Shraddha Patel A20499171

CSP554—Big Data Technologies

Assignment #1 (Modules 01a & 01b, 10 points)

1. (5 points) Answer each of the following questions about the article in just one to three sentences each:

1) What was the problem with the Google flu detection algorithm?

- GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC).
- As per paper GFT was overestimating flu occurrence [Graph in the Article]

2) **What is big data hubris?**

- Big data hubris is an assumption that big data/huge amount of data can substitute or replacement for traditional data collection and analysis.
- Its belief that huge amount of data always leads to better result.
- Article states that big data has great potential but still we cannot ignore validity, reliability and dependencies among data while analyzing it.

3) **What approach could have been used to improve the Google flu detection algorithm?**

- Google flu detection algorithm could have been improved by combining GFT data with near real time health data.
- For example, by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, can substantially improve on the performance of GFT or the CDC alone. [Graph in the Article shows less error]

4) **What is "algorithm dynamics?"**

- Algorithm dynamics are the changes made by engineers to improve the commercial service and by consumers in using that service.
- Constant development/changes of google search algorithms to be the main reason for failure of GFT.

5) **What aspect of algorithm dynamics impacted the Google flu detection algorithm?**

- "Blue team" dynamic- where algorithm producing the data has been modified by the service provider in accordance with their business model
  - Google reported modifications in June 2011 and February 2012.
- To improve services to customer, Google also changing the data-generating process like providing useful information quickly, promote more advertising revenue
- GFT uses the relative prevalence of search terms in its model, modifications in the search algorithm can adversely affect Google flu detection algorithm.

2. (5 points) Set up an Amazon Web Services (AWS) cloud account, if you don't already have one (see below for details), and then follow the tutorial about how to work with a storage service called S3. Since we will do most of our assignments using AWS, this will get you started. In a while we will come to understand S3 as one critical element of a big data processing architecture known as the "data lake."

   a. To receive credit for this question, provide a screen shot showing the S3 bucket you have created. The bucket name should be named something like "YourIITId-CSP554", for example: "A1234567_CSP554"