Especially, under 4.A:

> *"You may export or extract non-programmatic, GUI-driven Twitter Content as a PDF or spreadsheet by using "save as" or similar functionality. Exporting Twitter Content to a data store as a service or other cloud based service, however, is not permitted."*

Instructions on creating a new tweet archive and viewing the pie charts of the basic analysis

Step 1. After creating an account at the sign up page, please login by typing in your User Name and Password and then click the 'Log In' button.



Login

(Don't have an account? Sign up)

User Name

Seungwonvt

Password

••••••••••

Remember me: □

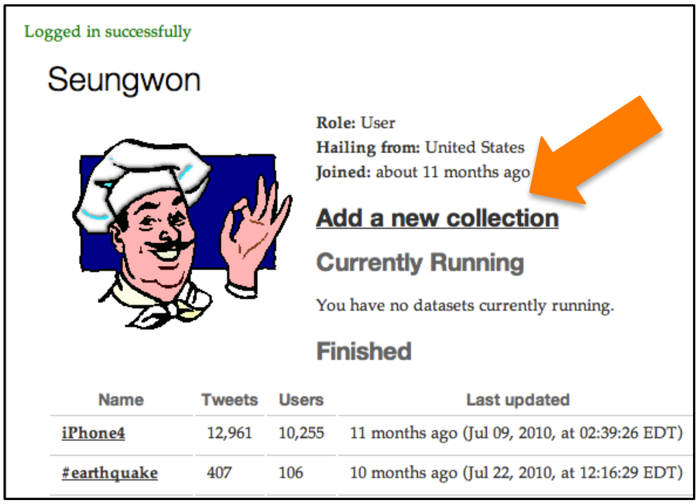
Forgot your password?

HEY: we're just getting started with this site. If you see anyt
here: TICKETS

(cc-by-nc-sa) Web Ecology Project, 2011 | Contact | TOS

Step 2. Click the 'Add a new collection' link on your main page.

[6] http://140kit.com/



Logged in successfully

Seungwon

Role: User
Hailing from: United States
Joined: about 11 months ago

Add a new collection

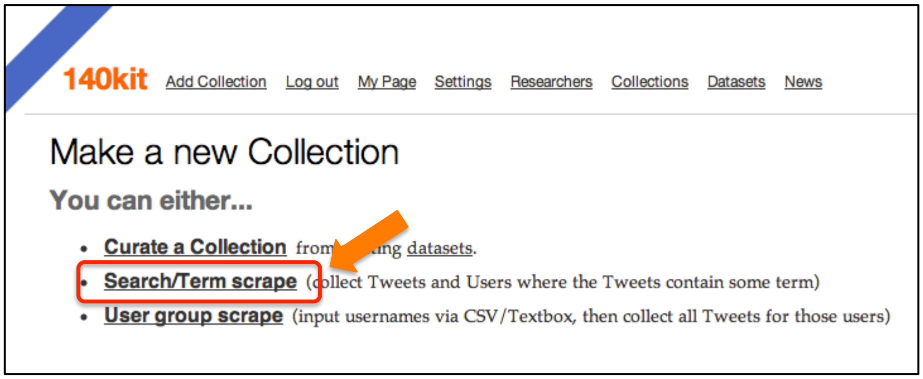
Currently Running

You have no datasets currently running.

Finished

Name	Tweets	Users	Last updated
iPhone4	12,961	10,255	11 months ago (Jul 09, 2010, at 02:39:26 EDT)
#earthquake	407	106	10 months ago (Jul 22, 2010, at 12:16:29 EDT)

Step 3. Click 'Search/Term scrape' link to collect tweets using a hashtag or terms.

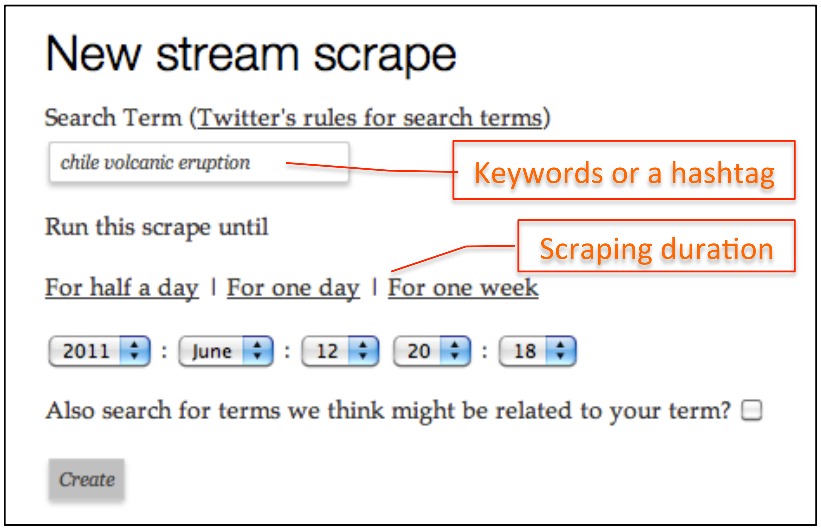


Make a new Collection

You can either...

- Curate a Collection from existing datasets.
- Search/Term scrape (collect Tweets and Users where the Tweets contain some term)
- User group scrape (input usernames via CSV/Textbox, then collect all Tweets for those users)

Step 4. Enter a search term or a hashtag (e.g., "#japanearthquake") to start archiving tweets, which contain them. Click 'Create' button on the bottom of the page to begin archiving tweets.



New stream scrape

Search Term (Twitter's rules for search terms)

chile volcanic eruption

Keywords or a hashtag

Run this scrape until

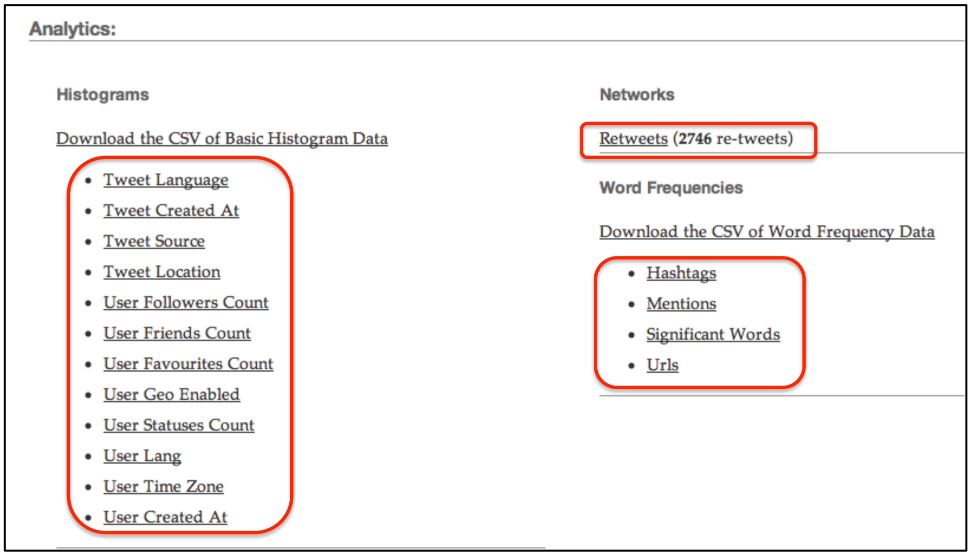
Scraping duration

For half a day | For one day | For one week

2011 : June : 12 20 : 18

Also search for terms we think might be related to your term? □

Step 5. Once archiving the tweet is completed, users can view pie charts of basic analysis results by clicking links under 'Histograms' and 'Word Frequencies'. Also the retweet network graph is provided. Please see the sections enclosed by rounded rectangles in the image below.



Analytics:

Histograms

Download the CSV of Basic Histogram Data

- Tweet Language
- Tweet Created At
- Tweet Source
- Tweet Location
- User Followers Count
- User Friends Count
- User Favourites Count
- User Geo Enabled
- User Statuses Count
- User Lang
- User Time Zone
- User Created At

Networks

Retweets (2746 re-tweets)

Word Frequencies

Download the CSV of Word Frequency Data

- Hashtags
- Mentions
- Significant Words
- Urls

Note: it usually takes some time to see the results from the 140kit tool.

(You need to make an account at 140kit.com to do this exercise.)

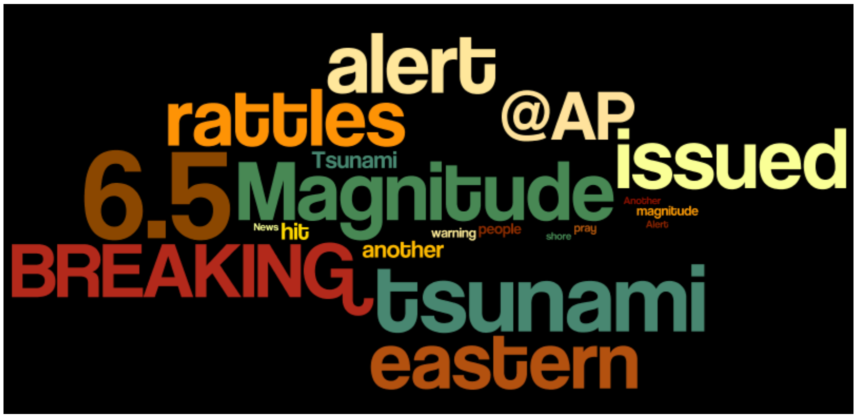
Exercise

1. Create an archive using 140kit tool. Use your keyword or key phrase as the search term. You may add a hash '#' to make your keyword a hashtag (Search Twitter or The Archivist Web version with your hashtag and see if it returns a reasonable result).
2. Leave the tool to archive tweets.
3. Select 'Collections' link on top of the 140kit homepage.
4. Select one of the collections and click the name.
5. Examine various histograms and word frequency graphs, as well as the retweet graph.

2. Visualizing Tweets

Word cloud visualizations show frequently appearing words from input text. More frequent words appear as bigger fonts. Colors are used to visually distinguish words and make word clouds visually appealing.

2.1. Wordle.net

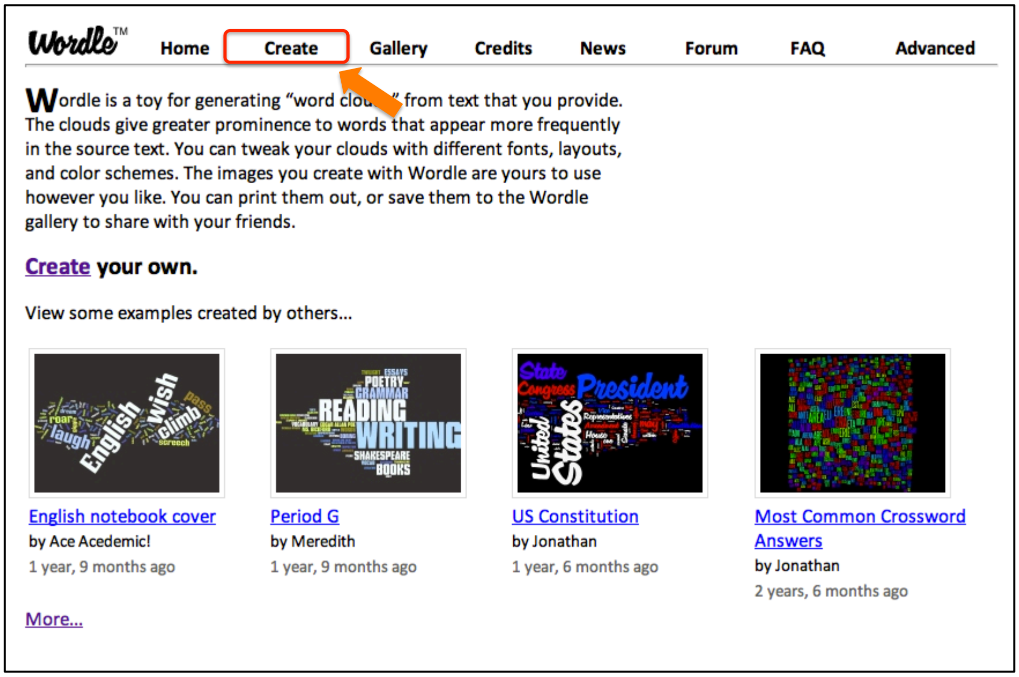


The image above shows a word cloud created by the Wordle[7] Web service using 100 tweets about the Japan earthquake disaster that occurred in March, 2011. Users can create word clouds and refine them by using the features:

- Remove uninteresting and common words directly in the GUI
- Limit the maximum words
- Change font, upper/lower cases, color schemes, layouts
- Assign weights on words, change background color

Instructions to create a word cloud

Step 1. Click the ‘Create’ button on the task bar to start the process.



Wordle is a toy for generating “word clouds” from text that you provide. The clouds give greater prominence to words that appear more frequently in the source text. You can tweak your clouds with different fonts, layouts, and color schemes. The images you create with Wordle are yours to use however you like. You can print them out, or save them to the Wordle gallery to share with your friends.

Create your own.

View some examples created by others...

