Tweet Categorization

CS 4624 Multimedia/Hypertext/Information Access

Virginia Tech, Blacksburg, VA 24061

4.27.2016

Author: Stephen Won

Client: Sunshin Lee

Table of Contents

Table of Tables

Table of Figures

Executive Summary

One of the main goals of Integrated Digital Event Archiving and Library (IDEAL) is to collect tweets and archive them in collection bases based on keyword. The Tweet Categorization project is to discover the suitable categorization schemes so that the users of the tweet collections will be able to understand each collection's general description such as what the keyword is, when and where it happened.

The project has been refined several times. First, the categorization scheme has been refined. In the beginning, the categorization scheme was to use a taxonomy scheme based on the event types. Also the GUI was to change the original static table that shows all tweet collections, to have search bar, and column ordering functions. Then the categorization scheme was changed to use a tag system. It would contain an event tag that describes the event type, place tag that shows the place where the event happened, and date tag that displays the date the event occurred. After that there was also a refinement that changed back to use a taxonomy scheme again but with a better GUI system that will show all the categories. Clicking the category will filter the tables to only show the related tweet collections. After that there was final refinement that uses a tag system, but also contain a taxonomy scheme for each tag. During the final refinement, the project was shifted to focus on creating a categorization scheme and applying it to the data file.

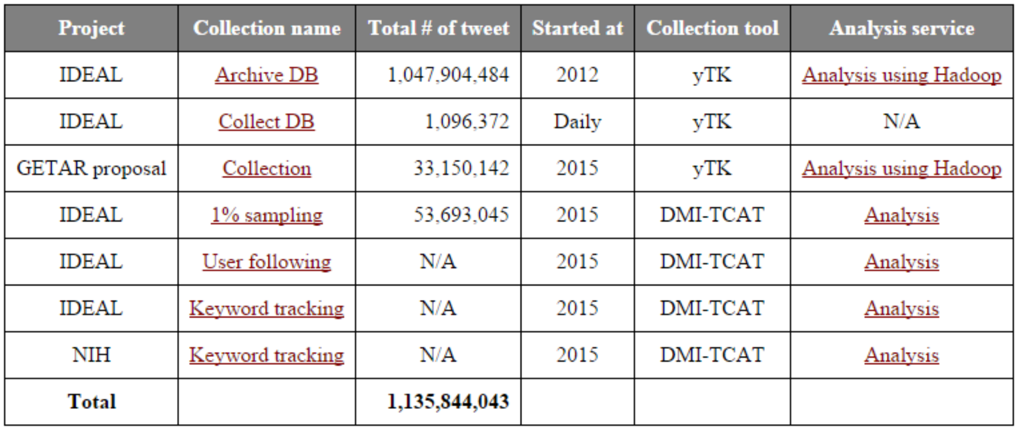
In this report, we discuss all the ideas and work undertaken these refinements. While others explain just descriptions, some are actual implementations.

User Manual

Project Description

Project background

One of the major goals of the Integrated Digital Event Archiving and Library (IDEAL) project is to collect tweets and Web-based content from social media and the general Web. As well as collecting data, the IDEAL project team also archives these materials permanently, and ensures access to these archives. The IDEAL team has collected tweets and Web collections about many events for many years, and archived Web collections using Internet Archive (IA) software, and tweet collections on local servers. These collections have been stored, organized, indexed and made available for searching, browsing, and other services. On the local server, currently there are 1,135,844,043 tweets archived. These tweets are divided into several databases. In these databases, the tweets are grouped into collections based on keyword or hashtags of general events (e.g., accidents, community activities), specific events (e.g., California shooting, chemical spill in West Virginia), places (e.g., California, Virginia Tech) and so on.



Project	Collection name	Total # of tweet	Started at	Collection tool	Analysis service
IDEAL	Archive DB	1,047,904,484	2012	yTK	Analysis using Hadoop
IDEAL	Collect DB	1,096,372	Daily	yTK	N/A
GETAR proposal	Collection	33,150,142	2015	yTK	Analysis using Hadoop
IDEAL	1% sampling	53,693,045	2015	DMI-TCAT	Analysis
IDEAL	User following	N/A	2015	DMI-TCAT	Analysis
IDEAL	Keyword tracking	N/A	2015	DMI-TCAT	Analysis
NIH	Keyword tracking	N/A	2015	DMI-TCAT	Analysis
Total		**1,135,844,043**			

Table 1 1,135,844,043 of tweets archived

Currently the database of these tweet collections contains seven fields: Archive ID, Keyword / Hashtag, Description, Tags, Screen Name, Count, and Create Time. The Archive ID field shows the ID number of a collection. In “Archive DB” the ID ranges from 1 to 705. The Keyword / Hashtag field shows the keyword or hashtag used to collect the tweets. The Description field gives specific details about the tweet collection. The tags field shows what type of collection it is. The count field shows the number of tweets in a collection. The create

Time field shows the date a collection was created, or when collection began. Overviews of collections are made into Tables (Table 2) in the website, http://hadoop.dlib.vt.edu/.

id	keyword	description	tags	screen_name	count	create_time
1	#egypt	Tweets for Egyptian revolution		Sslee77	13,136,418	07/10/2012
2	#libya			dlrl	2,799,862	07/10/2012
3	#blacksburg			dlrl	212,139	07/10/2012
4	#jan25			dlrl	1,141,723	07/10/2012
5	#bahrain			dlrl	22,010,557	07/10/2012
6	#yemen			dlrl	3,190,421	07/10/2012
7	japan earthquake			dlrl	1,273,463	07/10/2012
8	#syria			dlrl	16,860,208	07/10/2012
9	OccupyWallStreet			dlrl	1,006,610	07/10/2012
10	#nrv	new river valley (blacksburg) related tweets		dlrl	1,006,610	07/10/2012

…

Table 2 tweet collection table

However with so many collections of tweets, it may take a long time to search/browse for a specific collection, especially since few have text in the Description. Also, looking at collections by keywords or hashtags makes it hard to recognize connections with other collections. For example, “storm” collections and “earthquake” collections are both natural disasters, but from hundreds of collections, it is hard to recognize that in one look.

Also the user interface of the table is not made to be interactive so that the data can be organized in such a way the user wants. For example, from the table, even if the user wants to alphabetically order the collections by Keyword / Hashtag, there is no way of doing that.

Objectives

The major goal of the Event Based Categorization of Tweet Collections project is to help the IDEAL team’s research by accessing tweet collections easily. Client information is listed below.

Client name	Email
SunShin Lee	sslee777@vt.edu

Categorizing over 1,000 collections will make accessing a lot easier. Therefore from this project we will categorize the collections. The original method for categorization was to use a taxonomy scheme, but that was refined to use a tag system. This way the users will be able to see all the collections in organized categorizations. Although the object of the project was refined to focus on categorization of tweet collections, in the original planning, in addition to the categorizing, we planned to implement a user interface of the table so that it becomes more interactive, which will help the users' searching and browsing.

With the Event Based Categorization of Tweet Collections project, we will undertake three specific objectives: analyzing and identifying tweets collections, researching the most suitable categorization scheme for the collections, and categorizing them in a consistent fashion.

Analyzing & Identifying	As mentioned above, there are currently 1,047,904,484 tweets and 705 collections in "Archive DB" alone. We will analyze each tweet collections by its keywords, and study each keyword to find what they are. (for example, if the keyword is an event, we would search the event online and learn what incident occurred)
Researching	Using the information gathered from the analysis we will research and identify suitable categorization scheme that will fit well with the data and will best help the IDEAL team.
Categorizing	Using the categorization schema we will categorize the collections and a suitable user interface for the web application will be developed.

Table 3 Three Objectives

Target Audience

End Users

The project will serve mainly the IDEAL team. It will help them search and browse through tweet collections. When they need all the collections for a specific category, they could select the collections using a category that we have created, making the process easier and faster.

Not only the team, but also other viewers of the tweet collection table will have an easier time searching and browsing the collections by using the categories. In addition to that, they will have easier and more custom ways of organizing the table using interactive orderings of table columns.

People Maintaining

When the new tweet collections get ingested, the people who are maintaining them could fill the category fields when they are adding collections to the database.

People extending the project

One way the developers can extend upon the project is to add more layers to the taxonomy categorizations for each tag to make it more specific, or take out a layer to make it more broad and simple.

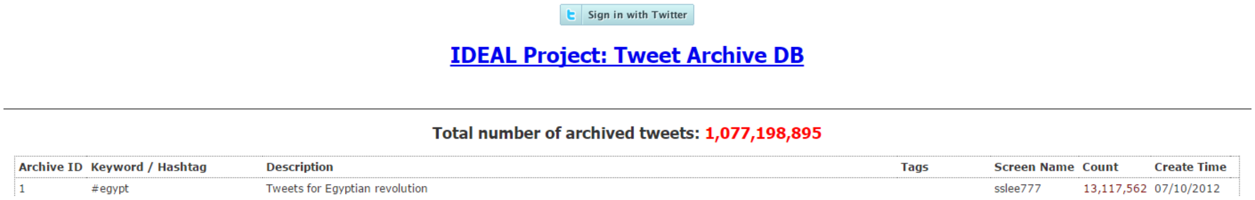
Developer's manual

Design

Current Design

Collections

- Tweets are grouped into collections based on keyword/hashtag, which were used for collecting.
- Currently there are seven fields in the collections database
  - Archive ID: shows numbered ID of each collections. This gets incremented as new collections are created.
  - Keyword / Hashtag: shows keyword or hashtag used to collect the tweets. Generally the keyword or hashtags indicate events, or places.
  - Description: shows brief descriptions of each keyword or hashtag. For example, the description of tweet collections of keyword "#PrayForKorea" is "North and South Korea exchanged fire, Aug 2015".
  - Tags: shows category of collections. For example, tags of "#Tunisia" and "wdbj7 shooting" collections are both "shooting".
  - Screen Name: shows the screen name of each collection.
  - Count: shows the number of tweets in a collection.
  - Create Time: shows the date of the collection creation.
- Although the tags somewhat categorizes the collections, only few of the collections have the field filled.



