Using *Omeka*'s Two Distant Reading Plugins to Explore the Language of Death and Mourning in the September 11 Digital Archive

Alyssa Fahringer, December 2017

Summary

The Ngram and Text Analysis plugins offer users the ability to use computational methods on a large corpus of user-generated items and perceive patterns contained within the text of those items that would remain difficult to see when reading one at a time. An ngram is a sequence of words, phrases, or letters found across a body of texts or collections. *Omeka*'s Ngram Plugin enables users to generate uni, bi, and trigrams of words and phrases. Then, the plugin calculates the frequencies of the ngrams produced. Once ngrams are created, the Text Analysis Plugin produces lists of entities, keywords, categories, and concepts using Watson Natural Language Understanding (NLU). NLU is a field of natural language processing that can examine structured and unstructured data to extract metadata and calculate sentiment and emotion counts for that data. Together, these plugins are useful for understanding broad trends in language use throughout a user's collections, for informing a user's understanding of an *Omeka* collection's data, and for generating meaningful research questions.

The goal of this case study is to explore how the Ngram and Text Analysis plugins can be used as exploratory tools with the robust collection of user-generated material in the September 11 Digital Archive to explore how people talk about death--in what words or terms--and how the discourse of death and mourning changes over time.

Using the Text Analysis and Ngram plugins I found the following:

- Visitors to the Smithsonian's National Museum of American History "Bearing Witness to History" who submitted online reflections and reactions (available in the September 11 Digital Archive) frequently referenced or reflected on Pearl Harbor when discussing or remembering the events of September 11.
- When discussing or mentioning people, visitors frequently described them as innocent, heroic, courageous, or as victims.
- The category--which, in this instance, is a term not specifically used within the corpus but is one that describes the collection of items in a broader sense--most appropriate for this corpus is society, unrest and war, which demonstrates that NLU is able to appropriately extrapolate a larger concept not overtly mentioned within the analyzed corpus.
- Visitors spoke of the idea that people who died on September 11 should be remembered, and that there is a collective responsibility for all Americans to do so.

These reflections and analyses offer additional support that complements the current historiography and scholarship on September 11.

Ultimately, this case study demonstrates that *Omeka*'s distant reading plugins can effectively be used as an exploratory research tool in the first step of a larger process of understanding and gaining insight into a large collection of text-based items, and can be used as a starting point for asking scholarly questions of that collection. In order to be able to more fully answer my initial research question, I would need to perform a close reading of the items within the selected collection.

Using the Text Analysis and Ngram Plugins

Step 1: Choosing a collection

After reviewing the collections and keeping in mind the limitations of the plugins--they can only read plain text, not PDFs or Word documents, for example--I determined that the collection I should focus on is the "September 11: Bearing Witness to History" Stories Submitted Online collection. This is a collection of 7421 items that were submitted by visitors to the Smithsonian National Museum of American History's "Bearing Witness to History" online exhibit. Visitors were prompted to reflect on how they witnessed history on September 11, how their lives have changed since, and what they think should be remembered about that day.

Step 2: Configuring the Ngram Plugin

On the Ngram Plugin configuration page I configured the text element to NMAH Story: Remembered because that text element correlates to the question "What do you think should be remembered about September 11th?" that visitors were asked in the Bearing Witness to History exhibit. This was the most appropriate text element for me to select because I am interested in exploring the ways people thought this event should be remembered--specifically, the ways in which people described an event that produced mass death, and the words they used to describe those who died.

Step 3: Creating a corpus

When adding a corpus I did not specify a specific search term, instead electing to search the entirety of that particular text element. I did not in any way want to narrow the results that the plugin would return by searching for a specific term. I selected Date Entered (911DA Item) as the sequence element, date by year as the sequence type, and 2001-2007 as the sequence range. I selected that specific range because visitors only contributed to the collection during that span of time. I named this corpus September 11: Bearing Witness Online Submissions - remembered. Out of a pool of 7421 items, the corpus was 7420 items--one item was considered out of range. I validated the 7420 items.

Step 4: Using the Text Analysis Plugin

I clicked the green add a corpus button, and selected September 11: Bearing Witness Online Submissions - remembered. I did not limit the features to analyze, so entities, keywords, categories, and concepts were all checked. There is an option to only calculate the item cost and to not generate any textual analysis.

