# Global Pattern Search at Scale

R. Jordan Crouser, Matthew C. Schmidt[*], Stephen Kelley, Benjamin Miller, Daniel Hook, Lauren Edwards,
Maja Milosavljevic, Elizabeth Michel, Elizabeth Ferme[†], Robert Carrington, Albert I. Reuther

MIT Lincoln Laboratory

Lexington, Massachusetts 02421

Email: jordan.crouser@ll.mit.edu

*Abstract*—In recent years, data collection has far outpaced the tools for data analysis in the area of non-traditional GEOINT analysis. Traditional tools are designed to analyze small-scale numerical data, but there are few good interactive tools for processing large amounts of unstructured data such as raw text. In addition to the complexities of data processing, presenting the data in a way that is meaningful to the end user poses another challenge. In our work, we focused on analyzing a corpus of 35,000 news articles and creating an interactive geovisualization tool to reveal patterns to human analysts. Our comprehensive tool, Global Pattern Search at Scale (GPSS), addresses three major problems in data analysis: free text analysis, high volumes of data, and interactive visualization. GPSS uses an Accumulo database for high-volume data storage, and a matrix of word counts and event detection algorithms to process the free text. For visualization, the tool displays an interactive web application to the user, featuring a map overlaid with document clusters and events, search and filtering options, a timeline, and a word cloud. In addition, the GPSS tool can be easily adapted to process and understand other large free-text datasets.

## I. INTRODUCTION

As increased computing power and data storage become more ubiquitous, analysts have access to information at unprecedented scales. As a result, the large unstructured text corpora that result from these massive aggregation efforts present a significant opportunity for advances in situational awareness and intelligence. In particular, geotemporal and other semantic trends buried within this unstructured text can help analysts to form a complete picture of the changing analytic landscape within a region of interest. Understanding these trends is of critical importance in areas such as epidemiology, international relations, and national security.

Unfortunately, traditional methods for exploration demand the analyst sift through unmanageable amounts of data manually, often restricting their analysis to a narrow subset of the larger corpus and rendering many of the broader and more subtle trends largely undiscoverable. In addition, these trends often span multiple data sources, types and granularities, with no single source capturing the entire story. Analysts must therefore be able to interactively fuse and transform these disparate data sources in order to construct a cohesive narrative.

To begin to address these challenges, we collaborated with domain experts to design and develop a scalable visual analytics platform to support the analysis of unstructured document corpora. With GPSS (Global Pattern Search at Scale), analysts can interactively explore the document corpus at multiple resolutions, identifying patterns that cut across various data dimensions as well as uncover key events in both space and time. Our comprehensive tool enables automated free text analysis on high volumes of data, as well as interactive data visualization to support exploratory analysts. To process the free text, GPSS uses a matrix of word counts and event detection algorithms. This is supported by a powerful Accumulo [1] database for high-volume data storage. The tool includes interactive visualization featuring a map overlaid with document clusters and events, search and filtering options, a timeline, and a word cloud. In addition to providing an interactive exploratory visualization for free-text news articles, the GPSS preprocessing pipeline can be readily adapted to process other large free-text datasets.

## II. DOMAIN CHARACTERIZATION

Through data mining and statistical modeling on large, unstructured document corpora, geospatial intelligence analysts try to uncover the latent patterns in movement, behavior, and communication that indicate areas and activities of interest. A deeper understanding of these patterns and how they evolve over time can dramatically improve situational awareness and better inform the decision-making process.

### A. State of the Practice

In geospatial intelligence analysis as in many other fields, the rate of data collection is rapidly outpacing the capacity of traditional analytical methods. Because manual analysis and synthesis of individual text documents into topics and themes would require countless hours and be subject to significant noise, the data must first be aggregated and cleaned to conserve time and energy. Although automated analysis through statistical modeling of the raw text is possible, it often proves insufficient; the generated anomalies require expert analysis to interpret, and the end user is rarely an expert in statistics.

