DETECTING ENVIRONMENTAL DISASTERS IN DIGITAL NEWS ARCHIVES

A regional flood database case study

Amelia Yzaguirre, Michael Smit, Robert Warren
Dalhousie University, Halifax, Nova Scotia

October 29, 2015

2015 IEEE International Big Data Conference
Big Data in the Geosciences workshop

CONTRIBUTORS

Initiated by Nova Scotia Environment's Climate Change Adaptation Project

Project Details

FLOOD DATABASE CASE STUDY

NS Environment's Climate Change Adaptation Project

- History of floods in Nova Scotia
- Build a database economically
- Web-based visualization for users

The Chronicle Herald (data source)

- Two decades of digitized newspaper stories (1992 - Present Day)
- 2 million articles in XML format
- International, regional, and local news coverage

Background

Text mining and GIR

- Goal: convert unstructured text into “interesting” data
- Tools: Natural language processing (NLP), analytical methods, gazetteers
- Typical tasks: text categorization, sentiment analysis, document summarization, semantic similarity

Framing the problem

RELEVANT AND INTERESTING INFORMATION

··· No major traffic accidents were reported in the Valley, but the storm resulted in hundreds of cancellations of church services, Sunday schools, bingos and flea markets. Moncton got the worst of the snow, with 52 centimetres reported by late Sunday afternoon. But along the Atlantic coast, in eastern Nova Scotia counties and in Cape Breton, residents had to contend with freezing rain and rain. Most of those areas received five to 10 centimetres of snow before the precipitation changed to rain. The mess left catch basins clogged, flooding streets in many communities. "We got a lot of rain, a lot of flooding in the streets," said Sgt. Barry Gordon of Cape Breton regional police in Sydney. "Public works . . . have been out early clearing catch basins and manhole covers that are ice-covered." Sgt. Gordon said with the snow being followed by rain, some roads were extremely slippery. "You couldn't even stand up on some of them." Yarmouth got off relatively easy, with six centimetres of snow followed by 1.8 millimetres of rain. The Halifax area received ···

Figure: Sample text from a news article.