IDEAL Project: Tweet Archive DB

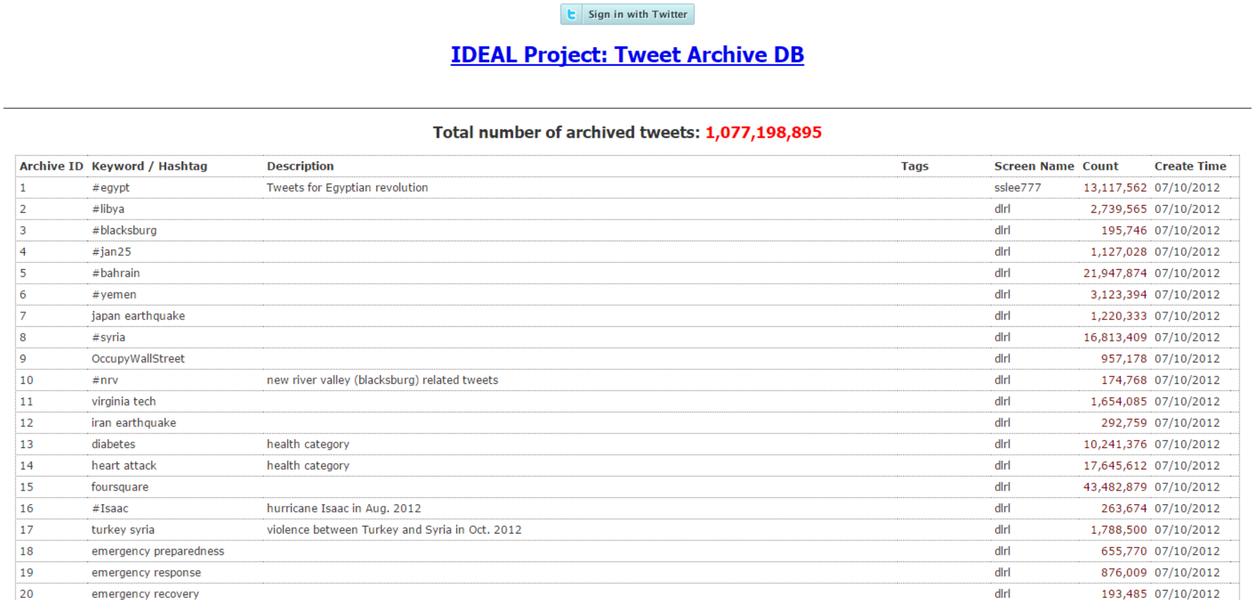
Total number of archived tweets: 1,077,198,895

Archive ID	Keyword / Hashtag	Description	Tags	Screen Name	Count	Create Time
1	#egypt	Tweets for Egyptian revolution		sslee777	13,117,562	07/10/2012

Figure 1 fields in current design of tweet collections

Web Table

- One of the table of tweet collections can be found on http://jingluo.dlib.vt.edu/twitter/
- At the top of the page, it shows the project name and the database name the tweets are archived in.
- The header of the table shows the total number of archived tweets.
- The tweet collections are shown in a table.
    - Columns: displays all fields of the collections.
    - Rows: displays all the collections in the database.
- Currently, the table is not interactive, so it is only for displaying the table (no ordering, no searching).
- The tool that the current design is using is "yourTwapperKeeper" (https://github.com/540co/yourTwapperKeeper).



IDEAL Project: Tweet Archive DB

Total number of archived tweets: 1,077,198,895

Archive ID	Keyword / Hashtag	Description	Tags	Screen Name	Count	Create Time
1	#egypt	Tweets for Egyptian revolution		sslee777	13,117,562	07/10/2012
2	#libya			dlrl	2,739,565	07/10/2012
3	#blacksburg			dlrl	195,746	07/10/2012
4	#jan25			dlrl	1,127,028	07/10/2012
5	#bahrain			dlrl	21,947,874	07/10/2012
6	#yemen			dlrl	3,123,394	07/10/2012
7	japan earthquake			dlrl	1,220,333	07/10/2012
8	#syria			dlrl	16,813,409	07/10/2012
9	OccupyWallStreet			dlrl	957,178	07/10/2012
10	#nrv	new river valley (blacksburg) related tweets		dlrl	174,768	07/10/2012
11	virginia tech			dlrl	1,654,085	07/10/2012
12	iran earthquake			dlrl	292,759	07/10/2012
13	diabetes	health category		dlrl	10,241,376	07/10/2012
14	heart attack	health category		dlrl	17,645,612	07/10/2012
15	foursquare			dlrl	43,482,879	07/10/2012
16	#Isaac	hurricane Isaac in Aug. 2012		dlrl	263,674	07/10/2012
17	turkey syria	violence between Turkey and Syria in Oct. 2012		dlrl	1,788,500	07/10/2012
18	emergency preparedness			dlrl	655,770	07/10/2012
19	emergency response			dlrl	876,009	07/10/2012
20	emergency recovery			dlrl	193,485	07/10/2012

Figure 2 current design of collection table

New Design

Collections

- We kept all the fields that are originally there.
- We added new fields for event type, place and date tags.
- Each tag contains taxonomic layers.

Event Type:

Event Type 1	Most general event types. Ex., Man-Made Disaster, Natural Disaster
Event Type 2	More specific event types. Ex., Shooting, Climate Change
Event Type 3	Most specific event types. Ex., School Shooting, Storm

Place:

Country	Country the event has occurred
State	State the event has occurred
City	City the event has occurred

Date:

Year	Event occurred year
Month	Event occurred month
Day	Event occurred day

Web Table

Although the project has been refined to focus on categorization, our already created GUI can be used later.

- We use “yourTwapperKeepper” tool.
- Using yourTwapperKeepper tool, we added search boxes for searching by categories.
- Each columns will be made clickable for ordering of the table.
- We created a view that will show all the categories, and by clicking the category will filter the table to display only the related collections.

Tools

- yourTwapperKeeper
- MySQL Database

- PHP
- JQuery
- DataTables JQuery plug-in
- JavaScript
- CSS

Implementation

Overview

For the implementation there are three major phases: Research and Design, System Implementation, and Testing. The Research and Design phase includes meetings and discussions with the IDEAL team to find the right solution for categorization schemes. The system Implementation phase includes implementation of the database and web application. Testing includes testing and verifying.

Description

The first step of Implementation of the Tweet Categorization project involves researching for the suitable categorization schemes based on requirements and conveniences of the IDEAL team, especially Mr. Lee. Therefore listening to the IDEAL team's opinions is required. During the IDEAL meetings, we discussed alternative categorization schemes. From the outcomes of the discussion, the right categorization scheme was selected, and the system implementations proceeded. The system implementation includes a database of over 1,000 tweet collections. The collections were manually analyzed and assigned into the correct categorizations. Then the yourTwapperKeeper tool, which the current Virginia Tech server is using for tweet collections, was modified. First the new search function was added to enable searching. Not only that, an order by columns function was added to the application. Also, the separate view was added that displayed all the categories, and we implemented so that clicking on the category will filter the table to show related collections.

Major Tasks

This subsection will show detailed tasks for the implementations.

Phase 1. Research and Design:

There are over 1,000 tweet collections that are assigned to us. The topics of these collections are various. Also there are many ambiguous topics that could belong to several or no categorization. Meetings and discussions was required to resolve these issues based on the needs and convenience of IDEAL team. Then the Excel file containing all the tweet collections was implemented based on the chosen categorization as a prototype and reference.

yourTwapperKeeper is the current tool that displays tweet collections to the web application. However, yourTwapperKeeper contains already set fields and provides limited functionalities. This tool was implemented so that searching functionality and ordering functionality are added. However, before making an actual implementation on the Virginia Tech server, it needs to be installed on a local server. The design will be implemented on the local server first. This will protect the current application from possible bugs.

- Preparation of several initial rough categorization schemes
    - These rough categorizations will later be shown to the IDEAL team and Mr.Lee, so if there is one from the choices that suits best for the IDEAL team's goal, that scheme will be used.
- Meetings and Discussions to pick out optimal scheme
- Implementation of the tweet collections Excel file
- Installation of yourTwapperKeeper to local server
    - creation of local dummy server
    - creation of local dummy database
    - integration of yourTwapperKeeper to dummy server and database

Phase 2. System Implementation:

After the research and design of the categorization scheme is finalized, the next step is actual implementation to the system. The first implementation is to the database. The new categorization fields will be added to the database. Finally, yourTwapperKeeper tool will be implemented. First it will be implemented in the local server. yourTwapperKeeper tool implementation will include implementing a search algorithm to add search functionality, and implementing a sorting algorithm to add sorting by column functionality.

- Implementation of yourTwapperKeeper
    - Implementation of GUI
    - Implementation of search functionality
    - Implementation of sort functionality
    - Adding separate view that shows all clickable categories that will filter the table.

Phase 3. Testing:

Although the categorization scheme has been changed, and our focus changed to categorize the tweet collections Excel file, the already created GUI was tested using the dummy database. It was tested and verified in a local server, so that if any bug or misbehaving functionality is occurred it will be fixed. Then the implementations and modifications could bedocumented.

- Testing and Verification
- Documentation

Refinements

Throughout the several meetings and discussions, there have been some refinements made to the project. The refinements are organized in table 4.

Original Plan	Categorization: Use topology scheme to the event type only. GUI: Create interactive table that will filter the collections from the search box, and that will enable users to use column ordering.
Refinement 1	Categorization: Use tag system for event type, place and date.
Refinement 2	Categorization: Use topology scheme to event type. GUI: Create a categorization view that will display all categories, and enable clicking the category will filter the table to only show related collections.
Refinement 3	Categorization: Use tag system for event type, place and date. And each tag will contain topology scheme. GUI: Focus on categorization implementation.

Table 4 refinements

Refinement 1

The original implementation plan of the tweet categorization was to create a taxonomic structure so that the user can go down through the taxonomic order to search and browse. However through the IDEAL group meeting, the topological categorization scheme plan was changed to call for using a tag system.

Prototype

Purpose

The goal of our project, Tweet Categorization, is to help IDEAL team's research. Therefore through multiple discussions and meetings, the project's goal has been refined multiple times. However, in this section we would like to show the prototype made for refinement 1.