English notebook cover
by Ace Acedemic!
1 year, 9 months ago

Period G
by Meredith
1 year, 9 months ago

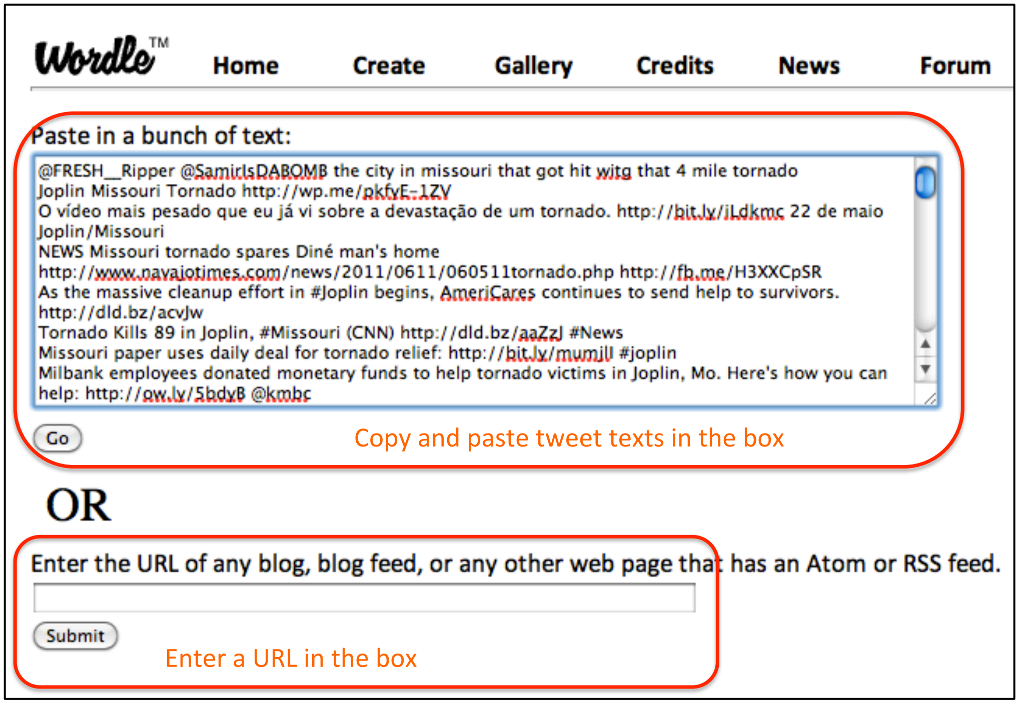
US Constitution
by Jonathan
1 year, 6 months ago

Most Common Crossword Answers
by Jonathan
2 years, 6 months ago

More...

[7] http://www.wordle.net/

Step 2. Copy and paste tweet texts from an excel file into the text box. Or type in a URL of the data file or a Web page. You can make multiple word clouds by each date or each hour of a specific date to see the different topical words emerge.



Paste in a bunch of text:

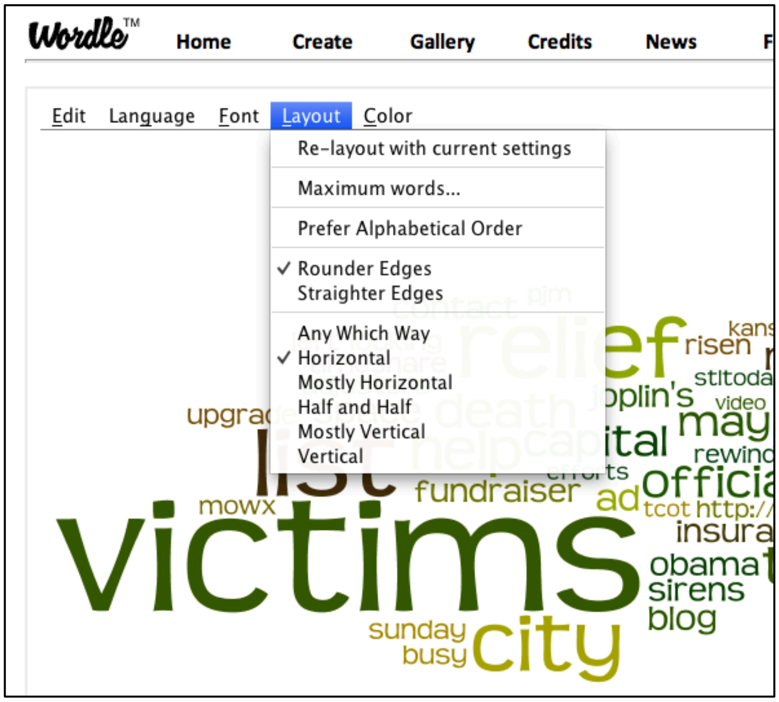
@FRESH__Ripper @SamirIsDABOMB the city in missouri that got hit witg that 4 mile tornado
Joplin Missouri Tornado http://wp.me/pkfyE-1ZV
O vídeo mais pesado que eu já vi sobre a devastação de um tornado. http://bit.ly/iLdkmc 22 de maio
Joplin/Missouri
NEWS Missouri tornado spares Diné man's home
http://www.navajotimes.com/news/2011/0611/060511tornado.php http://fb.me/H3XXCpSR
As the massive cleanup effort in #Joplin begins, AmeriCares continues to send help to survivors.
http://dld.bz/acvJw
Tornado Kills 89 in Joplin, #Missouri (CNN) http://dld.bz/aaZzJ #News
Missouri paper uses daily deal for tornado relief: http://bit.ly/mumjll #joplin
Milbank employees donated monetary funds to help tornado victims in Joplin, Mo. Here's how you can
help: http://ow.ly/5bdyB @kmbc

Copy and paste tweet texts in the box

Enter the URL of any blog, blog feed, or any other web page that has an Atom or RSS feed.

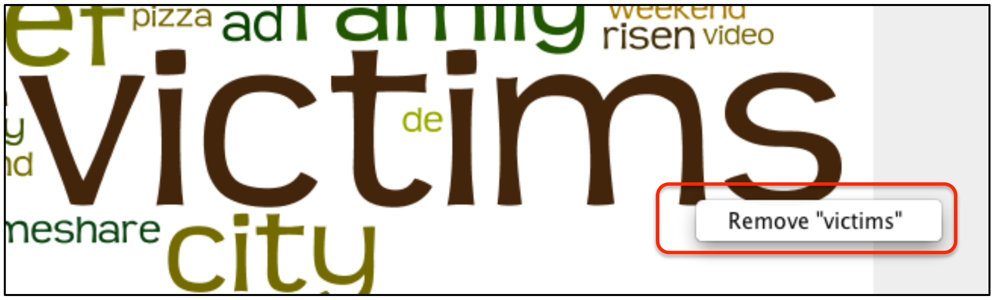
Enter a URL in the box

Step 3. Adjust layout and remove commonly repeated words.
Usually, ‘Horizontal’ layout is comfortable to read words in the cloud. To change the organization of the words, select ‘Re-layout with current settings’ under ‘Layout’.



The key phrase that was used to collect tweets can be removed since we know that the tweets are about that key phrase. For example, if a keyword ‘victims’ was used to collect tweets, it is shown with a bigger font size due to its frequency. By removing this dominating word, we can examine other interesting words closely.

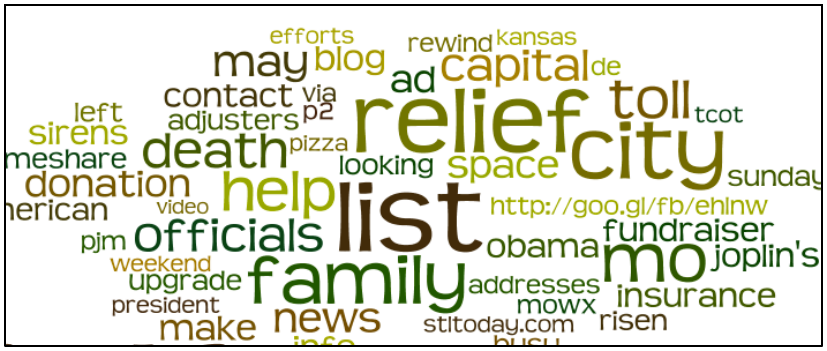
Right click on the word to remove and select ‘Remove “<word>”’ option to delete it from the visualization canvas.



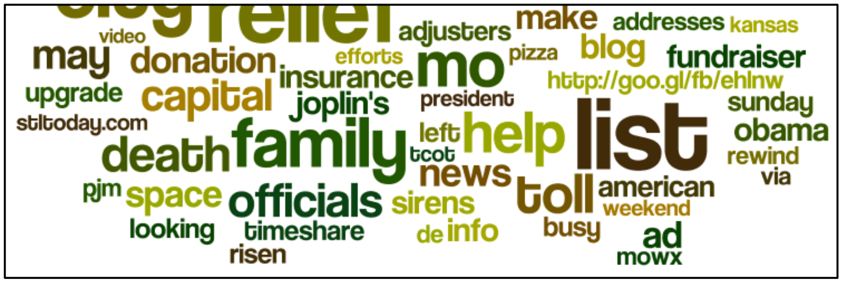
Step 4. Adjust the number of words to visualize. If there are too many words to visualize, the font of the less frequent words becomes too small to read. For this, please select ‘Maximum words’ under ‘Layout’ menu in the tool bar.

Step 5. Change fonts. Teen and Coolvetica fonts usually make a nice word cloud. For this, select ‘Font’ menu in the tool bar.

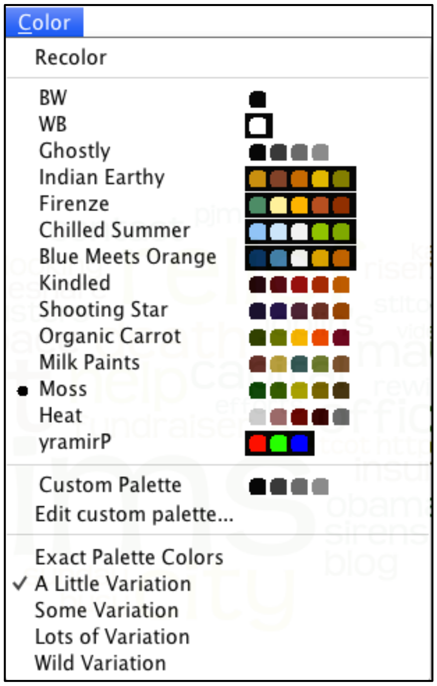
(Teen font)



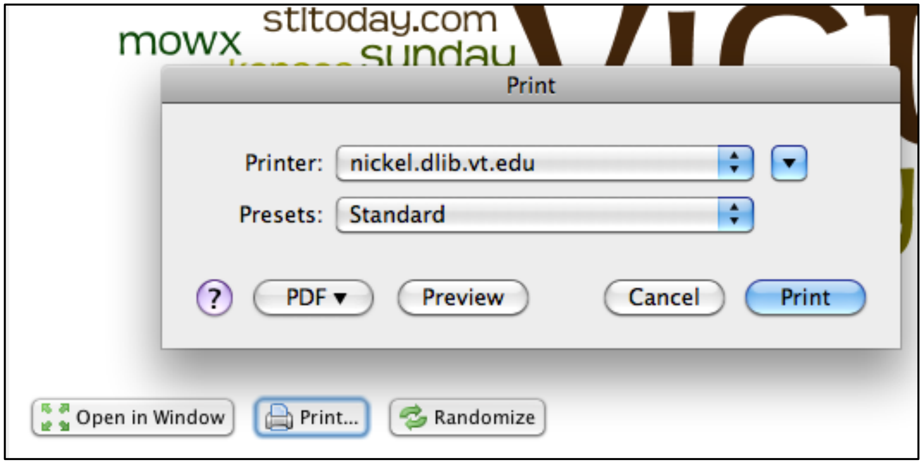
(Coolvetica font)



Step 6. Change color settings.
Try different color settings to find the one you like. Sometimes dark backgrounds make the words in the cloud more visible.



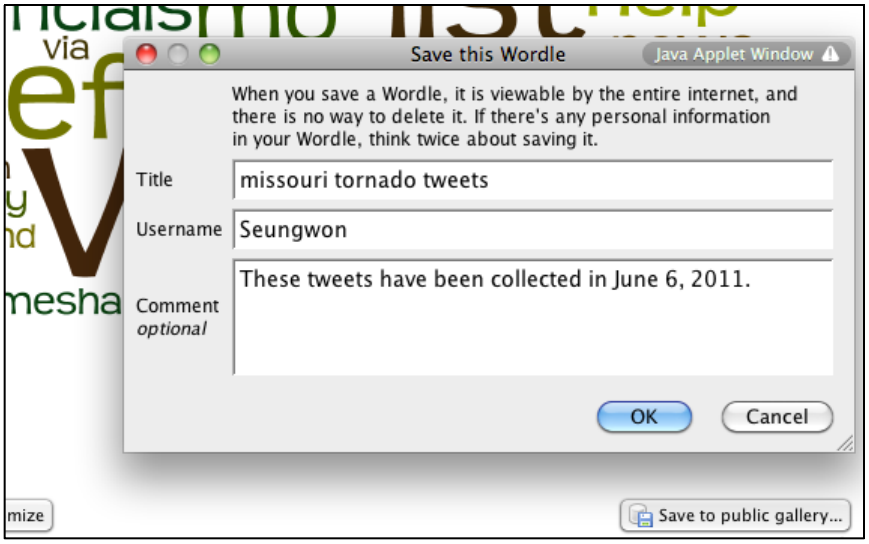
Step 7. Capturing the created word cloud.
You can use ‘Print’ feature to make a PDF file of the cloud or capture the screen.


Print

Printer: nickel.dlib.vt.edu

Presets: Standard

Step 8. Sharing the cloud with others.
You can make your word cloud public so that others can view it, too. Please click ‘Save to public gallery…’ on the bottom right corner of the interface. Then, enter the title, username and comments, and hit ‘OK’ button.



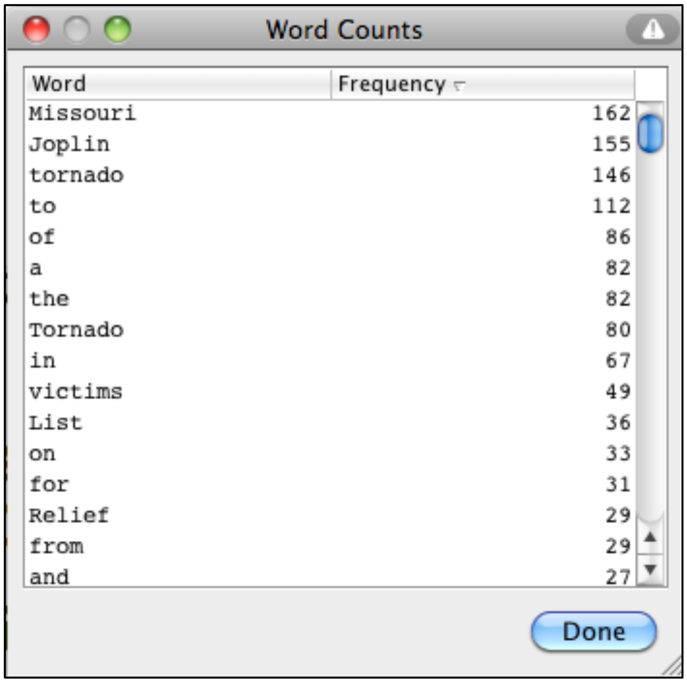
When you save a Wordle, it is viewable by the entire internet, and there is no way to delete it. If there's any personal information in your Wordle, think twice about saving it.