RELEVANT AND INTERESTING INFORMATION

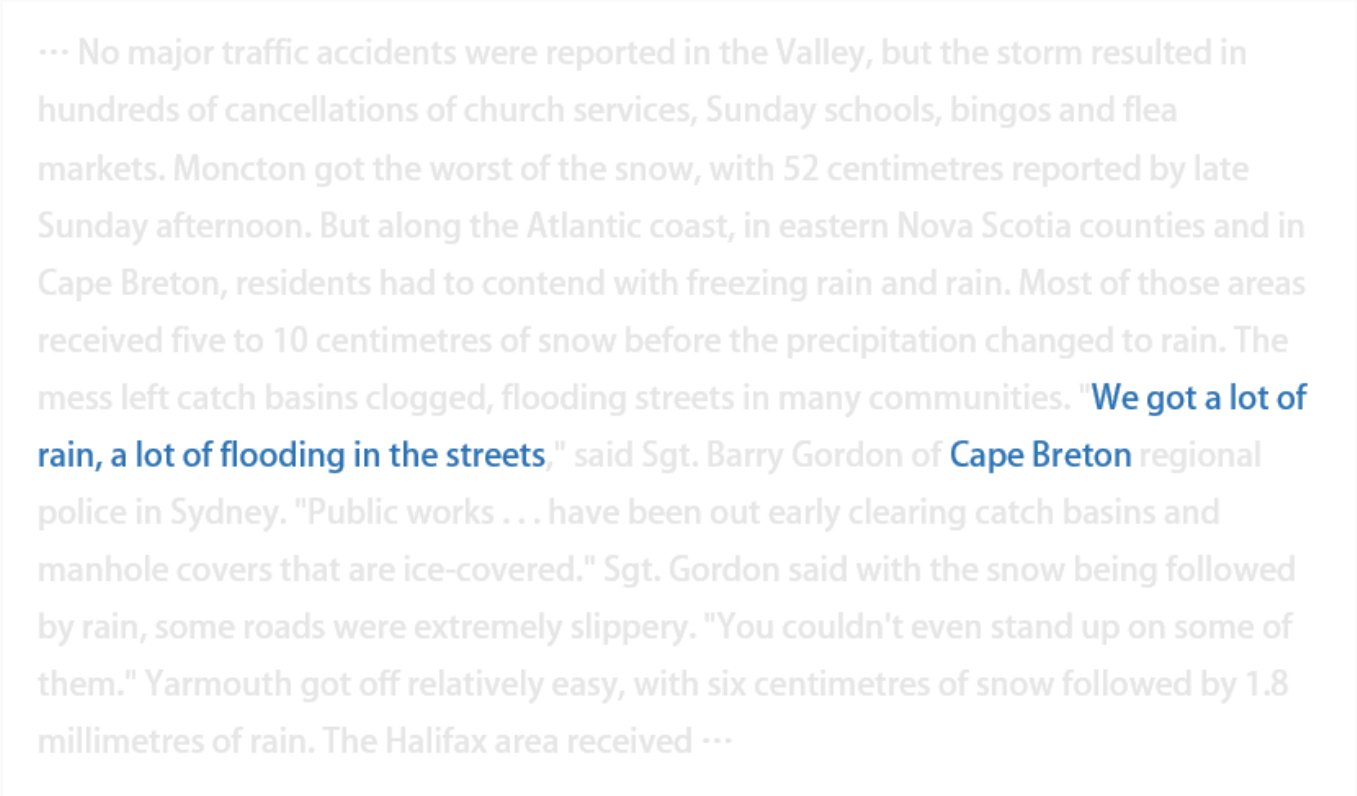


Figure: Extracting relevant details for our database.

CHALLENGES

Opportunities:

- Working with proprietary data
- Large corpus of articles [1]
- Extracting interesting details or summaries from each article

[1] 2 million newspaper articles, 6.54 GB

Methods

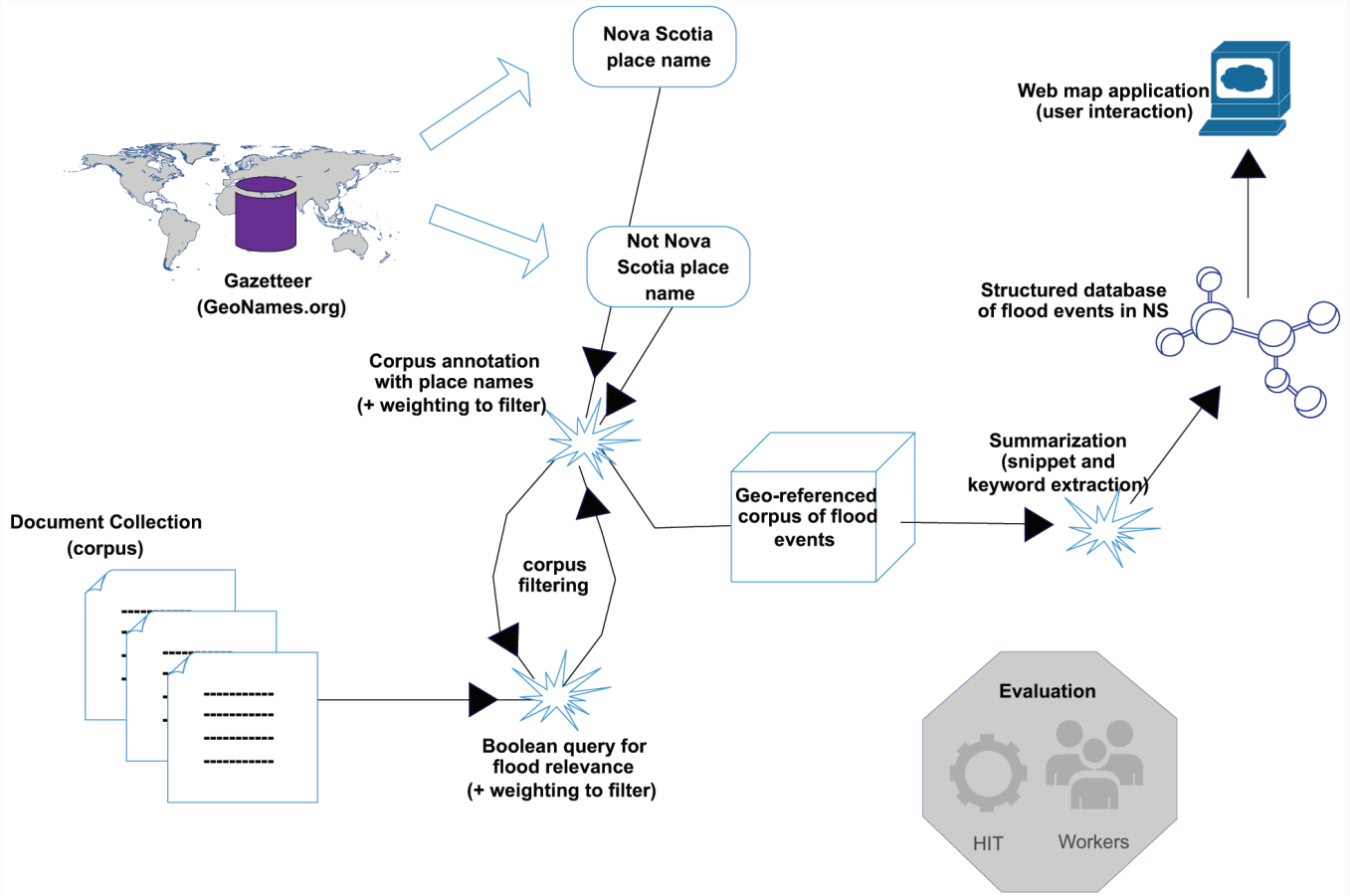
THE BASICS OF OUR INFORMATION EXTRACTION PROCESS