The two main parts of our project, for refinement 1, consist of categorizing tweet collections, and implementation of a User Interface for the table of yourTwapperKeeper. The categorization of tweet collections will enable users to group them based on the categorization so that the users can search for desired keywords or terms and show the results more easily, while the refined user interface of the table will make the features of the categorization feasible. The section will focus on the prototypes of the two main parts of our project: Database Implementation prototype and User Interface Implementation prototype.

Modelling the prototype

Tweet Categorization

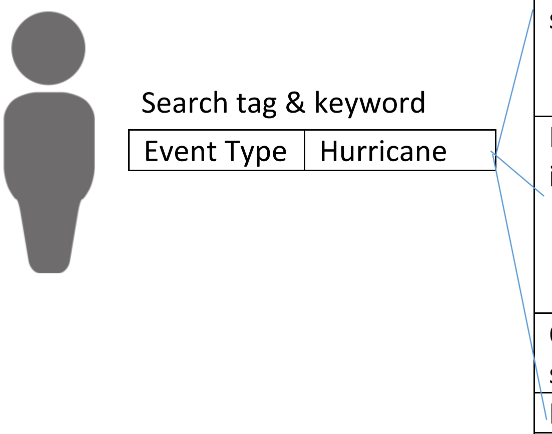
The tag system will consist of three new tags: event type, place, and date. Each tag may contain zero to many keywords that are related to each tag type.

Event Type	Types of event
Place	Place where the event happened, or related places of the event
Date	Date the event has occurred

Through these tags, the users may search for collections. When the user searches for a keyword of a tag, all the collections that contain the keyword in the tag will be gathered, which will help the users see the related collections of the specific keyword.

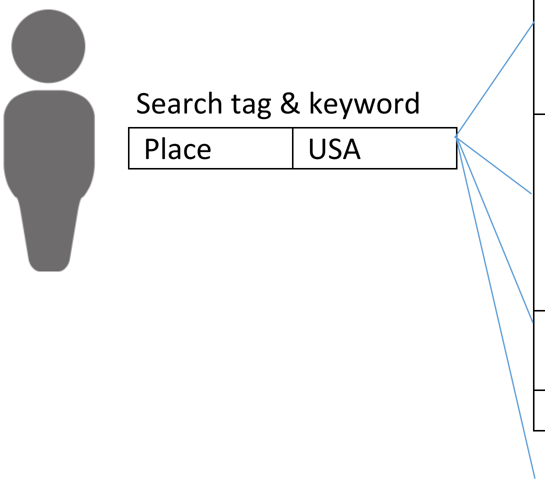
Diagram (Tweet Categorization)

Searching for keyword in event type tags



Keyword	Event type	Place	Date
Hurricane sandy	Hurricane	Carolinas, Virginia, Washington D.C, USA …	October, November, 2012
Hurricane issac	Hurricane	Arkansas, Mississippi, Alabama, Louisiana, USA …	August, September, 2012
Oregon shooting	Shooting	Clackamas, Oregon, USA	December, 2012
Hurricane	Hurricane		
#Tucson shooting	Shooting	Tucson, Arizona, USA	January, 2011
…			

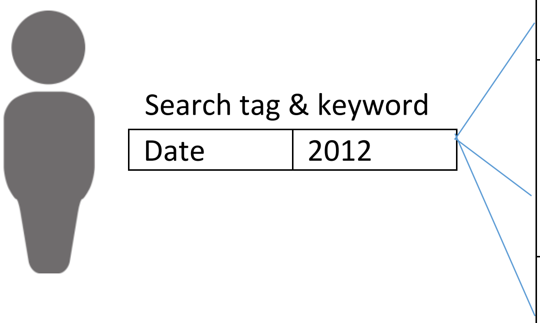
Searching for keyword in place tags



Keyword	Event type	Place	Date
Hurricane sandy	Hurricane	Carolinas, Virginia, Washington D.C, USA …	October, November, 2012
Hurricane issac	Hurricane	Arkansas, Mississippi, Alabama, Louisiana, USA …	August, September, 2012
Oregon shooting	Shooting	Clackamas, Oregon, USA	December, 2012
Hurricane	Hurricane		

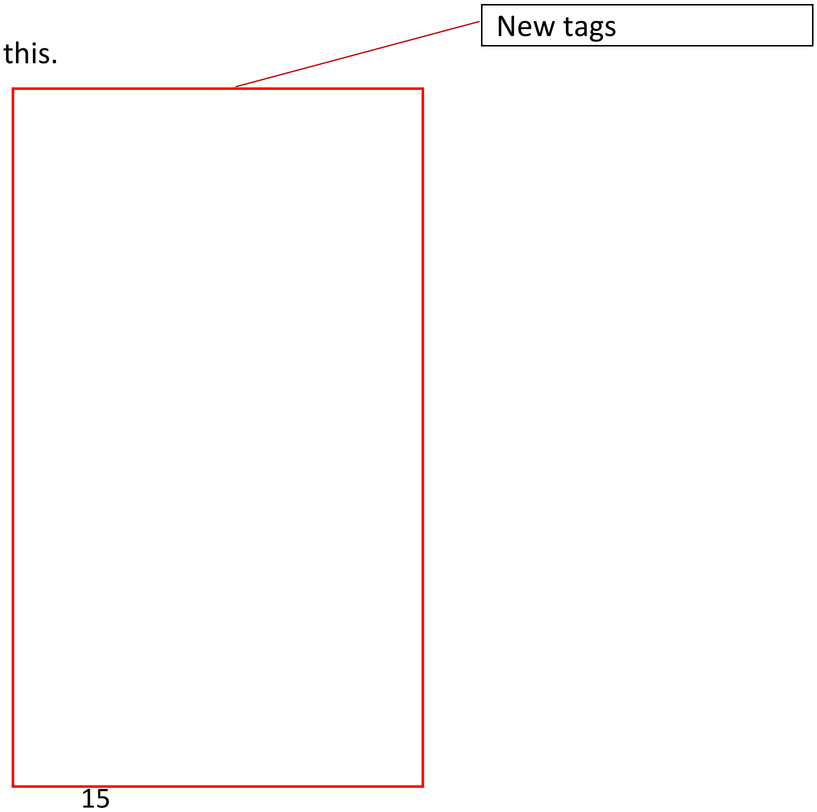
#Tucson shooting	Shooting	Tucson, Arizona, USA	January, 2011
…			

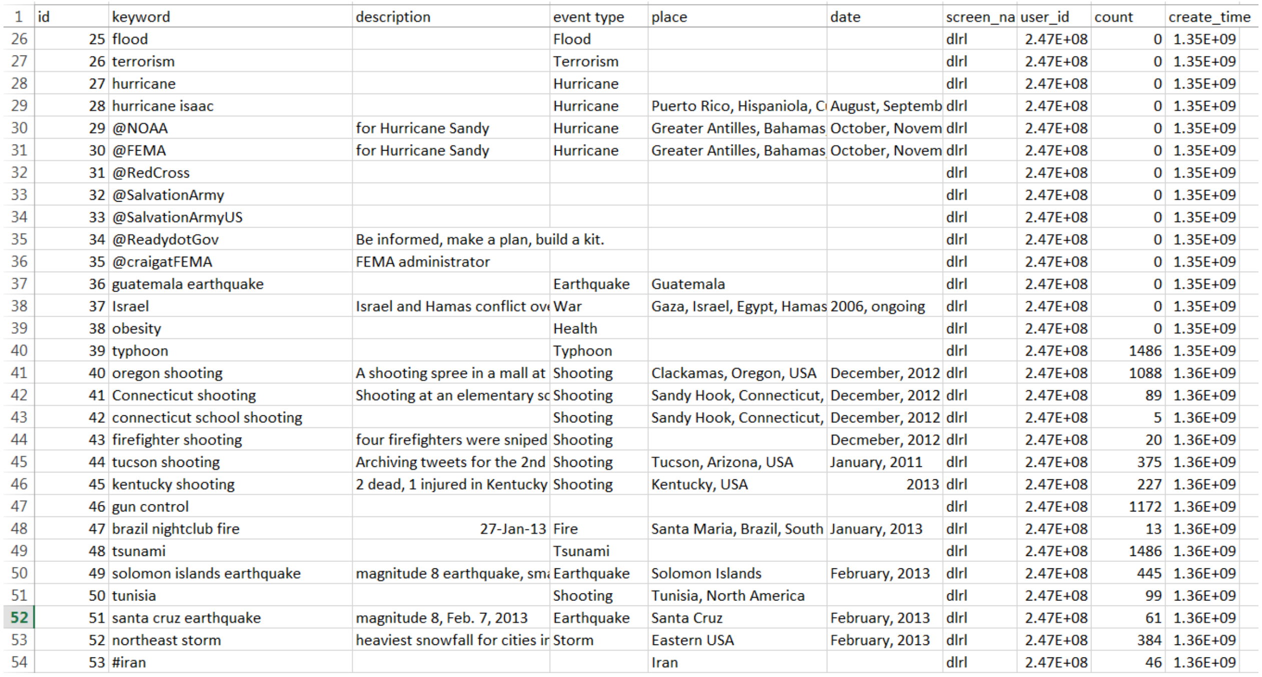
Searching for keyword in date tags



Keyword	Event type	Place	Date
Hurricane sandy	Hurricane	Carolinas, Virginia, Washington D.C, USA …	October, November, 2012
Hurricane issac	Hurricane	Arkansas, Mississippi, Alabama, Louisiana, USA …	August, September, 2012
Oregon shooting	Shooting	Clackamas, Oregon, USA	December, 2012
Hurricane	Hurricane		
#Tucson shooting	Shooting	Tucson, Arizona, USA	January, 2011
…			

The prototype of the database looks like this.





