Incorporating Big Data Tools for Social Media Analytics in a Business Analytics Course

Amir H. Zadeh
Information Systems and Supply Chain Management Department
Wright State University
Dayton, OH 45435, USA
amir.zadeh@wright.edu

Hamed M. Zolbanin
MIS, Operations & Supply Chain Management, Business Analytics Department
University of Dayton
Dayton, OH 45469, USA
hmzolbanin@udayton.edu

Ramesh Sharda
Management Science and Information Systems Department
Oklahoma State University
Stillwater, OK 74078, USA
ramesh.sharda@okstate.edu

ABSTRACT

The age of big data drives the need for emerging technologies to enable scalable analytics on massive, rapidly generated, and varied data. It requires "data scientists" with deep knowledge of managing the six Vs of big data: volume, velocity, variety, volatility, veracity, and value. As a result of this trend, new analytical tools are being taught in business analytics (BA) programs to foster students' development of this critical competency. In the era of big data, social media analytics has become an increasingly important topic. In this article, we present a social media analytics exercise that can be easily added to any analytics course or to any course in which students gain exposure to social media and big data technology. The case scenario is built upon using flu activity data on Twitter to extend the monitoring of flu outbreaks. Our analytics framework comprises temporal, spatial, and text mining. We demonstrate the use of IBM InfoSphere BigInsights – a Hadoop-based platform – for implementing our framework. The exercise guides students through a big data social media analytics journey that enables them to understand different aspects of social media as the main source for big data and develop skills in this emerging area. Our framework and the exercises are general enough to be used even if the instructor opts to use a different technology stack. The described approach was used in a class at a large university. At the end of the exercise, 27 students majoring in business analytics participated in a survey and expressed satisfaction with their learning process.

Keywords: Big data, Social media, Business analytics, Data visualization, IBM BigInsights

1. INTRODUCTION

During the last 10 years, the analytics field has seen unprecedented growth, and analytic architectures have gone through a substantial transition from enterprise data warehouses to big data platforms such as Hadoop (Chambers and Dinsmore, 2014). The paradigm shift is driving the need for data scientists to have a deep understanding of the volume, variety, velocity, and veracity of data that is processed in these platforms (Russom, 2011). Many vendors such as IBM, Teradata Aster, Hortonworks, Cloudera, and others, have established open-source, Hadoop-based analytics platforms to help businesses address their big data needs. Responding to the industry need for business professionals to employ big data technologies for data-driven decision making, the Association to Advance Collegiate Schools of Business (AACSB) recognized the need for students to understand big data across all business disciplines (Janvrin and Watson, 2017), and several academic

programs have included big data-related courses in their curricula to keep pace with changes in the big data era (Guthrie, 2013; Wixom et al., 2014). With social media as a public source for big data, social media analytics content is being developed and used in big data-related courses to introduce students to big data technologies to collect, process, analyze, and visualize big data and create business value through social media (Khan, 2018; Shang, 2019). Such instructional effort should be focused on adopting open-source projects due to the merits they offer, and its use in hands-on training and experiential learning, in order to achieve broader educational objectives in big data analytics education. While curriculum development and course design have been the center of big data education over the past several years, little attention has been paid to developing and implementing instructional and pedagogical resources that can be used in the classroom to help students gain expertise in big data analytics through practical hands-on training (Podeschi and DeBo, 2019). This paper presents a robust framework on big data and social media analytics that can be easily added to any analytics course or any course in which students are exposed to social media and big data technology.

The framework provides an end-to-end learning experience for students and enables them to gain skills in critical areas related to big data – collection, storage, preprocessing, integration, modeling, and visualization. IBM InfoSphere BigInsights was adopted as an open-source Hadoop framework with NoSQL database as a data storage (to store large volumes of unstructured data from Twitter) and MapReduce techniques for parallel processing to help the students develop the skills necessary to process large data sets, visualize trends, and make insights that drive decision making. The framework was implemented in a data mining course within a business analytics program where 27 business students participated voluntarily in the study. Besides providing a detailed, hands-on exercise that implements our pedagogical framework for big data-related courses, we cover important concepts and tools in big data and social media analytics and discuss students' feedback and learning outcomes for the exercise using a survey that was designed to rate the effectiveness and usefulness of the proposed pedagogical framework. This manuscript, therefore, offers several contributions. First, a pedagogical framework is developed for big data-related courses which consists of multiple phases related to big data, including collection, storage, preprocessing, integration, modeling, and visualization, and offers an end-to-end teaching and learning experience for both students and instructors. Second, using the framework, a step-by-step exercise is presented to illustrate various concepts and tools for big data and social media analytics. During the exercise, data from various sources, such as Twitter, Centers for Disease Control and Prevention (CDC), and Google, are collected, cleaned, integrated, analyzed, and visualized to track the spread of the flu across the U.S.

This teaching case also features IBM InfoSphere BigInsights (and its associated components) which is a big data analytics platform based on the open-source Apache Hadoop and can be used for efficiently analyzing massive volumes of data from various sources, such as social media platforms, web, or sensors. BigInsights uses the MapReduce programming model for parallel execution of various built-in and user-defined applications across a Hadoop cluster. In addition to the standard Apache Hadoop software, BigInsights provides additional technologies and programming languages with built-in accelerators to efficiently perform specialized operations over big data. JSON Query Language (JAQL) is designed to better support manipulation and analysis of semi-structured JavaScript Object Notation (JSON) data. BigSheets is a spreadsheet-style tool that supports scalable data exploration and visualization directly on a Big SQL table residing on Hadoop Data Storage File Systems (HDFS). Annotation Query Language (AQL) provides built-in libraries for advanced text analytics across vast amounts of semi-structured and unstructured documents. BigR is a platform for large-scale analytics on Hadoop that enables accessing, manipulating, analyzing, and visualizing data residing on HDFS from the R's user interface.

The remainder of this work is structured as follows. The next section provides a literature review on the topic, followed by our pedagogical framework on big data and social media analytics. The fourth section describes our experiment in detail followed by students' responses and assessment in section 5. The paper ends with a conclusion section.

2. LITERATURE REVIEW

As big data analytics continues to gain importance and prominence within organizations, the need to develop well-trained business professionals equipped with the necessary skills in big data analytics becomes more crucial (Provost and Fawcett, 2013; Khan, 2018; Jeyaraj, 2019). To overcome this challenge, many educational institutions have adapted to this increasing need by offering new courses for undergraduate and graduate programs aimed at developing student's big data skills to prepare them for the workplace. Big data analytics courses that provide students with soft and technical skills in acquiring, processing, analyzing, and visualizing large volumes of data are critical for the success of business professionals (Waller and Fawcett, 2013; Jeyaraj, 2019).

Modern business analytics are built upon the layer of big data available within most business organizations (Parks et al., 2018). Although big data analytics still follows the well-known CRISP-DM methodology for the most part, big data skills are very different from the traditional analytics skills (e.g., Java, SQL, R, Python, SAS, and Tableau). Big data analytics requires competencies in MapReduce, Hadoop, Hive, Pig, Flume, Mahout, and Spark for large-scale distributed data processing and analysis, which many business professionals may not possess today (Cegielski and Jones-Farmer, 2016; Verma et al., 2019). Other important big data skills are NoSQL databases, cloud and in-memory computing, and data virtualization. In addition, making sense of big data requires critical thinking, interpersonal, and communication skills that are key to the success of data-driven business professionals (Asamoah et al., 2017; Stanton and Stanton, 2020).

Although business schools are increasingly aware of the importance of big data skills for today's workforce, there is still little (but increasing) focus on big data within their curricula (Parks et al., 2018). Guided by AACSB standards and industry best practices, the curriculum design and course development efforts have been underway in educational institutions to offer big data courses and programs in order to address the industry's need for skilled big data professionals across all business disciplines. For example, Sledgianowski et al. (2017) reported

the use of the competency integration framework to integrate big data into the accounting curriculum. Schoenherr and Speier-Pero (2015) highlighted IT skills desired for successful supply chain professionals and illustrated how big data analytics can be implemented in the OM/SCM curriculum. Miah et al. (2020) used a design science-based approach to course development for data science programs in business schools and concluded that big data courses need to familiarize students with modern open-source big data platforms and empower them with real-world data in order to increase the likelihood of student success in the educational environment.