The work performed within the Text Analysis Plugin is done via IBM's Watson Developer Cloud Natural Language Understanding, which analyzes text within a corpus and generates concepts,

entities, keywords, categories, and sentiment using natural language understanding. To use this plugin, you will need to create an account on IBM Bluemix and then enter in your username and password on the plugin configuration page. Check their website for information on their pricing plans and how they calculate the item cost. Typically, they charge $0.003 per item. For my corpus, the projected item cost is approximately 615, which means that I would be charged $1.85 to analyze the entire corpus.

As the computational portion of the text analysis is done via IBM, the only information users can glean about their process is that which is found in their documentation. IBM is very much like a "black box"--they do not reveal the algorithms or computational methods used to produce the results users see on the plugin page. At times this can be problematic, as we will see below.

Functionality of the Text Analysis Plugin: Overview

Users can only view the results of the NLU analysis based on the sequence range of the corpus. For my corpus, I had to select a specific year to see results. The plugin divides the results into specific tabs--overview, entities, keywords, categories, and concepts. Within the overview tab, the user is presented with the year that the results are from, as well as links to the previous and next years. Users have to return to this tab to toggle between years, as the plugin is not designed to compare results between different years or display change over time. The results for each year are exportable as a .JSON file.

Functionality of the Text Analysis Plugin: Entities

The entities tab lists place names, people, events, and organizations specifically mentioned within the text of the corpus and then provides a type, emotion, sentiment, count, and relevance for each of those named entities. For a complete list of entities recognized by IBM, see their documentation, including this list of entity types and subtypes. Entity refers to the specific named place, person, event, or organization, while type lists the type of entity that is being referenced, such as location, facility, person, organization, or company. The emotion column provides sentiment analysis of the entities named within the corpus. The emotions that are quantitatively analyzed are pre-selected by IBM, and include sadness, joy, fear, disgust, and anger. These emotions are scored from a range of 0 to 1, with 0 meaning that the text does not convey that emotion while 1 means the text definitely conveys that emotion. The sentiment of the entity is also given a numeric value that ranges from -1 to 1. Negative scores demonstrate negative sentiment, and positive scores indicate positive sentiment. The count is the number of times that entity is named within the text, and the relevance of that entity is ranged from 0 to 1, with 0 meaning that entity is not relevant, while a score of 1 means that entity is highly relevant. There is no clear documentation on how exactly IBM calculates these results.

The entities found for this corpus included the geographic places in which the main events of the day happened--World Trade Center, Manhattan, New York, Pentagon, Washington, Pennsylvania--as well as the key political actors, including George Bush and Osama Bin Laden. There were several interesting results: in 2003, Iwo-Jima was a named entity; pain and murder were other frequently named entities for a number of years; and in 2003, there were several

named entities relating to the Irish Republican Army and Ireland. Because I am interested in death and memory, I focused on the entity Pearl Harbor.

Year	Entity	Type	Emotion	Sentiment	Count	Relevance
2002	Pearl Harbor	Facility - Location - GeographicFeature - BodyOfWater	Sadness: 0.133674 Joy: 0.269886 Fear: 0.084554 Disgust: 0.0864 Anger: 0.066609	0	1	0.294806
2003	Pearl Harbor	GeographicFeature - Location - BodyofWater	Sadness: 0.28914 Joy: 0.273405 Fear: 0.080749 Disgust: 0.053341 Anger: 0.045503	0.477931	1	0.345147
2004	Pearl Harbor	GeographicFeature - Location - BodyofWater	Sadness: 0.185225 Joy: 0.044854 Fear: 0.354236 Disgust: 0.144329 Anger: 0.357483	0	2	0.326017
2005	Pearl Harbor	Facility - Location - GeographicFeature - BodyOfWater	Sadness: 0.3475 Joy: 0.099293 Fear: 0.299677 Disgust: 0.137908 Anger: 0.161431	0.510529	1	0.287748
2005	Pearl Harbor	Location	Sadness: 0.3475 Joy: 0.099293 Fear: 0.299677 Disgust: 0.137908 Anger: 0.161431	-0.520308	1	0.284391
2007	Pearl Harbor	Location	Sadness: 0.120447 Joy: 0.193406 Fear: 0.31384 Disgust: 0.311847 Anger: 0.294161	0	1	0.814814