Title: missouri tornado tweets

Username: Seungwon

Comment *optional*: These tweets have been collected in June 6, 2011.

Step 9. Word counts.
To get a list of word frequency, please go to 'Language' and select 'Show word counts…'. You can reorder the words in either ascending or descending order by clicking 'Frequency' on the Word Counts pop up window.



Word	Frequency
Missouri	162
Joplin	155
tornado	146
to	112
of	86
a	82
the	82
Tornado	80
in	67
victims	49
List	36
on	33
for	31
Relief	29
from	29
and	27

(You can use your own tweet archive developed using yourTwapperKeeper at http://virginia.cc.vt.edu/ or any archives at http://mule.dlib.vt.edu/)

Exercise

1. Export tweets in the archive as an excel file (You can either include or exclude retweets when exporting).

2. Group tweets in the excel file by dates (if you use an existing archive with lots of tweets collected for many days), by the hour, or by 10 minutes.
3. Create at least 4 word clouds using the grouped tweet data in the previous step.
4. Be sure to remove the key phrase, which was used to archive tweets.
5. Compare the word clouds.
6. Do you see the differences? What do you think the word clouds as a visualization tool for tweets?

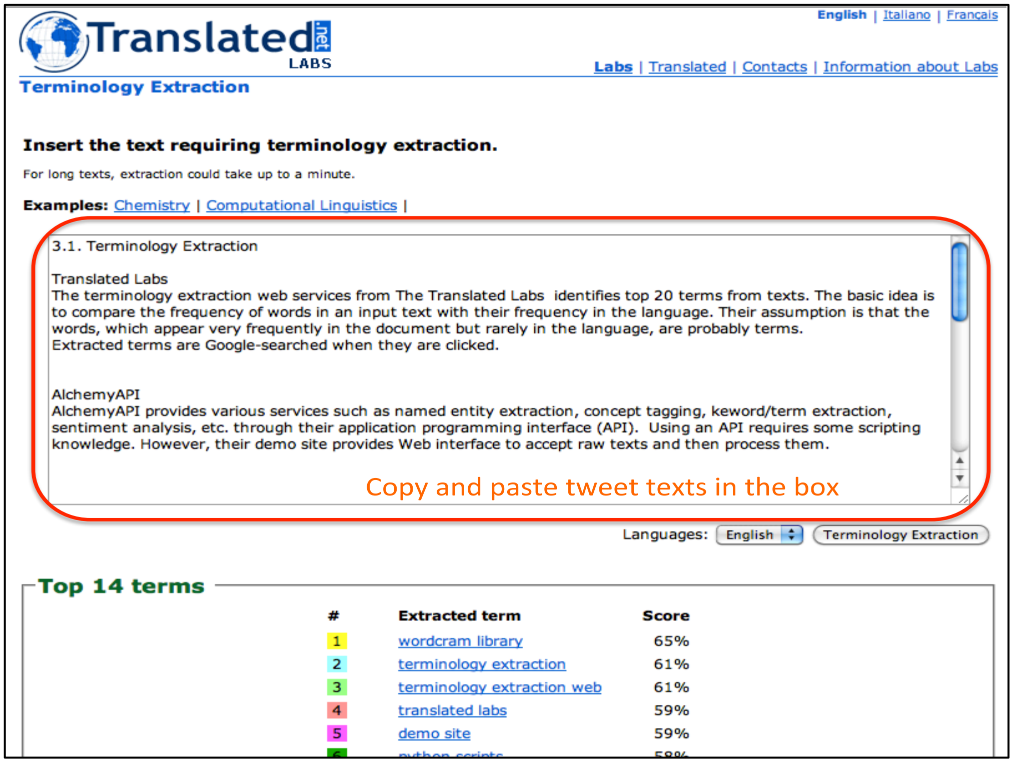
3. Analyzing Tweets Using Web-Based Tools

The basic algorithm for text analysis in word cloud creation is to count word frequencies in input texts. For more meaningful content analysis, terms and phrases can be extracted using web services and Natural Language Toolkit.

3.1. Terminology Extraction

Translated Labs

The terminology extraction web service from The Translated Labs[8] identifies terms from texts. The basic idea is to compare the frequency of words in an input text with their frequency in the language. Their assumption is that the words, which appear very frequently in the document but rarely in the language, are probably terms. Extracted terms are Google-searched when they are clicked.


Translated.net
LABS
English | Italiano | Français
Labs | Translated | Contacts | Information about Labs
Languages: English
Terminology Extraction

Terminology Extraction

Insert the text requiring terminology extraction.

For long texts, extraction could take up to a minute.

Examples: Chemistry | Computational Linguistics |

3.1. Terminology Extraction

Translated Labs
The terminology extraction web services from The Translated Labs identifies top 20 terms from texts. The basic idea is to compare the frequency of words in an input text with their frequency in the language. Their assumption is that the words, which appear very frequently in the document but rarely in the language, are probably terms.
Extracted terms are Google-searched when they are clicked.

AlchemyAPI
AlchemyAPI provides various services such as named entity extraction, concept tagging, keword/term extraction, sentiment analysis, etc. through their application programming interface (API). Using an API requires some scripting knowledge. However, their demo site provides Web interface to accept raw texts and then process them.

Copy and paste tweet texts in the box

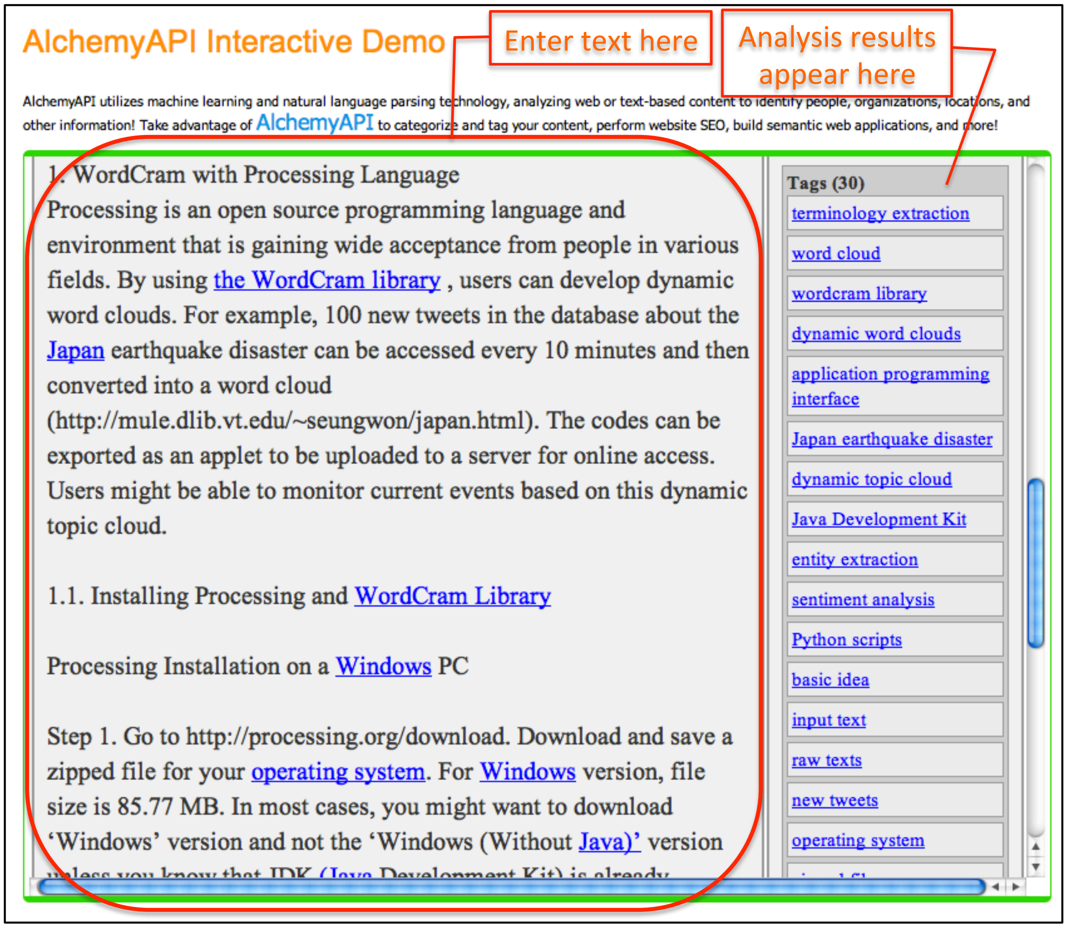
Top 14 terms

#	Extracted term	Score
1	wordcram library	65%
2	terminology extraction	61%
3	terminology extraction web	61%
4	translated labs	59%
5	demo site	59%
6	python scripts	58%

[8] http://labs.translated.net/terminology-extraction/

AlchemyAPI

AlchemyAPI provides more detailed analyses of the input text such as named entity extraction, concept tagging, keyword/term extraction, sentiment analysis, etc. Its demo site provides Web interface to accept raw texts and then process them.



AlchemyAPI Interactive Demo

AlchemyAPI utilizes machine learning and natural language parsing technology, analyzing web or text-based content to identify people, organizations, locations, and other information! Take advantage of AlchemyAPI to categorize and tag your content, perform website SEO, build semantic web applications, and more!

1. WordCram with Processing Language
Processing is an open source programming language and environment that is gaining wide acceptance from people in various fields. By using the WordCram library , users can develop dynamic word clouds. For example, 100 new tweets in the database about the Japan earthquake disaster can be accessed every 10 minutes and then converted into a word cloud (http://mule.dlib.vt.edu/~seungwon/japan.html). The codes can be exported as an applet to be uploaded to a server for online access. Users might be able to monitor current events based on this dynamic topic cloud.

1.1. Installing Processing and WordCram Library

Processing Installation on a Windows PC

Step 1. Go to http://processing.org/download. Download and save a zipped file for your operating system. For Windows version, file size is 85.77 MB. In most cases, you might want to download 'Windows' version and not the 'Windows (Without Java)' version unless you know that JDK (Java Development Kit) is already

Tags (30)

- terminology extraction
- word cloud
- wordcram library
- dynamic word clouds
- application programming interface
- Japan earthquake disaster
- dynamic topic cloud
- Java Development Kit
- entity extraction
- sentiment analysis
- Python scripts
- basic idea
- input text
- raw texts
- new tweets
- operating system

(Please use the grouped tweet data, which was used in the previous exercise.)

Exercise

1. Enter each of the grouped data into the text box in AlchemyAPI demo site.
2. Copy and Paste the analysis results to a separate sheet in the tweet excel file.
3. You will have four groups of results for each tweet group.
4. Compare the terms and concepts extracted, with the corresponding word clouds you previously created.
5. Do they match well? Which one summarizes the group of tweets better for human understanding?

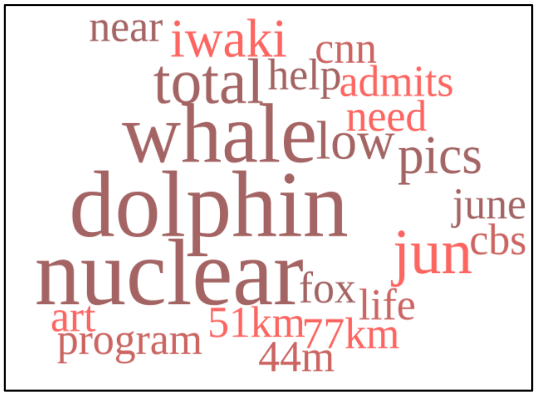
Part 2. Tools with Scripting

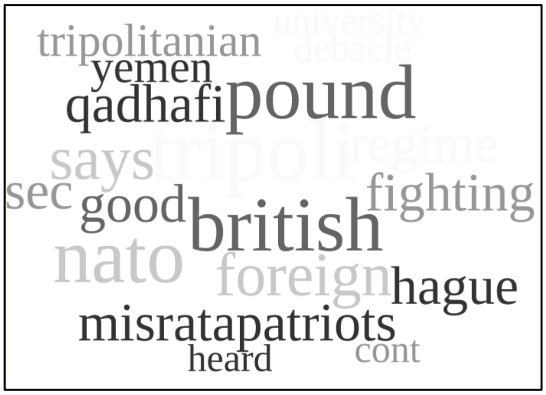
1. WordCram with Processing Language

Processing[9] is an open source programming language and environment that is gaining wide acceptance from people in various fields. By using the WordCram library[10] with Processing, users can develop dynamic word clouds. For example, new tweets for the last 10 minutes in the database about the Japan earthquake disaster can be accessed and then converted into a word cloud. The codes can be exported as an applet to be uploaded to a server for online access. Users might be able to monitor current events based on this dynamic word cloud.