1	id	keyword	description	event type	place	date	screen_na	user_id	count	create_time
26	25	flood		Flood			dlrl	2.47E+08	0	1.35E+09
27	26	terrorism		Terrorism			dlrl	2.47E+08	0	1.35E+09
28	27	hurricane		Hurricane			dlrl	2.47E+08	0	1.35E+09
29	28	hurricane isaac		Hurricane	Puerto Rico, Hispaniola, C	August, Septemb	dlrl	2.47E+08	0	1.35E+09
30	29	@NOAA	for Hurricane Sandy	Hurricane	Greater Antilles, Bahamas	October, Novem	dlrl	2.47E+08	0	1.35E+09
31	30	@FEMA	for Hurricane Sandy	Hurricane	Greater Antilles, Bahamas	October, Novem	dlrl	2.47E+08	0	1.35E+09
32	31	@RedCross					dlrl	2.47E+08	0	1.35E+09
33	32	@SalvationArmy					dlrl	2.47E+08	0	1.35E+09
34	33	@SalvationArmyUS					dlrl	2.47E+08	0	1.35E+09
35	34	@ReadydotGov	Be informed, make a plan, build a kit.				dlrl	2.47E+08	0	1.35E+09
36	35	@craigatFEMA	FEMA administrator				dlrl	2.47E+08	0	1.35E+09
37	36	guatemala earthquake		Earthquake	Guatemala		dlrl	2.47E+08	0	1.35E+09
38	37	Israel	Israel and Hamas conflict ov	War	Gaza, Israel, Egypt, Hamas	2006, ongoing	dlrl	2.47E+08	0	1.35E+09
39	38	obesity		Health			dlrl	2.47E+08	0	1.35E+09
40	39	typhoon		Typhoon			dlrl	2.47E+08	1486	1.35E+09
41	40	oregon shooting	A shooting spree in a mall at	Shooting	Clackamas, Oregon, USA	December, 2012	dlrl	2.47E+08	1088	1.36E+09
42	41	Connecticut shooting	Shooting at an elementary sc	Shooting	Sandy Hook, Connecticut,	December, 2012	dlrl	2.47E+08	89	1.36E+09
43	42	connecticut school shooting		Shooting	Sandy Hook, Connecticut,	December, 2012	dlrl	2.47E+08	5	1.36E+09
44	43	firefighter shooting	four firefighters were sniped	Shooting		Decmeber, 2012	dlrl	2.47E+08	20	1.36E+09
45	44	tucson shooting	Archiving tweets for the 2nd	Shooting	Tucson, Arizona, USA	January, 2011	dlrl	2.47E+08	375	1.36E+09
46	45	kentucky shooting	2 dead, 1 injured in Kentucky	Shooting	Kentucky, USA	2013	dlrl	2.47E+08	227	1.36E+09
47	46	gun control					dlrl	2.47E+08	1172	1.36E+09
48	47	brazil nightclub fire	27-Jan-13	Fire	Santa Maria, Brazil, South	January, 2013	dlrl	2.47E+08	13	1.36E+09
49	48	tsunami		Tsunami			dlrl	2.47E+08	1486	1.36E+09
50	49	solomon islands earthquake	magnitude 8 earthquake, sma	Earthquake	Solomon Islands	February, 2013	dlrl	2.47E+08	445	1.36E+09
51	50	tunisia		Shooting	Tunisia, North America		dlrl	2.47E+08	99	1.36E+09
52	51	santa cruz earthquake	magnitude 8, Feb. 7, 2013	Earthquake	Santa Cruz	February, 2013	dlrl	2.47E+08	61	1.36E+09
53	52	northeast storm	heaviest snowfall for cities ir	Storm	Eastern USA	February, 2013	dlrl	2.47E+08	384	1.36E+09
54	53	#iran			Iran		dlrl	2.47E+08	46	1.36E+09

Figure 3 Prototype of the database

User Interface

The purpose of the user Interface portion of the project was to enable the users to interact with the data table, so that the user can sort the table by column or search by keyword. If the new categorization (new tags) shows what features could be added, then the user interface part is making those features feasible.

Diagram (User Interface)

User sort by keyword

Keyword	Event type	Place	Date
#Tucson shooting	Shooting	Tucson, Arizona, USA	January, 2011
Hurricane	Hurricane		
Hurricane issac	Hurricane	Arkansas, Mississippi, Alabama, Louisiana, USA …	August, September, 2012
Hurricane sandy	Hurricane	Carolinas, Virginia, Washington D.C, USA …	October, November, 2012
Oregon shooting	Shooting	Clackamas, Oregon, USA	December, 2012

…

User sort by Event type

Keyword	Event type	Place	Date
Hurricane	Hurricane		
Hurricane issac	Hurricane	Arkansas, Mississippi, Alabama, Louisiana, USA …	August, September, 2012
Hurricane sandy	Hurricane	Carolinas, Virginia, Washington D.C, USA …	October, November, 2012
#Tucson shooting	Shooting	Tucson, Arizona, USA	January, 2011
Oregon shooting	Shooting	Clackamas, Oregon, USA	December, 2012
…			

User search “shooting”

Keyword	Event type	Place	Date
#Tucson shooting	Shooting	Tucson, Arizona, USA	January, 2011
Oregon shooting	Shooting	Clackamas, Oregon, USA	December, 2012
…			

User search “USA”

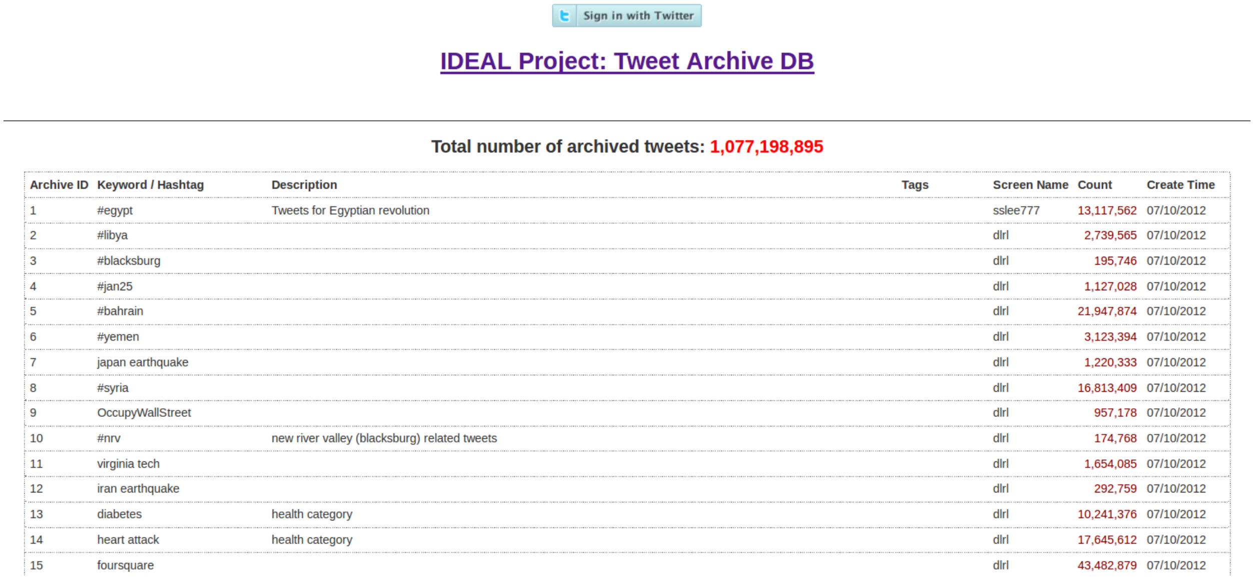
Keyword	Event type	Place	Date
Hurricane issac	Hurricane	Arkansas, Mississippi, Alabama, Louisiana, USA …	August, September, 2012
Hurricane sandy	Hurricane	Carolinas, Virginia, Washington D.C, USA …	October, November, 2012

#Tucson shooting	Shooting	Tucson, Arizona, USA	January, 2011
Oregon shooting	Shooting	Clackamas, Oregon, USA	December, 2012
…			

Prototype Process

In order to make the table interactive, we used the DataTables plug-in. DataTables is a plug-in for jQuery Javascript library. It is very flexible and allows developers freedom of modifying table to match their purposes.

In order to integrate DataTables, we modified the existing index.php file from yourTwapperKeeper sources. Figure 5 shows the prototype of DataTables integrated into yourTwapperKeeper. Notice how column headers include marks besides them which shows if a column was sorted. Also there is search bar at the top right corner. Figure 6 shows how it only shows collections that contains the keyword “shooting”.