Pearl Harbor was mentioned only once or twice for each year--with the exception of 2006, in which it seems that Pearl Harbor was not mentioned at all--based on the count column. The year in which Pearl Harbor was most relevant was 2007, and the year in which it was least relevant was 2005. These results also demonstrate a shortcoming of this plugin--for 2005, Pearl Harbor was identified as both a facility and a location, and in other years it was also identified as a geographic feature. It is evident that IBM could not determine that both mentions of Pearl Harbor in 2005 are, presumably, referencing the attack on Pearl Harbor in 1941. Surprisingly, IBM determined that the sentiment in regard to Pearl Harbor were mostly positive or neutral. For the first two years of this corpus, the predominating emotion--by which I mean the emotion for a specific year that scored the closest to 1--were joy in 2002 (0.269886) and 2003 (0.28914). Both of those results demonstrate that the text does not *definitely convey that emotion*, but it is the highest ranking emotion based on their analysis. For 2004, 2005, and 2007, the predominating emotions were fear and anger, sadness, and fear, respectively. However, none of these

emotions scored a 0.4 or higher, so those results are inconclusive as to the emotions conveyed in the named entities. Only in 2005, for the Pearl Harbor location entity, was Pearl Harbor associated with negative sentiment. Based solely on these results, it does not seem as though there was any significant change over time when visitors referenced Pearl Harbor when describing the ways they think that September 11 should be remembered.

After viewing these results for the named entity Pearl Harbor, I had several questions that would require a close reading of the items within the collection, such as: Was Pearl Harbor only mentioned once or twice in each year, and was Pearl Harbor not mentioned at all in 2006? How and in what ways were people using or referencing Pearl Harbor when discussing the events of September 11? Does it seem that visitors were referencing Pearl Harbor with positive, not negative, sentiment?

Functionality of the Text Analysis Plugin: Keywords

The keywords tab identifies the important keywords in the text. This tab also includes a quantitative analysis of emotion, sentiment, and relevance. Like the entity tab, the analyzed emotions are sadness, joy, fear, disgust, and anger. The scores range from 0 to 1, with 0 meaning the text does not convey the emotion and 1 meaning the text definitely carries the emotion. Similarly, sentiment is scored from -1 to 1, with negative scores for keywords that convey negative sentiments, and positive scores for keywords that convey positive sentiment. Relevance is also scored on a scale from 0 to 1, with 0 being no relevant to the corpus and 1 meaning it is highly relevant. This tab does not include a count for the number of times a keyword has been detected within the corpus.

It is important to note that the keywords analysis distinguishes between keywords that are essentially the same thing, such as firefighters and firemen, and their quantitative scores for emotion, sentiment, and relevance are different--in some cases, vastly different (the corpus demonstrates a measure of 0.13 of disgust at firefighters, while for firemen disgust measures a 0.43), and in other cases, fairly similar (the relevance of firefighters is 0.58, while for firemen it is 0.55).

I chose the following keywords to analyze because they are, presumably, descriptions of people who died in the attacks, experienced the attacks firsthand, or took part in rescue efforts on September 11. Initial results demonstrate that commonly found keywords across the corpus include adjectives and their derivative forms such as innocence, heroism, bravery, and courage. Resilience and human spirit provide insight into how people felt, experienced, and described the immediate aftermath of September 11.

Year	Keyword	Emotion	Sentiment	Relevance
2002	Innocent people	Sadness: 0.732546 Joy: 0.058496 Fear: 0.124857 Disgust: 0.57707 Anger: 0.145959	-0.0519344	0.740151