Tightly coupled human and machine analysis is critical to the successful and timely utilization of these rich data sources. Systems for supporting these analytical processes must present

---

[*] Dr. Schmidt is now at North Carolina State University.
[†] Ms. Ferme is now at University of California, Berkeley.

the data in an intuitive, domain-appropriate context while high-lighting trends and anomalous behavior, as well as allow for deep exploration of the raw data while maintaining scalability.

### B. Utilizing Open-Source Datasets

Intelligence analysts are increasingly interested in leveraging open-source or publicly available datasets to supplement their analysis using more sensitive information. In this demonstration, we collected a sample dataset consisting of approximately 35,000 published news articles from Reuters, CNN, the Guardian, NPR, and the New York Times collected over a period of three months. These articles are published in English, and their coverage spans much of the globe. To collect these articles, we set up a collection pipeline to scrape articles from these news outlets' published RSS feeds using the Python `feedparser`[1] library. From each entry, we then crawled the links to the original articles and scraped the raw text and related metadata using the Python `BeautifulSoup`[2] library.

## III. DESIGN CONSIDERATIONS

Preliminary discussion with our collaborators at the National Geospatial-Intelligence Agency revealed that their analysts face two main challenges in their analysis of the geotemporal trends in unstructured text corpora: maintaining geospatial and temporal context when working with massive datasets, and identifying patterns over time and space to detect the occurrence of significant events.

### A. Maintaining Analytic Context

When faced with massive datasets, one of the main challenges analysts face is maintaining analytic context when exploring multiple facets of the dataset. As an analyst moves from one area of the data to another, it is easy to lose track of the logical path they have followed thus far, leading to the loss of supporting evidence and an increase in redundant or duplicated effort. Tightly coupled human and machine analysis is critical to the successful and timely utilization of these rich data sources. Systems for supporting these analytical processes must present the data in an intuitive, domain-appropriate context while highlighting trends and anomalous behavior, as well as allow for deep exploration of the raw data while maintaining scalability.

### B. Event Detection

Another challenge facing analysts as they try to fuse nontraditional intelligence sources is detecting the occurrence of significant events. The ability to detect events is required for other important analytic tasks such as identifying subtle patterns in the data and making accurate evidence-based predictions. In the context of our data analytics on a corpus of documents containing news reports, we define an event as a set of documents that have a similar enough vocabulary (i.e. they are talking about the same thing) and share other common characteristics. Events can be manifested in the data as, for example, a surge in the number of reports coming from a certain region or regarding a certain topic.

## IV. CHARACTERIZING SIMILARITY BETWEEN TEXT DOCUMENTS

In order to be able to draw comparisons between text documents, a few more assumptions are needed in order to compute a numerical distance. We chose a simple bag-of-words model for our text documents, where each document is represented only by the number and frequency of the distinct terms it contains, and the ordering of the terms is disregarded. In this way, each document is represented by a vector of length $V$, where $V$ is the number of distinct terms in the corpus. Each document's vector will capture the essence of the document's meaning through the relative term frequencies it contains as a point in "topic space".

Some terms, however, are less informative than others. Suppose we have a set of English text documents and we want to identify the document most related to the phrase "a big explosion". We might start by eliminating any documents that do not contain all three of the words "a", "big", and "explosion", but depending on the size of the collection this may still leave us with many documents to look through. To further distinguish them, we could count the number of times each term occurs in each document and add them all up, then select the document that gets the highest total score, or *frequency*.

However, because some terms appears more often in the English language, relying on frequency alone will tend to incorrectly emphasize documents that use common words. In this case, too much weight would be give to documents using the word "a" more frequently, without giving enough weight to the more meaningful terms "big" and "explosion". Common words, such as "a" and "the", are not good keywords to distinguish relevant and non-relevant documents and terms. To mitigate this, an inverse document frequency factor is incorporated, which reduces the impact of terms that occur very frequently in the document set and increases the impact of terms that occur less often. This is known as Term Frequency-Inverse Document Frequency, or `tf-idf` [2].

