Weather Severity Prediction Model
Project Proposal

Alex Kopp ARK236@CORNELL.EDU
Cornell Tech, New York, NY

Andrew Li AL543@CORNELL.EDU
Cornell Tech, New York, NY

Abstract

We propose a method for predicting the severity of an upcoming storm event using the free-text descriptions and instructions inside public weather alerts issued by the US National Weather Service. In particular, we show that given an alert structured in the Common Alerting Protocol (CAP) format, we can we can relate certain keywords and patterns to the severity of the storm by predicting a severity metric based on a combination of injuries, deaths, and property damage. This measure may be used to improve alert content and public awareness of an upcoming storm.

1. Introduction

The US National Weather Service issues weather alerts in the event of an upcoming severe weather event. These alerts trigger alerting systems that can launch news, internet, and mobile messages to warn the public of threats to life and property. These alerts however do not have a structured measure to indicate the severity of the weather event. As a result, a storm that passes by and causes little or no damage may have the same Severe Thunderstorm Warning as the 2012 derecho that killed 22 people across seven states and left 4.2 million without power for an extended period.

The goal of this project is to build a machine learning model that can predict the severity of an upcoming storm event based on the free-text descriptions and instructions inside the public alerts.

2. Research Questions

1. Is there a relationship between certain keywords or patterns in the public alerts text descriptions and the resulting impact of a weather event?
2. Can we come up with a useful structured measure to indicate the severity of an upcoming weather alert that the NWS can use to improve alert content?

3. Related Work

Although somewhat dated and mainly about classification, there is valuable information in (Joachims, 1998). Specifically, we are considering the use of *term-frequence inverse document frequency (TFIDF)* instead of pure frequency counts when dealing with text processing.

According to (skl, 2013), the best algorithm in the Scikit-Learn package, given our relatively large amount of data, is the stochastic gradient descent regressor. Other algorithms that may fit our needs include ElasticNet Lasso and Ridge Regression SVR.

4. Data Sets

For this project, we already have access to two datasets (discussed below). While we would like to have additional information on topics such as school closings and power outages, it does not appear as if this data is readily available. As such, we will work with what we have.

4.1. National Weather Service Alert Database

Google has given us access to a database of hundreds of thousands of alerts issued by the National Weather Service. The alerts span from November 2011 to February 2013 and are using the Common Alert Protocol (CAP) format. Useful information can be found inside each alert including geographic locations, dates/times, and free text describing the alert. An example CAP Report taken from the National Weather

Table 1. Example entry from NOAA's Storm Event Database. Bolded fields are the ones that we believe are most important to the project.

Field Name	Value
LAST_DATE_MODIFIED	05/16/2007 12:31:12
LAST_DATE_CERTIFIED	05/26/2007 14:06:08
EPISODE_ID	3749
EVENT_ID	22222
state	CALIFORNIA
state_fips	6
YEAR	2007
MONTH_NAME	March
event_type	Funnel Cloud
CZ_TYPE	C
cz_fips	73
cz_name	SAN DIEGO
WFO	SGX
begin_date_time	03/27/2007 15:03:00
cz_timezone	PST-8
end_date_time	03/27/2007 15:03:00
injuries_direct	0
injuries_indirect	0
deaths_direct	0
deaths_indirect	0
damage_property	0.00K
DAMAGE_CROPS	0.00K
SOURCE	Trained Spotter
MAGNITUDE	*(null)*
MAGNITUDE_TYPE	*(null)*
FLOOD_CAUSE	*(null)*
CATEGORY	*(null)*
TOR_F_SCALE	*(null)*
TOR_LENGTH	*(null)*
TOR_WIDTH	*(null)*
TOR_OTHER_WFO	*(null)*
TOR_OTHER_CZ_STATE	*(null)*
TOR_OTHER_CZ_FIPS	*(null)*
TOR_OTHER_CZ_NAME	*(null)*

Service's website is displayed in figure 1 (end of document).

4.2. NOAA Storm Events Database

The Storm Events Database is a collection of storm reports maintained by the National Oceanic and Atmospheric Administration (NOAA) (sto, 2012). Each report documents a particular weather-related event and includes valuable information such as geographic locations, dates/times, injuries, deaths, and property damage. Table 1 displays an example entry from this database.