2003	Innocent people	Sadness: 0.733687 Joy: 0.075683 Fear: 0.134673 Disgust: 0.452294 Anger: 0.138388	-0.0889349	0.742744
2004	Innocent people	Sadness: 0.770779 Joy: 0.016252 Fear: 0.144666 Disgust: 0.459848 Anger: 0.159	-0.682278	0.568773
2005	Innocent people	Sadness: 0.743672 Joy: 0.044277 Fear: 0.109687 Disgust: 0.499401 Anger: 0.511203	0.20324	0.637829
2006	Innocent people	Sadness: 0.722597 Joy: 0.03518 Fear: 0.086218 Disgust: 0.484707 Anger: 0.520946	-0.542099	0.633805
2002	Heroes	Sadness: 0.673566 Joy: 0.557382 Fear: 0.071303 Disgust: 0.431435 Anger: 0.059664	0.0140902	0.56763
2004	Heroes	Sadness: 0.650439 Joy: 0.550626 Fear: 0.068995 Disgust: 0.473761 Anger: 0.094164	0.0767499	0.468169
2004	Heroic people	Sadness: 0.354535 Joy: 0.433587 Fear: 0.107533 Disgust: 0.034002 Anger: 0.162639	-0.374939	0.440568
2002	Courageous people	Sadness: 0.306031 Joy: 0.223203 Fear: 0.07928 Disgust: 0.316657 Anger: 0.110639	0.842254	0.547792
2006	Couragious (sic) people	Sadness: 0.215821 Joy: 0.257432 Fear: 0.084548 Disgust: 0.190669 Anger: 0.043764	-0.489516	0.571103
2002	Brave people	Sadness: 0.258485 Joy: 0.473523 Fear: 0.049476 Disgust: 0.282559 Anger: 0.044548	0.817203	0.540521

2004	Brave people	Sadness: 0.500175 Joy: 0.223087 Fear: 0.073039 Disgust: 0.346415 Anger: 0.016836	0.765227	0.47467
2005	Brave people	Sadness: 0.387235 Joy: 0.476964 Fear: 0.052619 Disgust: 0.150666 Anger: 0.018825	-0.732673	0.567374
2006	Brave people	Sadness: 0.337607 Joy: 0.076824 Fear: 0.083207 Disgust: 0.633405 Anger: 0.119117	0.18228	0.59206
2003	Good innocent people	Sadness: 0.750238 Joy: 0.132718 Fear: 0.049889 Disgust: 0.109968 Anger: 0.075005	0	0.596469
2005	Innocent victims	Sadness: 0.62327 Joy: 0.031394 Fear: 0.028309 Disgust: 0.496707 Anger: 0.045745	0.600264	0.569529
2006	Innocent heros (sic)	Sadness: 0.769088 Joy: 0.067199 Fear: 0.016195 Disgust: 0.126869 Anger: 0.125124	-0.649557	0.543636

The keyword innocent people is used from 2002 through 2006, and for each year it is associated with negative sentiment with the exception of 2005. This keyword declines in relevance during those years, but even in 2006 it maintained a relatively high relevance at 0.633805. This suggests that even five years after September 11, innocent people was still a commonly used keyword to describe those who died.

The keywords heroes and heroic people were found in 2002 and 2004. Heroes is associated with a slightly positive sentiment, while heroic people is associated with negative sentiment. In 2002 and 2004, the emotions most strongly associated with heroes are sadness and joy, while the emotion most strongly associated with heroic people is joy. These results are puzzling, as they suggest that while an overall sentiment for a keyword such as heroic people is negative, the emotion most strongly found with that keyword is joy.

Other keywords of interest are courageous people and brave people. Both keywords are fairly relevant to the corpus. In 2002, courageous people had a positive sentiment, while in 2006 they had a negative sentiment. In neither year were any emotions particularly prevalent. In 2002, 2004, and 2006 brave people had a positive sentiment, while in 2005 it had a negative

sentiment. The predominant emotion associated with the keyword brave people varied widely: In 2002, the predominant emotion was joy, in 2004 it was sadness, in 2005 it was joy, and in 2006 it was disgust.

It is clear that in order for these results to be meaningful, a close reading of the individual items within the collection is required.