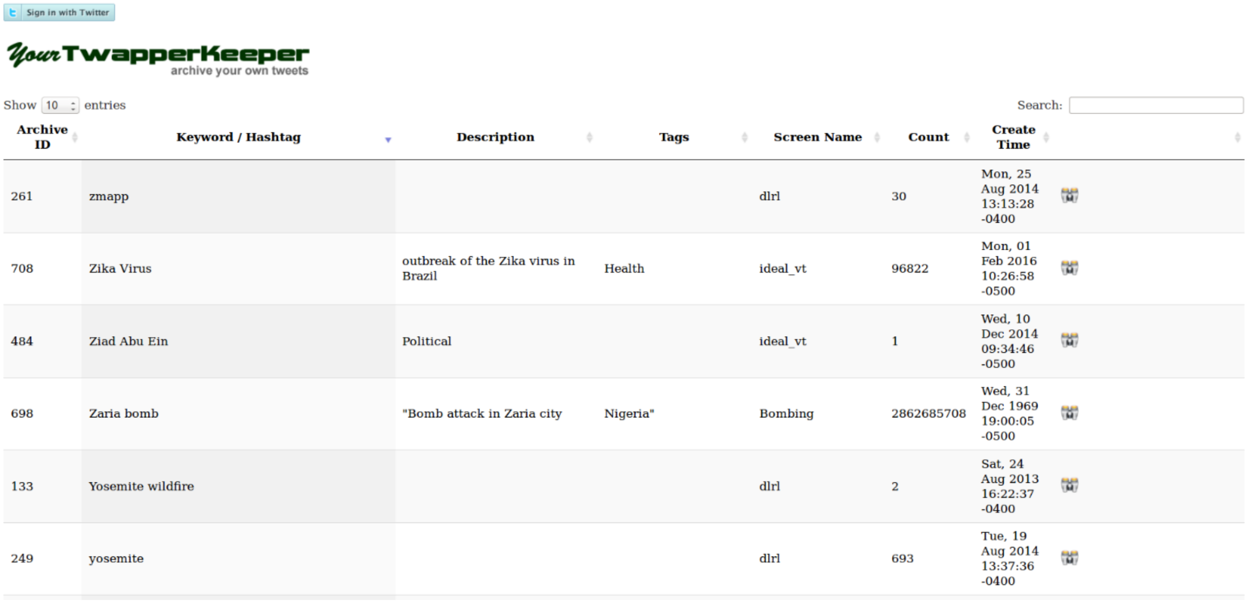


IDEAL Project: Tweet Archive DB

Total number of archived tweets: 1,077,198,895

Archive ID	Keyword / Hashtag	Description	Tags	Screen Name	Count	Create Time
1	#egypt	Tweets for Egyptian revolution		sslee777	13,117,562	07/10/2012
2	#libya			dlrl	2,739,565	07/10/2012
3	#blacksburg			dlrl	195,746	07/10/2012
4	#jan25			dlrl	1,127,028	07/10/2012
5	#bahrain			dlrl	21,947,874	07/10/2012
6	#yemen			dlrl	3,123,394	07/10/2012
7	japan earthquake			dlrl	1,220,333	07/10/2012
8	#syria			dlrl	16,813,409	07/10/2012
9	OccupyWallStreet			dlrl	957,178	07/10/2012
10	#nrv	new river valley (blacksburg) related tweets		dlrl	174,768	07/10/2012
11	virginia tech			dlrl	1,654,085	07/10/2012
12	iran earthquake			dlrl	292,759	07/10/2012
13	diabetes	health category		dlrl	10,241,376	07/10/2012
14	heart attack	health category		dlrl	17,645,612	07/10/2012
15	foursquare			dlrl	43,482,879	07/10/2012

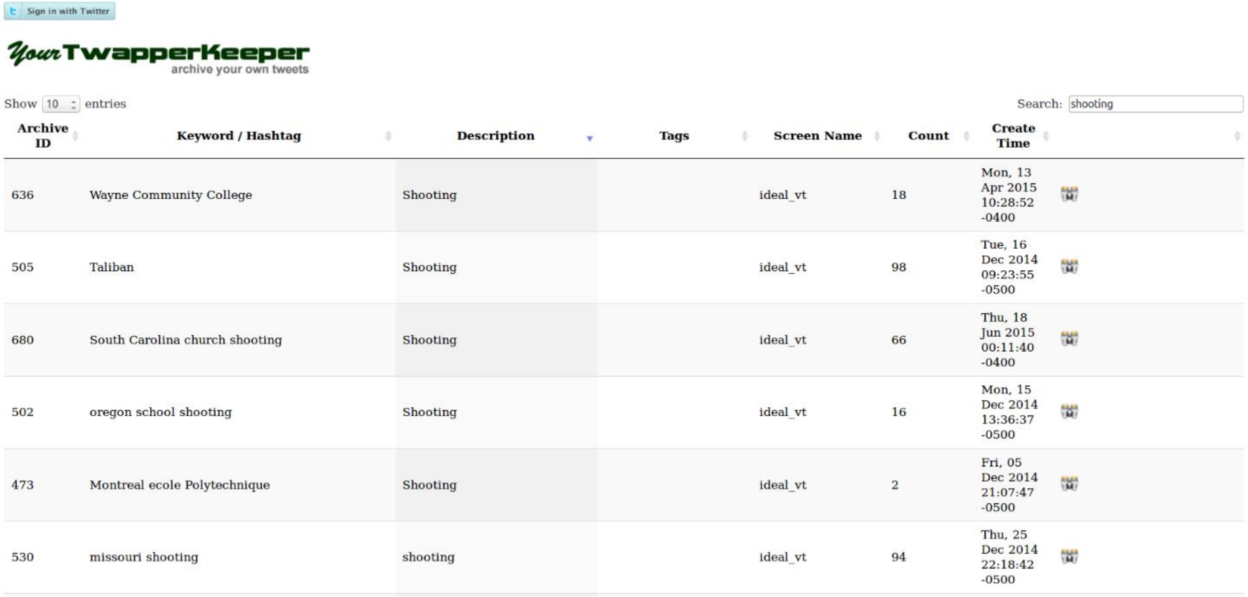
Figure 4 Current page using original yourTwapperKeeper UI



Show 10 entries Search:

Archive ID	Keyword / Hashtag	Description	Tags	Screen Name	Count	Create Time
261	zmapp			dlrl	30	Mon, 25 Aug 2014 13:13:28 -0400
708	Zika Virus	outbreak of the Zika virus in Brazil	Health	ideal_vt	96822	Mon, 01 Feb 2016 10:26:58 -0500
484	Ziad Abu Ein	Political		ideal_vt	1	Wed, 10 Dec 2014 09:34:46 -0500
698	Zaria bomb	"Bomb attack in Zaria city	Nigeria"	Bombing	2862685708	Wed, 31 Dec 1969 19:00:05 -0500
133	Yosemite wildfire			dlrl	2	Sat, 24 Aug 2013 16:22:37 -0400
249	yosemite			dlrl	693	Tue, 19 Aug 2014 13:37:36 -0400

Figure 5 Table using DataTables jQuery Plug-in

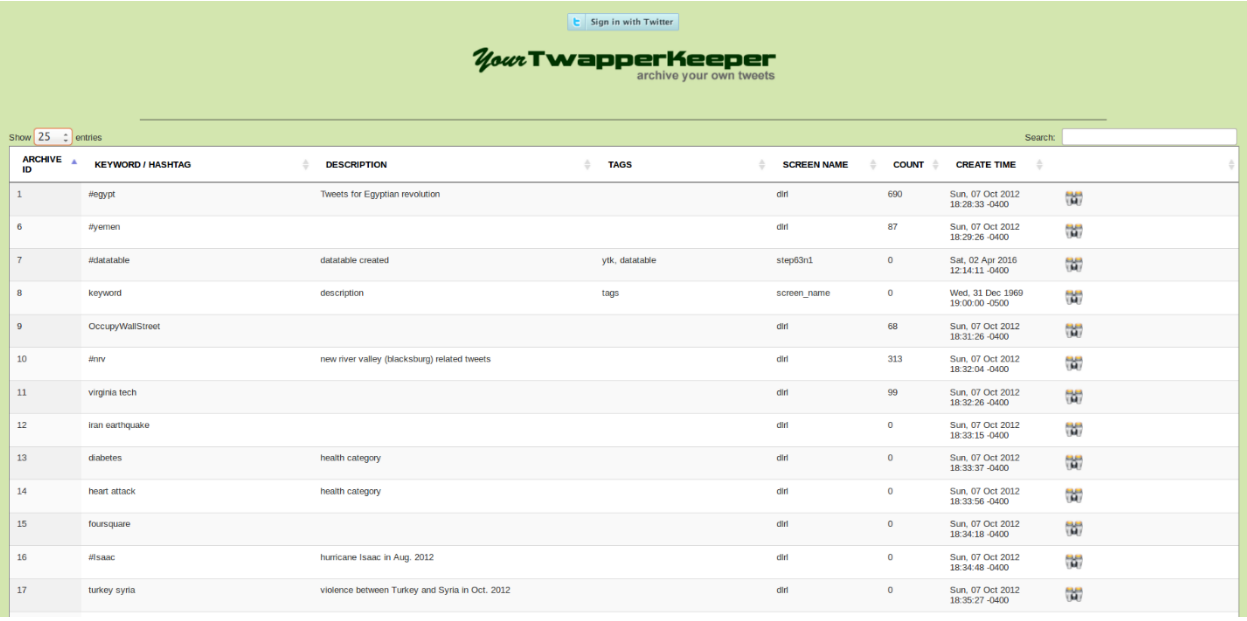


Show 10 entries Search: shooting

Archive ID	Keyword / Hashtag	Description	Tags	Screen Name	Count	Create Time
636	Wayne Community College	Shooting		ideal_vt	18	Mon, 13 Apr 2015 10:28:52 -0400
505	Taliban	Shooting		ideal_vt	98	Tue, 16 Dec 2014 09:23:55 -0500
680	South Carolina church shooting	Shooting		ideal_vt	66	Thu, 18 Jun 2015 00:11:40 -0400
502	oregon school shooting	Shooting		ideal_vt	16	Mon, 15 Dec 2014 13:36:37 -0500
473	Montreal ecole Polytechnique	Shooting		ideal_vt	2	Fri, 05 Dec 2014 21:07:47 -0500
530	missouri shooting	shooting		ideal_vt	94	Thu, 25 Dec 2014 22:18:42 -0500

Figure 6 Table search with keyword "shooting"

After that the DataTables was integrated with the modified yourTwapperKeeper CSS file so that it becomes consistent with other pages in yourTwapperKeeper. Figure 7 shows the data table with yourTwapper background and containing sorting, and searching features.