Several papers have reported experiments in which big data courses can be successfully designed and delivered to undergraduate and graduate students from multiple disciplines. Asamoah et al. (2017) discussed concepts, tools, and applications covered in a big data course. Some of the concepts included the three Vs of big data, Hadoop eco-systems, MapReduce, Distributed File system, and Hadoop sub-projects, such as Hbase, Hive, and Pig. Summaries of applications covered in the course were outlined as social media analytics, network analysis, and stream mining using a verity of big data tools, such as IBM BigInsights, Teradata Aster, Hortonworks, and Cloudera. Parks (2020) presented a pedagogic experience in which a big data analytics course in healthcare was designed and delivered. IBM Watson analytics platform was used to analyze large volumes of clinical, medical, and social media data for different health applications. Dinter et al. (2017) detailed course content for a big data management course and included summaries of several tutorials on Apache Hadoop, NoSQL databases, IoT applications using IBM Bluemix, and log analytics using Apache Splunk. Fowler et al. (2016) emphasized the increasing importance of NoSQL databases in today's environment and developed a teaching case to introduce NoSQL in relational database courses.

Other important big data technologies include cloud and in-memory computing, big data warehousing, and data virtualization (Song and Zhu, 2016). For instance, McLeod et al. (2017) examined SAP HANA as an in-memory database management system to cover big data concepts. Zadeh et al. (2020) included a big data lab in an ERP course to demonstrate how ERP systems utilize big data to create predictive and prescriptive decision-based models. Podeschi and DeBo (2019) evaluated two different methods of providing undergraduate students with exposure to Hadoop via cluster or virtual machines and suggested that Cloudera Quick Start is the most effective and efficient platform for classroom big data labs.

Social media data is an important part of the big data spectrum (Goh and Sun, 2015; Phillips-Wren et al., 2015), and social media analytics can be taught as a subcomponent in a big data course due to the abundance of unstructured data that online social networks offer. Most courses on social media analytics have been relying on existing tools such as NodeXL, and Gephi (Gruzd et al., 2016); however, big data solutions have become more common in today's business environment for handling social media data and correlating it with transactional data for greater insights. This paper presents a pedagogical framework on how big data analytics tools can be utilized for social media analytics. Guided by the proposed framework, a comprehensive exercise was designed and implemented in a data mining course within a business analytics program to enable the students to gain the skills necessary to work with large data sets, visualize trends, and make insights that drive decision-making.

More specifically, our exercise implements a big data analytics framework to build an application that uses Twitter and CDC data to track influenza activities. Prior studies have shown that Twitter can be used as a surveillance method for early detection of influenza outbreaks (Achrekar et al., 2011; St Louis and Zorlu, 2012; Achrekar et al., 2013; Broniatowski et al., 2013; Lamb et al., 2013; Lee et al., 2013; Li and Cardie, 2013; Zadeh et al., 2019). Although a social media-based surveillance system is limited to people who seek health-related information on the Internet and use social media to share their thoughts, feelings, and experiences publicly with their friends, we can sense the simultaneous presence of real and "digital" outbreaks from such data streams. We can read the city's public health status in real time, determine infectious hotspots, and highlight "must-know" facts for people visiting these areas. With such geo-located and time-stamped social media data streams at hand, consumer-generated social media data can be synthesized and mashed up with large sets of clinical and medical data to gain valuable insights (for patients and health care providers) into complex infectious disease conditions and anticipate public health crises. These data-driven applications can be used by public health officials involved in vaccination campaigns and in resource allocation and strategy implementation activities to combat the spread of diseases.

Our big data analytics framework comprises temporal, spatial, and text analytics. In the temporal analysis, students analyze whether Twitter data could indeed be adapted for the nowcasting of influenza outbreaks. In the spatial analysis, students map flu outbreaks to the geo-spatial property of Twitter data to discover patterns such as influenza hotspots. In the text mining phase, students extract and summarize useful information from textual data and identify patterns, such as popular types, symptoms, and treatments of the flu discussed in tweets.

3. FRAMEWORK

Our framework follows a big data analytics methodology to demonstrate various phases of social media analytics. As shown in Figure 1, the framework underlying this exercise consists of six modules: data collection, data storage, data preprocessing, data integration, data modeling, and data visualization. In the following, we will discuss each of these steps in detail.

3.1 Data Collection

The first stage of the framework is data collection (acquisition) which focuses on how to collect data from one of the well-known sources for big data, Twitter. A Twitter crawler collects tweets using the open-source Flutrack Twitter Streaming Application Program Interface (API) (Talvis et al., 2014) and forwards it to Hadoop. The crawler is written in JAQL (see the Appendix). The JAQL program perpetually creates JSON files from the incoming tweet stream and stores them in the Hadoop Distributed File Systems (HDFS). The monitoring words used as tags were "influenza" and "flu." The Flutrack API filters out tweets with false or nonexistent location coordinates. Only geo-

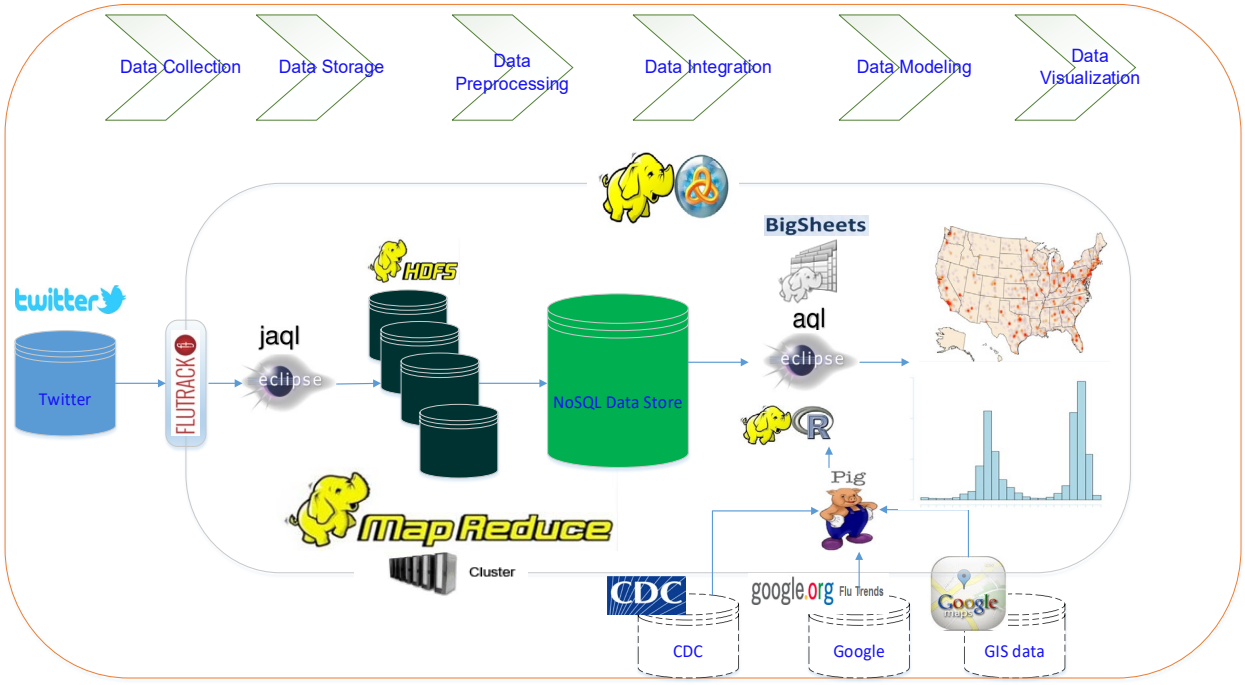

Figure 1. Pedagogical Framework

located tweets and tweets whose geolocations can be extracted from user profile locations are returned to the Hadoop database.