Functionality of the Text Analysis Plugin: Categories

The categories tab categorizes the words of the corpus using a classification hierarchy determined by IBM. You can view a complete listing of their categories hierarchy on their website. The plugin gives each category--termed label--a score ranging from 0 to 1, with 1 indicating confidence in the categorization, and 0 indication hardly any confidence in the categorization. For the September 11: Bearing Witness Online Submissions - remembered corpus, the following results were generated:

Label	2002	2003	2004	2005	2006	2007
Society/unrest and war	0.521325	0.481436	0.530454	0.493537	0.602129	0.344506
Family and parenting	0.42752		0.418507	0.522187		
Law, govt and politics/espionage and intelligence/terrorism	0.401234	0.446089	0.442996			
Law, govt and politics/law enforcement/police		0.401747			0.337358	
Family and parenting/children				0.347885	0.335897	
Health and fitness/disorders/mental disorder/panic and anxiety						0.503465
Science/social science/history						0.495969

The category society/unrest and war is the most predominant in the entire corpus, as it is included in the results for each year of study. The next most commonly found categories are family and parenting, and law, govt and politics/espionage and intelligence/terrorism. None of these results are particularly surprising or unexpected--it is not unusual for results to focus on unrest and war, government and politics, terrorism, and panic and anxiety with this specific corpus--and again, the results are much less confident than not in these categorizations.

I was left with the following questions that would require a close reading of the items within the collection: Family and parenting, or family and parenting/children, is a commonly found category within this corpus--how frequently are parents and/or children specifically mentioned, and in what contexts? What do the results for 2007 specifically look like, and why was health and fitness/disorders/mental disorder/panic and anxiety returned as a category for that year?

Functionality of the Text Analysis Plugin: Concepts

The concepts tab identifies concepts that might not be specifically referenced within the corpus. Each concept is given a relevance score on a range from 0 to 1, with 0 meaning that concept is not relevant, and 1 meaning it is highly relevant. Some of the concepts are only understandable after having viewed the results on the keywords tab. For example, race and Africa are two concepts that were generated for 2002, which is understandable given that some of the keywords identified included black people and brown people. For other years, the concepts do not make sense given my prior knowledge of the corpus. For example, in 2005, the concepts include 2008 albums, 1995 albums, English-language films, and Aerosmith.

Concept	2002	2003	2004	2005	2006	2007
September 11 attacks	0.958201	0.957608	0.974361	0.98077	0.945388	
World Trade Center	0.672397	0.759123	0.51063	0.469902	0.860791	
English-language films	0.512286		0.497657	0.541227	0.709231	
United States	0.476602	0.548944				
White people	0.442029					
Other People's Lives	0.42471			0.684684		
Race	0.393085					
Africa	0.380084					
1995 albums		0.517959		0.509833		
Twin Towers		0.50364				
2008 albums		0.490047	0.407696	0.531898	0.800602	
Attack!		0.477476				0.839117
Twin towers		0.453003				
September 11			0.537067	0.61011		
Remembrance Day			0.465781			
Veterans Day			0.404729			
The Nation			0.382261			
Aerosmith				0.471054	0.728064	
American way					0.749609	
Personal life					0.66996	
United Airlines Flight 93					0.65149	

Attack						0.92115
2001 albums						0.896501
Attack on Pearl Harbor						0.778605
2002 singles						0.709799
2005 albums						0.707526
Harbor						0.678074
World War II						0.674963

Based on these results, it appears that the concepts tab might be the least useful of the four features for this particular corpus and this particular research question. Unsurprisingly, September 11 attacks and World Trade Center are the most prevalent concepts identified throughout the corpus, and other anticipated concepts include United Airlines Flight 93, United States, attack, and the nation. Like the keywords tab, some results that are identified as separate concepts are essentially the same thing, such as September 11 attacks and September 11, or World Trade Center and Twin Towers. Other concepts actually are the same thing, like Twin towers and Twin Towers.

Step 5: Using the Ngram Plugin

The Ngram Plugin page contains a table of the existing corpora within your *Omeka* site. To access the Ngram functionality, users have to navigate to the specific corpus they wish to analyze by clicking on the title of the corpus. This directs the user to a page that provides an overview of the parameters of that corpus that were set when the corpus was generated: whether or not it is public, the search query, sequence element, sequence type, and sequence range. This page is where users can edit or delete their corpus, view what text element the corpus was configured with, and the item count of both the pool and corpus. This is also where users can generate unigrams, bigrams, and trigrams, and where they can navigate to the ngram search and ngram frequencies.