5. Technology Summary

5.1. Python 2.7

5.1.1. NumPy

NumPy is a highly efficient and robust scientific computing Python package (num, 2012). We will use NumPy to store our large data sets

5.1.2. NWS-CAP-Parser

NWS-CAP-Parser is a Python class that parses the XML of a Common Alerting Protocol (CAP) message issued by the National Weather Service (NWS)(nws, 2012).

5.1.3. Natural Language Toolkit (NLTK)

NLTK is a leading platform for building Python programs to work with human language data (nlt, 2012). We will be using it extensively for its tokenization, stopword removal, stemming, and n-gram functions.

5.1.4. Scikit-Learn

Scikit-learn is a simple and efficient machine learning package for Python (sci, 2013). It will be used to train various models after the feature engineering and extraction stages are complete.

5.2. Google Compute Engine

Google Compute Engine allows large-scale computing workloads on the same infrastructure that runs Google Search, Gmail and Ads (com, 2012). Due to the size of the data sets that will be discussed in section 4, and the fact that our work may be used in their products, Google was kind enough to donate ample computing resources via this new service.

6. Proposed Milestones

Week 1 (02/25 - 03/03)

Draft formal project proposal
02/28 - Project Proposal Due

Week 2 (03/04 - 03/10)

Match NWS alerts with events in Storm Event Database.

Week 3 (03/11 - 03/17)

03/12 - Revised Project Proposal Due

Extract features from alerts. We intend on using most of the fields within the alert as a separate feature. The free text fields will likely be tokenized, have stopwords removed, stemmed, and have bigrams generated.

Week 4 (03/18 - 03/24)

Spring Break

Continue extracting features from alerts. Depending on the total number of alerts and time constraints, we may opt to go back and extract trigrams instead of bigrams.

Week 5 (03/25 - 03/31)

03/27 - Monthly Meeting with Project Advisor (Alice)

Build property damage model and test. We want to see if we can make a model that says "given a CAP alert, predict the amount of property damage that the storm will cause"

Week 6 (04/01 - 04/07)

Build death prediction model and test. Similar to the damage model, we would like to predict the number of deaths that will result from the storm.

Week 7 (04/08 - 04/14)

04/08 - Monthly Meeting with Project Advisor (Alice)

Build injury prediction model and test. Again, similar to the damage model, one of the factors that goes into a storm's severity is the number of injuries.

Week 8 (04/15 - 04/21)

Build framework that takes a CAP alert, extracts features, runs them through three prediction models, and computes a final storm severity.

Week 9 (04/22 - 04/28)

Gather all results and begin drafting final report.

Week 10 (04/29 - 05/05)

05/02 - Polished Draft of Project Report Due

Start polishing the final report.

Week 11 (05/06 - 05/12)

05/10 - Final Project Report Due

Place finishing touches on the final project report.

References

Google compute engine - computation in the cloud, 2012. URL https://cloud.google.com/products/compute-engine. [Online; accessed 27-February-2013].

Natural language toolkit, 2012. URL http://nltk.org/. [Online; accessed 26-February-2013].

Scientific computing tools for python - numpy, 2012. URL http://www.numpy.org/. [Online; accessed 26-February-2013].

morrissimo/nws-cap-parser - github, 2012. URL https://github.com/morrissimo/NWS-CAP-parser. [Online; accessed 26-February-2013].

Storm events database — national climatic data center, 2012. URL http://www.ncdc.noaa.gov/stormevents/. [Online; accessed 27-February-2013].

scikit-learn: machine learning in python, 2013. URL http://scikit-learn.org/stable/. [Online; accessed 26-February-2013].

Machine learning cheat sheet, 2013. URL http://peekaboo-vision.blogspot.com/2013/01/machine-learning-cheat-sheet-for-scikit.html. [Online; accessed 12-March-2013].