Big data is about analyzing massive amounts of data that are often collected from multiple sources. We used other publicly available data sources to make our scenario as realistic as possible and to address the serious issues with respect to data integration and analytics in big data education. The actual flu data was obtained from the CDC (http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html) using the R-package cdcfluview (https://cran.r-project.org/web/packages/cdcfluview/index.html). The CDC is the leading public health agency in the U.S. which monitors illness activities by collecting data from medical institutions, collating reports, and publishing them weekly. See the Appendix for the R script.

3.2 Data Storage

In the big data world, data is stored in HDFS where queries are parallelized through MapReduce (Dean and Ghemawat, 2008). As a data storage system, we use HDFS, which is also a part of BigInsights, to store the data collected from Twitter. BigInsights includes an integrated console that shows the status of individual Hadoop components such as NameNode, DataNode, JobTracker, Task Tracker, and Zookeeper.

3.3 Data Preprocessing

Data preparation and preprocessing is the most crucial process of mining massive datasets and can take up to 80 percent of the time and effort involved in a project (Piramuthu, 2003; Zolbanin et al., 2015). This process includes understanding the volume, variety, and velocity aspects of the data to be processed (Douglas, 2001), understanding what the data represent, binding the data streams, transforming the data for storing in proper format or structure for querying and analysis purposes, filtering, and reducing the data.

Twitter's API delivers data in the JSON format, a simple and human-readable way of storing data. JSON objects contain the tweet text along with metadata, such as the time, the sensor-based geo-location (longitude and latitude) associated with the tweet, and user profile information, which includes self-reported user information such as the user's real name and location. These metadata can be used to geo-locate the tweets (Dredze et al., 2013; Kumar et al., 2014). Using JAQL, a query language for JSON on big data and an "SQL-like" component of the BigInsights Hadoop platform, the unstructured JSON data is transformed into a structured format. The new dataset consists of multiple data points of individual tweets. Next, using appropriate JAQL commands, JSON files are converted into a simpler structure and stored in HDFS as a comma-delimited file. For each tweet, information such as username, time stamp, geographic location, and text were recorded. A list of the variables along with their description is provided in Table 1. Note that the JAQL plugin for Eclipse, which is incorporated within BigInsights, guides students through the entire process.

Variable	Description
Username	Name associated with the user who posted this tweet.
Time stamp	Long integer that represents the number of seconds between the Unix Epoch and the time of tweet generation.
Longitude	Indicates longitude of the tweet's location.
Latitude	Indicates latitude of the tweet's location.
Text	A short text message limited to 140 characters posted by a user.

Table 1. List of Variables Included in the Twitter Flu Dataset

- List all functions provided by the current version of BigR.
- List files on the BigInsights file system.
- Locate tweets file in the HDFS.
- Move the data from the Hadoop server (which is a *BigR.frame*) into a local machine as a *data.frame*.
- Explore the structure and dimension of the dataset — for example, how many rows and columns are in the dataset.
- Use the *BigR.sample* function in BigR to draw a random sample from the *tweets* data.
- Examine the class and the dimension of the random sample.
- Use random sampling support in BigR to split the tweets data into training set (~70%) and test set (~30%).
- Understand the R *data.frame* objects are held in memory and may exhaust memory if the data is very large.

Table 2. Basic R Tasks to Explore the Twitter Data

Twitter text is noisy with linguistic errors and idiosyncratic style (Ghiassi et al., 2013) and needs some transformations, such as changing letters to lower case; stemming words; and removing punctuations, URLs, numbers, stop words, and exotic characters, among others. In this phase, several natural language processing (NLP) algorithms, such as parsing, tokenization, stemming, synonyms, and parts of speech, are used to build a term-document frequency matrix. First, common text processing techniques, such as stop words removal and word stemming, are applied to the tweet corpus. The general English stop words list is used to remove common words. Also, the tweet dataset is cleaned from the "exotic" characters within the tweets. As a way to remove such noise, the tweet texts are converted into Unicode (utf-8). This results in a few "NA" entries associated with words that could not be handled. Then, the "NA" entries need to be removed from the entire collection.

Finally, students use word stemming techniques to consolidate various word forms derived from an identical stem. In this step, they deploy a combination of R and BigR functions to clean up their Tweets data located in the BigInsights/Hadoop cluster. BigInsights includes IBM's BigR feature, which allows the R algorithms to execute across all nodes of the Hadoop cluster. BigR's mission is to provide large-scale analytics on Hadoop using R (Yejas et al., 2014). First, students are guided to install BigR atop the Hadoop cluster. Once they connect their R session to the InfoSphere BigInsights server, they can browse the HDFS and explore their Twitter data from their R session. A list of the tasks along with their descriptions are provided in Table 2. More information about the installation and use of BigR library is provided through the IBM sources (public.dhe.ibm.com/cloud/bluemix/analyticsforhadoop/BigR.docs_SocialGoodChallenge.pdf).

3.4 Data Integration

According to the Intel IT Center's Big Data Analytics survey (2012), data integration is the third most challenging aspect of big data. Data integration is the process of matching and combining data residing at different sources and providing a unified view of the data. Big data sources have widely differing structures and qualities with substantial differences in the coverage, accuracy, relevance, and timeliness of the data provided (Dong and Srivastava, 2013).

The data sources used in the exercise need to be consolidated in terms of time and geo-location. For example, the CDC contained flu incidences in the U.S. only from 1997 to 2015; however, the Twitter dataset included tweets from all over the world from 2013 to 2015, and the Google Trends Data provided aggregated Google search data from 2008 to 2015. Students had to create a subset of each dataset so that all datasets related to the same time period and location.

3.5 Data Modeling

We define three approaches in our analytics initiative: temporal, spatial, and text mining. In the temporal analysis, students examine whether Twitter data could indeed be adapted for nowcasting of influenza outbreaks. They track and compare clinical flu incidences from the CDC and flu-related activities on Twitter during the outbreaks. Students use Pearson correlation, which assumes a zero lag between the online and real-world activities, and a time-series analysis approach to obtain the temporal cross-correlations between the two trends.

The goal of temporal analysis is to help students perform the basic analysis around time-stamped data. To this end, students examine whether online activities on Twitter/Google reflect the occurrence of a flu outbreak. The assumption is that people post tweets about the flu more often when the flu outbreak is imminent. Students perform a cross-correlation test between actual flu activity and Twitter flu activity and observe that clinical flu encounters lag one month behind online posts, concluding that the number of unique users posting about flu per month can be a good measure of the number of patients who visited hospitals for Influenza-like Illness (ILI) symptoms and were diagnosed with the flu based on the CDC data.

In the next stage students perform spatial analysis. The goal of the spatial analysis is to track the geographic spread of influenza activities using information gathered from microblogging websites such as Twitter. The Twitter dataset contains all the flu-related tweets that have U.S. geographic coordinates. Students excluded all the tweets that originated from outside the U.S. and ignored those users who generated flu tweets with invalid geo-location information. In order to determine areas with a high risk of flu-related complications, i.e., hotspots, students are guided to use Google Maps API tools to zoom in to identify public locations where many flu-related

tweets were posted over time. To find places from which people tweeted more, they aggregated data points by latitude and longitude, filtered by locations with greater than 1,000 tweets, and plotted the results on a U.S. map as shown in Figure 2.