Example dynamic word clouds can be accessed at:

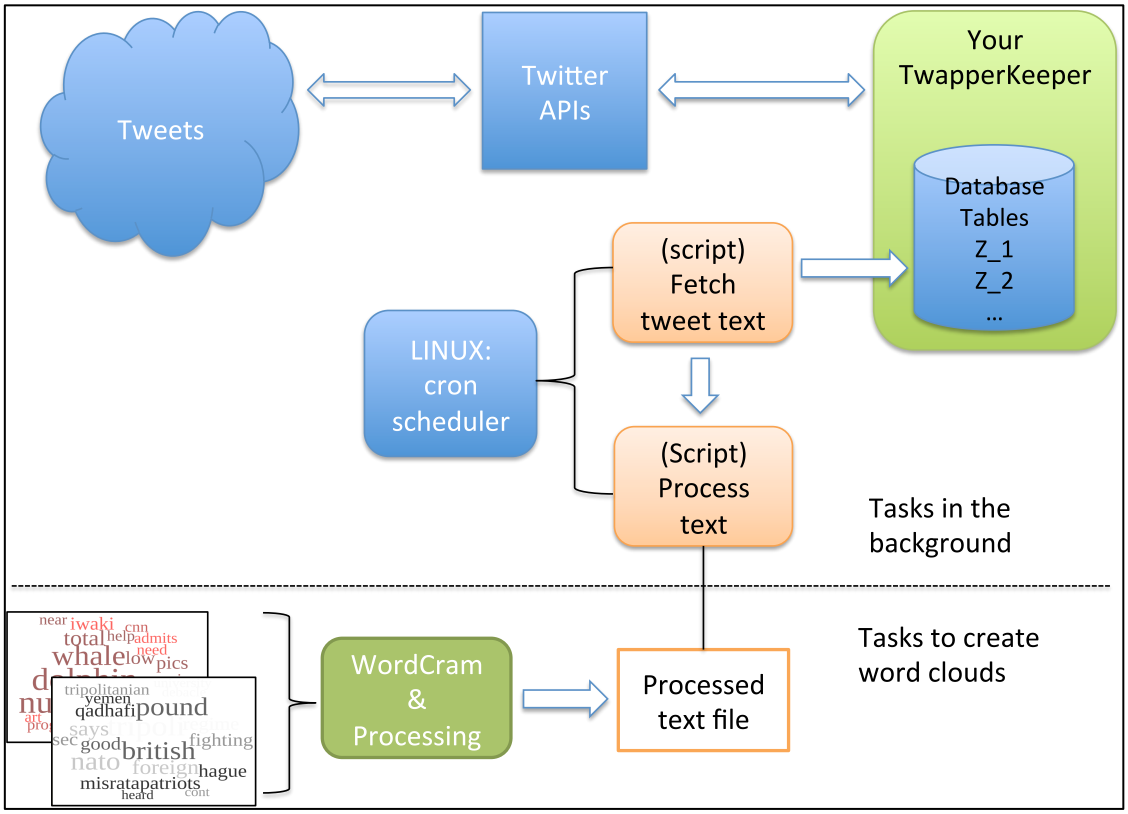
(Left) Japan Earthquake Disaster: http://mule.dlib.vt.edu/~seungwon/japan.html
(Right) Libya Revolution: http://mule.dlib.vt.edu/~seungwon/libya.html





1.1. Procedure

[9] Processing programming language and environment. http://processing.org
[10] http://code.google.com/p/wordcram/



Tasks running in the background

1. YourTwapperKeeper archives tweets into MySQL database tables.
2. For each time period (e.g., every 5 minutes), a script accesses one of database tables and fetches text in tweets. Then it writes the raw text into a file.
3. Following execution of the script at step 2, another script processes the raw text file to remove stop-words (e.g., 'the', 'a', 'about', 'it', etc.) and symbol characters (e.g., '$', '!', '?', etc.), and creates a processed text file.

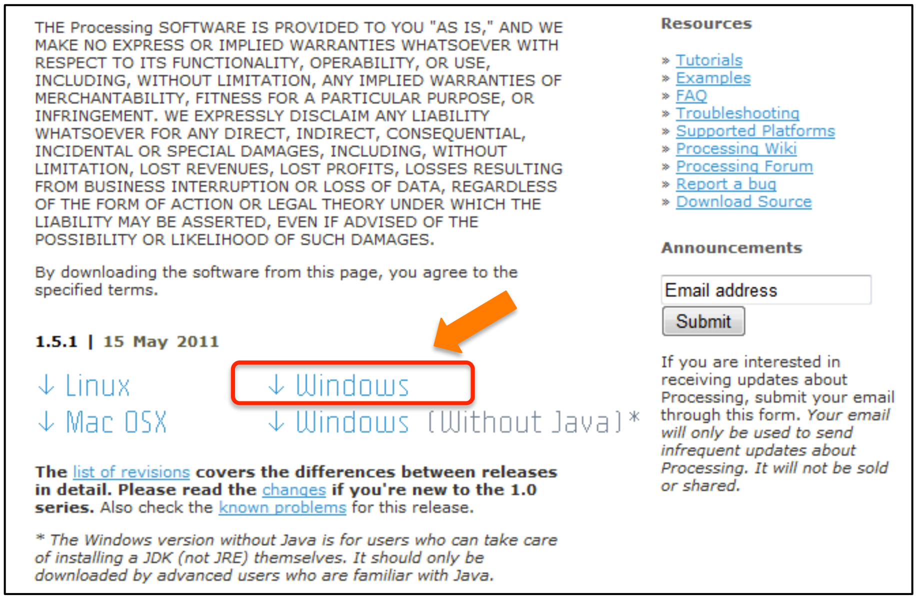
Tasks for creating word clouds

4. Processing sketch (i.e., code) with WordCram library accesses the processed text file developed at step 3.
5. Methods to change the layouts, color settings, canvas size, maximum and minimum font sizes, etc. are applied to make the words on the canvas comfortable to read and visually appealing.
6. Export Processing sketch to an applet, which can be inserted in a Web site.

1.2. Installing Processing and WordCram Library

Processing Installation on a Windows PC

Step 1. Go to http://processing.org/download. Download and save a zipped file for your operating system. For Windows version, file size is 85.77 MB. In most cases, you might want to download 'Windows' version and not the 'Windows (Without Java)' version unless you know that JDK (Java Development Kit) is already installed on your computer.



THE Processing SOFTWARE IS PROVIDED TO YOU "AS IS," AND WE MAKE NO EXPRESS OR IMPLIED WARRANTIES WHATSOEVER WITH RESPECT TO ITS FUNCTIONALITY, OPERABILITY, OR USE, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR INFRINGEMENT. WE EXPRESSLY DISCLAIM ANY LIABILITY WHATSOEVER FOR ANY DIRECT, INDIRECT, CONSEQUENTIAL, INCIDENTAL OR SPECIAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST REVENUES, LOST PROFITS, LOSSES RESULTING FROM BUSINESS INTERRUPTION OR LOSS OF DATA, REGARDLESS OF THE FORM OF ACTION OR LEGAL THEORY UNDER WHICH THE LIABILITY MAY BE ASSERTED, EVEN IF ADVISED OF THE POSSIBILITY OR LIKELIHOOD OF SUCH DAMAGES.

By downloading the software from this page, you agree to the specified terms.

1.5.1 | 15 May 2011

↓ Linux ↓ Windows

↓ Mac OSX ↓ Windows (Without Java)*

The list of revisions **covers the differences between releases in detail. Please read the** changes **if you're new to the 1.0 series.** Also check the known problems for this release.

** The Windows version without Java is for users who can take care of installing a JDK (not JRE) themselves. It should only be downloaded by advanced users who are familiar with Java.*

Resources

- » Tutorials
- » Examples
- » FAQ
- » Troubleshooting
- » Supported Platforms
- » Processing Wiki
- » Processing Forum
- » Report a bug
- » Download Source

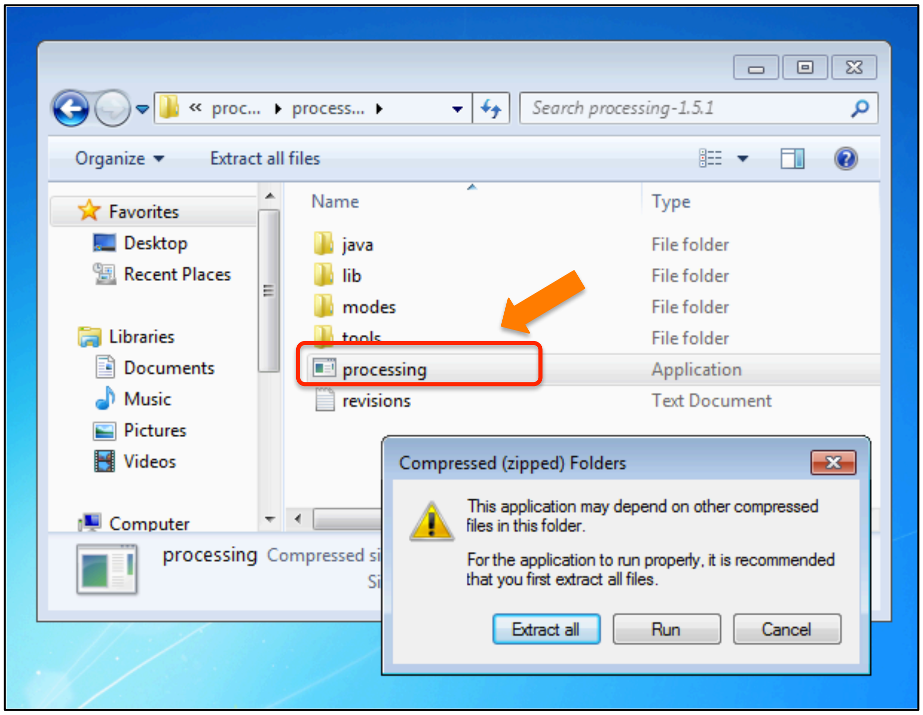
Announcements

Email address

Submit

If you are interested in receiving updates about Processing, submit your email through this form. *Your email will only be used to send infrequent updates about Processing. It will not be sold or shared.*

Step 2. Double-click (total twice) the zipped file to see its content. You will see folders such as ‘java’, ‘lib’, ‘modes’, and ‘tools’ as well as files such as ‘processing’ and ‘revisions’. Double-click ‘processing’ and follow the instruction to extract the content onto your computer.



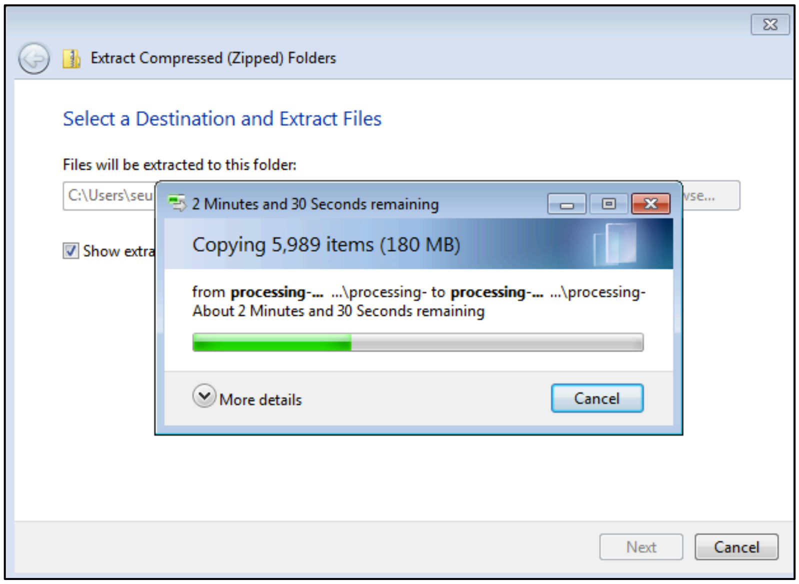
Name	Type
java	File folder
lib	File folder
modes	File folder
tools	File folder
processing	Application
revisions	Text Document

Compressed (zipped) Folders

This application may depend on other compressed files in this folder.

For the application to run properly, it is recommended that you first extract all files.

Extract all | Run | Cancel



Extract Compressed (Zipped) Folders

Select a Destination and Extract Files

Files will be extracted to this folder:

C:\Users\seu

☑ Show extra

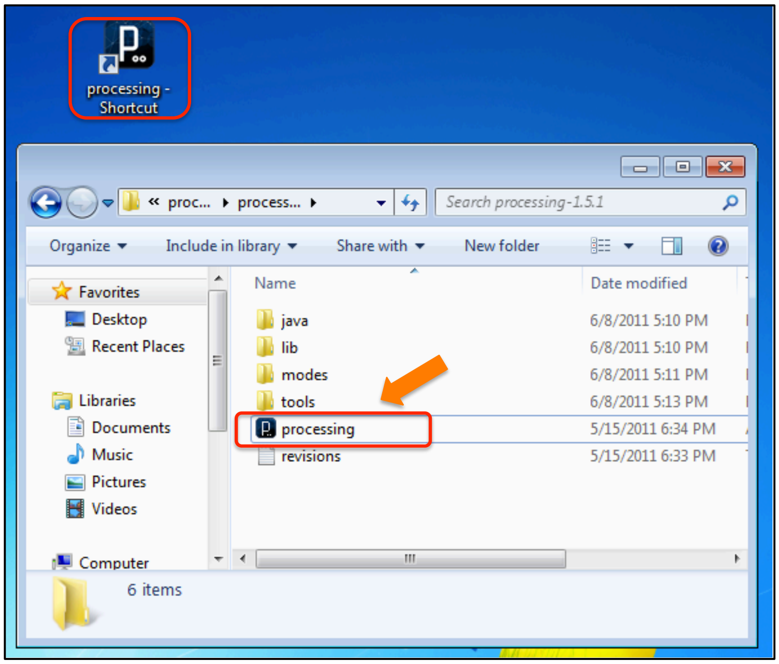
2 Minutes and 30 Seconds remaining

Copying 5,989 items (180 MB)

from **processing-...** ...\processing- to **processing-...** ...\processing-

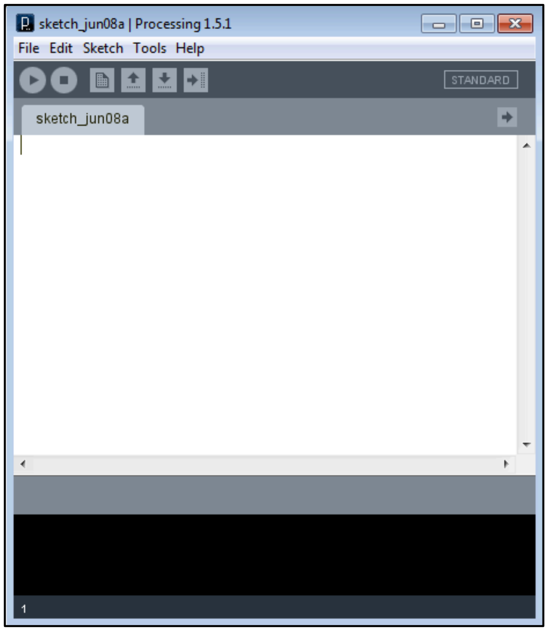
About 2 Minutes and 30 Seconds remaining

Step 3. After extraction is done, make a shortcut icon in the Desktop screen for easy access.