"Off-the-shelf" text mining tools

- Build a dictionary using NLP [2]
- Rule-based filtering using NLP and ML [3]
- Location tagging using Named Entity Recognition and high-precision lists of named entities (*gazetteer*) [4]

[2] Natural Language Toolkit - http://www.nltk.org/
[3] scikit-learn - http://scikit-learn.org/
[4] Geonames - http://www.geonames.org/

SYSTEM ARCHITECTURE



SUMMARIZATION

Set a seed snippet.

> The weekend's heavy rain caused a sewer backup which flooded the basements of an apartment building and about 10 area homes on Perron Street.

SUMMARIZATION

Sentences that contain dictionary words.

"... Wild winds overnight Saturday caused a power outage in about 22,500 Halifax households. ... A wire fastened to the pole broke free in the wind, connected with another wire and shorted out, cutting power to homes in Clayton Park, Fairview and Spryfield. Heavy rain that accompanied the high winds also caused problems for some metro residents by flooding their basements. ... A store clerk said customers had been coming in all morning, complaining of basements flooded by cracks in foundations, roof leaks or sewer backups. Several people also had frozen pipes in their homes burst. ..."

SUMMARIZATION

Most similar to seed snippet.

"... Wild winds overnight Saturday caused a power outage in about 22,500 Halifax households. ... A wire fastened to the pole broke free in the wind, connected with another wire and shorted out, cutting power to homes in Clayton Park, Fairview and Spryfield. Heavy rain that accompanied the high winds also caused problems for some metro residents by flooding their basements. ... A store clerk said customers had been coming in all morning, complaining of basements flooded by cracks in foundations, roof leaks or sewer backups. Several people also had frozen pipes in their homes burst. ..."

Random seeds cultivate variety.

SNIPPET EXAMPLE

Odie.

> “Odie the puppy forlornly stared out over the flooded front lawn he used to romp through, perhaps trying to figure out where it went.”