To compute this value for a term $t$ in a document $d$, first let $\text{tf}(t, d)$ be the frequency of term $t$ in document $d$. Then calculate the inverse document frequency over a collection of $N$ documents:

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

where $|\{d \in D : t \in d\}|$ is the number of documents where the term $t$ appears (i.e., $\text{tf}(t, d) \neq 0$). Finally, the weighted value is calculated as:

$$\text{tf}(t, d) \times \text{idf}(t, D)$$

This normalization step amplifies terms that better characterize certain documents, and washes out terms that occur commonly across the corpus.

## V. EVENT DETECTION

Leveraging this method for characterizing text documents, we can now begin to identify interesting events described in the raw data. We approached event detection from two angles, resulting in the use of two algorithms to perform

---

[1] https://pypi.python.org/pypi/feedparser
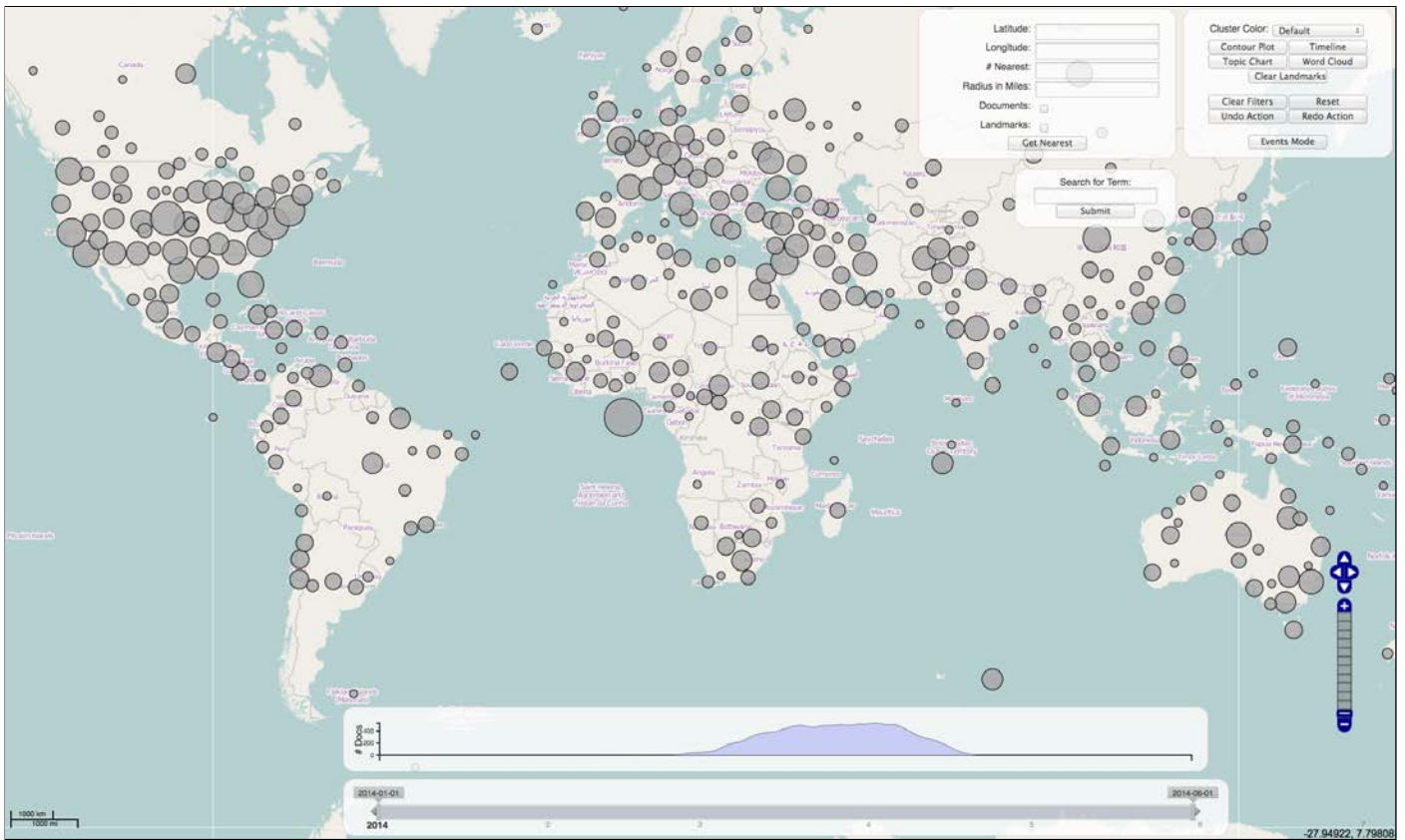[2] https://pypi.python.org/pypi/BeautifulSoup/3.2

