Capstone Project Milestone Report

Bernard Fraenkel

December 29, 2015

Synopsis

Objective

This report addresses two important questions related to storm data provided by the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database:

1. Across the United States, which types of events are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

The raw data is available here: Storm Data [47Mb]

Findings

- Tornado is by far the event that causes the most fatalities and injuries, causing 37% and 65% of fatalities and injuries, respectively.
- The total contribution of all heat-related events causes 20% of Fatalities and 2/3 of Injuries
- The top 10 types of storms combine to over 91% of the total damage. Most of the damage is caused to Property (compared to Crops)

Processing Pipeline

The data is processed according to the processing pipeline:

- The original source data is "tokenized" and stored in tokenized form, to be used by the following stages in the pipeline

Findings

To state the obvious, the complexity of NLP is immense - even when limiting oneself to "next word prediction". The following are ideas that jumped out by simply dealing with the data:

- Each of the data sets contains 30-40 Million words. As a consequence, writing efficient code is critical to not only experiment, but simply to process the data in a timely manner: i.e 2-3 hours per processing step
- Current implementation uses a simple tokenizer, which breaks up content based on punctuation. Obvious improvements would include:
-

Plans for Shiny App