Evaluation

EVALUATION VIA AMAZON'S MECHANICAL TURK

Thematic and spatial extraction evaluation

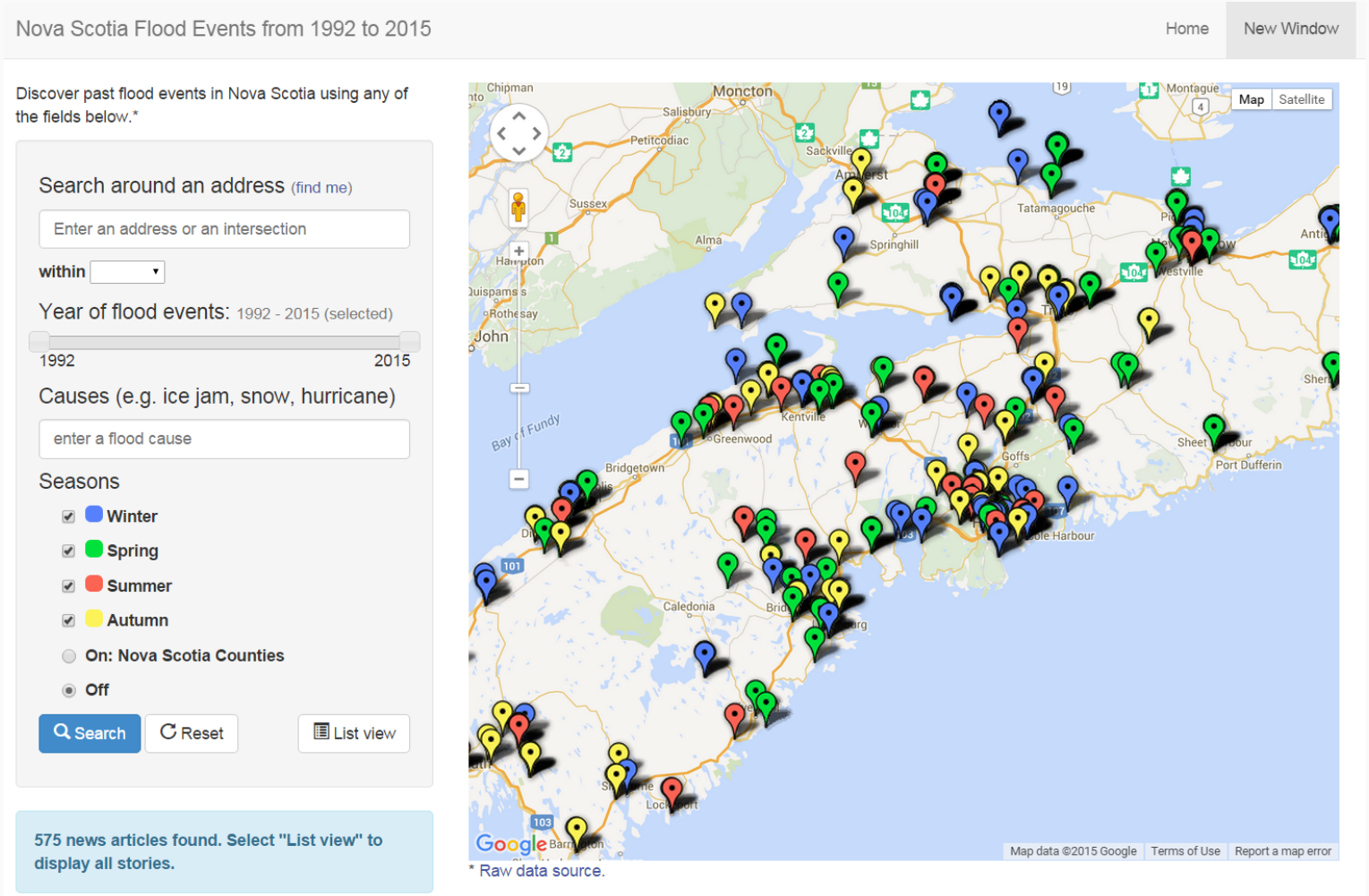
Event Comparator	Accuracy (%)	Geo-location Comparator	Accuracy (%)
Positive	92%	MTurk	52%
Non-negative	95%	Manual ($n = 20$)	60%
		MTurk (consistent)	64%

Location extraction is difficult to assess even for human readers.