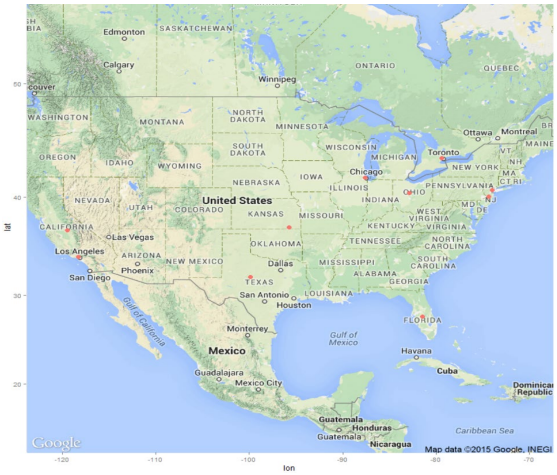

Figure 2. Top 10 Locations with More than 1,000 Tweets about the Flu

Our pedagogical framework also engages students in text mining using big data tools. Students are encouraged to use text mining techniques to extract and summarize useful information from textual data, and to identify patterns, such as popular symptoms and treatments of the flu, that were discussed in Twitter messages. For example, Figure 3 shows a word cloud of the most common terms in the tweets. Among the tweets that mention at least one of the influenza-like illnesses (ILI) symptoms in our bag of words, "cough," "sore throat," "fever," "runny nose," and "headache" were the most common symptoms. Figure 4 displays a heat map of the frequency of each term in each tweet in the corpus. Terms that discriminate effectively will be orange for tweet posts with flu symptoms. When the colors are faded to a darker color for a term, the term is not discriminating effectively to detect the presence of flu in the tweets.

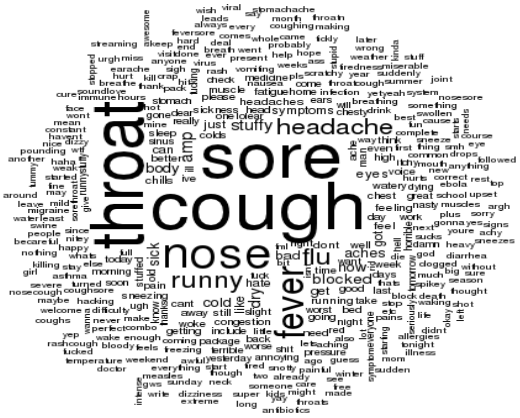

Figure 3. Word Cloud of Most Frequent Flu Symptoms

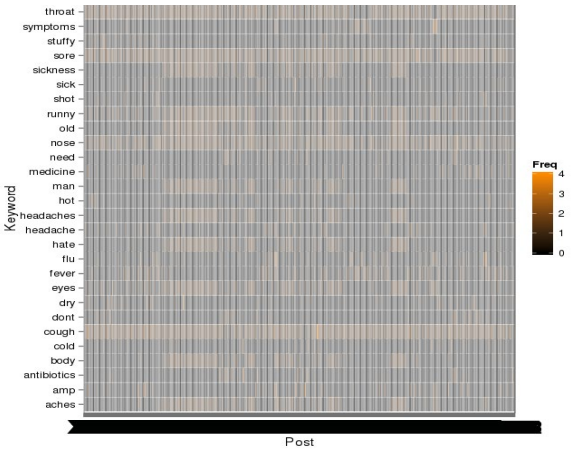

Figure 4. Frequency of the Words in the Tweet Collection

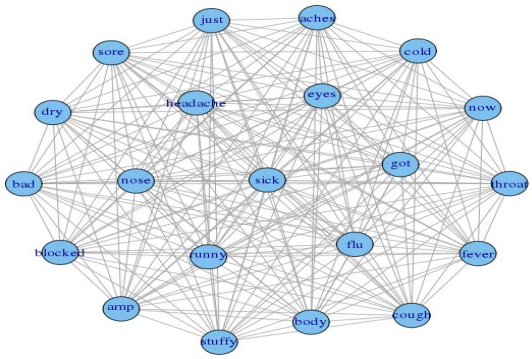

Figure 5. Network Representation of Popular Flu Symptoms

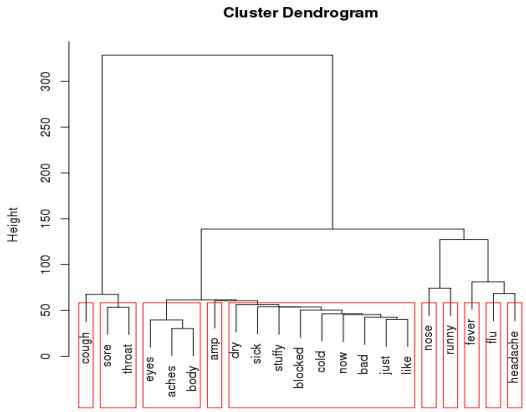

Figure 6. Symptom Clustering in Flu

Finally, in the last part of this framework, students use the BigInsights tool for Text Analytics and write Annotation Query Language (AQL) statements to perform entity recognition in order to extract structured information from unstructured and semi-structured documents. Students define dictionaries/bags of words or phrases to identify matching terms of interest across the tweets' text through extract statements or predicate functions. See the Appendix for an AQL snippet that extracts an anti-flu medication feature that was discussed in tweets.

3.6 Data Visualization

Visualizing big data by converting data and information into graphical representation is indispensable for discovering geographic patterns. Tools with a rich palette of visualizations, such as Tableau, R, Aster Lens, Gephi, and BigSheets, can supply the appropriate graphics over big data. In this exercise, with geo-located and time-stamped Twitter and CDC data at hand, students can develop real-estate mashup applications to derive valuable insights from such diverse data sources.

In the visualization step, students import the tweets data into BigSheets. Because the original data is in JSON format, they must use Comma Separated Value (CSV) Line Reader to edit their data and build a Big SQL table. Next, they use geo-location coordinates (longitude and latitude) for mapping the tweets on both the world map and the U.S. map. Students use BigSheets visualization functionality to create a map to show the spread of flu in the U.S. For example, from the 2013-14 flu outbreak map, students could observe that people from different places in the U.S. tweet about their flu, and many of these tweets originate from the eastern U.S., while the largest number of tweets in the western U.S. are from California. The mid-eastern states, such as New Jersey, Virginia, North Carolina, and Delaware, make up a high density of flu tweets, whereas western states, such as Oregon, Nevada, Idaho, Montana, Wyoming, and North and South Dakota, do not tweet about flu as much. Southern states, such as Texas and Florida, also show a high number of tweets. The distribution of tweets in other continents is roughly even.

4. EXPERIMENT

As mentioned earlier, our big data analytics framework consists of temporal, spatial, and textual analytics. In the temporal phase, students analyze whether Twitter data could be used to forecast flu outbreaks. In the spatial phase, students use location data to map the spread of flu and discover patterns such as public flu hotspots. In the textual phase, students apply text mining algorithms to extract insights, such as popular types, symptoms, and treatments of the flu, discussed in tweets.

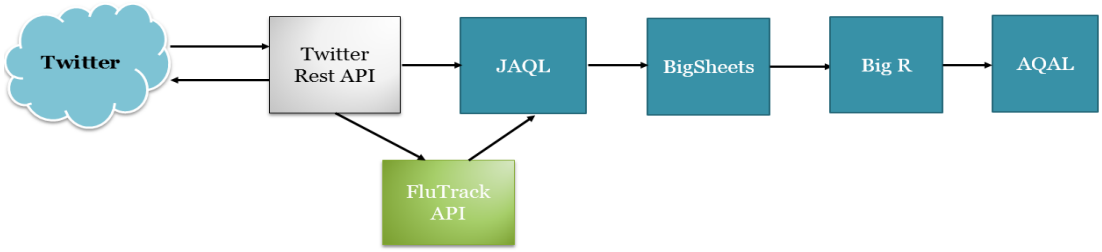

Figure 7. Exercise's Workflow

- Develop knowledge, skills, and understanding about a wide range of subjects in the field of big data technologies.
- Understand techniques for storing and processing large amounts of structured and unstructured data.
- Build big data applications through highly scalable systems capable of collecting, processing, storing, and analyzing large volumes of rapidly generated, varied data.
- Interpret and communicate their ideas through written and oral reports.

Table 3. Course Objectives

By the end of the case, students will have:

- Gained experience working within a big data environment.
- Applied current modeling and visualization techniques to big data.

