Watch Flu Spread

Big Data for Social Good Challenge

Brian Norris; Alec McGail; Matt Gilliam

Bob Boehnlein; John Springer; Wei Kao

Anuja Rayarikar; Yanan Tao; Yuanhsin Huang

03-03-2015

Flu seasons are unpredictable and can be severe. According to the Centers for Disease Control and Prevention (CDC), it is estimated that more than 200,000 people are hospitalized from seasonal flu-related complications in the United States each year. In our project, we attempt to monitor and predict flu trends in real time, using query-based flu estimates as well as social media data. Our analysis is done at big data scale and empowered by IBM's analytics for Hadoop.

A flu surveillance system can make enormous social impact. According to CDC, between 1976 and 2006, estimates of flu-associated deaths in the United States range from a low of about 3,000 to a high of about 49,000 people. Seniors, young children, pregnant women, and people with certain health conditions can be more susceptible to the flu virus. Our project aims to help people who are at high risk for serious flu complications by keeping them informed when there is high flu activity at places where they live or plan to go. Furthermore, our goal is to make the information as accessible as possible so they can stay constantly vigilant.

During the process of our project we gained insight to the factors that influence the spread of flu as well as some preconceived assumptions that did not contain correlations. In addition to the insights gained from the research of our chosen domain, we discovered a coding issue within the Node-RED platform that needed to be resolved in order for it to work with scaled HDFS data. By identifying this issue we were able to improve this system for our use in the Big Data for Social Good Challenge as well as future users of this service.

A version of our finished application can be found at http://watchfluspread.mybluemix.net/. On this site is the representation of our application, which is a map displaying the model output historic and forecasted impact of flu in the US distributed by county. In addition, there contains basic information of our project, a list of the tools used, a visual representation of the model in Node-RED, and information of the team that completed is project. Our team consists of students of Purdue University and data professionals from Perscio, a data consultancy based in Indianapolis.

Our motivation for this project was to create a visualization of the flu that could be displayed on a map of the US in a way that relates to a weather map. By doing this the end user can benefit in seeing both the historic and forecasted dispersion of flu incidents in the US to gain a better understanding of risk in their area. An additional motivation was to improve upon current models of flu prediction in identifying the weakness in variables and diversify that risk by including additional variables.

As it has been shown in past years, Google Flu Trend can be inaccurate during peak flue season based on a flood of individuals who are searching about the flu, but may not in fact have symptoms. Our hypothesis was that by incorporating additional variables from other correlating data we could offset the variation seen in GFT to actual CDC reported cases. In addition to GFT, we incorporated several other sources of data into our model as well as utilized for exploratory data analysis. These include: Weather Underground data, Daily Global Weather Measurements, Twitter, National Household Travel Survey, and Census data.

By incorporating these additional data sources we were able to discover the relationship of social media, geographic proximity, travel, ambient temperature, and population density to the spread of flu.

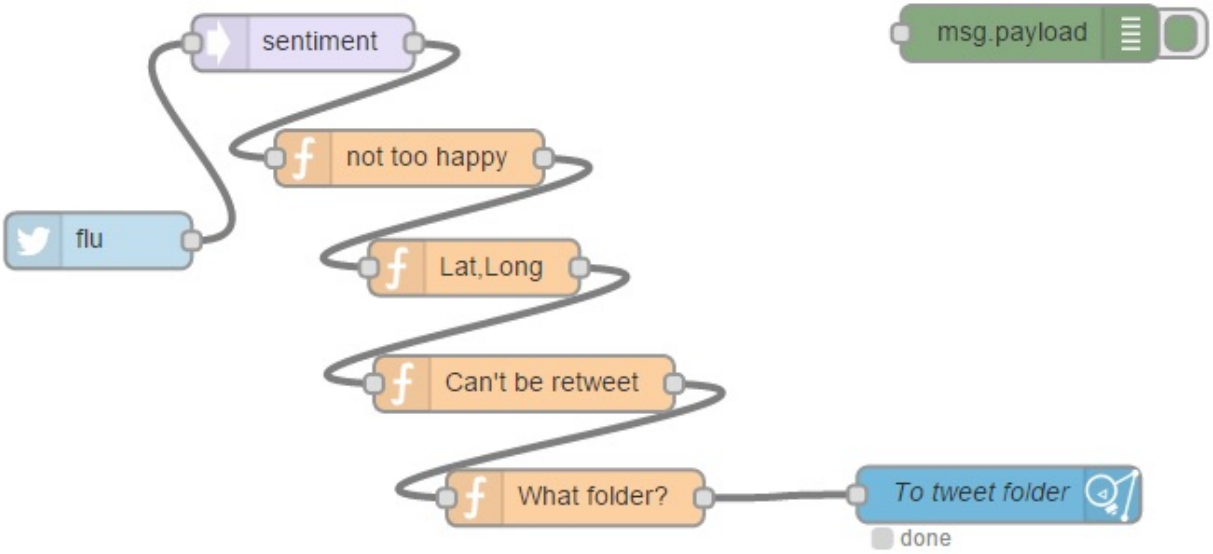
The tools we used in both the exploratory phase and building the application included R, SAS, HDFS, IBM Analytics for Hadoop, Node-RED, JavaScript, and D3.