ARCHIVE ID	KEYWORD / HASHTAG	DESCRIPTION	TAGS	SCREEN NAME	COUNT	CREATE TIME	
1	#egypt	Tweets for Egyptian revolution		dlrl	690	Sun, 07 Oct 2012 18:28:33 -0400	
6	#yemen			dlrl	87	Sun, 07 Oct 2012 18:29:26 -0400	
7	#datatable	datatable created	ytk, datatable	step63n1	0	Sat, 02 Apr 2016 12:14:11 -0400	
8	keyword	description	tags	screen_name	0	Wed, 31 Dec 1969 19:00:00 -0500	
9	OccupyWallStreet			dlrl	68	Sun, 07 Oct 2012 18:31:26 -0400	
10	#nrv	new river valley (blacksburg) related tweets		dlrl	313	Sun, 07 Oct 2012 18:32:04 -0400	
11	virginia tech			dlrl	99	Sun, 07 Oct 2012 18:32:26 -0400	
12	iran earthquake			dlrl	0	Sun, 07 Oct 2012 18:33:15 -0400	
13	diabetes	health category		dlrl	0	Sun, 07 Oct 2012 18:33:37 -0400	
14	heart attack	health category		dlrl	0	Sun, 07 Oct 2012 18:33:56 -0400	
15	foursquare			dlrl	0	Sun, 07 Oct 2012 18:34:18 -0400	
16	#Isaac	hurricane Isaac in Aug. 2012		dlrl	0	Sun, 07 Oct 2012 18:34:48 -0400	
17	turkey syria	violence between Turkey and Syria in Oct. 2012		dlrl	0	Sun, 07 Oct 2012 18:35:27 -0400	

Figure 7 DataTables integrated with yourTwapperKeeper css

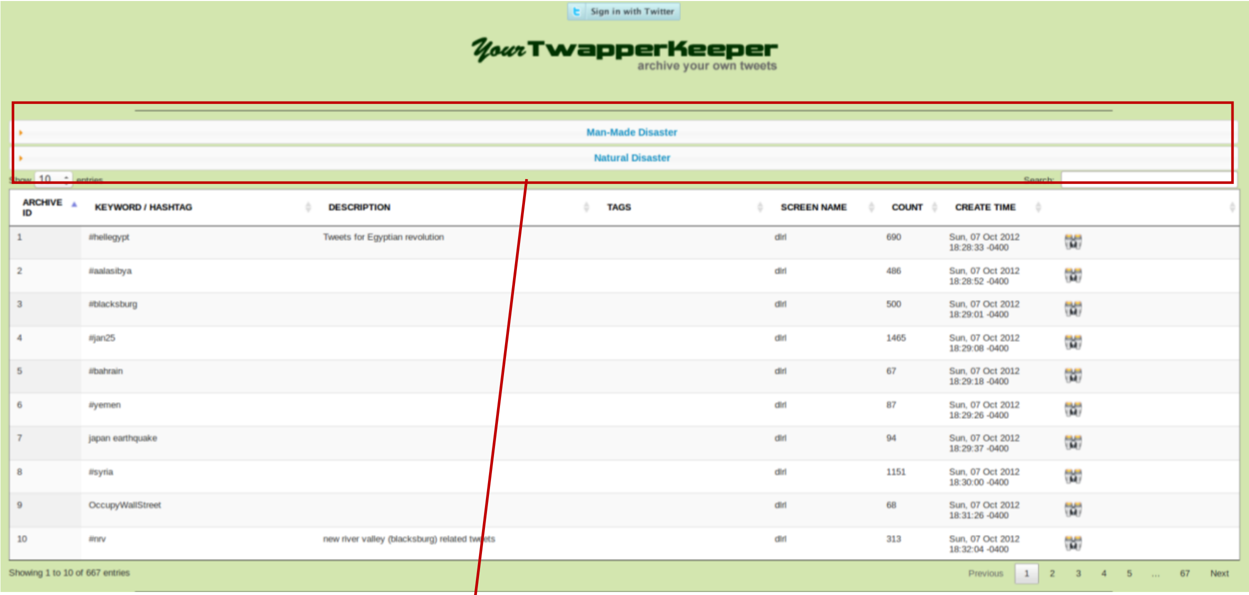
Refinement 2

From the meetings, the client wanted to have a database with categorization that uses the taxonomy scheme from the original project plan. Also in addition to the interface that allows users to search and order columns from the table, the client wanted an interface that will also allow users to view what types of categorizations there are. Clicking on that category will show all related tweet collections.

Interface

For refinement 2, we have created an interface that shows the users with categorizations that allows users to view what categories are available. Also each category in the view is made clickable, and clicking the category will show the related collections.

The interface will first show the most general categorization, as shown in Figure 8 and Figure 9.



Sign in with Twitter

YourTwapperKeeper
archive your own tweets

Man-Made Disaster

Natural Disaster

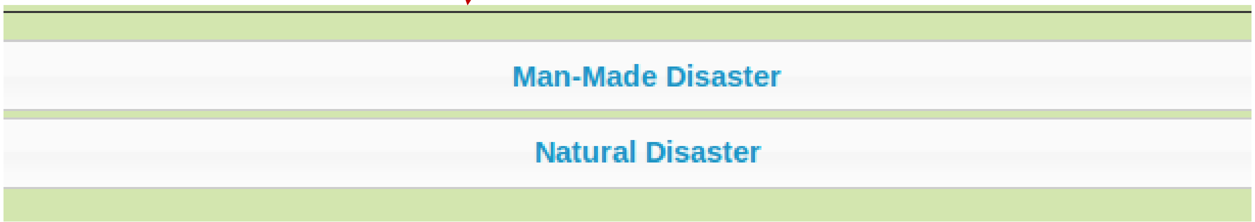
Show 10 entries

Search:

ARCHIVE ID	KEYWORD / HASHTAG	DESCRIPTION	TAGS	SCREEN NAME	COUNT	CREATE TIME
1	#hellegypt	Tweets for Egyptian revolution		dlrl	690	Sun, 07 Oct 2012 18:28:33 -0400
2	#aalasibya			dlrl	486	Sun, 07 Oct 2012 18:28:52 -0400
3	#blacksburg			dlrl	500	Sun, 07 Oct 2012 18:29:01 -0400
4	#jan25			dlrl	1465	Sun, 07 Oct 2012 18:29:08 -0400
5	#bahrain			dlrl	67	Sun, 07 Oct 2012 18:29:18 -0400
6	#yemen			dlrl	87	Sun, 07 Oct 2012 18:29:26 -0400
7	japan earthquake			dlrl	94	Sun, 07 Oct 2012 18:29:37 -0400
8	#syria			dlrl	1151	Sun, 07 Oct 2012 18:30:00 -0400
9	OccupyWallStreet			dlrl	68	Sun, 07 Oct 2012 18:31:26 -0400
10	#nrv	new river valley (blacksburg) related tweets		dlrl	313	Sun, 07 Oct 2012 18:32:04 -0400

Showing 1 to 10 of 667 entries

Figure 8 The new interface

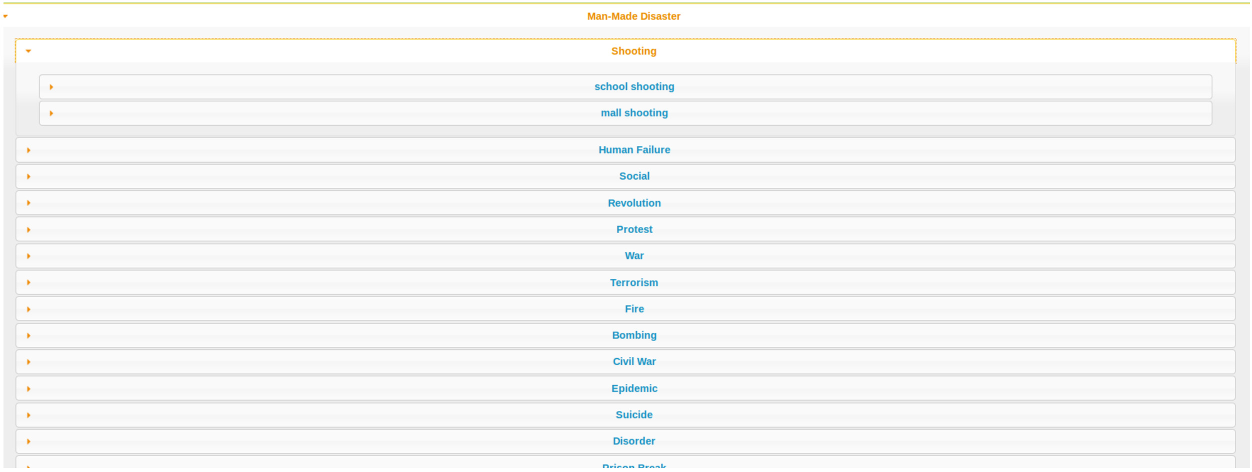


Man-Made Disaster

Natural Disaster

Figure 9 Most general categorization view in the Interface

When the category is clicked the sub-categories will be shown to the users. Using the same method, when the sub-category is clicked the more specific categorizations will appear.



- Man-Made Disaster
  - Shooting
    - school shooting
    - mall shooting
  - Human Failure
  - Social
  - Revolution
  - Protest
  - War
  - Terrorism
  - Fire
  - Bombing
  - Civil War
  - Epidemic
  - Suicide
  - Disorder
  - Prison Break