Fig. 1. The overview screen of the Global Pattern Search at Scale (GPSS) tool. Each grey dot represents a document or cluster of documents referencing a particular geospatial region. The selection tools in the upper righthand corner enable the analyst to search along various facets of the data to refine the collection of documents on the screen. The timeline views along the bottom of the screen enable the analyst to narrow the temporal window from which documents are displayed. The line graph above the time selection bar shows that the documents currently displayed were generated between March and June of 2014.

event detection on the corpus (described in detail below). Each algorithm takes in the document corpus and returns a set of related documents that are likely to be describing the same event. Given a set of documents generated by the algorithm, the analyst can then explore the documents to discover their underlying relationship.

### A. Spatiotemporal Event Detection

Our first algorithm adopts a foreground independence approach to event detection. The articles are given an arbitrary ordering, $d_1, d_2, ..., d_n$ where $n$ is the number of articles in the corpus. Each article is then associated with a 0-1 vector, where the $j$th entry of article $i$'s vector is 1 if article $i$ and article $j$ are close enough, for some measure of distance and 0 otherwise. For each article, we can then construct a commutative distribution function of the spatial and temporal distribution of the article. We then analyze these distribution functions to obtain a loose bound on which articles textually close to article $i$ are also close enough in space and time to constitute an event. These sets of articles are pre-events. The pre-events are then scanned for duplication events, which are merged based on a similarity clustering into events.

### B. Term-spike Event Detection

Our second event detection algorithm approaches the problem slightly differently. Here, documents are analyzed a week at a time. For each week of documents, we compute a `tf-idf` vector where the inverse document frequency data is taken from the previous five weeks and the term frequency is taken from the given week. In this way, we ensure that a spike in the use of certain words in the week of interest is noted. With this modified `tf-idf` matrix on the documents from the week of interest, we can then compute the textual similarity between documents. Our pre-events here are the sets of documents that are within this week time frame and are textually similar up to a predefined threshold. Like in the first algorithm, the pre-events are passed through certain filters. Here, we ensure that the documents in an event are not too spatially spread.