By the end of the case, students will have learned how to complete all the following tasks:

- Create a big data project on InfoSphere BigInsights.
- Build a crawling application to access Twitter data using the Eclipse InfoSphere BigInsights tool.
- Create and populate a JAQL program with application logic to deal with JSON files using the Eclipse JAQL InfoSphere BigInsights tool.
- Publish an application to the InfoSphere BigInsights catalog and deploy and run it on the Hadoop cluster.
- Build a Big SQL table in BigSheets.
- Work with big data without writing scripts.
- Create visualizations of Twitter data in BigSheets.
- Use BigR functions to analyze data located on the InfoSphere BigInsights server within an R-enabled environment.
- Perform temporal-spatial analyses using time-series and spatial data analysis packages in R.
- Perform text analytics using R packages.
- Create visualizations in R.
- Create Mashup applications using Google's maps API.
- Use text analytics module in Eclipse to write AQL statements to extract the information that they want to extract.

Table 4. Key Case Objective

The exercise addresses all the course learning objectives but focuses specifically on student objectives listed in Table 4. The complete set of course objectives is listed in Table 3.

The instructors recorded videos that cover concepts and tools pertinent to the exercise and take the students through a step-by-step process to complete each task. Students watched these videos and prepared the tutorials prior to attending the class. In the class time, we were less devoted to introductory concepts and simple point-and-click software questions covered in the videos but were more focused on the technical and practical sides of the exercise. Students were expected to demonstrate some level of proficiency in class. This method represents a more blended, active learning methodology in which a balanced mix of face-to-face and online learning is used to achieve learning outcomes (Lim and Morris, 2009). The exercise consists of two parts, spanning over two weeks: a class assignment to be completed during the class time in the first week and a homework assignment to examine students' knowledge and to encourage them to reflect on what they learned. Through the class assignment, students replicated what was performed in the videos and gained experience and confidence working in a big data environment. At this phase, the codes and instructions were explicitly provided to retain students' attention and interest. Through the homework, students were encouraged to go beyond the class assignment; take the lead and develop their own ideas, questions, and problem-solving strategies; and apply them to the flu-related activities on Twitter. The homework was designed to engage students in a challenging learning experience in which they became developers of content.

As mentioned earlier, our exercise followed the big data analytics framework outlined in the previous section to demonstrate social media analytics using big data tools. In the following, we will discuss each of these steps in detail. The steps can be completed all at once or in parts, as per the discretion of the instructor, in accordance with the course time frame.

During the data collection phase, we familiarized students with the basic concept of JAQL by referring them to the IBM sources to learn how to develop and execute a JAQL application in the BigInsights Hadoop cluster. Also, we provided them with the basic syntax of JAQL through the JAQL Wikipedia link (http://en.wikipedia.org/wiki/Jaql). At this stage, students had to demonstrate that they had created a BigInsights project within Eclipse, created a JAQL program with the given code, tested it, published it in the BigInsights applications catalog, and executed it on the Hadoop cluster. In this phase, students extensively worked within the IBM InfoSphere BigInsights tool for Eclipse which involves command-line tools. In addition to the Eclipse environment, JAQL could be accessed from the command-line interface: JAQL shell. Students learned how to execute their JAQL scripts from the JAQL shell. Furthermore, students were shown how to deploy a Java library for the Twitter API (twitter4j.org/) as an alternative way to collect data from Twitter. As mentioned before, Twitter data was integrated with data from other sources to demonstrate the variety of big data and the value it can produce when leveraged with traditional data sources. Therefore, students were guided to obtain the actual flu data from Centers for Disease Control and Prevention (CDC) (http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html) using the R-package 'cdcfluview.' The R script was provided to students to automate the crawling process of the CDC platform (see the Appendix).

During the storage stage, we walked students through the HDFS and MapReduce components of Hadoop. Students explored the HDFS layer of Hadoop as well as the NameNode,

JobTracker, DataNode, and TaskTracker and their responsibilities through the BigInsights Web Console. This helped students see how the data is split into blocks and distributed throughout the Hadoop cluster.

As is common with most data mining initiatives, data extraction, preparation, and cleaning steps constitute a significant amount of this exercise. During the data preprocessing phase, students were provided with code snippets to parse out their tweets in JSON format and store them in a structured form. Also, they installed the BigR package to interface R with the BigInsights platform. They used R packages such as 'tm' to remove punctuation, numbers, and stop words from the tweet dataset and build a term-document frequency matrix.

The data sources used in the exercise needed to be consolidated in terms of time and geo-location. For example, the CDC contained flu incidences in the U.S. only from 1997 to 2015; however, the Twitter dataset included tweets all over the world from 2013 to 2015, and the GFT provided aggregated Google search data from 2008 to 2015. During the data integration phase, students created a subset of each dataset so that all datasets related to the same time period and location. This allowed them to compare the two data sets based on time and geo-location.

During the data modeling and visualization, students performed temporal, spatial, and text analytics. For example, they were instructed to perform cross correlation analysis to determine whether the actual flu activity from CDC data was leading or lagging the twitter flu activity. They verified that clinical flu encounters lag one month behind online posts. In this phase, we emphasized that this result does not imply that the same individuals who tweet about their flu symptoms would visit a hospital after one month; rather, it simply points out the lag between the trends that are observed in these two worlds. Next, they also performed location analytics to identify several locations from which a majority of flu-related tweets initiated. They leveraged Google's API to visualize their results in a way similar to Figure 2. Using Google's API, students were able to zoom in to identify and inspect these locations. For example, Forest Park, New York, New York 10007, is a place from where many flu-related tweets originated. As can be seen in Figure 8, this is a non-residential area with many shops and offices. Looking at the word cloud of the text of the tweets originating from this place revealed that most of the tweets are posted by people affected by the flu. There was no evidence of tweets advertising flu remedies and medications. Figure 9 represents the details of the location of Disney Adventure Park in California which was among the top ten flu hotspots in the U.S. during the outbreak season.

Students dug into the top ten locations with suspicious flu activities. They discovered that such non-residential places as parks, restaurants, hotels, and stores had the most flu-related tweets. Because of the number of people who visit these places every day, they play a significant role in the spread of the flu virus. Therefore, public health agencies could benefit from these findings by tracking and locating high-risk places and by highlighting "must-know" facts during the outbreaks. This information can help people who visit such locations to take appropriate measures.

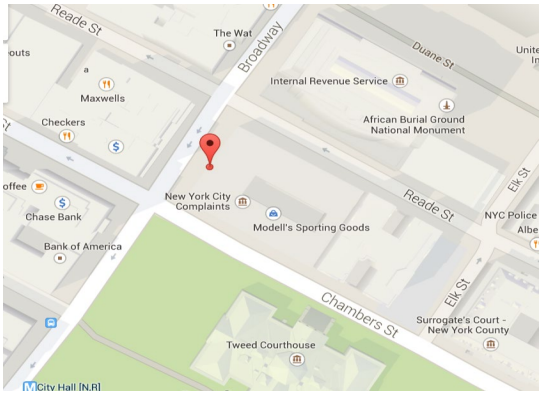

Figure 8. The Highest Tweet Activity Place: Forest Park, New York

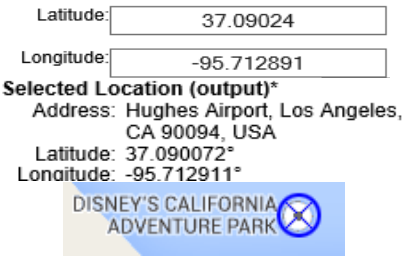

Figure 9. Disneyland among the Top Ten Flu-Infected Places

In the last part of this exercise, students used the BigInsights tool for Text Analytics, i.e., Annotation Query Language (AQL), and developed codes to extract meaningful information from the textual data using entity recognition methods. Students were referred to the IBM sources to learn the basic syntax of AQL and how to develop and execute a JAQL application on the BigInsights Hadoop cluster. The Eclipse tool for Text Analytics provides a step-by-step wizard through the Text Analytics development process. Students watched the demo of the Eclipse tool for Text Analytics to get familiar with this tool as well as the AQL syntax. After enough exposure to the tool, students followed the step-by-step wizard to define dictionaries/bags of words or phrases to identify matching terms of interest across the tweets' text through extract statements or predicate functions.