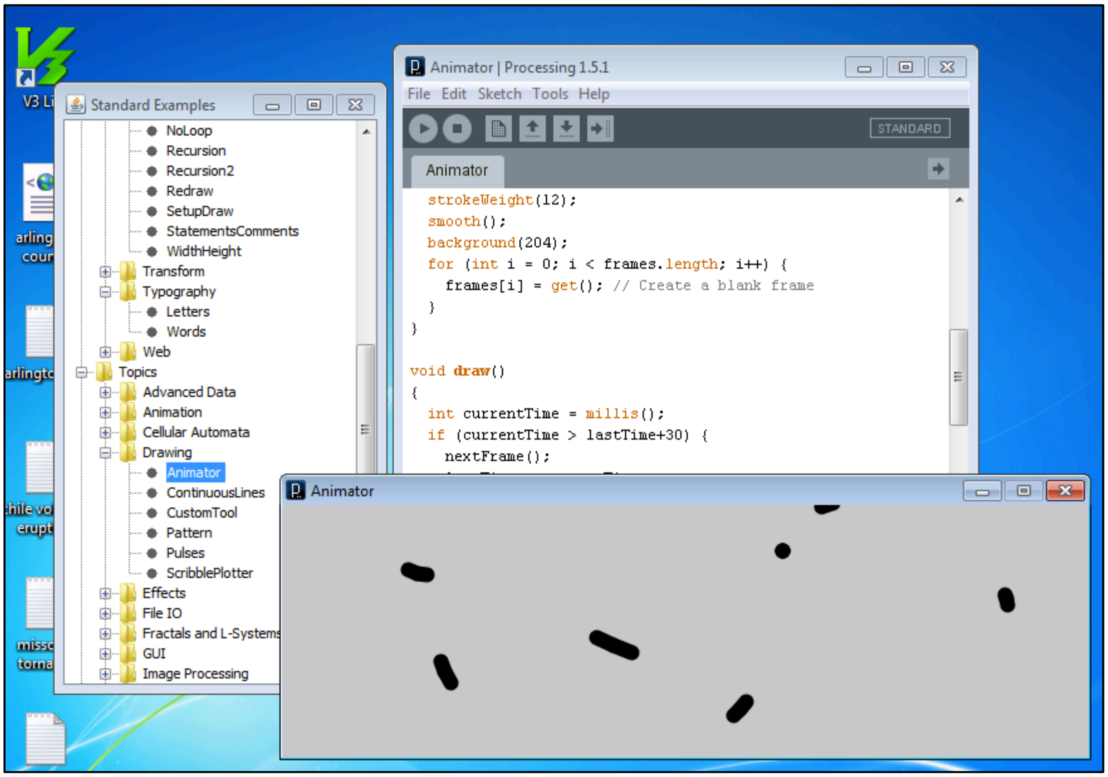


Name	Date modified
java	6/8/2011 5:10 PM
lib	6/8/2011 5:10 PM
modes	6/8/2011 5:11 PM
tools	6/8/2011 5:13 PM
processing	5/15/2011 6:34 PM
revisions	5/15/2011 6:33 PM

Step 4. Run Processing by double-clicking the shortcut icon. Its interface window appears.



Step 5. Test with an example sketch. Go to ‘File’ → ‘Examples’. It will open a ‘Standard Examples’ window. Select an example sketch by double-clicking one, and then run it by clicking the play button below the task bar, or by going to ‘Sketch’ → ‘Run’. The image below shows an example sketch, ‘Animator’.

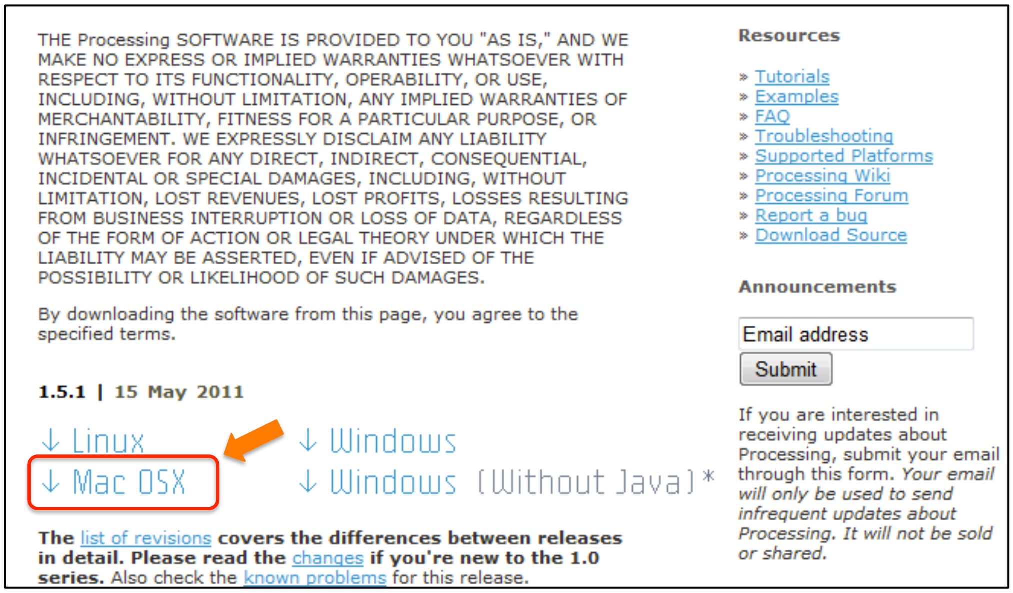


```
  strokeWeight(12);
  smooth();
  background(204);
  for (int i = 0; i < frames.length; i++) {
    frames[i] = get(); // Create a blank frame
  }
}

void draw()
{
  int currentTime = millis();
  if (currentTime > lastTime+30) {
    nextFrame();
```

This completes an installation of Processing on a Windows PC.

Processing Installation on a Mac Machine

Step 1. Go to http://processing.org/download. Click 'Mac OSX' link to download a zipped file. Processing-1.5.1-macosx.zip is 50.5 MB in size.



THE Processing SOFTWARE IS PROVIDED TO YOU "AS IS," AND WE MAKE NO EXPRESS OR IMPLIED WARRANTIES WHATSOEVER WITH RESPECT TO ITS FUNCTIONALITY, OPERABILITY, OR USE, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR INFRINGEMENT. WE EXPRESSLY DISCLAIM ANY LIABILITY WHATSOEVER FOR ANY DIRECT, INDIRECT, CONSEQUENTIAL, INCIDENTAL OR SPECIAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST REVENUES, LOST PROFITS, LOSSES RESULTING FROM BUSINESS INTERRUPTION OR LOSS OF DATA, REGARDLESS OF THE FORM OF ACTION OR LEGAL THEORY UNDER WHICH THE LIABILITY MAY BE ASSERTED, EVEN IF ADVISED OF THE POSSIBILITY OR LIKELIHOOD OF SUCH DAMAGES.

By downloading the software from this page, you agree to the specified terms.

1.5.1 | 15 May 2011

↓ Linux ↓ Windows
↓ Mac OSX ↓ Windows (Without Java)*

The list of revisions **covers the differences between releases in detail. Please read the** changes **if you're new to the 1.0 series.** Also check the known problems for this release.

Resources

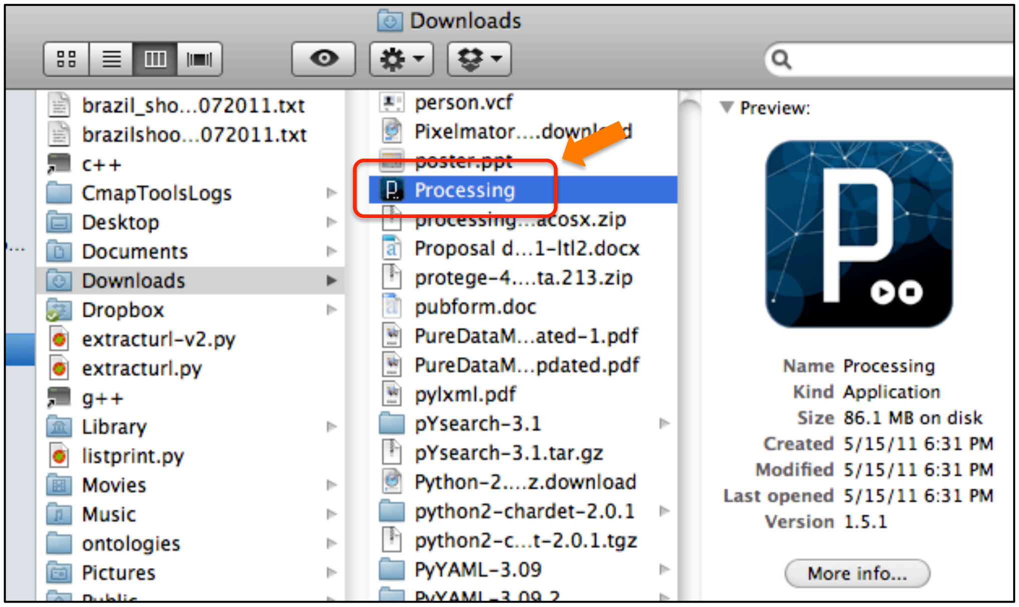
» Tutorials
» Examples
» FAQ
» Troubleshooting
» Supported Platforms
» Processing Wiki
» Processing Forum
» Report a bug
» Download Source

Announcements

Email address
Submit

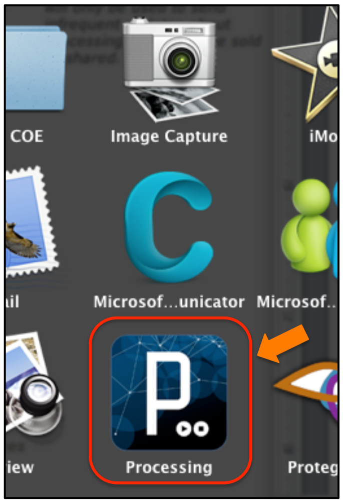
If you are interested in receiving updates about Processing, submit your email through this form. *Your email will only be used to send infrequent updates about Processing. It will not be sold or shared.*

Step 2. Double-click the file to unzip it. Then an application with the Processing logo appears in Downloads folder.

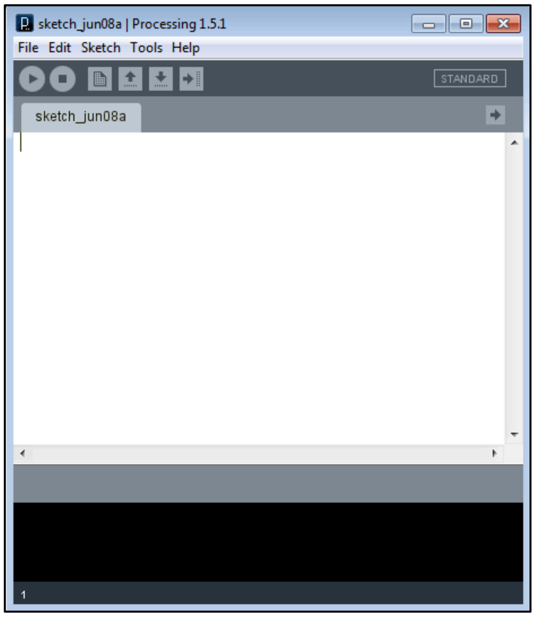


Name Processing
Kind Application
Size 86.1 MB on disk
Created 5/15/11 6:31 PM
Modified 5/15/11 6:31 PM
Last opened 5/15/11 6:31 PM
Version 1.5.1

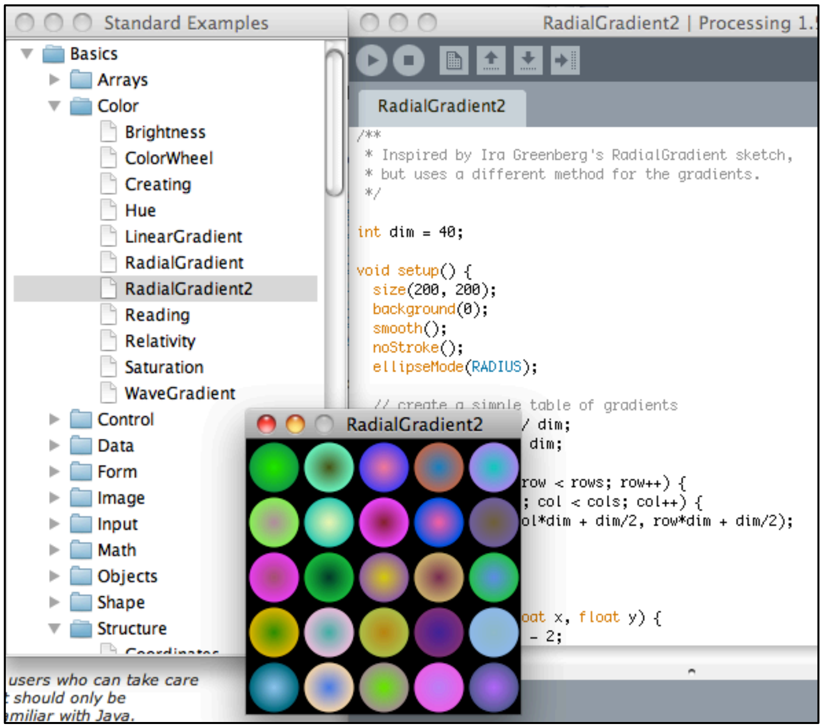
Step 3. Move the Processing application from Downloads folder to Applications folder using drag-and-drop.