Joachims, Thorsten. Text categorization with support vector machines: Learning with many relevant features. *European Conference on Machine Learning*, 1398:137–142, 1998. URL http://www.cs.cornell.edu/People/tj/publications/joachims_98a.pdf.

```
<alert xmlns="urn:oasis:names:tc:emergency:cap:1.1">
...
  <sent>2013-03-11T20:23:00-05:00</sent>
  <status>Actual</status>
  <msgType>Alert</msgType>
  <scope>Public</scope>
  <note>Alert for Hancock; Pearl River (Mississippi) Issued by the National Weather
     Service</note>
  <info>
    <category>Met</category>
    <event>Flood Warning</event>
    <urgency>Expected</urgency>
    <severity>Moderate</severity>
    <certainty>Likely</certainty>
...
    <effective>2013-03-11T20:23:00-05:00</effective>
    <expires>2013-03-13T02:23:00-05:00</expires>
    <senderName>NWS New Orleans (Southeastern Louisiana)</senderName>
    <headline>Flood Warning issued March 11 at 8:23PM CDT until further notice by NWS
       New Orleans</headline>
    <description>...THE FLOOD WARNING CONTINUES FOR THE FOLLOWING RIVERS IN
LOUISIANA...MISSISSIPPI...
THE PEARL RIVER NEAR BOGALUSA AFFECTING ST. TAMMANY...WASHINGTON...
HANCOCK AND PEARL RIVER COUNTIES/PARISHES
THE PEARL RIVER NEAR PEARL RIVER AFFECTING ST. TAMMANY...HANCOCK AND
PEARL RIVER COUNTIES/PARISHES
THE FLOOD WARNING CONTINUES FOR
THE PEARL RIVER NEAR PEARL RIVER.
* UNTIL FURTHER NOTICE.
* AT 7:00 PM MONDAY THE STAGE WAS 14.1 FEET.
* MINOR FLOODING IS OCCURRING AND MINOR FLOODING IS FORECAST.
* FLOOD STAGE IS 14.0 FEET.
* FORECAST...THE RIVER IS EXPECTED TO TO FALL TO A STAGE AT OR NEAR
12.0 FEET ON EARLY THURSDAY MORNING MARCH 14TH THEN RISE TO A STAGE
NEAR 15.0 FEET ON WEDNESDAY MARCH 20TH.
* IMPACT...AT 14.0 FEET...SECONDARY ROADS TO THE RIVER AND THROUGHOUT
HONEY ISLAND SWAMP ARE INUNDATED. PROPERTY IN THE VICINITY OF THE
GAGE IS FLOODED THREATENING ABOUT 20 HOMES ALONG THE LEFT BANK.
* IMPACT...AT 13.0 FEET...THE EAST AND WEST CHANNELS OF THE RIVER WILL
BEGIN TO MERGE. HONEY ISLAND SWAMP TRAILS WILL BE UNDER WATER AS
INUNDATION OF THE SWAMP BEGINS.</description>
    <instruction>FORECAST CRESTS ARE BASED UPON RAINFALL THAT HAS OCCURRED ALONG WITH
ANTICIPATED RAIN FOR THE NEXT 12 HOURS. ADJUSTMENTS TO THE FORECASTS
WILL BE MADE IF ADDITIONAL HEAVY RAINFALL OCCURS.
DO NOT DRIVE CARS THROUGH FLOODED AREAS. REMEMBER...TWO FEET OF
RUSHING WATER CAN CARRY AWAY MOST VEHICLES INCLUDING PICKUPS. TURN
AROUND AND DON`T DROWN.
A FOLLOWUP PRODUCT WILL BE ISSUED LATER. STAY TUNED TO NOAA WEATHER
RADIO...LOCAL TV AND RADIO STATIONS...OR YOUR CABLE PROVIDER...FOR
THE LATEST INFORMATION. THE LATEST GRAPHICAL HYDROLOGIC INFORMATION
CAN ALSO BE FOUND AT WEATHER.GOV.</instruction>
...
    <area>
      <areaDesc>Hancock; Pearl River</areaDesc>
      <polygon />
      <geocode>
        <valueName>FIPS6</valueName>
        <value>028045</value>
      </geocode>
      <geocode>
        <valueName>FIPS6</valueName>
        <value>028109</value>
      </geocode>
...
    </area>
  </info>
</alert>
```


Figure 1. Example National Weather Service Alert from http://alerts.weather.gov/cap/us.php?x=1