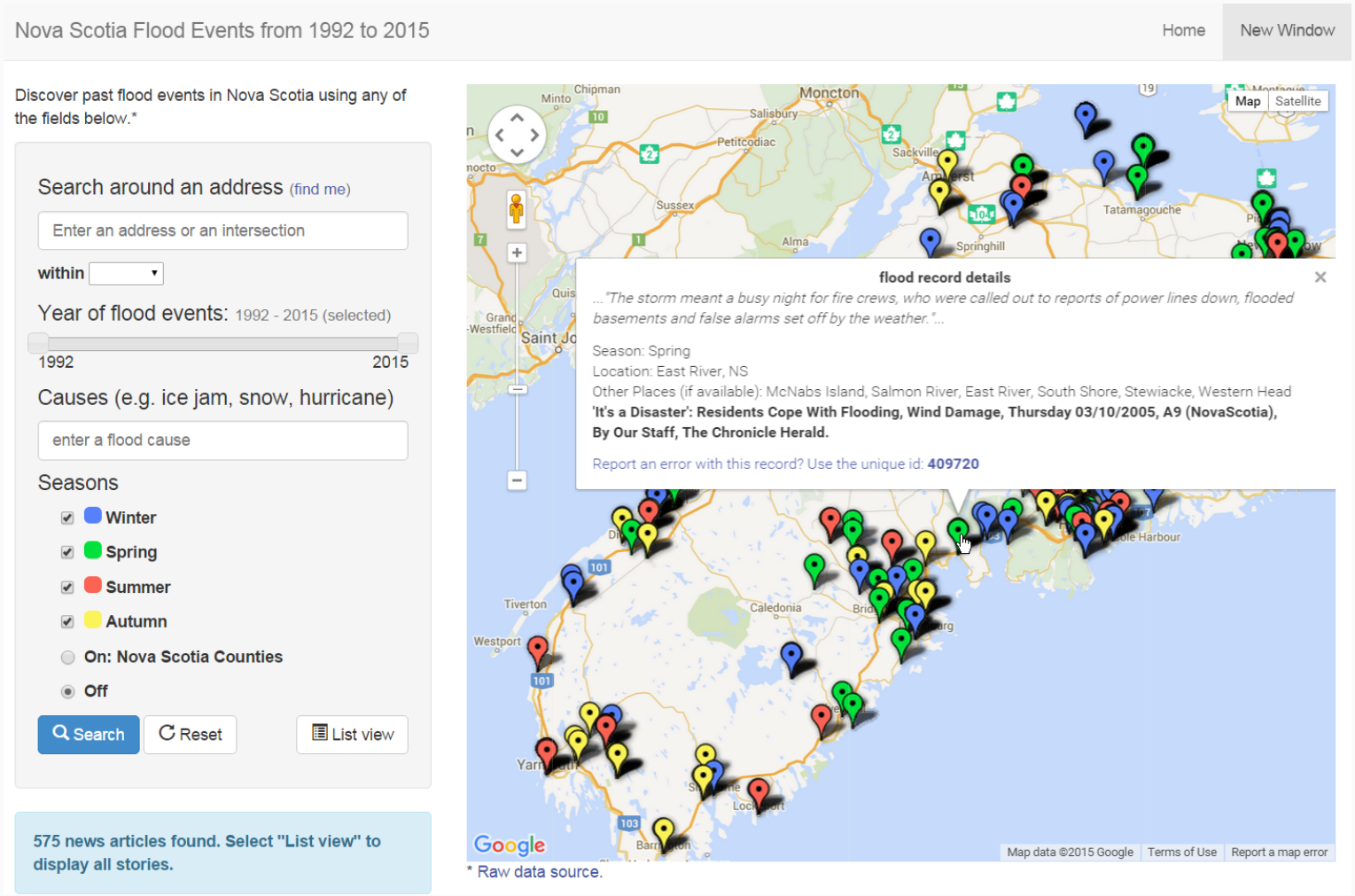
The Web Application

Early Stage

INTERACTIVE MAP



INTERACTIVE MAP



* Raw data source.