As homework, students were encouraged to develop their own ideas and look for ways to validate them. For example, Table 5 contains a list of other ideas that might come up in the homework section.

- Examine how people reacted to treat their sickness; whether they used home remedies or drugs.
- Explore tweets that show whether a patient's condition has deteriorated or improved.
- Explore if a patient's condition has deteriorated.
- Explore the frequency and popularity of major flu-related terms such as flu types, symptoms and treatments over time.
- Develop histograms/bar charts that show the popularity of the treatments used.
- Test whether people who tweeted about flu also suffered from other diseases.
- Examine what systems/organs in the body can be affected by flu.
- Consider what treatment is being highly recommended for each type of flu.
- Develop bar charts to show changes in the number of daily tweets about different flu symptoms.
- Develop bar charts to show changes in the number of daily tweets about different types of flu.
- Develop bar charts to show changes in the number of daily tweets about different types of treatment.
- Test whether there was a correlation between flu-related Twitter data and the actual flu data across different regions within the U.S.

Table 5. Students' Ideas and their Development in the Homework Section

5. ASSESSMENT

We surveyed students to assess the overall effectiveness and usefulness of the framework/exercise described in this paper. Twenty-seven students majoring in business analytics participated in the survey. We used a 5-point Likert scale as shown in Table 6. The results are summarized in Figure 10.

The main objective of the exercise was to provide students with hands-on experience with big data tools and techniques for social media analytics. Almost all students agreed that they gained knowledge through the exercise and that the big data platform used for doing social media analytics was reasonable, relevant, and challenging. Even though the exercise was very technically and analytically oriented, most students appreciated learning such content through experiential learning.

Students expressed that the exercise showed them how to develop an end-to-end big data application (4.67), improved their understanding of big data components in a hands-on setting (4.63), and identified trends and patterns for business intelligence (4.63). Altogether, this implies that students benefited from the hands-on experience in class, the real-world application of big data tools and techniques, and the potential of big data in decision making. Additionally, monitoring influenza trends through mining social media using big data technology received very strong positive feedback from the students (4.56). This suggested that students gained an understanding of the valuable insights that can be derived from social media data. The exercise even inspired students to go beyond the in-class assignment by employing more advanced techniques and performing additional analyses to boost their analytical skills in big data.

Q	Statement	Strongly Agree (5)	Agree (4)	Neutral (3)	Disagree (2)	Strongly Disagree (1)
1	Helped me understand what big data analytics is and how it works.	14	11	1	1	0
2	Improved my understanding of different components of a big data eco-system.	17	10	0	0	0
3	Demonstrated how to perform social media analytics in a big data platform.	16	10	1	0	0
4	Improved my understanding of big data architecture (Hadoop, MapReduce) without delving too much into the technical complexities of it.	14	8	2	2	1
5	Demonstrated how to develop end-to-end big data application.	19	7	1	0	0
6	Demonstrated how various tools (e.g., BigSheets, JQAL, AQL and BigR) in BigInsights can help process big data, identify trends/patterns and generate business intelligence.	17	10	0	0	0
7	The goals of the exercise were clearly stated and consistently pursued.	24	3	0	0	0
8	The step-by-step handout helped me go through the exercise.	22	4	0	1	0
9	Was reasonable and useful.	19	6	1	1	0
10	I gained new knowledge from this exercise.	21	5	0	1	0

Table 6. Exercise Evaluation (N = 27)

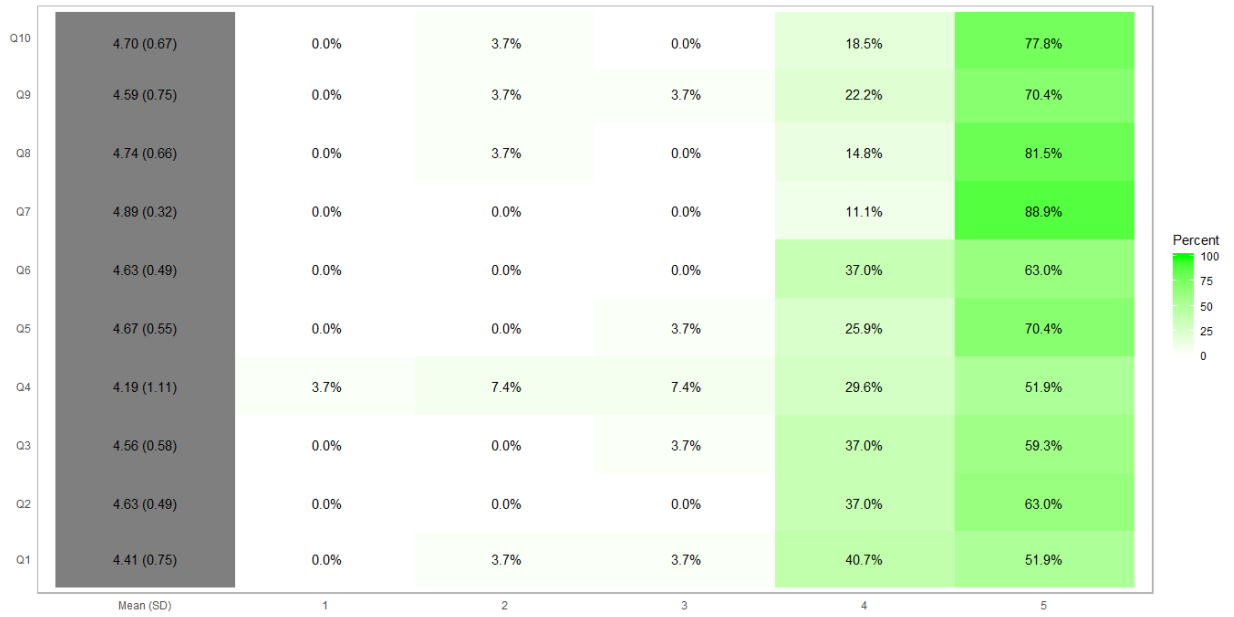

Figure 10. Summary Statistics

Scale:1 = Strongly disagree; 2 = somewhat disagree; 3 = neutral; 4 = somewhat agree; 5 = Strongly agree

Statement	Strongly disagree	Somewhat disagree	Neutral	Somewhat agree	Strongly agree	Average rating
The instructor was available for consultation.	0	0	1	5	13	4.63
Student responsibilities for this course were well defined.	0	1	0	7	11	4.47
Class time was well spent.	1	2	2	2	12	4.16
I learned a lot from the instructor in this course.	0	0	1	4	14	4.68
Course materials contributed to my learning.	0	0	2	3	14	4.63
I was challenged in this course.	1	0	2	8	8	4.16
Coming into this course I was motivated to learn this subject.	0	1	2	3	13	4.47

Table 7. Instructor and Course Evaluation (N = 19)

Similar feedback and observation were also recorded in the formal course evaluation. Table 7 shows the result of the 5-point Likert scale survey for the course where 19 students filled out the survey. For example, students rated the instructors' overall performance at 4.46 out of 5. They overwhelmingly expressed that they learned a lot and were challenged in this course.

Several students commented on the hands-on nature of the course and appreciated it as a refreshing change from conventional courses. For instance, one student commented: "With our hands-on activities it really interested me in my learning; it proves that I can apply the material in a real-world basis in which evaluates data, which is really important." Another student stated:

> I thought this was a very useful course that will be beneficial for me moving forward. I liked that I did not really have to study for this course (very nice break from other classes!), yet I still absorbed a surprisingly large amount of information. I liked the hands-on component of the course. SAS and R labs were very useful in learning how to operate such analytical tools.

In addition, providing timely in-depth feedback is very important for students to maintain progress and improve their work. A number of students commented on the pace of feedback and its critical importance for continuing the learning process. For example, one student noted: "The instructor routinely provided timely feedback each week, added additional materials for continued learning, and provided necessary assistance for our timely progress in class." Another one commented:

> The instructor was a swift with email responses and thorough with feedback. Though the software we used, especially big data tools, were very complex, the instructor made sure to go out of his way to record videos, write instructions, and coach via email to get it right.

