# Integrating Interactive and Computational Approaches for News Understanding

Liang Wu, Fred Morstatter, Daniel Baird, Justin Sampson, Kai Shu, Huan Liu

Arizona State University
Tempe, AZ, USA

## 1 Introduction

Event corpora such as GDELT[1] provide detailed summaries of events as they happen day-to-day. These datasets are unprecedented in their ability to extract machine-readable information from news data. However, this immense amount of data is nearly impossible to understand at the corpus level, requiring immense amounts of time in order to understand the meaning of, and relationships in between events [2]. The immense and rich data available in GDELT enables social scientists to ask questions previously unavailable. For example, the prominence of different actors, and the geographic distribution of different stories can be answered at a greater scale. While this analysis is suddenly possible, it is difficult to process this large amount of data. Reading the individual news stories, or even the meta data of them, is still impossible. Thus, we will integrate the power of topic modeling with the giant news source of GDELT to facilitate fast understanding of news stories.

Topic modeling [3] presents an unprecedented opportunity to take a large corpus of data and reduce it to a few core "topics". These topics are often presented as lists of words which are then presented as a summary of the data. Topic modeling algorithms enable researchers and practitioners to quickly understand a huge corpus just by reading the few topics that are extracted from the corpus. The key issue in generating topics of interest is to identify the relevant data and suitable parameters. However, since domain experts, such as social scientists and policy makers, are not necessarily experts of computational tools, a new challenge arises for us as researchers to develop better ways to quickly understand event data by combining this rich data source with established methods like topic modeling.

In this demonstration we propose to integrate user interaction with computational tools, instead of generating results and then visualizing them. In particular, we will build a software tool which enables researchers to select/remove/modify topics that are being generated. The tool can help researchers and practitioners to quickly and accurately understand the immense data available in the GDELT dataset and find topics of interest effectively. Moreover, we will allow the users to specify both time intervals, as well as geographic and regional queries, and display the topics from the results of those queries.

[1] http://gdeltproject.org/

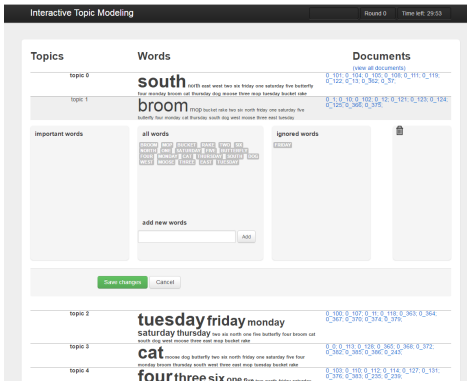


Fig. 1: An example of a topic visualization, taken from the system developed by [1]. We will build upon this to incorporate the special features of the GDELT corpus.

## 2 System Features

We are not the first to build a system to visualize statistical topics. Hu et. al [1] has built a successful tool that easily shows the top terms in different topics, shown in Figure 1. While these types of systems are useful, they lack the capability to visualize the unique properties of the GDELT dataset, namely the relationships between entities and the special types of events and actors that are annotated within the corpus.

The system will have a visualization tool which will help the user to filter, and select the data that they want to visualize. The user will be able to interactively select and filter GDELT news stories by selecting stories through a text entry field, they will also be able to select actors which are of interest to them.

Once the query is entered, we will show the topics derived from this dataset. Furthermore, we will show the popularity of the topic over time, which will help give a sense of *when* the topic was important. Moreover, we will show a map that displays *where* the topic is being discussed. In addition to this data, we will show a network of which actors appear alongside each other in stories, which can help to understand the importance of the actor via certain centrality measures as well as a richer understanding of the connections of the actors in the network.

## References

1. Hu, Y., Boyd-Graber, J., Satinoff, B., Smith, A.: Interactive topic modeling. Machine learning 95(3), 423–469 (2014)
2. Kumar, S.: Social Media Analytics for Crisis Response. Ph.D. thesis, Arizona State University (2015)
3. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on Machine learning. pp. 977–984. ACM (2006)