These two methods of event detection complement each other nicely. The first algorithm finds more temporally spread events and "larger" events; i.e. each event contains a larger (approx. 20-50) list of documents. The second finds events that are temporally tight and of a smaller size (approx. 5-10 documents). Both algorithms found about an equal number of total events in the corpus, and there was no overlap of events detected. On a corpus of 35,000 documents, the first event detection algorithm took approximately 16 hours to run, whereas the second algorithm only took a few minutes. Thus,
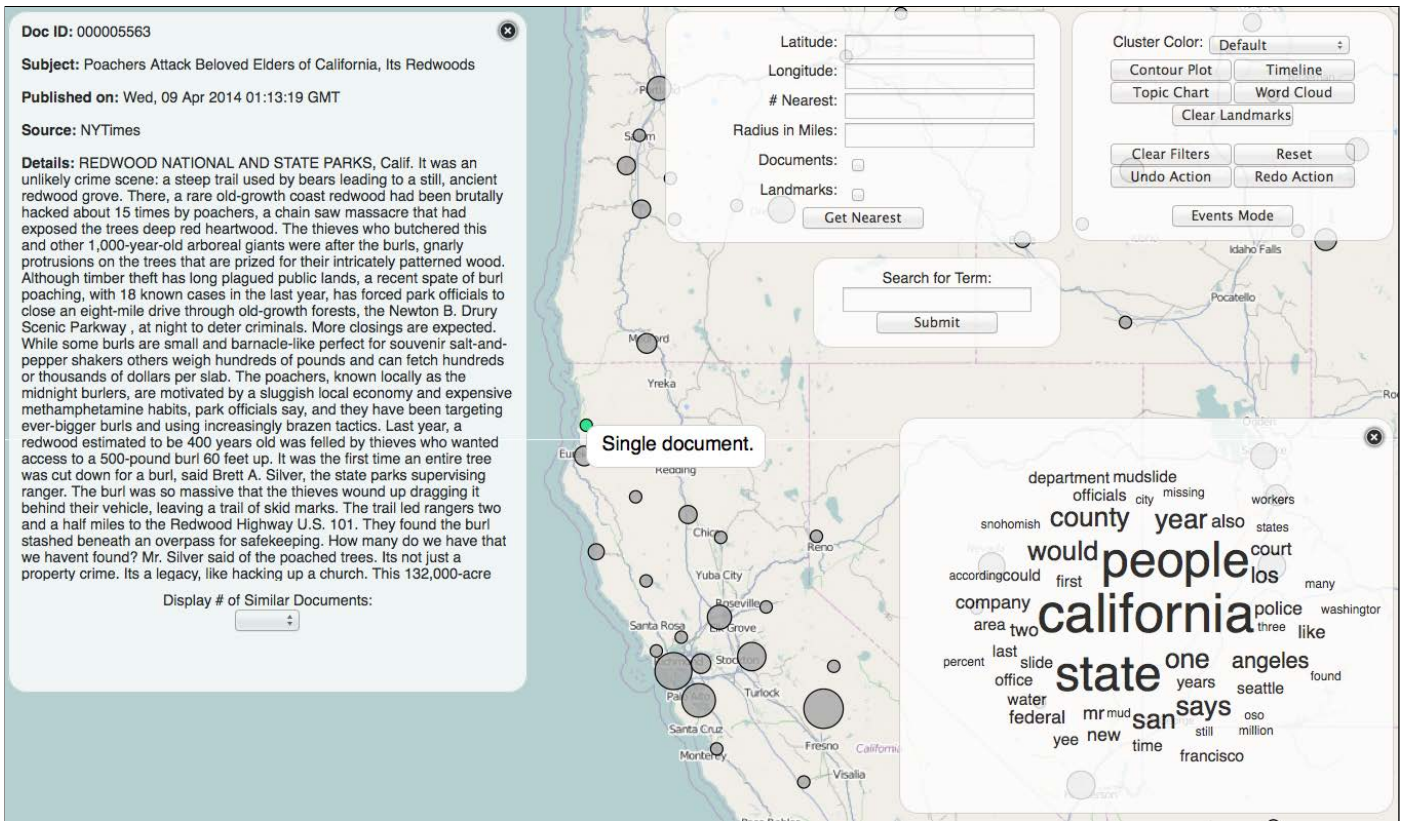
Fig. 2. The GPSS tool zoomed in to a region of interest (California, USA), with a single document selected. The detail view of the selected document is visible in the upper left corner of the screen, displaying information about the subject of the article, when the article was written, the reporting agency, and the full text. In addition, a word cloud visible on the righthand side has been generated from the collection of articles currently visible in the zoomed region.

the second algorithm is likely scalable to much larger datasets.

## VI. VISUAL ANALYTICS TOOL

To enable the analyst to explore these events within their larger geotemporal context, we designed and implemented Global Pattern Search at Scale (GPSS), a secure web-based visual analytics platform. To support the exploratory analysis of geotemporal trends in large, unstructured text corpora, this tool employs a coordinated multiple views (CMV) architecture [3], and is supported by a powerful Accumulo data store to deliver data to the visual front-end at interactive rates. Each of the coordinated views that GPSS offers will be discussed in detail later in this section.

The front-end visualization (see Fig. 1) is built using javascript tools such as D3.js [4]. By deploying our tool on the web, we are able to provide the end-user with cross-platform compatibility and a centralized upgrade procedure; new features are implemented on the server and automatically delivered to the user. The tool is straightforward to deploy and manage, and requires minimal resource allocation on the client side. Finally, HTML5 natively facilitates rich interactive environments in any compatible web browser.