INTERACTIVE MAP

Nova Scotia Flood Events from 1992 to 2015

Home | New Window

Discover past flood events in Nova Scotia using any of the fields below.*

Search around an address (find me)

Argyle Street, Halifax, NS, Canada

within 2 km

Year of flood events: 2005 - 2014 (selected)

1992 2015

Causes (e.g. ice jam, snow, hurricane)

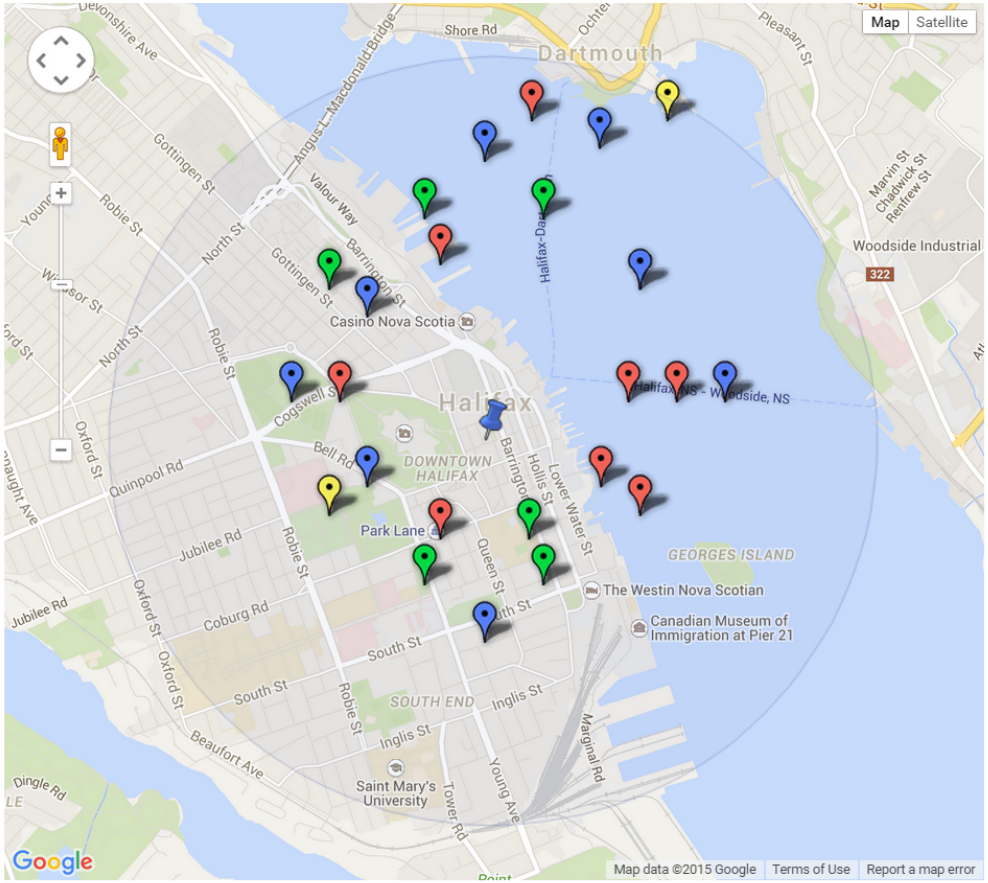
enter a flood cause

Seasons

- [x] Winter
- [x] Spring
- [x] Summer
- [x] Autumn
- [ ] On: Nova Scotia Counties
- [x] Off