Finally, the miscellany of content in big data can be overwhelming and confusing. For example, one student mentioned: "There is an abundance of content that can get overwhelming; however, I don't think it could be organized any better." Therefore, organizing the course content in a structured

manner according to the students' need and interest allows students to easily follow the development of ideas and to see the connections between them without feeling overwhelmed and frustrated.

Such comments by students offer indirect evidence that the framework/hands-on exercise outlined in this paper was useful in student learning and enhanced their skills and knowledge in data mining, big data, and business analytics.

6. DISCUSSION

The positive feedback we received from the students through the survey suggests that using the pedagogical framework presented in this paper and its application in social media health analytics to help them learn about big data technologies is effective. Based on the mean response to each question, students agreed that applying big data technology to a new application helped them not only understand the technology better, but also to develop novel and innovative solutions using big data technologies. In addition, the quality of ideas and work that students developed when asked to use the technology to perform additional analyses provided further evidence that they were able to utilize the technology and apply their newfound skills and knowledge successfully. Overall, we believe that this exercise increased both students' understanding of big data technology and their ability to think innovatively about how big data technology can be applied to business and societal opportunities. While previous research lacks a pedagogical framework for developing and implementing hands-on exercises in big data education, our proposed framework guides instructors to develop exercises that enable students to experience end-to-end learning on the activities related to big data analytics. Using the framework, students deal with various stages of big data analytics life cycle (e.g., data collection, storage, preprocessing, integration, modeling, and visualization) and can gain knowledge and skills on various big data tools and techniques.

Big data is often created by aggregating multiple data sources that may or may not have the same structure or format and, as previous research indicated, the real challenge lies in managing the variety of big data. While most instructor resources include exercises with a single source for big data, thus emphasizing the volume of data, this paper focuses on the unstructured data, a major source of the variety in big data. Our framework and the exercise enable students to get a great level of exposure to multiple sources of data and understand the role of data integration in big data environments.

While we found that our framework is useful and effective, it can be tweaked and applied to other big data applications. The context for the exercise could be different from the flu outbreak. For example, data on music, movies, or the stock market can be integrated with social media data to understand customer preferences and behavior. While we use R primarily for data analysis and visualization, instructors can give students the option to use any or a combination of software packages to develop innovative solutions based on the research questions of interest. Other R packages for text mining can also be used to provide students with additional skills and a more in-depth learning experience with unstructured data. For instance, while we used R-package 'tm' to perform stemming, one can use package 'textstem' to perform both stemming and/or lemmatization.

Although we designed and implemented our exercise with the intention that students work independently, we observed that they may reap benefit from interactions with other classmates to understand how to get the software to do what they need to do, and then complete the homework on their own. We found that this will greatly enhance buy in and stimulate students' learning enthusiasm and increase the joy of learning.

Even with the strong design of our framework and the exercise, the implementation was not without its challenges. First, it may be possible that not all students watch the demonstration videos prior to class, despite our urging. We suggest addressing this issue by inserting pop-up quiz questions in the videos and selectively showing portions of the video tutorials in the first week of the exercise. Second, teaching big data analytics is a double-edged sword. While exposing students to a wide array of new software and platforms helps them gain and enhance their technical knowledge and skills, it may have unintended consequences of making students overly focused on the technology itself at the expense of not learning as much of the underlying concepts. Therefore, providing students with a balanced view of big data is important and it is advised that instructors be mindful of not overwhelming students with unnecessary details of big data tools but helping them understand the process and its underlying concepts to solve problems. Third, as can be expected in a technology intensive class, tools and technologies are constantly evolving. Indeed, some of the tools may no longer be freely available as consolidation occurs in a rapidly evolving industry. However, our framework and the exercises described are general enough to be usable in a class setting with another collection of tools.

Finally, business analytics is an interdisciplinary field that draws students from a variety of disciplines or backgrounds (e.g., business, statistics, and engineering) and applicable to different fields of study (e.g., MIS, marketing, accounting, and finance) (Jeyaraj, 2019). Depending on the backgrounds of the students in statistics, programming, data management, and data-driven decision-making, instructors may need to provide additional coverage to ensure that students are all on the same page. For example, one of the areas we found most students are lacking in is the point before they import the data into statistical software such as R. Particularly, non-MIS students may lack the knowledge of how to gather, clean, and transform data; therefore, it is important for instructors to identify areas in which students need additional guidance and adjust their instruction based on student need. It may require additional or alternate tools or techniques to be used in order to enhance the learning experience for students.

7. CONCLUSION

This effort is guided by Sigman et al. (2014) on how to teach about big data in the classroom and how to find the right balance between teaching and providing hands-on experience with big data. Our paper presents a pedagogical framework along with a robust big data analytics exercise that helps students experience the development of a real-world, end-to-end, big data application and gain related knowledge and skills in big data, including data collection, storage, preprocessing, integration, modeling, and visualization. Our analytics framework

encompasses temporal, spatial, and text mining. We used IBM's InfoSphere BigInsights platform to provide our students with hands-on experience with big data technologies. The teaching case was carried out at a major Midwestern university in the U.S. The class included 27 business students who enrolled in a data mining course as part of their degree program. After doing the exercise, students expressed more curiosity and confidence in doing big data and social media analytics.

Although this manuscript offers several contributions to the literature, five deserve special mention and are specifically highlighted. First, this paper proposes a comprehensive, yet streamlined, process framework for teaching, explaining, and demonstrating the development of big data applications. Second, the paper creates a practical, efficient, and effective "learning by doing" experience within the proposed pedagogical framework to satisfy individual learning needs and curiosity. Third, unlike previous research that often used a single source for big data, this paper includes multiple sources of data to demonstrate the role of data integration in big data environments. Fourth, the same applies to the technology itself. Unlike previous studies that covered one or two big data tools, this paper includes the use of multiple big data technologies to demonstrate their capabilities. Last, but not least, the paper proposes an integrated application that intersects big data concepts – health and social network analysis – while providing a pedagogical framework that is flexible enough for other big data applications with different technology stacks.

8. REFERENCES

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., & Liu, B. (2011). Predicting Flu Trends using Twitter Data. In *2011 IEEE Conference on Computer Communications*, 702-707.

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., & Liu, B. (2013). Online Social Networks Flu Trend Tracker: A Novel Sensory Approach to Predict Flu Trends. In Gabriel J. et al. (Eds.), *Biomedical Engineering Systems and Technologies. BIOSTEC 2012. Communications in Computer and Information Science*, 353-368.

Asamoah, D. A., Sharda, R., Zadeh, A. H., & Kalgotra, P. (2017). Preparing a Data Scientist: A Pedagogic Experience in Designing a Big Data Analytics Course. *Decision Sciences Journal of Innovative Education,* 15(2), 161-190.

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PloS one,* 8(12), e83672.

Cegielski, C. G. & Jones-Farmer, L. A. (2016). Knowledge, Skills, and Abilities for Entry-Level Business Analytics Positions: A Multi-Method Study. *Decision Sciences Journal of Innovative Education,* 14(1), 91-118.

Chambers, M. & Dinsmore, T. W. (2014). *Advanced Analytics Methodologies: Driving Business Value with Analytics*. Pearson Education, Inc.

Dean, J. & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51(1), 107-113.

Dinter, B., Jaekel, T., Kollwitz, C., & Wache, H. (2017). Teaching Big Data Management–An Active Learning Approach for Higher Education. *Proceedings of the Pre-ICIS SIGDSA Symposium.*

Dong, X. L. & Srivastava, D. (2013). Big Data Integration. In *IEEE 29th International Conference on Data Engineering (ICDE)*.

Douglas, L. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Research Note* 6(70).

Dredze, M., Paul, M. J., Bergsma, S., & Tran, H. (2013). *Carmen:* A Twitter Geolocation System with Applications to Public Health. *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI).*

Fowler, B., Godin, J., & Geddy, M. (2016). Teaching Case: Introduction to NoSQL in a Traditional Database Course. *Journal of Information Systems Education,* 27(2), 99-104.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter Brand Sentiment Analysis: A Hybrid System Using n-Gram Analysis and Dynamic Artificial Neural Network. *Expert Systems with Applications,* 40(16), 6266-6282.