### A. Geospatial View

Data points in these text-based non-traditional GEOINT sources have the potential to be associated with a geotemporal location through data enrichment and extraction. In order to leverage analysts' existing domain expertise in geospatial intelligence, it is critical to frame the analytical tools by embedding them in a geospatial context. To facilitate this, we built our interactive visualization on top of OpenLayers[3] to provide the georeferenced map layer. All articles and events are plotted within this georeferenced layer, and interaction is coordinated across all non-geospatial views of the data.

### B. Control panel

In order to populate the various views, the analyst can use the suite of selection and search tools provided within the system (see Fig. 1 (upper right)). These tools enable the analyst to refine their search along various facets of the data such as term or geospatial region. These refinements can be used independently, or can be used in conjunction with one another to produce a union or intersection. The current built-in iterators support intersection over a set of features, but they do not support union and thus do not support arbitrary boolean operations. Simply applying a filter to the entire set of data would solve the problem, but such a solution is far from optimal. In order to efficiently deliver the results of these queries, we developed a custom iterator for Accumulo that extends the Intersecting Iterator.

---

[3]openlayers.org

## C. Temporal Views

A timeline view enables the analyst to narrow the temporal window and subselect a collection of articles from a specific window of time. The analyst can also slide the selected window along the timeline to see how the distribution of documents in a region changes over time. The line graph above the time selection bar shows the temporal distribution of the current set of documents.

## D. Detail View

When an analyst has identified an area of interest, they may wish to drill down into the individual documents in that region to read the full text. In this tool, the analyst simply clicks on the document of interest, and a detail box is populated with information about the subject of the article, when the article was written, the reporting agency, and the full text (see Fig. 2 (left)). The analyst is also presented with a drop down menu for selecting similar documents, which can be used to further refine their search.

## E. Document Similarity

A standard question an analyst might pose is: given a document $d$, which other documents in the corpus are most "similar" to $d$? In a geospatial intelligence context, an analyst is most likely concerned with spatial similarity (how close two documents are to each other on the globe) and topic similarity (the degree to which the subject matter of two documents is the same).

*1) Geospatial Similarity:* In order to rank documents' similarity to the prototype document $d$, we first need to establish a pairwise distance function. In the case of spatial similarity, we approximate the Earth as being a sphere. The distance between two documents is then the great circle distance between them on the surface of this sphere. To calculate this distance, let $\phi_1, \lambda_1$ and $\phi_2, \lambda_2$ be the geographical latitude and longitude of two points 1 and 2, and $r$ the radius of the sphere; then $D$, the distance between them, is given by:

$$D = r * \arccos\big(\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos(|\lambda_1 - \lambda_2|)\big)$$

*2) Topic Similarity:* In the case of topic similarity, we consider the cosine of the angle between two documents' `tf-idf` vectors - if two documents contain roughly the same words in similar frequencies, their `tf-idf` vectors will point in roughly the same direction in topic space, and the cosine of the angle between them will be close to 1. Similarly, if the documents share very few terms, the cosine of the angle between them will be close to 0.

As a part of the tool we provide analysts with the capability to perform efficient nearest neighbor searches. We accomplish this with a combination of cached values and efficient spatial data structures called the R-Tree [5]. R-Trees require some computational effort to organize the documents according to their geographic positions as the tool is initialized, but facilitate nearest neighbor queries which are significantly more efficient than a brute-force approach.
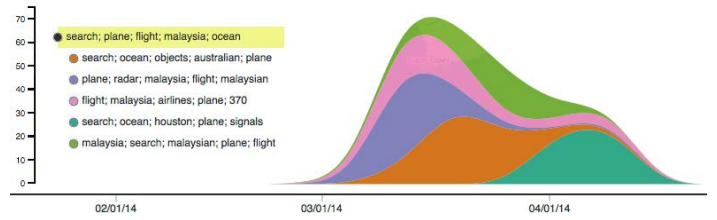
Fig. 3. The GPSS tool displays the topic cluster hierarchy in an expandable indented list (left), where each topic and subtopic is represented by a five term summary. Additionally, the temporal distribution of sibling topics is displayed using a stack graph (right).