Figure 10 how it shows in the interface when categories are clicked

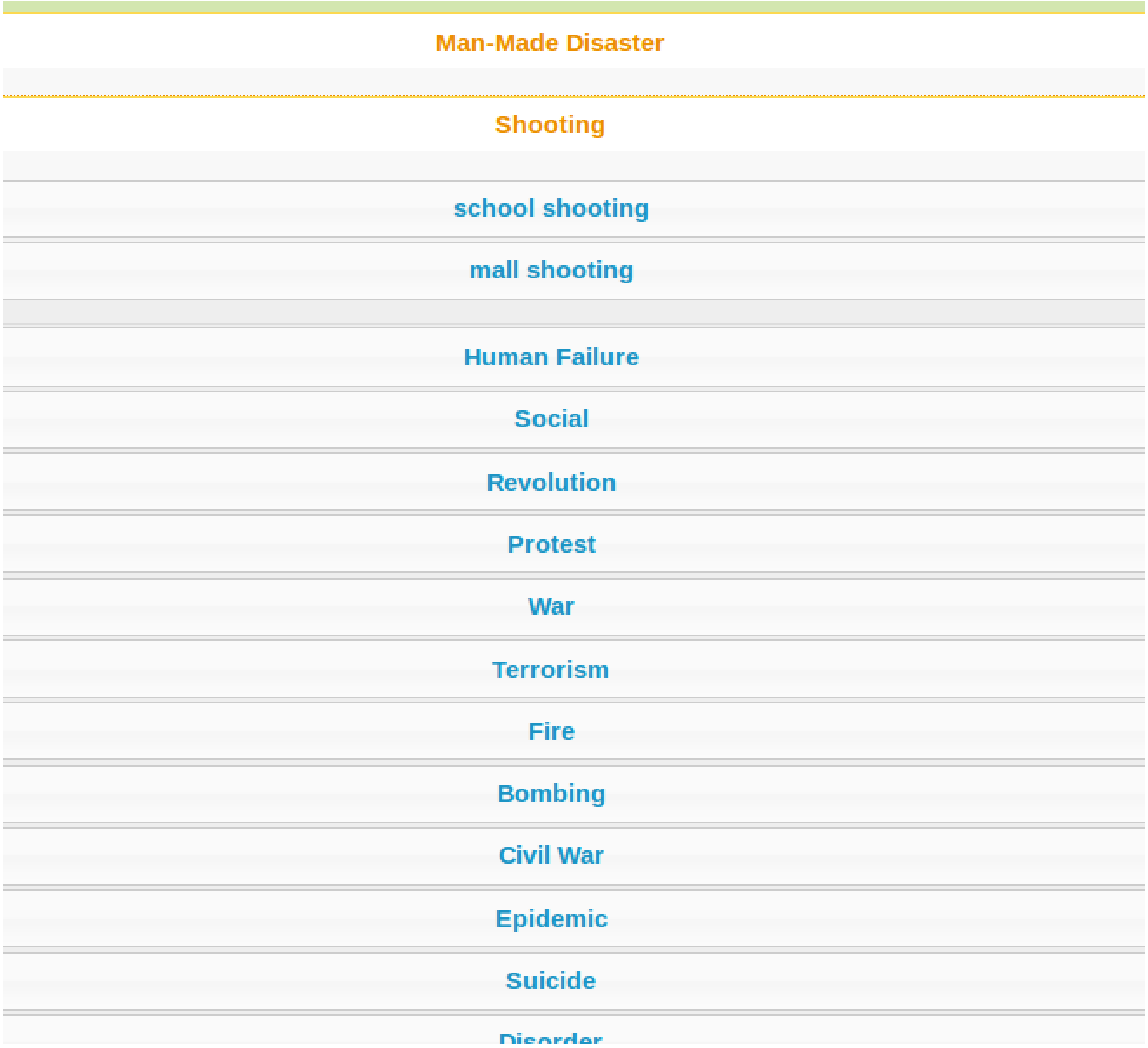
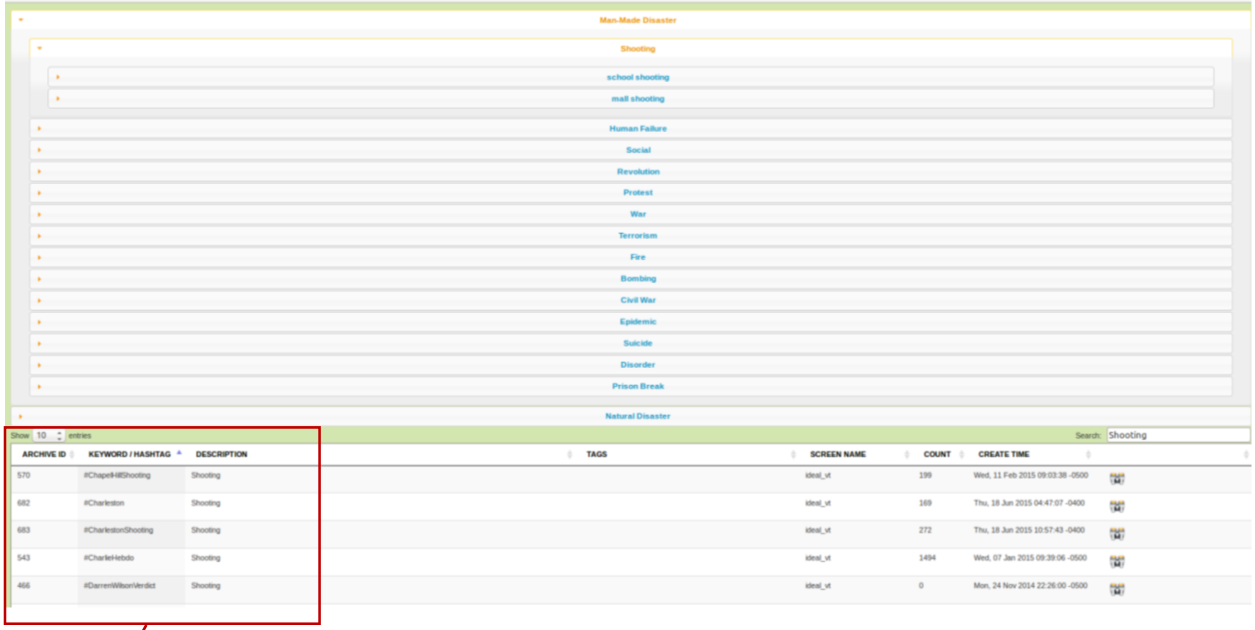


Figure 11 enlarged image of the interface

Also when each category is clicked it will list related collections. Figure 13 shows results of the table after clicking on the "Shooting" category.



Man-Made Disaster

Shooting

school shooting

mall shooting

Human Failure

Social

Revolution

Protest

War

Terrorism

Fire

Bombing

Civil War

Epidemic

Suicide

Disorder

Prison Break

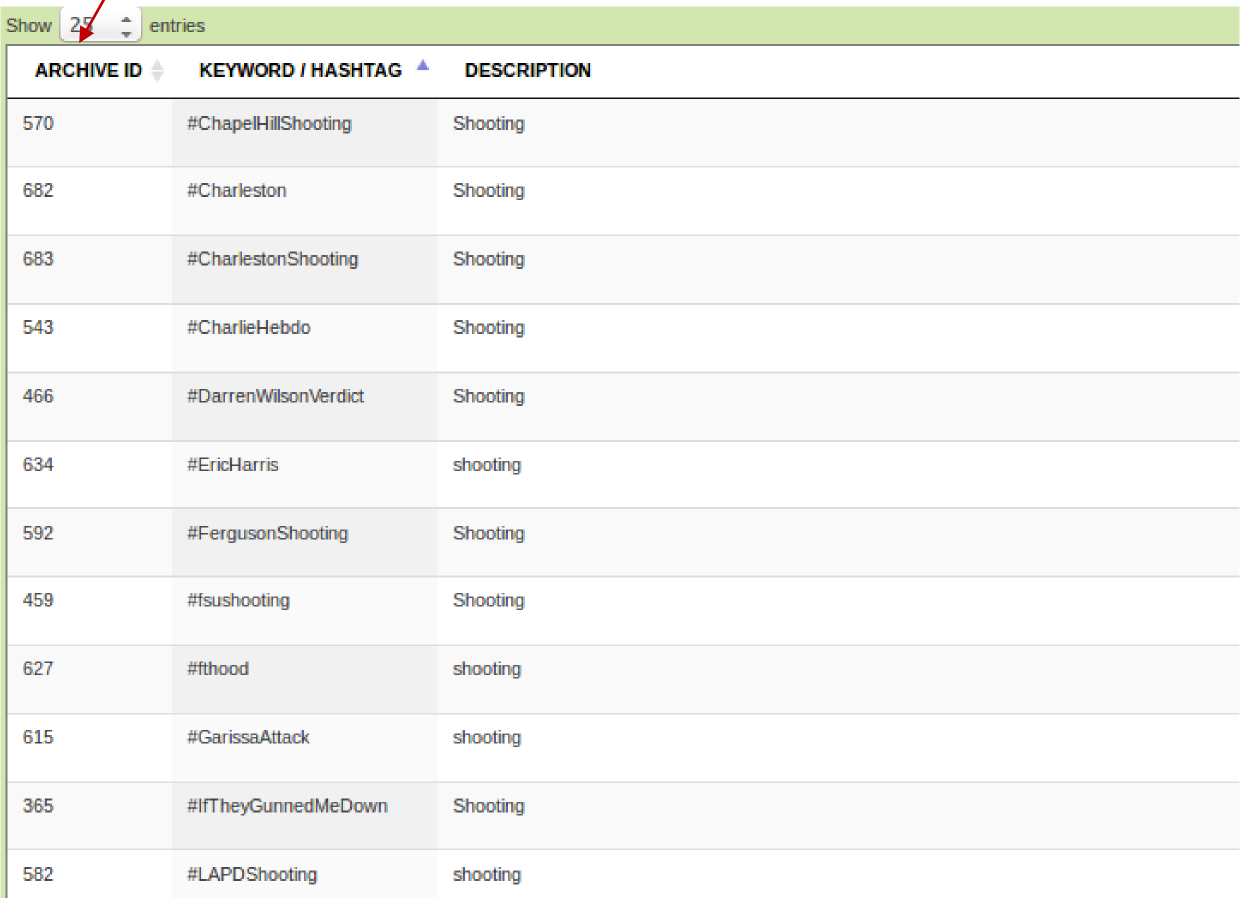
Natural Disaster

Show 10 entries

Search: Shooting

ARCHIVE ID	KEYWORD / HASHTAG	DESCRIPTION	TAGS	SCREEN NAME	COUNT	CREATE TIME	
570	#ChapelHillShooting	Shooting		ideal_vt	199	Wed, 11 Feb 2015 09:03:38 -0500	
682	#Charleston	Shooting		ideal_vt	169	Thu, 18 Jun 2015 04:47:07 -0400	
683	#CharlestonShooting	Shooting		ideal_vt	272	Thu, 18 Jun 2015 10:57:43 -0400	
543	#CharlieHebdo	Shooting		ideal_vt	1494	Wed, 07 Jan 2015 09:39:06 -0500	
466	#DarrenWilsonVerdict	Shooting		ideal_vt	0	Mon, 24 Nov 2014 22:26:00 -0500	

Figure 12 interface after "shooting" category is clicked



Show 25 entries

ARCHIVE ID	KEYWORD / HASHTAG	DESCRIPTION
570	#ChapelHillShooting	Shooting
682	#Charleston	Shooting
683	#CharlestonShooting	Shooting
543	#CharlieHebdo	Shooting
466	#DarrenWilsonVerdict	Shooting
634	#EricHarris	shooting
592	#FergusonShooting	Shooting
459	#fsushooting	Shooting
627	#fthood	shooting
615	#GarissaAttack	shooting
365	#IfTheyGunnedMeDown	Shooting
582	#LAPDShooting	shooting

Figure 13 Enlarged view of table after "shooting" category is clicked

Refinement 3

However, from the later meetings the client wanted to change the categorization to use tagging system, and with the taxonomy for each tag. The detailed taxonomy explanation is explained later. Also from these meetings the project was shifted to put more weights on creating Excel files with the mentioned categorization system, instead of on the interface.