Step 4. Click the Processing icon in Applications folder to run it. The processing interface window appears.



Step 5. Test with an example sketch. Go to ‘File’ → ‘Examples’. It will open a ‘Standard Examples’ window. Select an example sketch by double-clicking one, and then run it by clicking the play button below the task bar, or by going to ‘Sketch’ → ‘Run’. The image below shows an example sketch, ‘RadialGradient2’ in ‘Color’ folder.

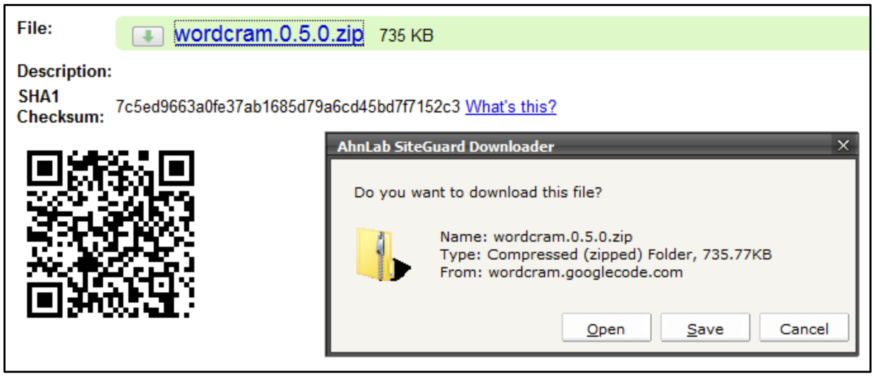


```
/**
 * Inspired by Ira Greenberg's RadialGradient sketch,
 * but uses a different method for the gradients.
 */

int dim = 40;

void setup() {
  size(200, 200);
  background(0);
  smooth();
  noStroke();
  ellipseMode(RADIUS);

  // create a simple table of gradients
                     / dim;
                      dim;

                     row < rows; row++) {
                     ; col < cols; col++) {
                     ol*dim + dim/2, row*dim + dim/2);




                     oat x, float y) {
                      - 2;
```

This completes an installation of Processing on a Mac machine.

WordCram Installation (after installing Processing)

Step 1. Go to WordCram installation page at http://code.google.com/p/wordcram/wiki/Installing. Download a zipped file 'wordcram.0.5.0.zip' by following the 'Download' link under 'Details'. Once downloaded, double-click it to unzip it. The unzipped folder name is 'wordcram.0.5.0'.


File: wordcram.0.5.0.zip 735 KB

Description:

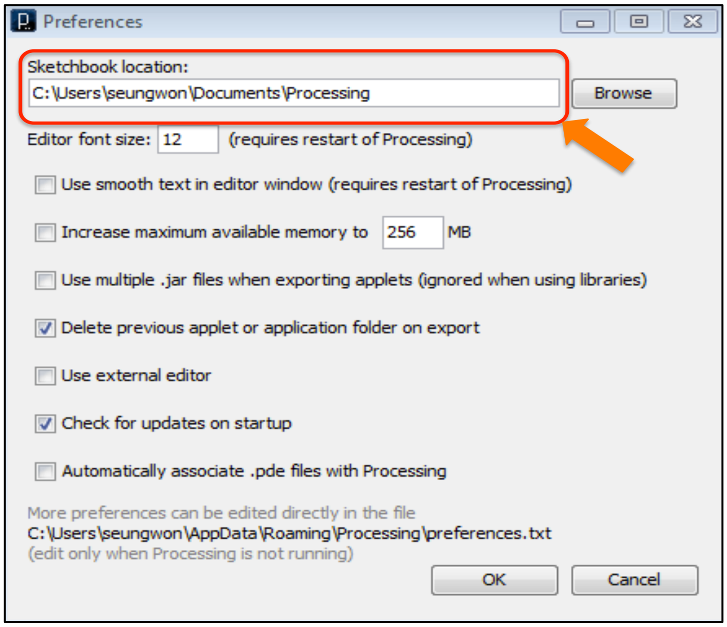
SHA1 Checksum: 7c5ed9663a0fe37ab1685d79a6cd45bd7f7152c3 What's this?

AhnLab SiteGuard Downloader

Do you want to download this file?

Name: wordcram.0.5.0.zip
Type: Compressed (zipped) Folder, 735.77KB
From: wordcram.googlecode.com

Step 2. Find out your Processing Sketchbook folder.
Go to 'File' → 'Preferences'. It will open a window. Processing Sketchbook folder location is specified below 'Sketchbook location'. The location of the Sketchbook folder in the example below is at 'C:\Users\seungwon\Documents\Processing'.



Sketchbook location:
C:\Users\seungwon\Documents\Processing

Editor font size: 12 (requires restart of Processing)

□ Use smooth text in editor window (requires restart of Processing)

□ Increase maximum available memory to 256 MB

□ Use multiple .jar files when exporting applets (ignored when using libraries)

☑ Delete previous applet or application folder on export

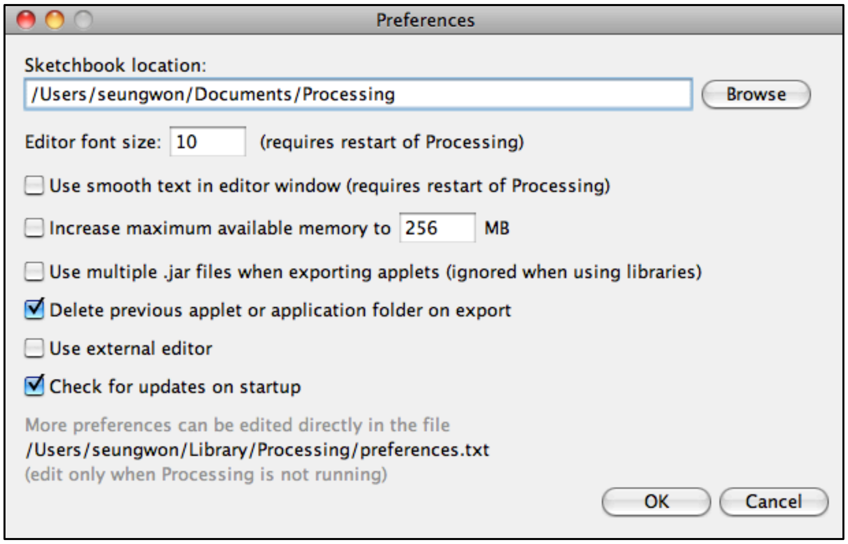
□ Use external editor

☑ Check for updates on startup

□ Automatically associate .pde files with Processing

More preferences can be edited directly in the file
C:\Users\seungwon\AppData\Roaming\Processing\preferences.txt
(edit only when Processing is not running)

For a Mac machine, the Sketchbook location looks, for example, "/Users/seungwon/Documents/Processing".



Sketchbook location:
/Users/seungwon/Documents/Processing

Editor font size: 10 (requires restart of Processing)

□ Use smooth text in editor window (requires restart of Processing)

□ Increase maximum available memory to 256 MB

□ Use multiple .jar files when exporting applets (ignored when using libraries)

☑ Delete previous applet or application folder on export

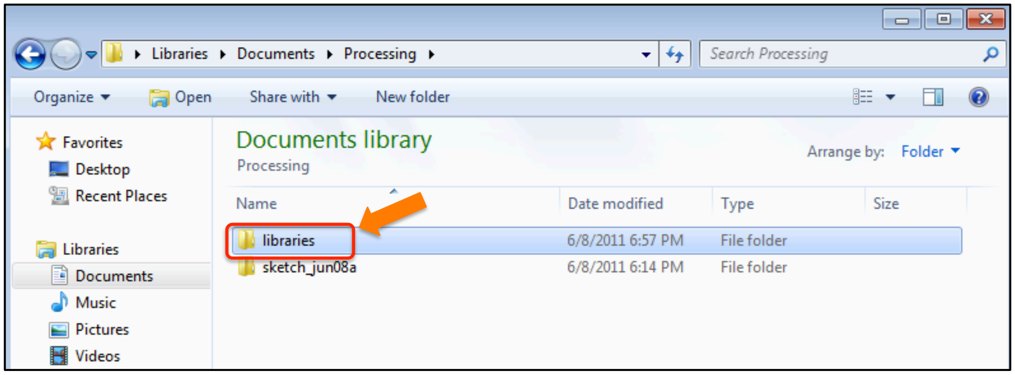
□ Use external editor

☑ Check for updates on startup

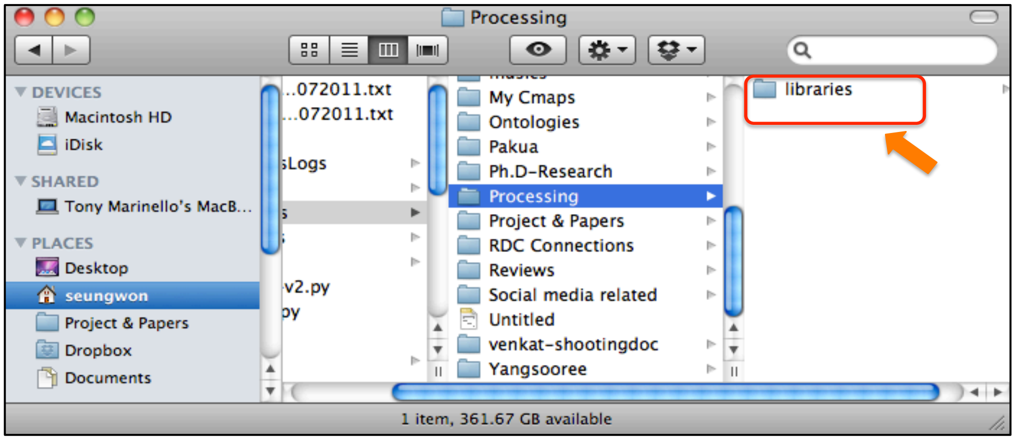
More preferences can be edited directly in the file
/Users/seungwon/Library/Processing/preferences.txt
(edit only when Processing is not running)

Step 3. Open a file browser and go to that Sketchbook folder. Then, create a folder named, 'libraries' if it does not exist there.

(Windows screen shot)



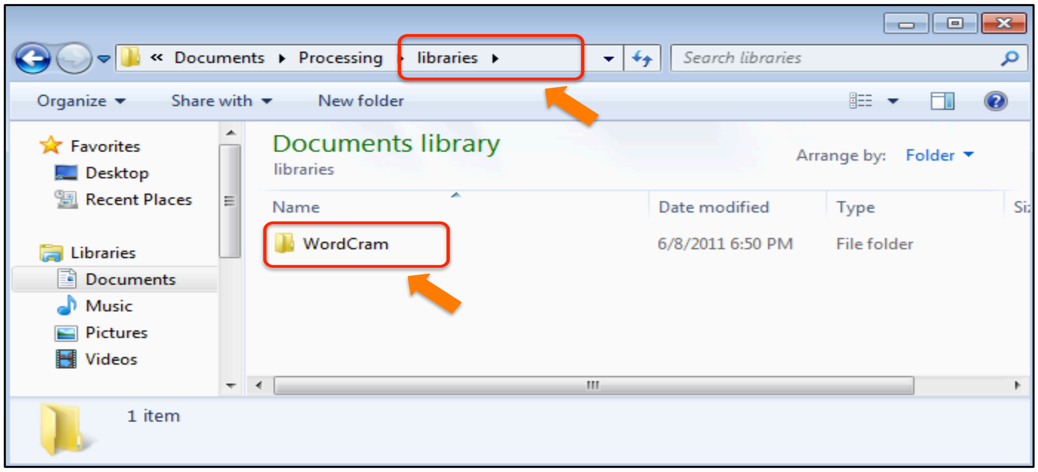
(Mac screen shot)



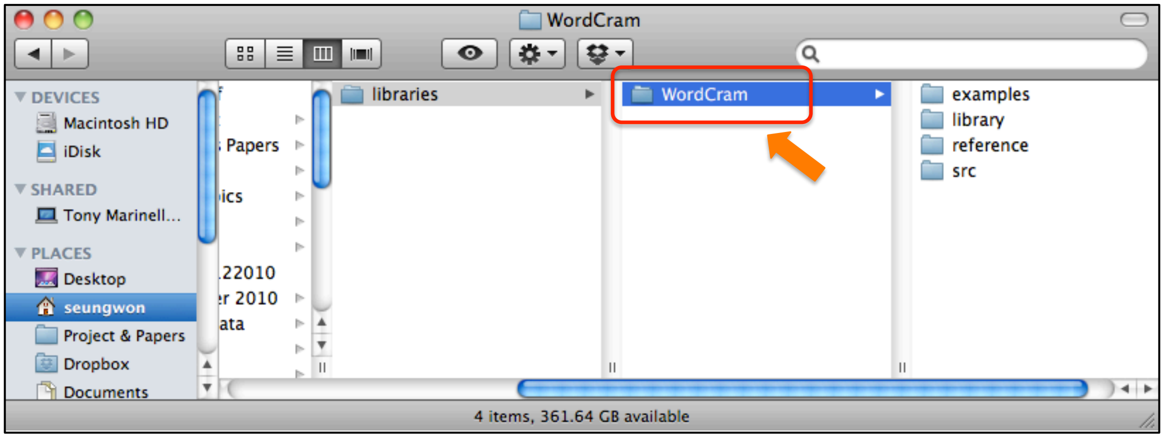
Step 4. Copy the folder named 'WordCram', which is located inside the unzipped folder, 'wordcram.0.5.0'. (For a Mac machine, 'WordCram' folder is created in the same folder, where wordcram.0.5.0.zip is located, once this zip file is unzipped.)