Model Structure

To begin the process of flu forecast over time, the current volume of flu cases and distribution among counties is needed. To do this the model utilizes Google Flu Trends and Twitter data, which contain geo location. If there is no Tweet data with location in a given county, the current flu volume is equally distributed. Tweets with the word "flu" in it are indicative of flu incidents at or around the location it was tweeted from. We assume a Gaussian falloff, such that at each coordinate in the country (lt, lon), the concentration of flu is proportional to

$$\sum_i e^{-\frac{(lt-lt_i)^2+(lon-lon_i)^2}{d_{\frac{1}{2}}}}$$

fig 1: Tweet ingest

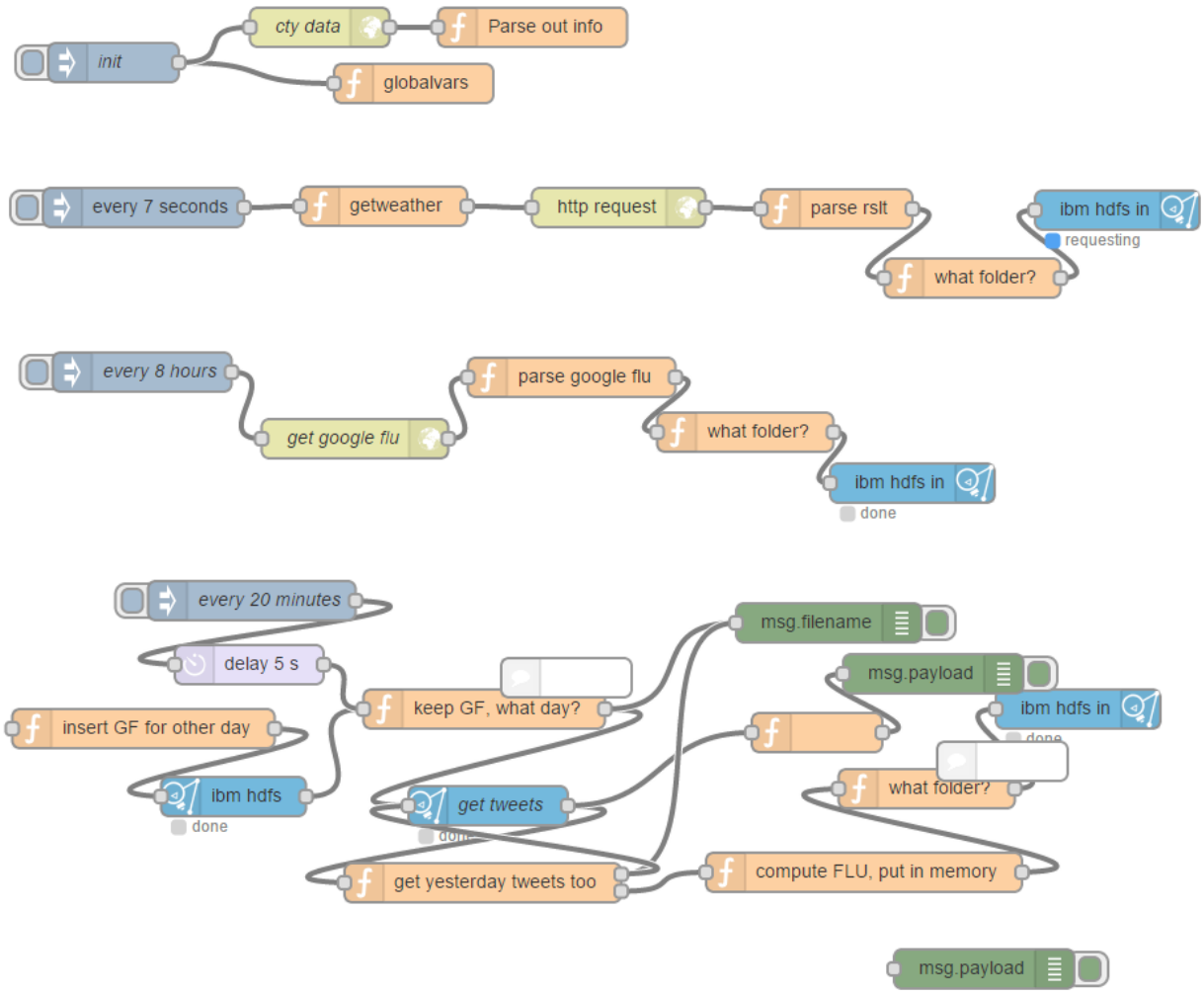


This sum iterates over all the tweets with the word "flu" in it. $(lt - lt_i)^2 + (lon - lon_i)^2$ is the squared distance between the coordinates (lt_i, lon_i) and (lt, lon) in degrees. $d_{\frac{1}{2}}$ is the distance in degrees at which the influence of a tweet is half (1/e to be precise) its influence at distance 0. We assume that the flu is equally prevalent in those who tweet about it and those who don't tweet about it, and that the number of flu cases in any state is equal to the estimations given by the latest Google Flu Trends numbers.

Once the current state of flu cases is determined among counties the forecast of flu over time can be generated to asses flu risk in any given area. To start we are given the current number of cases in each county in the United States and are tasked with producing the same numbers for tomorrow. This process includes weather data by county and population density by county. The weather data used is the "feels like" temperature as this contains adjustments such as wind chill and humidity. There are several incorporated assumptions that were derived from our research of how the flu impacts an individual al the exploratory data analysis that was done towards the beginning of the process. The first assumption is that the flu lasts seven days. While this varies from person to person, actual data on the time an individual is impacted by the flu is not currently available. So using an average figure for this is necessary. The second assumption is that people who have the flu in a given county are distributed evenly in the range of the disease. That is, 1/7 of the sick population

gets better each day. Again, while certain strains of the flu or individuals impacted may be concentrated and have prolonged symptoms, this data is not publicly available. The third assumption is the probability of a person spreading the flu to another person in a county is directly proportional to the population density of that county. The flu is spread by an individual being within six feet of someone who already is infected. Areas that have a higher population density will contain a higher probability of transfer based on proximity of individuals. The fourth assumption is that the probability of a person moving to another county is inversely proportional to the distance between the two counties, and proportional to the population of both. As the distance between geographical areas increases, the probability of an infected individual traveling between these areas decreases. Our initial exploratory data analysis on temperature and correlation to flu transfer was based on the current Google Flue data and Weather Underground measurements around the country. This displayed a positive relation between a warmer temperature and flu. As we dug deeper, we discovered that this correlation was spurious, and a historical look at all the weather measurements across the country back several years showed that it's quite hard to use temperature in a predictive model of the flu. This is not to say that temperature does not affect the spread of flu. As Yanyan explained to our team, each variety of the flu has a specific temperature in which it thrives. But that does not mean the flu hits at a specific temperature, for once a person has a specific strand of the flu, they won't get it again. All these factors considered, the best model without knowing the details of each strain would ignore temperature, a fundamental insight when creating this model.