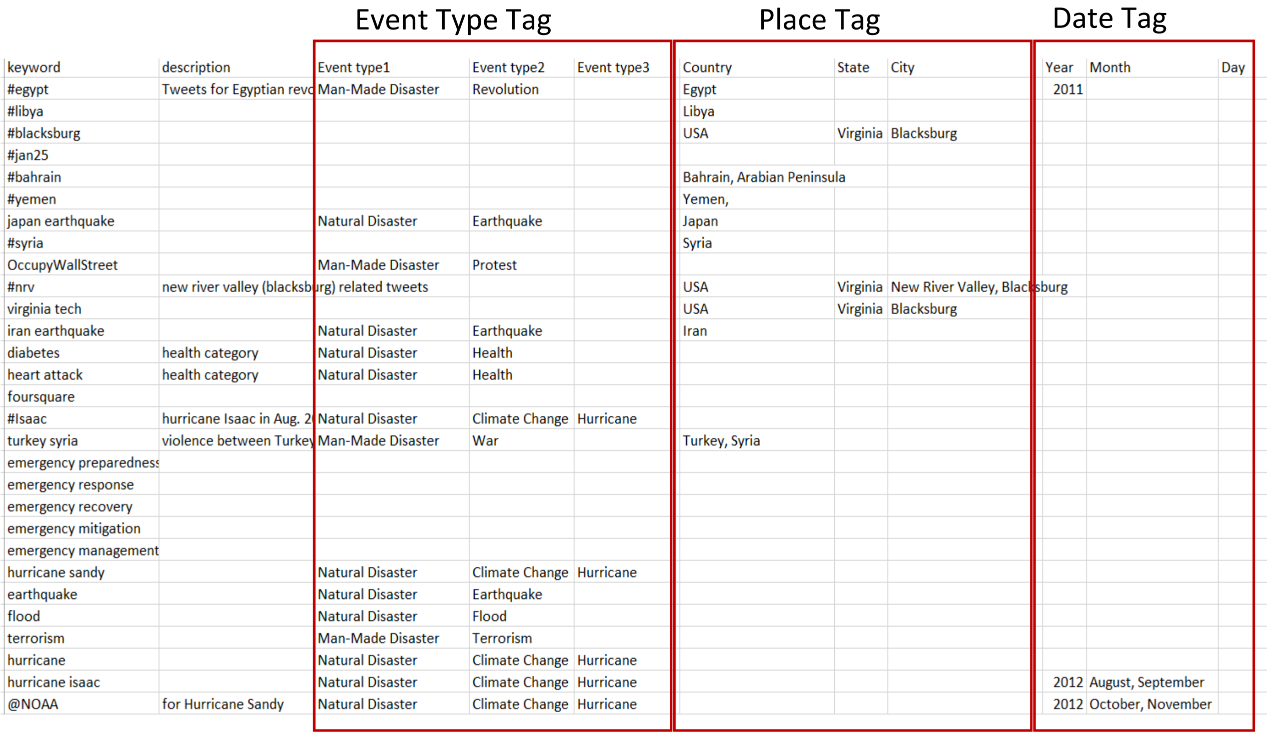
Categorization

Like in the previous tagging system introduced in refinement 1, there will still be three main tags, Event Type, Place, and Date. However, each tags will have its own topological system.

The Event Type tag uses a taxonomy scheme with three layers that is similar to the taxonomy scheme introduced in refinement 2. The first topological layer will be the general categories. The general categories will contain tags such as “Natural Event”, “Man-made Event”. The next layer will be the sub-categories with tags such as “Shooting”, “Human Failure”, and “Climate Change”. The last layer will be the specific categories that contain tags such as “School Shooting”, “Mall Shooting”, “Hurricane”, “Storm”, and so on.

The Place tag also contains three topological orders. The first field is country. If the event occurred across multiple countries, all the countries will be added in this field. The second field is state. Same as the country if the event occurred across multiple states, all the states will be included as tags in this field. Also if the country does not have states, then this field will be ignored. The last field is city. Also if the event occurred or happened among or across multiple cities, all cities will be added to this field. If the event happened in the entire country or state, this field may be ignored.

The last tag is Date. Date tag will contain three fields, year, month, and day. If the date tag is not required, for collections such as “Blacksburg” the tag may be left empty. If the event occurred in many days, months or years, all the appropriate years and months and days will be added to each field.



		Event Type Tag			Place Tag			Date Tag		
keyword	description	Event type1	Event type2	Event type3	Country	State	City	Year	Month	Day
#egypt	Tweets for Egyptian revo	Man-Made Disaster	Revolution		Egypt			2011		
#libya					Libya					
#blacksburg					USA	Virginia	Blacksburg			
#jan25										
#bahrain					Bahrain, Arabian Peninsula					
#yemen					Yemen,					
japan earthquake		Natural Disaster	Earthquake		Japan					
#syria					Syria					
OccupyWallStreet		Man-Made Disaster	Protest							
#nrv	new river valley (blacksburg) related tweets				USA	Virginia	New River Valley, Blacksburg			
virginia tech					USA	Virginia	Blacksburg			
iran earthquake		Natural Disaster	Earthquake		Iran					
diabetes	health category	Natural Disaster	Health							
heart attack	health category	Natural Disaster	Health							
foursquare										
#Isaac	hurricane Isaac in Aug. 2	Natural Disaster	Climate Change	Hurricane						
turkey syria	violence between Turkey	Man-Made Disaster	War		Turkey, Syria					
emergency preparedness										
emergency response										
emergency recovery										
emergency mitigation										
emergency management										
hurricane sandy		Natural Disaster	Climate Change	Hurricane						
earthquake		Natural Disaster	Earthquake							
flood		Natural Disaster	Flood							
terrorism		Man-Made Disaster	Terrorism							
hurricane		Natural Disaster	Climate Change	Hurricane						
hurricane isaac		Natural Disaster	Climate Change	Hurricane				2012	August, September	
@NOAA	for Hurricane Sandy	Natural Disaster	Climate Change	Hurricane				2012	October, November	

Figure 14 Excel file with new tag and taxonomy system

After creating an Excel file with the tagging system, this file will be moved into the database, and using search platforms such as SOLR, the interface and search function will be added in a different project.

Testing

Tested Functionalities

Functional testing for the following modules are in Scope of Testing

- searching
- column ordering
- category view expanding
- searching related tweet collections based on the clicked category from category view

Items Not Tested

Since the goal of the project has been altered, the integration of the new categorization database to the interface was not performed. However, when the interface testing was performed we used the Archive DB from the IDEAL project.

Types of Testing Performed

1. Smoke Testing
2. Integration Testing

1. Smoke Testing

Whenever new modules or functionalities are created, we made sure the major functionality is working fine. We made sure all the functionalities work correctly from the local Apache HTTP server before any database was integrated with the functionality. During this phase, we tested searching and, column ordering.

2. Integration Testing

We tested that the interface was working using the database "Archive DB" from http://hadoop.dlib.vt.edu/.

During this testing, we recognized that the database was not correctly showing on the page. We fixed the issue by fixing the CSS file that is imported to index.php file.

Results

From the tests, we found that the table works as intended. The search finds the keyword from the table. The column ordering orders the columns by its spelling or numbers. Also, the "Archive DB" was integrated correctly and shows on the table.

Lessons Learned

Throughout the semester long project, we have learned a great deal.

First I learned the importance of understanding the goal of the project. Although from the first meeting with the client we learned about the project, we were more focused on the tasks that are given to us than understanding the users' needs and the purpose of the project. This resulted in several refinements.

Another lesson learned was the requirements of constant communication. In the beginning, before we had any knowledge about the tools, we took many hours to try to use and fix the problems that occurred with the tools. However when we met our client Sun Shin Lee, it was easily fixed. If we had contacted him earlier about the problem, we would have saved much time.

Implementation Schedule

Date	Description
February 26	yourTwapperKeeper tool installation to our local machine in order to learn its functionalities, and user interface.
March 4	Preparation of several categorization schemes to present to the IDEAL team. If there is a best fit scheme, the scheme will be used in the implementation
March 11	Meeting with IDEAL team to show the prepared categorization and get feed backs.
March 20	Implement Database using the new categorization
March 27	Implement Graphical User Interface to allow users to search and order by column
April 3	Implement refinement 1 and Create prototype
April 15	Implement refinement 2 and Test
April 19	Implement excel file with categorization from refinement 3

Future work

For the future work of the project, there are several advances for categorization and GUI that could be added.

Categorization:

- It could be applied to SOLR for search and interfaces.
- More topological layers could be added to each tag to be more specific.
    - more detailed event types
    - more specific region
    - time of the day

GUI:

- Instead of SOLR, the implemented GUI page could be used to display the collections
- Columns for each tag could be added.

Acknowledgements

I would like to express gratitude to **Dr. Edward Fox** for all the help, guidance, encouragement, and suggestions. His moral support helped me get through the project well.

I am grateful for **Sun Shin Lee** for taking time and providing guidance through the project. I appreciate very much for all the help he tried to provide when there were problems.

I would also like to thank the **IDEAL team** for valuable opinions and feedbacks.

Thanks Also go to NSF for support of grant IIS – 1319578: III: Small: Integrated Digital Event Archiving and Library (IDEAL)

References

Sunshin Lee. "DLRL Hadoop Cluster." 2016. Web. 03 Apr. 2016. http://hadoop.dlib.vt.edu/.

IDEAL project team. *Collections page on Events Archiving website*. 2016. Web. 03 Apr. 2016. http://www.eventsarchive.org/node/18.

IDEAL project team. "IDEAL Project: Tweet Archive DB." *Your Twapper Keeper*. 2016. Web. 03 Apr. 2016. http://hadoop.dlib.vt.edu:82/twitter/.

IDEAL project team. "Integrated Digital Event Archiving and Library (IDEAL) Annual Report." *2014-07-09.* Web. http://vtechworks.lib.vt.edu/handle/10919/52853.