Then paste the entire content of the 'WordCram' directory into the 'libraries' folder that you created in Step 3. Restart Processing to make changes to be effective.

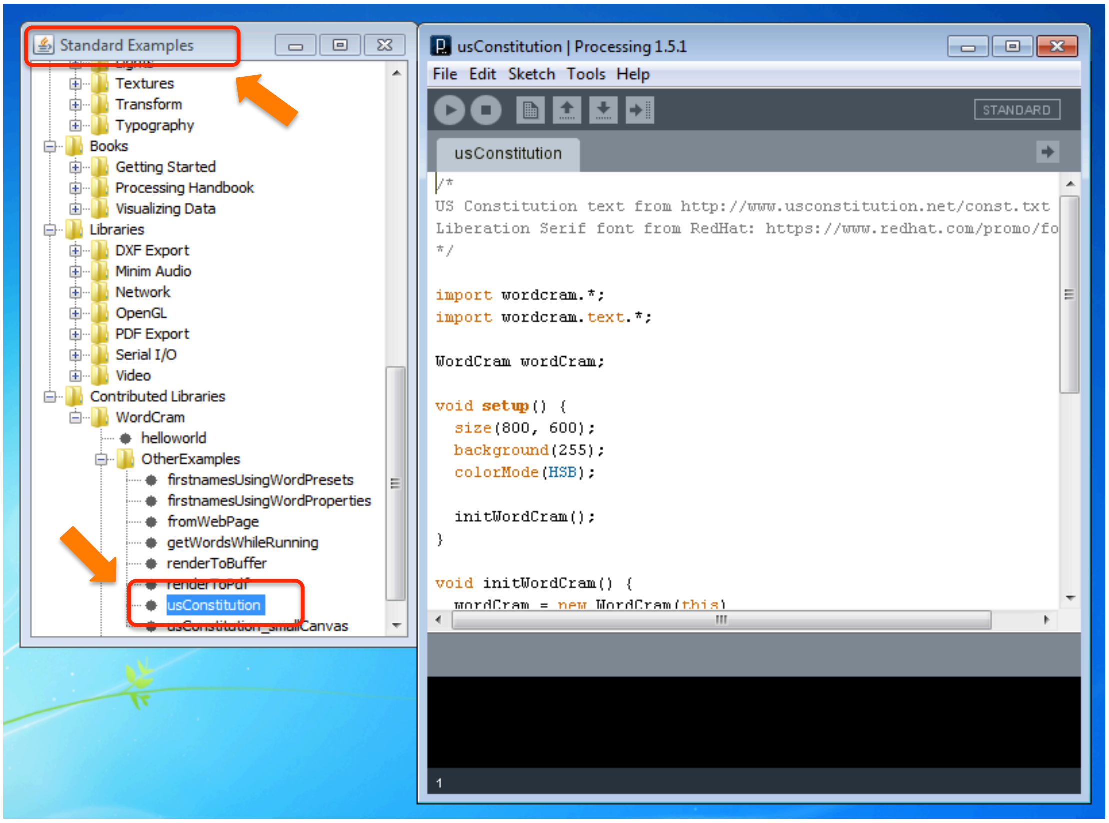
(Windows screen shot)



(Mac screen shot)

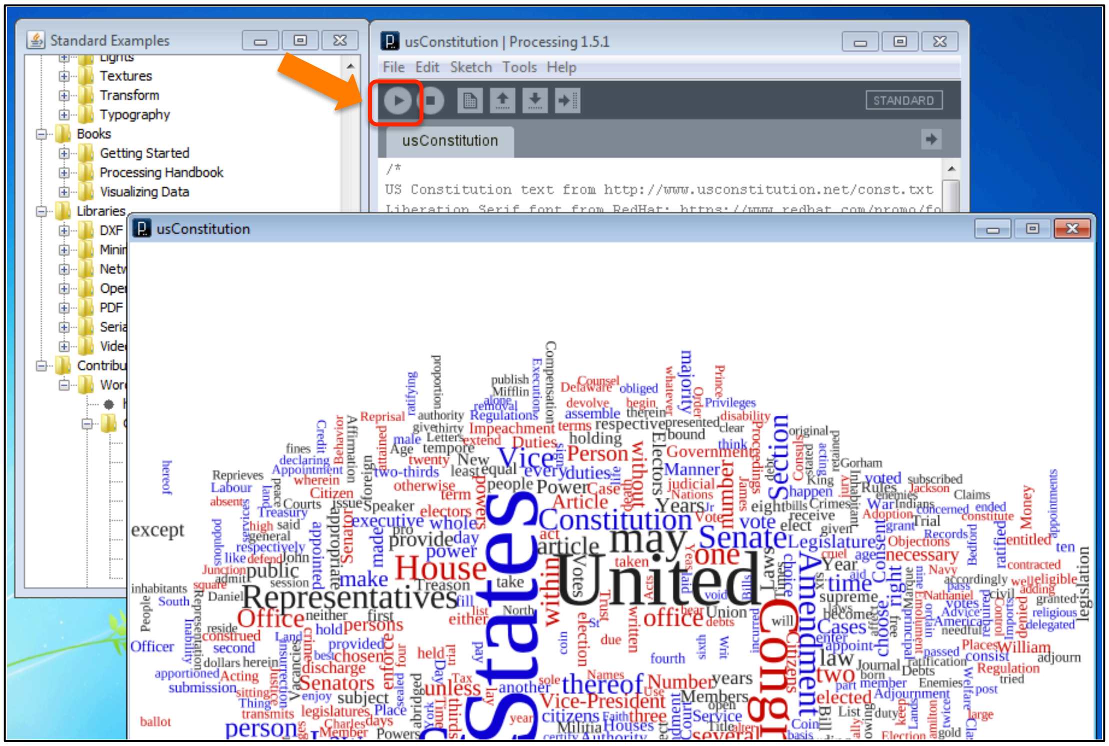


Step 5. Testing an example of the WordCram library.
Go to ‘File’ and select ‘Examples’. ‘Standard Examples’ window will open. Expand ‘WordCram’ folder under ‘Contributed Libraries’ folder on the bottom of the window. Expand ‘Other Examples’ folder and double-click ‘usConstitution’ sketch to open.



```
/*
US Constitution text from http://www.usconstitution.net/const.txt
Liberation Serif font from RedHat: https://www.redhat.com/promo/fo
*/

import wordcram.*;
import wordcram.text.*;

WordCram wordCram;

void setup() {
  size(800, 600);
  background(255);
  colorMode(HSB);

  initWordCram();
}

void initWordCram() {
  wordCram = new WordCram(this)
```

Step 6. Click the play button under the task bar to visualize a word cloud of the US constitution. If you see a beautiful word cloud opened up in a window, then your WordCram library is successfully installed.

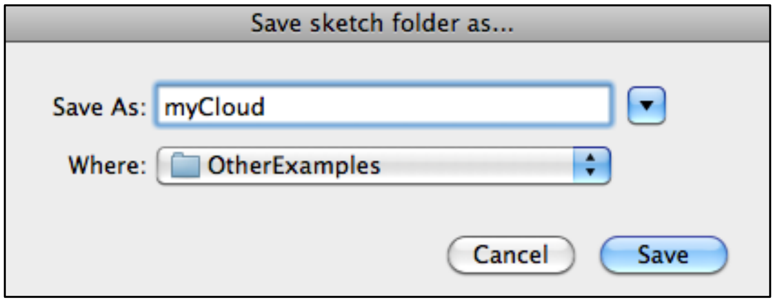


```
/*
US Constitution text from http://www.usconstitution.net/const.txt
Liberation Serif font from RedHat: https://www.redhat.com/promo/fo
```

At this point, both Processing and WordCram have been installed successfully. The next is to connect them to a dynamically updated tweet text file on a server.

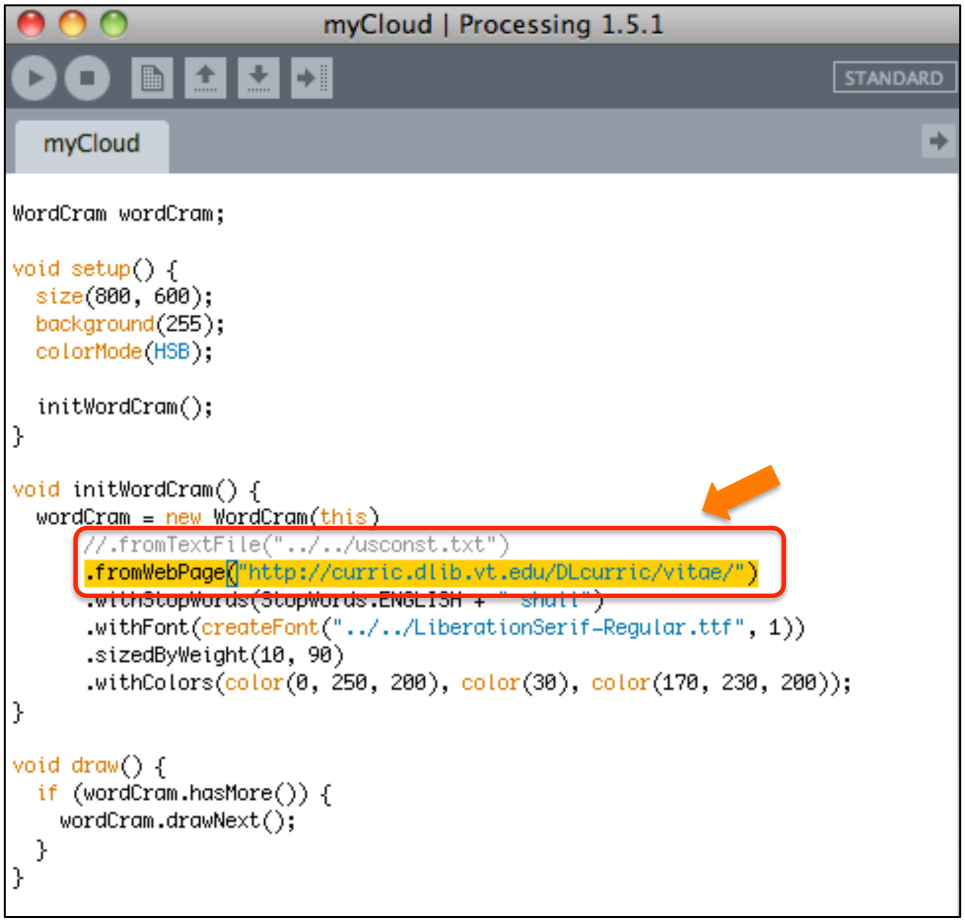
1.3. Connecting WordCram/Processing Sketch with Processed Tweet Text File

Step 1. Using usConstitution example as a template.
Go to ‘File’ and select ‘Save as’ to save the ‘usConstitution’ example as another name, for example, ‘myCloud’. You can save it under the folder, where usConstitution is saved.



Save As: myCloud
Where: OtherExamples

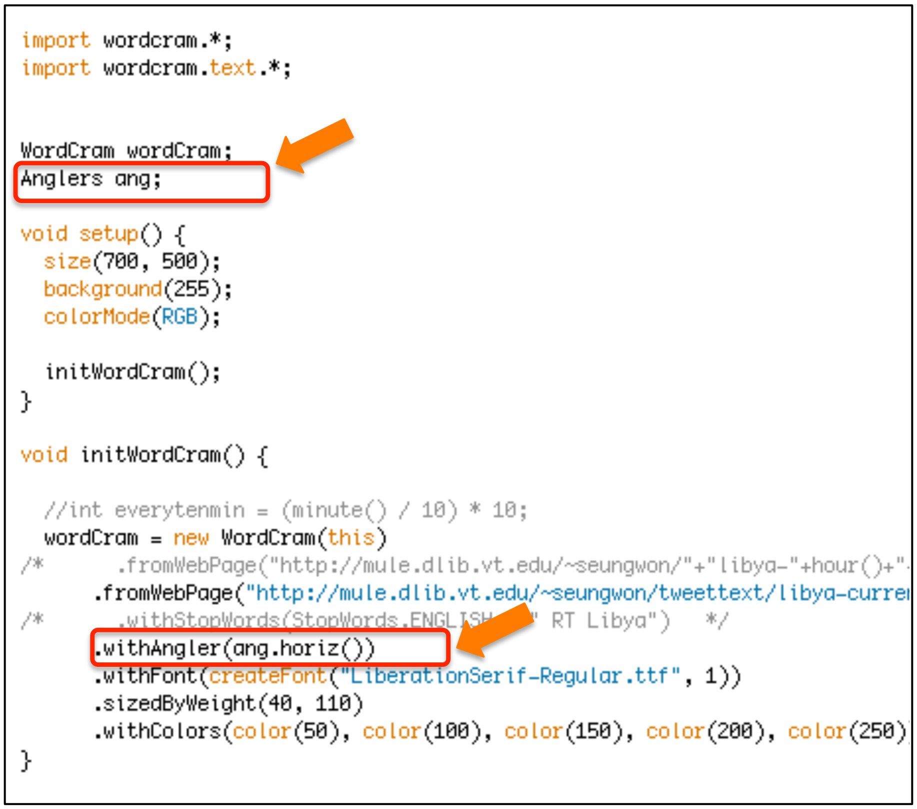
Step 2. Accessing a text file in a server.
Comment out ‘.fromTextFile(“../../usconst.txt”)’ line by adding two slashes in front of it, and add ‘.fromWebPage("<URL to the text file>")’ below the commented line. It allows the Processing sketch to use the text file in a server as an input. You can experiment with a URL to your homepage and see how it looks.



```
WordCram wordCram;

void setup() {
  size(800, 600);
  background(255);
  colorMode(HSB);

  initWordCram();
}

void initWordCram() {
  wordCram = new WordCram(this)
      //.fromTextFile("../../usconst.txt")
      .fromWebPage("http://curric.dlib.vt.edu/DLcurric/vitae/")
      .withStopWords(StopWords.ENGLISH + " shall")
      .withFont(createFont("../../LiberationSerif-Regular.ttf", 1))
      .sizedByWeight(10, 90)
      .withColors(color(0, 250, 200), color(30), color(170, 230, 200));
}

void draw() {
  if (wordCram.hasMore()) {
    wordCram.drawNext();
  }
}
```

