1 Abstract

This project focuses on a full data pipeline -collection, cleaning, analyzing and reporting of data related to a topic of our choosing. Given the nature of this project, an appropriate topic was the recent data scandal involving Facebook and Cambridge Analytic. Twitter and the New York Times were data mined using keywords "facebook" and "cambridge analytic".

2 Raw Data Collection

In order to collect the raw data, python programs were written to communicate with the Twitter and NY Times api's. Identifying tweet and article information was exported to a database for storage, written out to a .csv as a master list. Each tweet was also written out to an individual text file for use with the Hadoop 'hdfs'.

3 Pre-processing

The NYtimes .csv's in their raw form only contained an 'article ID', 'url' and 'date'. A python program was developed to use this information to scrape the url's and export the article content to individual text files, again for use with Hadoop.

4 MapReduce

Python was again used to write mapper and reducer programs for word counts and word co-occurence. The data for each keyword ("facebook", "cambridge analytic") and source (Twitter, New York Times) were fed into a virtual machine with access to Hadoop and processed with each map reduce combination.

5 Visualization

In order to visualize and present the results. A simple webpage was created using embedded d3.js code. This webpage includes interactive bar-charts to visualize word counts for each source and keyword, as well as interactive word clouds to demonstrate the same.

6 Moving Forward

As far as future use, this project was compartmentalized in such a way that it can easily be recycled and used with a different set of keywords. Each stage in the pipeline is modular and can be improved upon individually without having to re-code the whole project.