## F. Topic Views

An analyst may also wish to know what general "topics" are present among the documents in the corpus. The notion of a topic can be ambiguous; by topic we mean a cluster of vectors in `tf-idf` space. We compute a hierarchical clustering of topics using a recursively applied k-means clustering algorithm. The number of clusters and the recursive depth are determined heuristically, with the goal of making the collection of topics both meaningful and easy to navigate. We begin by clustering the full corpus into a set of high-level clusters, and then recursively applying the $k$-means algorithm on each of these clusters to produce sub-clusters.

We display the topic cluster hierarchy in an expandable indented list, where each topic and subtopic is represented by a five term summary. Additionally, the temporal distribution of sibling topics is displayed using a stack graph (see Fig. 3). Strackgraphs are a method of data visualization in which layers of data are stacked on top of one another which emphasizes legibility of individual layers [6].

## G. Word Clouds

Topic summaries are small special cases of word clouds, another view we provide in the tool. A word cloud is a summary of a subset of the corpus which displays the $k$ terms which best characterize that set of documents [7]. In our visualization, higher-weighted words are displayed with larger fonts or closer to the center of the word cloud correspond to words that are more significant to describing the data (see Fig. 2 (right)). We compute word clouds by summing the `tf-idf` vectors of all the documents in the subset. The $k$ terms with the highest aggregate `tf-idf` scores are displayed, with their size weighted by the magnitude of these scores. The word cloud is restricted to the documents in the user's current view and can aid the analyst in discovering latent topics covered by the current documents.

## H. Event View

Our tool employs a separate event layer to display the most significant events on the map (see Fig. 4). When viewing this layer, events are shown on the map as colored bubbles centered at the spatial center of the event (center of the smallest circle encapsulating the events). This map of events allows an analyst to easily see the spatial distribution of events throughout the zoomed region, and with careful use of the timeline feature, an analyst will also be able to detect temporal patterns associated with events. Each event will be one of three colors. Two colors
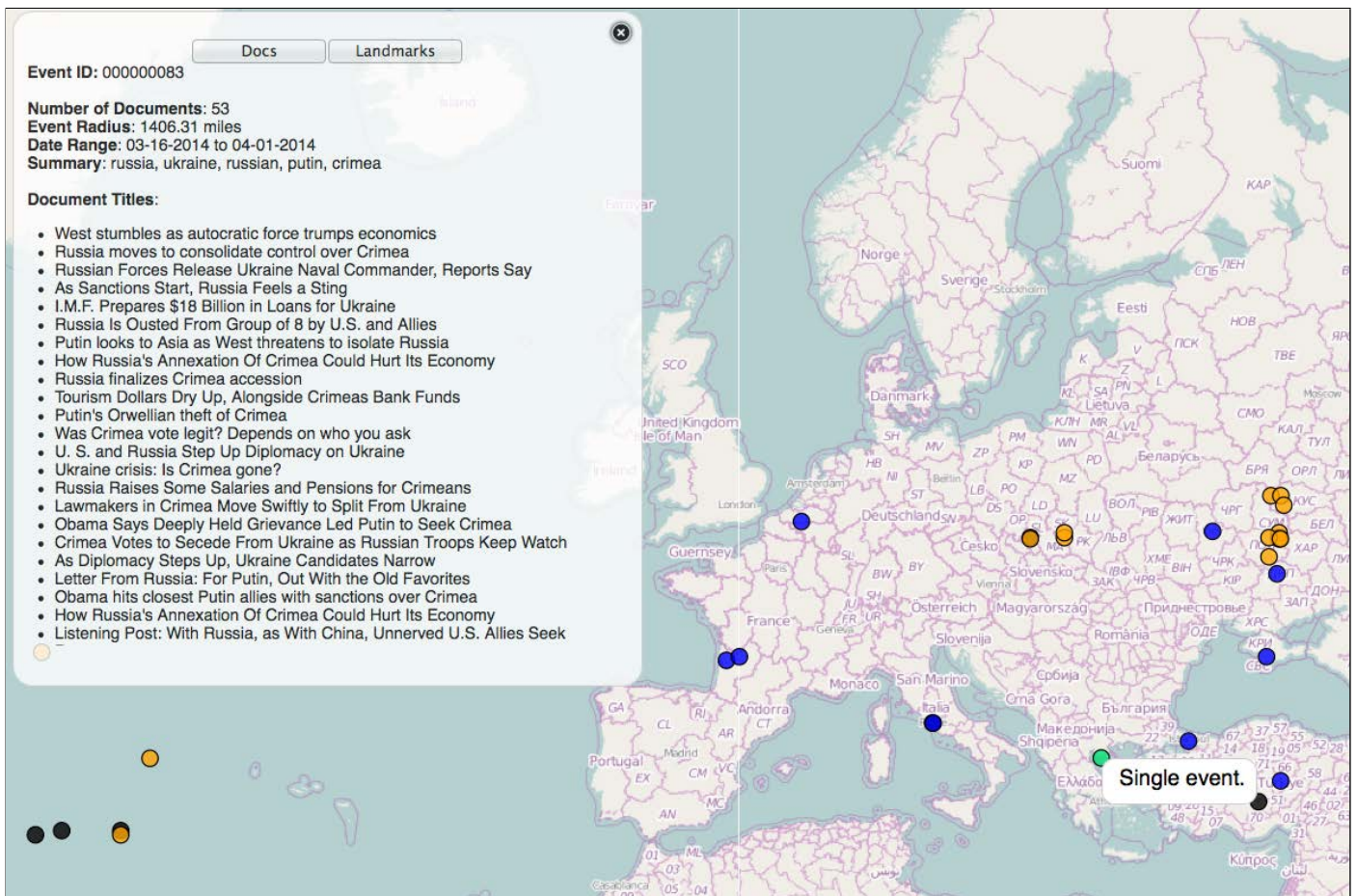
Fig. 4. The events mode overview screen of the Global Pattern Search at Scale (GPSS) tool, zoomed in to a region of interest (Europe). Each dot represents an event or cluster of events localized to a particular geospatial region. Events are colored according to the detection mode used to identify them. The details view in the upper left corner shows information about the number of documents in the event, the radius the event spans, the date range over which the event occurred, a brief summary of the keywords that describe the event, and the titles of the articles referenced.

