Using a Knowledge Graph to Discover Earth Science Information

Brian Freitag[1], JJ Miller[1], Kaylin Bugbee[1], Xiaoyi Duan[2], Jia Zhang[2], Rahul Ramachandran[3], Manil Maskey[3], Patrick Gatlin[3], Sundar Christopher[1]

1 – University of Alabama in Huntsville; 2 – Carnegie Mellon University; 3 – NASA Marshall Space Flight Center

IMPACT
Interagency Implementation and Advanced Concepts Team

Introduction

Knowledge graphs link key entities within a specific domain to other entities via relationships. Researchers are able to mine these relationships from numerous sources to infer new knowledge. Text extraction from peer-reviewed papers and scientific reports are untapped resources that can be leveraged by knowledge graphs to accelerate scientific discovery.

Methods

GOAL: Demonstrate the feasibility of applying natural language processing techniques used in knowledge graphs to support scientific discovery of Earth Science information.

Knowledge Graph*

A knowledge graph with a semantic entity recognition (SER) and a relationship model was developed for Earth Science application. Using the natural language processing (NLP) techniques developed for knowledge graphs, how can Earth Science information be extracted from large data sets?

NWS Area Forecast Discussions

The National Weather Service (NWS) issues area forecast discussions (AFD) about 4 times a day for each weather forecasting office (WFO). Forecast discussions include short-term and long-term forecasts with detailed discussion on upcoming weather events. We use the Iowa Environmental Mesonet site hosted by Iowa State University and their archive of NWS text reports to extract AFDs from 2001 to 2017 for CONUS (122 of 126 NWS WFOs). Select terms are extracted from AFDs to study spatial and temporal patterns in the use of terms.

AMS Glossary of Meteorology

Terms are selected from the American Meteorological Society (AMS) glossary of meteorology. The AMS glossary is a static database that contains more than 12,000 terms and definitions.

Data Analysis

More than 2M unique AFDs are compiled to create a database more than 50 GB in size. Such a database is difficult to parse without efficient text mining and NLP techniques like those used in knowledge graph development. A 10-year period from 2007 to 2017 is used for extracting 20 terms to assess the feasibility of using text mining techniques for meaningful Earth Science applications.

- accumulation, atmospheric river, bomb, bow echo, bright band, derecho, downburst, downslope wind, flooding, fog, freeze, gap wind, hard freeze, hurricane, microburst, sea breeze, smoke, snow, supercell, tornado

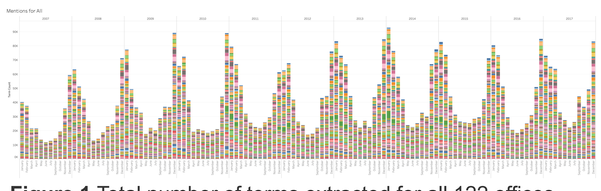

Figure 1 Total number of terms extracted for all 122 offices from 2007 – 2017.

Trends from specific terms can be used for several applications

1) Case identification from spatial and temporal anomalies
2) Trend analysis of term usage for climatological application
3) Evaluation of consistency in communication of scientific information to the general public

Results

We use the term “snow” and evaluate the time series from 2007 – 2017 over the southeastern U.S. (Figure 2). An anomalous peak can be seen in 2015 with more than 4000 extractions for “snow” from AFDs occurring in the month of February.

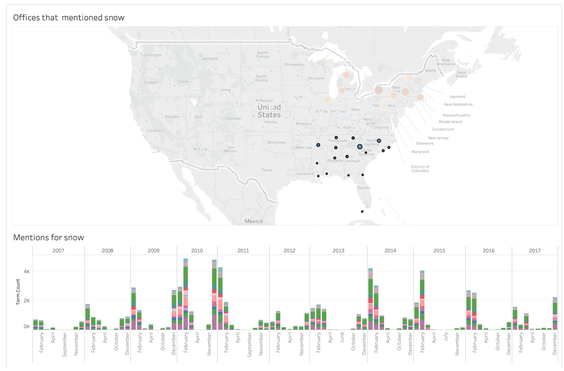

Figure 2 Total extractions for the term “snow” for 2007 – 2017 for the southeastern U.S.

A closer look at term extractions by day for February 2015 reveals two distinct peaks in counts of the term “snow” where term usage exceeds 250 counts per day (Figure 3). The peak on the 24th and 25th was associated with a record breaking snowstorm across the southeast when places in Alabama received up to 12” of snowfall.

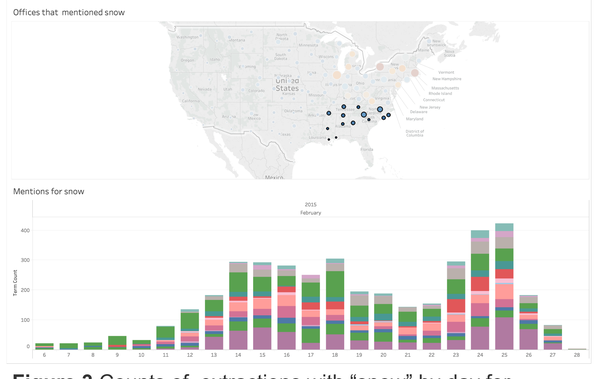

Figure 3 Counts of extractions with “snow” by day for February 2015 for the southeastern U.S.

Another approach useful for extracting meaningful information from the AFDs is to look at term usage as a function of time. With sea levels rising in response to global warming and melting land ice, coastal flooding events are projected to increase. Analysis from 2007-2017 of the term “flooding” from offices along the Gulf of Mexico indicates that coastal flooding events are possibly on the rise (Figure 4). In 2007, no month exceeded 200 mentions of flooding whereas each of the past 3 years have seen 7 months exceed 200 and 2 exceed 400.

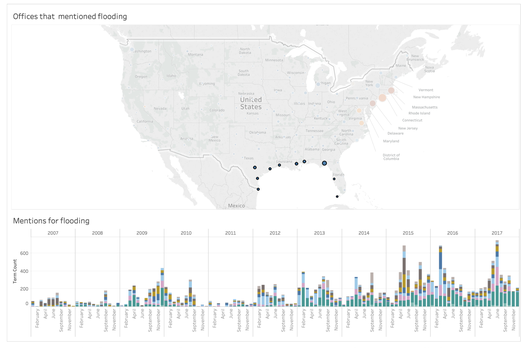

Figure 4 Counts of extractions with “flooding” from 2007 – 2017 for Gulf Coast NWS offices.

Conclusions

We demonstrate the ability of text mining and NLP techniques to perform data analysis on large amounts of text data that is too cumbersome for traditional methods. Potential applications of this approach include: 1) spatial and temporal patterns for case identification; 2) climate studies from extended time series; and 3) social science studies for improved communication of scientific information.

Contact: brian.freitag@nsstc.uah.edu

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

NASA

* For more information on the NASA science knowledge graph, see presentation IN53E-0669 in this poster hall on Friday afternoon.