Functionality of the Ngram Plugin: Ngram search

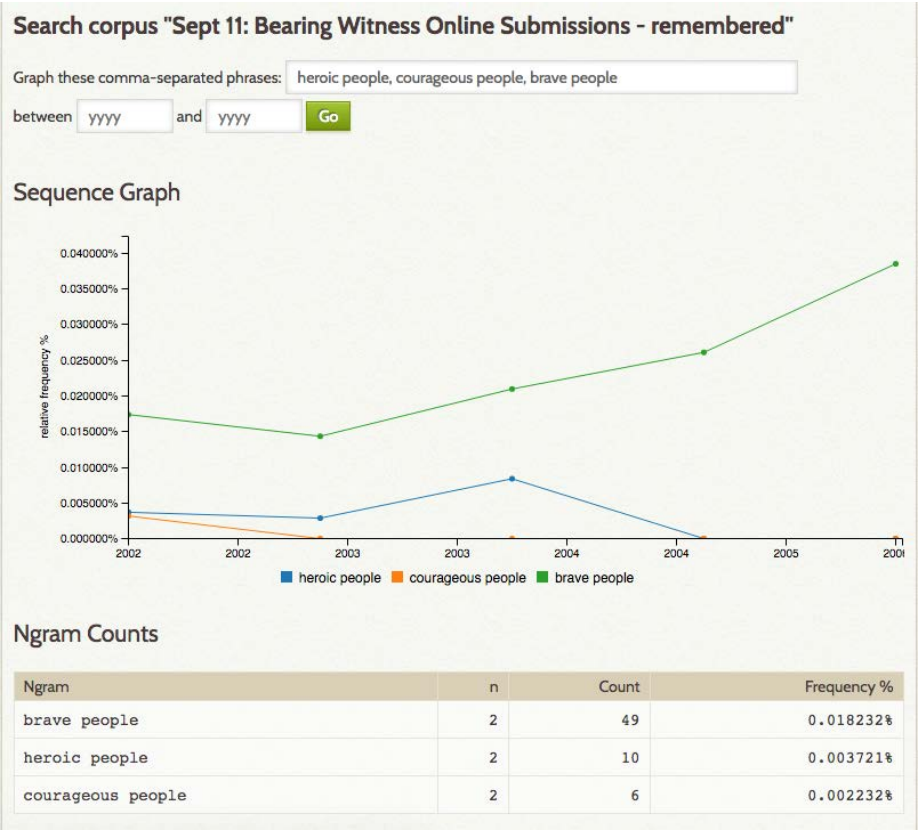
Within the ngram search feature, users are able to graph words or phrases in order to gain insight into how that ngram(s) is represented throughout the corpus. Users can graph numerous ngrams at one time, allowing them to compare results of various ngrams throughout the corpus. Users can also specify which dates they wish to view, but if left blank, results will appear for each year of the sequence range, which was configured when the corpus was created.

After inputting a phrase and the sequence range, users are presented with a sequence graph, ngram counts, and total ngram counts. The x-axis of the sequence graph includes the sequence range and the y-axis is a percentage of the relative frequency. Ngram counts provide the count and percentage of the frequency of the specific phrase(s) throughout the corpus, while total

ngram counts calculates the total count and total unique count of the entire corpus, not just for the specified ngram(s).

Based on the results generated by the Text Analysis Plugin, I specifically chose the following ngrams to use in the ngram search:

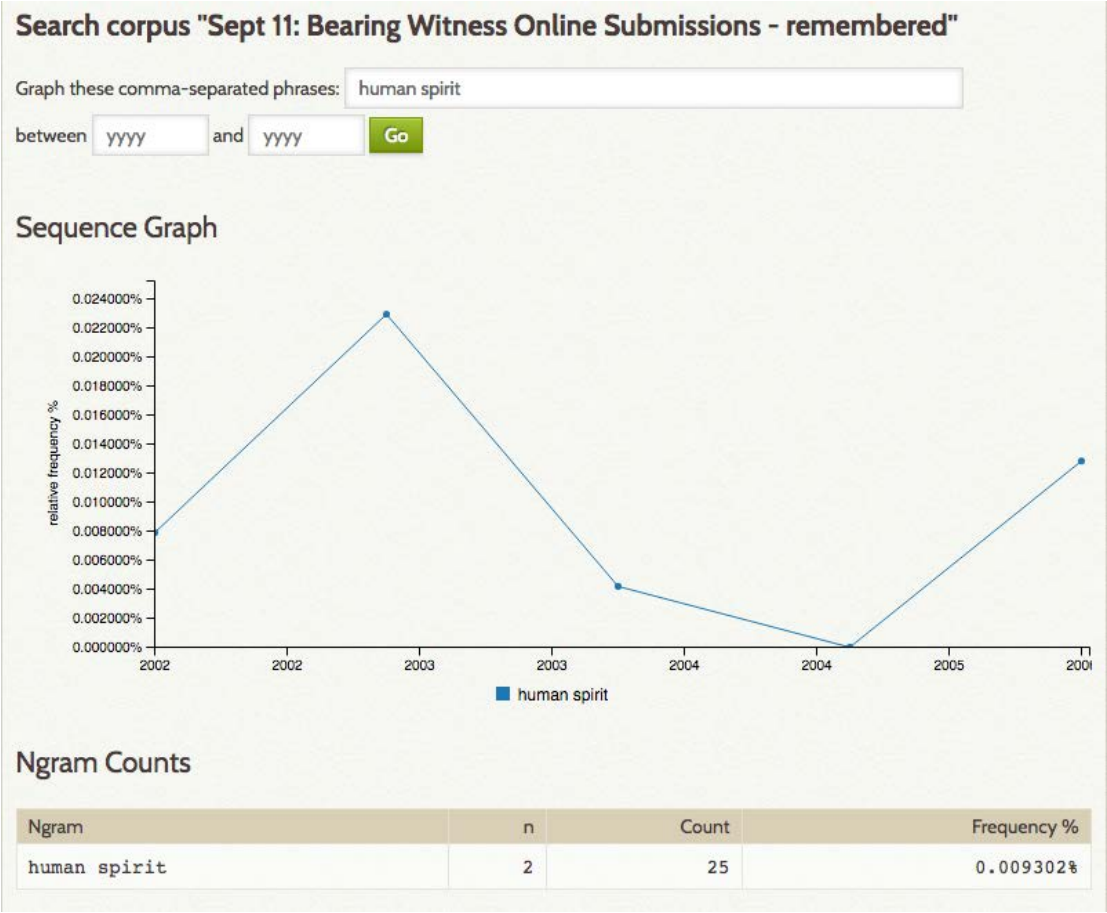
Heroic people, courageous people, brave people



Ngram	n	Count	Frequency %
brave people	2	49	0.018232%
heroic people	2	10	0.003721%
courageous people	2	6	0.002232%