Step 3. Changing to a horizontal layout.
Words that are placed vertically may not be comfortable to read. The default layout setting in WordCram is a mixed mode that places some words vertically and other words horizontally. To make all words horizontal:

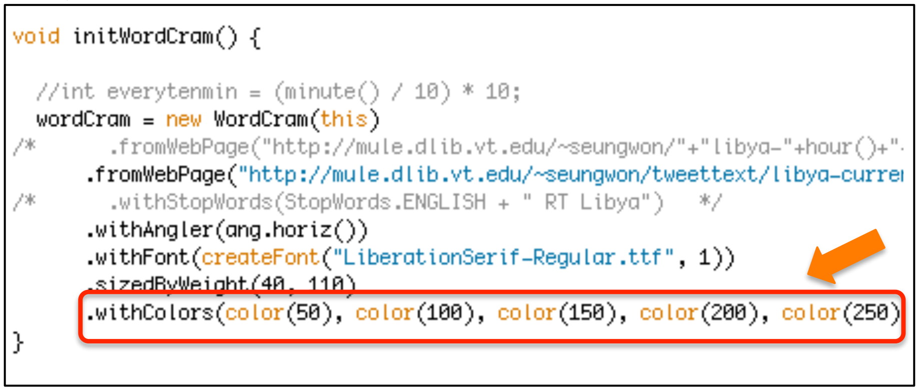
- Add ‘Anglers ang;’ (including the semicolon after ‘ang’) following the ‘WordCram wordCram’ line.
- Add ‘.withAngler(ang.horiz())’ line following the ‘.fromWebPage(…)’ line that we added in Step 2.
- Save the sketch and run it to see the changes



```
import wordcram.*;
import wordcram.text.*;


WordCram wordCram;
Anglers ang;

void setup() {
  size(700, 500);
  background(255);
  colorMode(RGB);

  initWordCram();
}

void initWordCram() {

  //int everytenmin = (minute() / 10) * 10;
  wordCram = new WordCram(this)
/*      .fromWebPage("http://mule.dlib.vt.edu/~seungwon/"+"libya-"+hour()+"-
      .fromWebPage("http://mule.dlib.vt.edu/~seungwon/tweettext/libya-curre
/*      .withStopWords(StopWords.ENGLISH + " RT Libya")   */
      .withAngler(ang.horiz())
      .withFont(createFont("LiberationSerif-Regular.ttf", 1))
      .sizedByWeight(40, 110)
      .withColors(color(50), color(100), color(150), color(200), color(250)
}
```

Step 4. Changing the color setting.

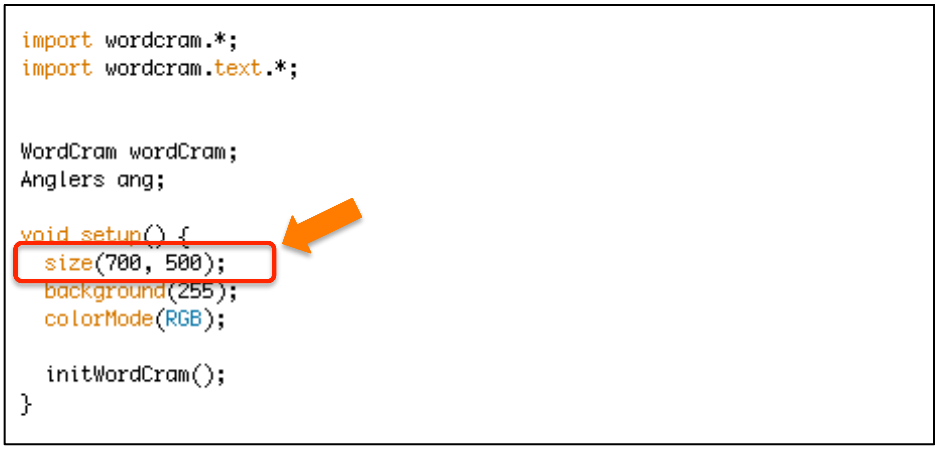
You can make the words either color or gray scale. If you add a single number inside parentheses of ‘color(…)’, it is interpreted as a gray scale. For RGB color mode, please add three numbers separated by commas (e.g., color(150, 255, 150) makes a bright green color). Numbers from 0-255 are allowed.



```
void initWordCram() {

  //int everytenmin = (minute() / 10) * 10;
  wordCram = new WordCram(this)
/*      .fromWebPage("http://mule.dlib.vt.edu/~seungwon/"+"libya-"+hour()+"-
      .fromWebPage("http://mule.dlib.vt.edu/~seungwon/tweettext/libya-curre
/*      .withStopWords(StopWords.ENGLISH + " RT Libya")   */
      .withAngler(ang.horiz())
      .withFont(createFont("LiberationSerif-Regular.ttf", 1))
      .sizedByWeight(40, 110)
      .withColors(color(50), color(100), color(150), color(200), color(250)
}
```

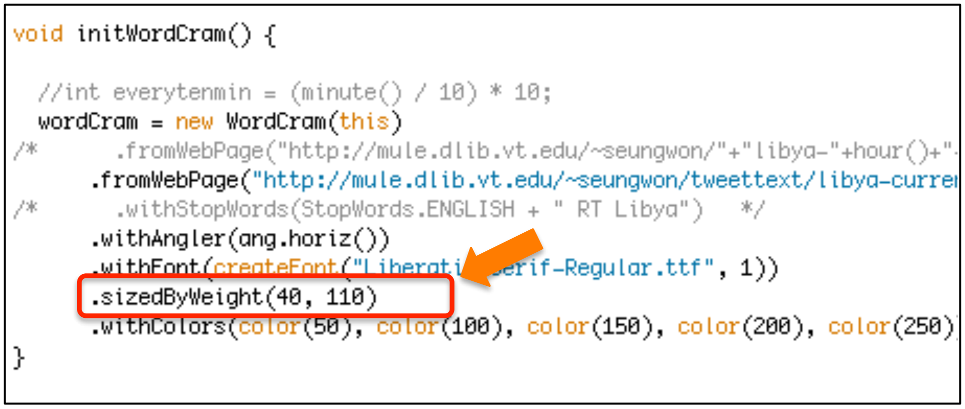
Step 5. Adjusting the canvas size, minimum and maximum font sizes.

Canvas size can be adjusted with ‘size(width, height)’.



```
import wordcram.*;
import wordcram.text.*;


WordCram wordCram;
Anglers ang;

void setup() {
  size(700, 500);
  background(255);
  colorMode(RGB);

  initWordCram();
}
```

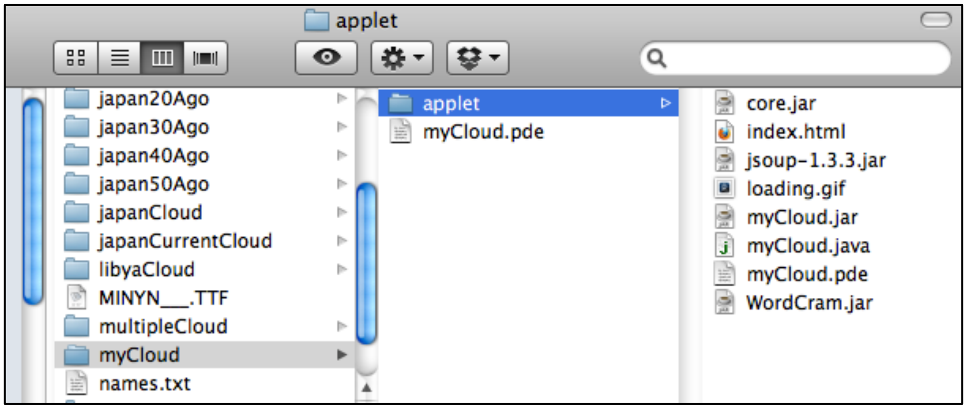
Try different numbers in ‘.sizedByWeight(minimum font size, maximum font size)’.



```
void initWordCram() {

  //int everytenmin = (minute() / 10) * 10;
  wordCram = new WordCram(this)
/*      .fromWebPage("http://mule.dlib.vt.edu/~seungwon/"+"libya-"+hour()+"-
      .fromWebPage("http://mule.dlib.vt.edu/~seungwon/tweettext/libya-curre
/*      .withStopWords(StopWords.ENGLISH + " RT Libya")   */
      .withAngler(ang.horiz())
      .withFont(createFont("Liberati erif-Regular.ttf", 1))
      .sizedByWeight(40, 110)
      .withColors(color(50), color(100), color(150), color(200), color(250)
}
```

Step 6. Exporting to an applet.

Applet is a program written in the Java programming language. It can be included in an HTML page and run on the browser of a client’s machine. In Processing, go to ‘File’ and select ‘Export Applet’. Then a pop up window appears. Upload the files in the ‘applet’ folder to your server. Then, the ‘index.html’ file will display the word cloud that you created.



Exercise

1. Create a word cloud using the New York Times web site (http://nytimes.com/) or your personal homepage URL.
2. Adjust settings such as layout and color based on the given instructions.
3. Another URL is given to each participant, which links to a pre-processed tweet text fetched every five minutes from your archive.
4. Replace the URL in your New York Times (or your homepage) word cloud with the provided one.
5. Compare your dynamic word cloud with the actual tweets (by searching with your key phrase at Twitter.com).
6. Does your word cloud give a snapshot of what is happening in your issue/event?

2. Analyzing Tweets with Python Scripts and Term Extraction API

The basic algorithm for text analysis in word cloud creation is to count word frequencies from input texts. For more meaningful content analysis, terms and phrases can be extracted using the term extraction application programming interface (API) and Natural Language Toolkit (NLTK).

2.1. Useful Unix/Linux Commands

There are some useful commands to use when working with scripts and APIs in either Linux or Mac OS X systems.

- ssh: Linux Secure Shell
  Provides access to a Linux system using password encryption in the login process.
  Usage
  shell> ssh seungwon@newyork.cc.vt.edu [hit enter key]
  Then, you will be prompted to enter a password

- scp: Secure copy.
  Scp allows to copy files across an ssh connection.
  Usage
  shell> scp [[user@]from-host:]source-file [[user@]to-host:][destination-file]
  shell> scp dlrl@georgia.dlib.vt.edu:/home/dlrl/testfile.txt /home/tmp
  You will be prompted to enter a password. Once verified, the file 'testfile.txt' is copied to the folder '/home/tmp' in the user's computer.

- ls: List directory contents
  ls lists information about the content of the current directory. With '-al' option, detailed information is displayed, too.
  Usage
  shell> ls → list content of the current directory

shell> ls –al /home/seungwon → list content details of /home/seungwon directory

- cd: Change directory
  Usage
  shell> cd /home/seungwon/tweettext → go to /home/seungwon/tweettext directory

- pwd: Print the current working directory
  Usage
  shell> pwd → this will show you the path to the directory that you are in.

- cat: Concatenate files and print on the standard output. It is often used to view the content of a file quickly.
  Usage
  shell> cat /home/seungwon/tweet.txt → it shows the content of 'tweet.txt' file on the standard output

- vi and vim editors: Command line-based Linux file editor
  Usage
  shell> vi <file name> → this will open the file specified in a vi interface.

  Once the file is open, you can edit/write when it is in input mode.
  Entering Input Mode

a	Add text after the cursor.
i	Insert text before the cursor.

  After finishing edits, push 'esc' key to enter command mode.
  Command Mode

:w	Write the file to disk.
:q	Quit vi.
dd	Delete an entire line where the cursor is located
yy	Copy an entire line where the cursor is located

- crontab: This utility creates, replace, or edit a user's crontab entry. The entry lists when and how often each task should be executed. It is used to schedule periodic background work.
  Usage
  shell> crontab –e [-e option is to edit the invoking user's crontab entry]

To run Python or PHP scripts:

- shell> python <script name>
- shell> php <script name>

2.2. Terminology Extraction

Yahoo! TermExtraction API[11]
This API allows extracting terms and key phrases from a textual input. This service is limited to 5,000 API calls per IP address per day for non-commercial use. The result is returned in XML format.

AlchemyAPI
It provides various services such as named entity extraction, concept tagging, keword/term extraction, sentiment analysis, etc. through their application programming interface (API). The rate limits for a registered account is 1,000 calls per day. It can increase to 30,000 if you get permission from the AlchemyAPI company.

Exercise

1. Login to server, virginia.cc.vt.edu, using the account provided.
2. Under your home directory, /home/<account name>/, you will create 4 data files named, input1.txt, input2.txt, input3.txt, and input4.txt. Each of them will include a group of tweets you used for previous exercises.
3. Run a Python script located in your home directory to extract terms from those four input files. For example, you may run the script:

```
Shell> python term_extractor.py input1.txt
```

2.3. Natural Language Toolkit (NLTK)

NLTK[12] is a powerful language analysis toolkit for Python. Exercises include identifying bigrams and trigrams using the provided Python script developed using NLTK.

Exercise

1. Login to server, virginia.cc.vt.edu, using the account provided.
2. Under your home directory, /home/<account name>/, a Python script "nltk_collocations.py" is located.
3. Run this script to find bigrams and trigrams from the input files, which were used in the previous exercise.

[11] http://developer.yahoo.com/search/content/V1/termExtraction.html
[12] http://www.nltk.org/