Healthcare twitter analysis.
Project breakdown.

Theofilos Papapanagiotou

July 15, 2014

1 Algorithms Research

Classification of the tweets based on stages of a disease, like diagnosis, treatment, etc. Filter out the spam, most probably generated by drug advertisments, etc. Train the clustering model using the Jan-May dataset, and test using the June dataset. Produce a vocabulary which might help the disease stage classification, like the sentiment analysis of Introduction to Data Science [1].

2 Technology / Framework Research

Use the experience gained in Octave from Andrew Ng's Machine Learning [2] and using R from Data Science Specialization [3], to learn another language libraries on Data Science, Python.

The given dataset contains date, user, url and tweet in a csv, which are sufficient for the analysis. It would be cool though, to load the full tweet metadata in a mongo calling the twitter API by tweetId. Mongolab free 0.5GB db might be sufficient for the training dataset.

If we need to scale more and mapreduce our algorithm, utilization of a few hadoop instances in ec2 might be cooler.

3 Business/ Domain Research

Search heathdata.gov, datahub.io, enigma.io, gapminder.org for open data on healthcare to correlate the disease related tweets as proposed by Pratik.

4 Visualization Research

Extend the gained ggplot2 knowledge in Python world, to produce high quality graphs and present the results of the project. Make sure that everything is reproducible and delivered in iPython Notebook.