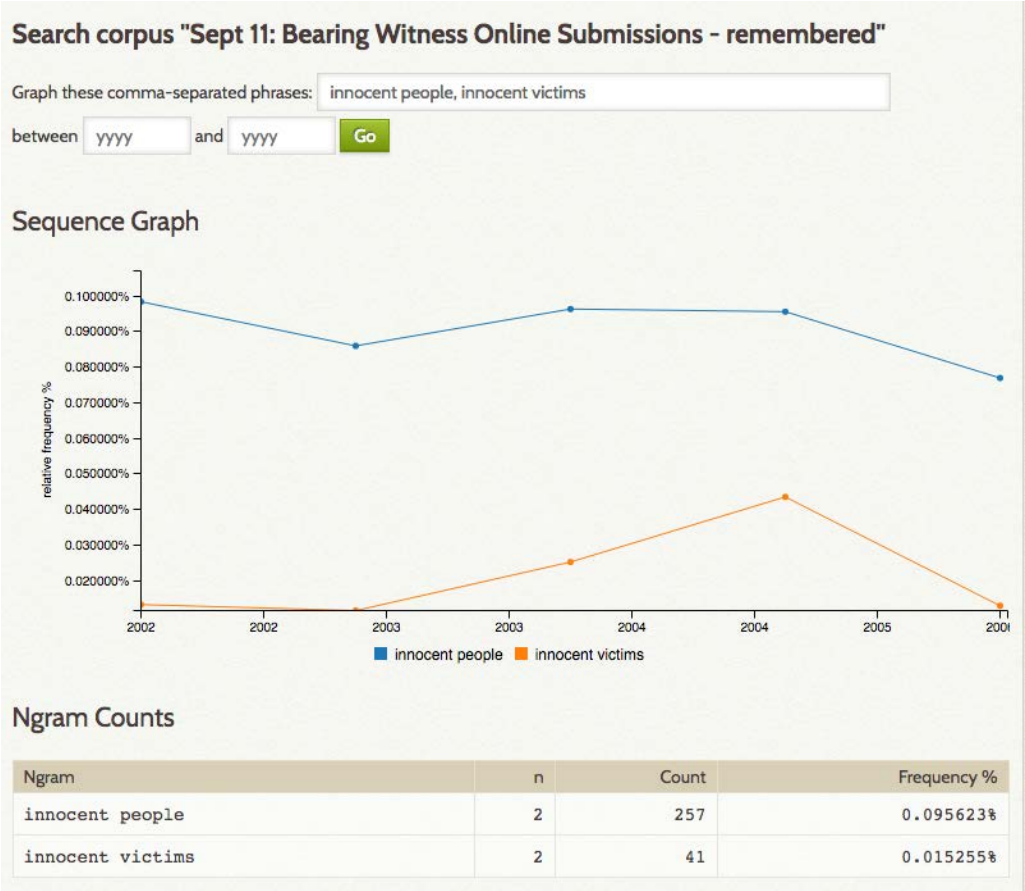
These results demonstrate that visitors used the word brave to describe those who experienced the events of September 11 more so than the words heroic or courageous. Usage of the word brave increased over time, which indicates that visitors were less likely in the year(s) immediately following September 11 to describe those who experienced the events as brave--only after a year or so had passed were they utilizing that description with a greater frequency. Heroic people was not particularly relevant, and the ngram courageous people was only used six times throughout the corpus.

Human spirit



The use of the ngram human spirit occurred 25 times throughout the corpus, although its usage varied from 2002 through 2007, with the term not being used in some years. It is not particularly relevant to the corpus given its low frequency. It seems most likely that visitors were more likely to *describe* human spirit, rather than using that particular phrase.

Innocent people, innocent victims



The ngram innocent people is highly relevant to this corpus and is used with consistent frequency from 2002 through 2007. The use of innocent victims is much less frequent, and there is a spike in usage between 2004 to 2005. It is evident that visitors identified those who experienced the events of September 11 as innocent, and that characterization continued to be utilized even as visitors remembered the events of September 11 in subsequent years.

Functionality of the Ngram Plugin: Ngram frequencies

The ngram frequencies pane allows users to determine how many unigrams, bigrams, or trigrams they wish to generate. Users will be presented with a table consisting of the ngram, total count of that ngram, and frequency of the occurrence of that ngram throughout the corpus. Additionally, the plugin states how many uni/bi/trigrams are contained within the corpus and how many unique uni/bi/trigrams are extant.

For the purposes of this case study, I examined trigrams.

Ngram	Total count	Frequency %
Should be remembered	1,878	0.715519%

The people who	1,201	0.457581%
We should remember	1,116	0.425196%
I think that	1,063	0.405003%
The people that	782	0.297942%
All the people	711	0.270891%
People who died	588	0.224028%
I think we	578	0.220218%
Think we should	570	0.217170%
All of the	528	0.201168%
Should remember the	452	0.172212%
I think the	439	0.167259%
Lost their lives	420	0.160020%
People that died	406	0.154686%
We need to	351	0.133731%
Think that the	344	0.131064%
That we should	325	0.123825%
Be remembered about	308	0.117348%
The twin towers	287	0.109347%
Should remember all	269	0.102489%

Given that the question visitors were answering was “What do you think should be remembered about September 11th?” it is not surprising that several of these trigrams contain some form of the verb remember. It is significant that several trigrams contain people--such as people who died or all the people--which led me to infer that many visitors were describing the individuals who died while also expressing how important they believed it was for them to be remembered. Additionally, it might be possible to argue that visitors believed acts of remembering those who died should be a collective experience, given that several trigrams use the pronoun we--”we should remember,” “I think we,” “think we should.” This action of remembering those who died was used in contexts that described it as a necessity or an obligation--should be remembered, we should remember, should remember the, we need to.