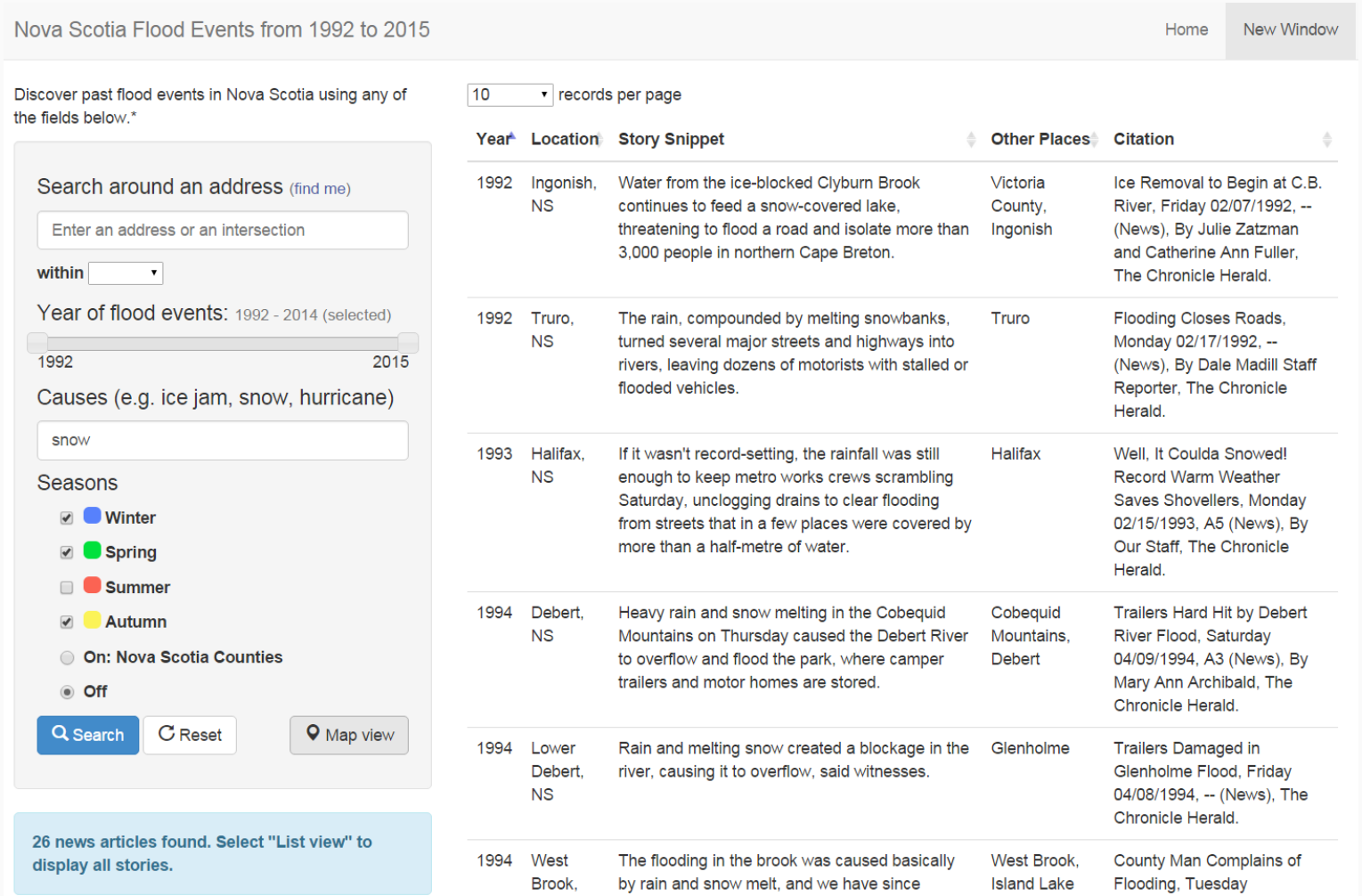
Search | Reset | List view

25 news articles found. Select "List view" to display all stories.