are used to distinguish events generated by the two event detection algorithms used. A third color is used to represent a cluster of events, which reduces clutter on the screen. Simply clicking a cluster will transform it into individual events on the map.

To learn more about the documents generating an event, an analyst can click on the event icon on the map. A window will then pop up displaying important information about the event such as the number of documents the event contains, the titles of these documents, a list of key words providing a summary of the event, the date range of the documents of the event and the radius of the documents that make up the event.

## VII. CONCLUSION

In this work, we present a comprehensive visual analytics tool, Global Pattern Search at Scale (GPSS), which addresses three major problems in data analysis: free text analysis, high volumes of data, and interactive visualization. We demonstrate its utility by analyzing a corpus of 35,000 news articles, revealing underlying patterns to human analysts via an inter-active web application. We posit that the GPSS tool can be easily adapted to process and understand other large free-text datasets, and hope that this work might be leveraged in support of other non-traditional GEOINT analytics in the future.

## REFERENCES

[1] A. Fuchs, "Accumulo–extensions to google's bigtable design," Technical report, National Security Agency, Tech. Rep., 2012.

[2] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the First Instructional Conference on Machine Learning*, 2003.

[3] J. C. Roberts, "State of the art: Coordinated & multiple views in exploratory visualization," in *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*. IEEE, 2007, pp. 61–71.

[4] M. Bostock, "D3.js," *Data Driven Documents*, 2012.

[5] T. Brinkhoff, H.-P. Kriegel, and B. Seeger, "Efficient processing of spatial joins using r-trees," *SIGMOD Rec.*, vol. 22, no. 2, pp. 237–246, Jun. 1993. [Online]. Available: http://doi.acm.org/10.1145/170036.170075

[6] L. Byron and M. Wattenberg, "Stacked graphs-geometry & aesthetics." *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1245–1252, 2008.

[7] T. Gottron, "Document word clouds: Visualising web documents as tag clouds to aid users in relevance decisions," in *Research and Advanced Technology for Digital Libraries*. Springer, 2009, pp. 94–105.