fig 2: reoccurring process of data ingest for current state of flu



Here we know what our data was telling us in the last N days, and we want to find out what the best parameters would have been to predict the flu well. First is to determine an error function. That is, given our prediction and the measured value of flu incidents in each county, how well did this perform? This can be accomplished by using a simple sum-of-squares error metric, as used in other areas of statistics. The computation is as follows:

$$\sum_c (predicted_c - actual_c)^2$$

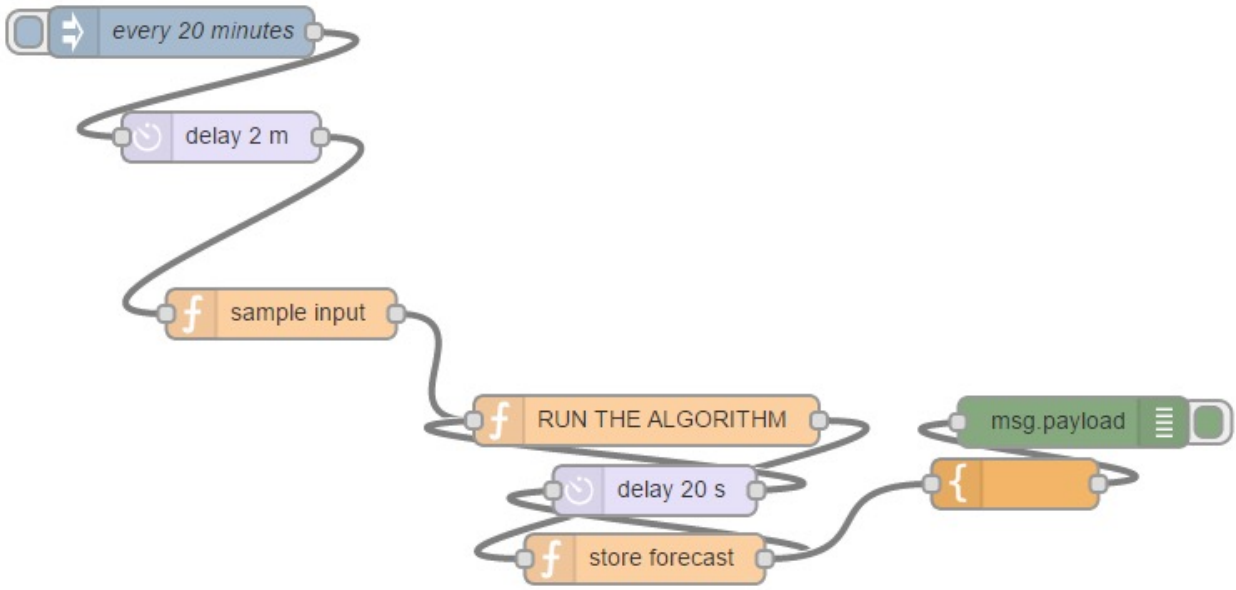
$predicted_c$ is the predicted flu number for county c.

$actual_c$ is the actual flu number for county c.

The counties with a higher population density require a greater priority in the distribution based on the probability of flu transfer. Thus each county's error should be weighted by the population of that county:

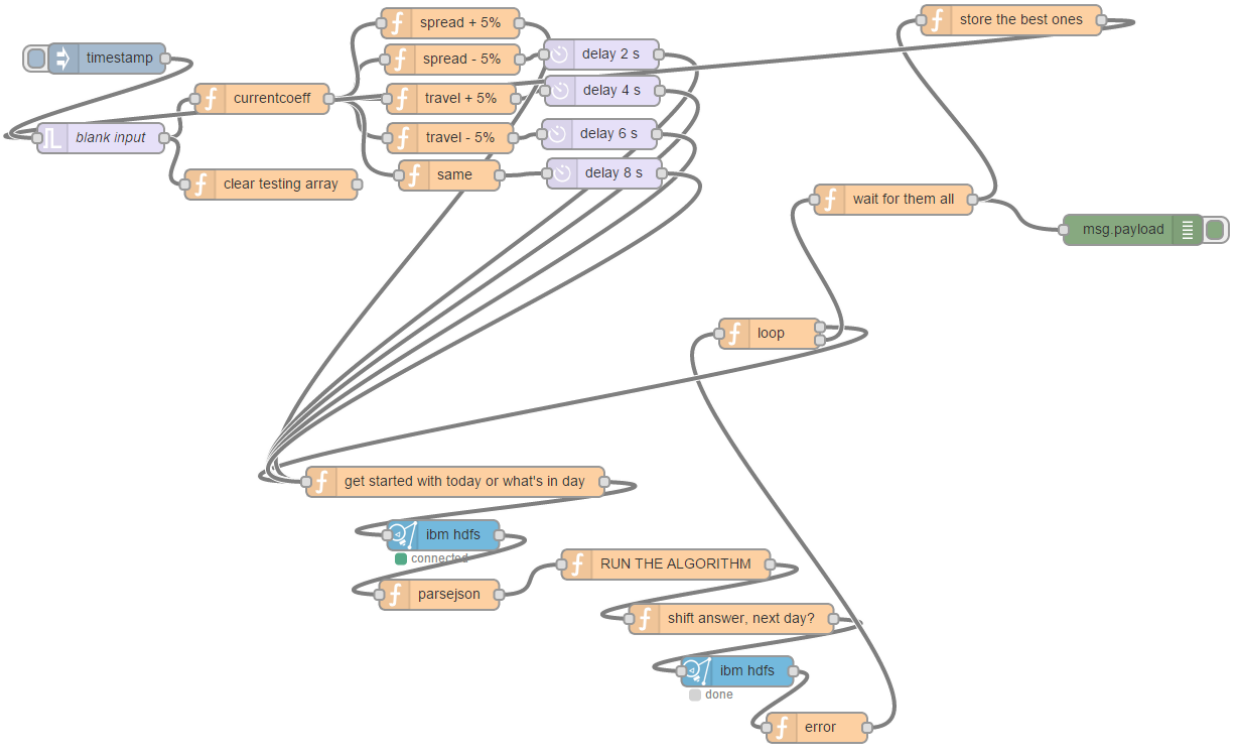
$$\sum_c population_c * (predicted_c - actual_c)^2$$

fig 3: flu forecast development



The final aspect of the model development is to optimize the regression so that improved prediction will take place as additional data is added. This will test what values for the parameters best fits the data. Essentially this algorithm will say "How should I slightly change the parameters to best account for the data we've seen?" It does this by taking the parameters one step in a direction, and running the model with those parameters. The optimization is done in varied directions, and based on error rate; decision on further optimization of parameters is decided. This process continues until performance of the outcome decreases. As additional data is imported into our set upon submittal, the application will continue to improve in performance. Visual representation of the model structure can be seen in images on the application website. This structure is graphical in nature as it was built in the Node-RED platform.

fig 4: parameter optimization



Bug Correction

During the implementation of our model into the graphical programming platform Node-RED, a bug in the source code was identified relating to file size transfer capabilities. Within IBM's HDFS connector to IBM Analytics for Hadoop, the method of receiving data is through an HTTP request. This HTTP request is broken into chunks when the file size becomes large, specifically above around 5kb. The issue that existed was that instead of appending each chunk to the returned message, it simply set the returned message to be the chunk. Therefore when the result came back, the data only contained the last chunk. This is an issue because when files sizes exceed 5kb the transfer between HDFS and IBM Analytics for Hadoop in Node-RED is unusable. Our developer Alec identified this issue and realized that with a relatively simple change to the source code this could be resolved. After pushing the change to the repository the change was accepted and this is now a very usable tool. The fix that was accepted can be reviewed at Github:

https://github.com/amcgail/iot-nodered/commit/9b56e8d0b0a47348afa27956d5769354e21dfdf0

Visualization

Once our predictive model was formulated it was necessary to visualize this in a gradient color scale on a map of the US for the end user to understand flue risk in a given area. This was built in a way where the user can select a button to change the historic or forecasted representation of flu impact that continues to update around the current day. The code for this visualization was made using D3 as it provides both the flexibility and quality we desired. More can be read about creating map visualizations' using D3 in the following links:

http://eyeseast.github.io/visible-data/2012/12/14/mapping-with-d3/

http://bl.ocks.org/mbostock/raw/3750900/us-counties.json