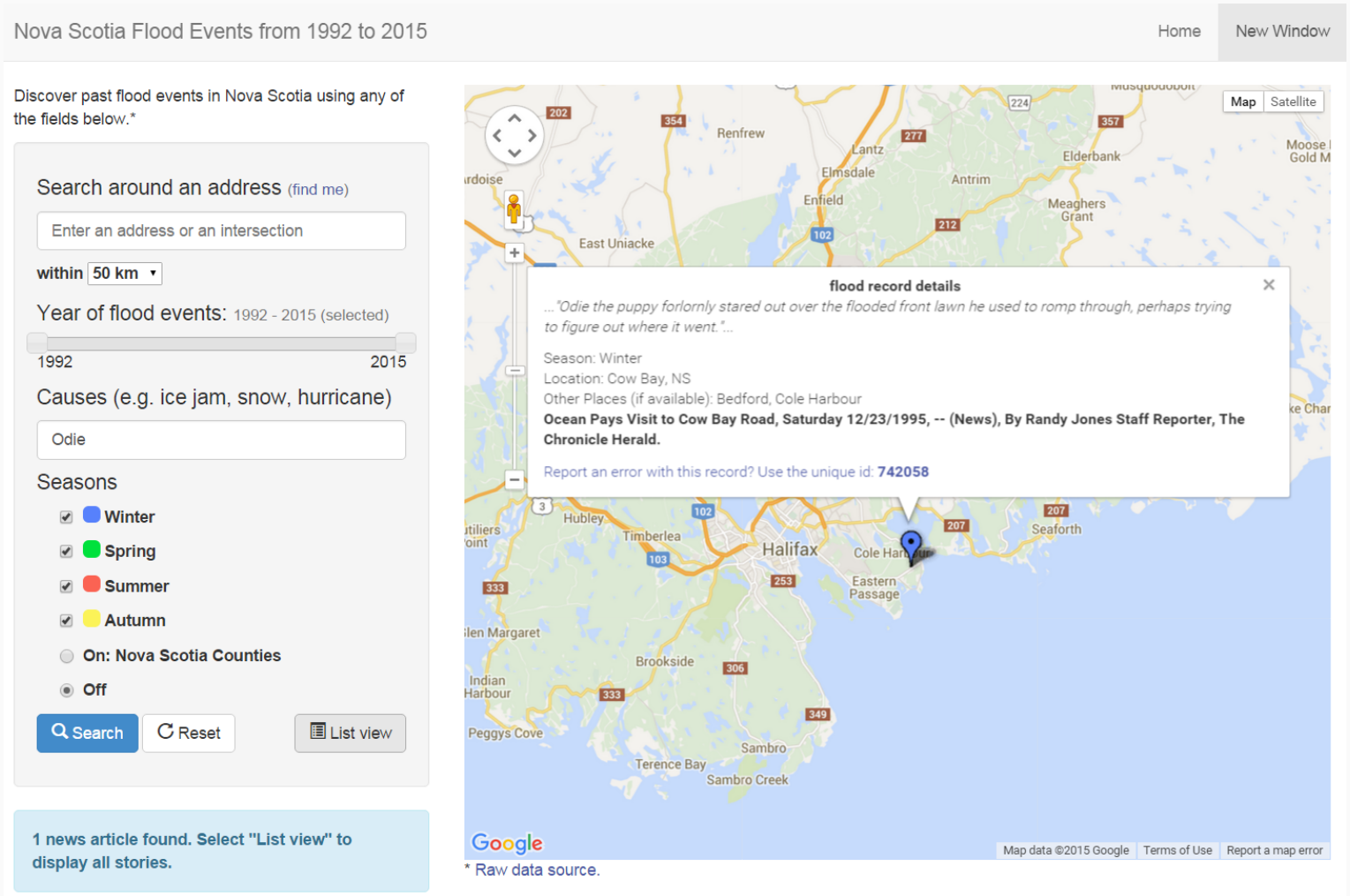


* Raw data source.

INTERACTIVE MAP



INTERACTIVE MAP



Current Focus

- Integrate multiple “high-res” gazetteers into geo-tagging
- Include other regional datasets, study other types of events
- Evaluation methods
- ??

Thanks!

ACKNOWLEDGEMENTS

Thank you, and to the organizers of the Big Data in the Geosciences workshop.

Contributors

- Nova Scotia Environment
- The Chronicle Herald
- The School of Information Management and The Institute for Big Data Analytics at Dalhousie University

Questions later?

email: ameliay@mathstat.dal.ca
Site demo: http://nsfloodhistory.management.dal.ca/

Text mining and GIR

- Topic Modeling - Blei, D. (2012). “Introduction to Probabilistic Topic Models”
- List of text mining software on Wikipedia
- Toponym Resolution - Leidner, J. (2007) “Toponym Resolution in Text”
- GIR System example - Adams, B., McKenzie, G., & Gahegan, M. (2015) “Frankenplace: Interactive Thematic Mapping for Ad Hoc Exploratory Searching”

TOOLS - BEHIND THE SCENES

Main site is built with HTML, CSS, and Javascript with the data stored in Google Fusion Tables.

Technologies used:

- jQuery
- Google Maps API V3
- Google Fusion Tables

OTHER RESOURCES

Create mobile-friendly interactive maps.

- Leaflet - http://leafletjs.com/index.html
- Neatline - http://neatline.org/
- CartoDB - https://cartodb.com/