Goh, T. T. & Sun, P.-C. (2015). Teaching Social Media Analytics: An Assessment Based on Natural Disaster Postings. *Journal of Information Systems Education,* 26(1), 27-36.

Gruzd, A., Paulin, D., & Haythornthwaite, C. (2016). Analyzing Social Media and Learning through Content and Social Network Analysis: A Faceted Methodological Approach. *Journal of Learning Analytics,* 3(3), 46-71.

Guthrie, D. (2013). *The Coming Big Data Education Revolution*. Retrieved August 25, 2021 from https://www.usnews.com/opinion/articles/2013/08/15/why-big-data-not-moocs-will-revolutionize-education.

Intel. (2012). Big Data Analytics, Intel's IT Manager Survey on How Organizations are Using Big Data. *Intel Report*.

Janvrin, D. J. & Watson, M. W. (2017). Big Data: A New Twist to Accounting. *Journal of Accounting Education,* 38, 3-8.

Jeyaraj, A. (2019). Pedagogy for Business Analytics Courses. *Journal of Information Systems Education,* 30(2), 67-83.

Khan, G. F. (2018). *Creating Value with Social Media Analytics: Managing, Aligning, and Mining Social Media Text, Networks, Actions, Location, Apps, Hyperlinks, Multimedia, & Search Engines Data*. CreateSpace Independent Publishing Platform.

Kumar, S., Morstatter, F., & Liu, H. (2014). *Twitter Data Analytics*. New York: Springer.

Lamb, A., Paul, M., & Dredze, M. (2013). Separating Fact from Fear: Tracking Flu Infections on Twitter. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 789-795).

Lee, K., Agrawal, A., & Choudhary, A. (2013). Real-Time Digital Flu Surveillance Using Twitter Data. *The 2nd Workshop on Data Mining for Medicine and Healthcare.*

Li, J. & Cardie, C. (2013). Early Stage Influenza Detection from Twitter. Retrieved August 25, 2021 from https://arxiv.org/abs/1309.7340.

Lim, D. H. & Morris, M. L. (2009). Learner and Instructional Factors Influencing Learning Outcomes within a Blended Learning Environment. *Educational Technology & Society,* 12(4), 282-293.

McLeod, A. J., Bliemel, M., & Jones, N. (2017). Examining the Adoption of Big Data and Analytics Curriculum. *Business Process Management Journal*, 23(3), 506-517.

Miah, S. J., Solomonides, I., & Gammack, J. G. (2020). A Design-Based Research Approach for Developing Data-Focussed Business Curricula. *Education and Information Technologies,* 25(1), 553-581.

Parks, R., Ceccucci, W., & McCarthy, R. (2018). Harnessing Business Analytics: Analyzing Data Analytics Programs in US Business Schools. *Information Systems Education Journal,* 16(3), 15-25.

Parks, R. F. (2020). A Pedagogic Experience in Designing a Healthcare Analytics Course: Lessons Learned. *Information Systems Education Journal,* 18(5), 4-15.

Phillips-Wren, G., Iyer, L. S., Kulkarni, U., & Ariyachandra, T. (2015). Business Analytics in the Context of Big Data: A Roadmap for Research. *Communications of the Association for Information Systems,* 37(23).

Piramuthu, S. (2003). On Learning to Predict Web Traffic. *Decision Support Systems,* 35(2), 213-229.

Podeschi, R. & DeBo, J. (2019). Integrating Big Data Analytics into an Undergraduate Information Systems Program using Hadoop. *Information Systems Education Journal,* 17(4), 42-50.

Provost, F. & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data,* 1(1), 51-59.

Russom, P. (2011). Big Data Analytics. *TDWI Best Practices Report, Fourth Quarter*, 19(4), 1-34.

Schoenherr, T. & Speier-Pero, C. (2015). Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential. *Journal of Business Logistics,* 36(1), 120-132.

Shang, D. R. (2019). Applying Social Media Analysis to Real World Business Problems: A Course Project. *Journal of Computing Sciences in Colleges,* 34(6), 35-42.

Sigman, B. P., Garr, W., Pongsajapan, R., Selvanadin, M., Bolling, K., & Marsh, G. (2014). Teaching Big Data: Experiences, Lessons Learned, and Future Directions. *Decision Line,* 45(1), 10-15.

Sledgianowski, D., Gomaa, M., & Tan, C. (2017). Toward Integration of Big Data, Technology and Information Systems Competencies into the Accounting Curriculum. *Journal of Accounting Education,* 38, 81-93.

Song, I. Y. & Zhu, Y. (2016). Big Data and Data Science: What should We Teach? *Expert Systems,* 33(4), 364-373.

St. Louis, C. & Zorlu, G. (2012). Can Twitter Predict Disease Outbreaks? *BMJ*, 344:e2353.

Stanton, W. W. & Stanton, A. D. A. (2020). Helping Business Students Acquire the Skills Needed for a Career in Analytics: A Comprehensive Industry Assessment of Entry-Level Requirements. *Decision Sciences Journal of Innovative Education,* 18(1), 138-165.

Talvis, K., Chorianopoulos, K., & Kermanidis, K. L. (2014). Real-Time Monitoring of Flu Epidemics through Linguistic and Statistical Analysis of Twitter Messages. *The 9th IEEE International Workshop on Semantic and Social Media Adaptation and Personalization* (pp. 83-87).

Verma, A., Yurov, K. M., Lane, P. L., & Yurova, Y. V. (2019). An Investigation of Skill Requirements for Business and Data Analytics Positions: A Content Analysis of Job Advertisements. *Journal of Education for Business,* 94(4), 243-250.

Waller, M. A. & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution that will Transform Supply Chain Design and Management. *Journal of Business Logistics,* 34(2), 77-84.

Wixom, B., Ariyachandra, T., Douglas, D., Goul, M., Gupta, B., Iyer, L., Kulkarni, U., Mooney, B. J. G., Phillips-Wren, G., & Turetken, O. (2014). The Current State of Business Intelligence in Academia: The Arrival of Big Data. *Communications of the Association for Information Systems,* 34(1).

Yejas, O. D. L., Zhuang, W., & Pannu, A. (2014). Big R: Large-Scale Analytics on Hadoop Using R. In *2014 IEEE International Congress on the Big Data,* 570-577.

Zadeh, A. H., Sengupta, A., & Schultz, T. (2020). Enhancing ERP Learning Outcomes through Microsoft Dynamics. *Journal of Information Systems Education,* 31(2), 83-95.

Zadeh, A. H., Zolbanin, H. M., Sharda, R., & Delen, D. (2019). Social Media for Nowcasting Flu Activity: Spatio-Temporal Big Data Analysis. *Information Systems Frontiers,* 21(4), 743-760.

Zolbanin, H. M., Delen, D., and Zadeh, A. H. (2015). Predicting Overall Survivability in Comorbidity of Cancers: A Data Mining Approach. *Decision Support Systems,* 74, 150-161.

AUTHOR BIOGRAPHIES

Amir H. Zadeh is an associate professor of information systems at Wright State University. He holds a Ph.D. in management science and information systems from Oklahoma State University. His research interests are in data-driven decision making and machine learning with applications in health, sports, cybersecurity, and social networks. His research has been published in journals such as *Decision Support Systems*, *Information & Management*, *Information Systems Frontiers*, and the *Journal of Business Analytics*. He received the 2020 Ranyard Medal from the Operations Research (OR) Society.

Hamed M. Zolbanin is an assistant professor of information systems at the University of Dayton. He had several years of professional experience as an IT engineer prior to receiving his Ph.D. in management science and information systems from Oklahoma State University. His research has appeared in journals such as *Decision Support Systems*, *Information & Management*, *Journal of Business Research*, and *Information Systems Frontiers*. His main research interests are healthcare informatics, data science, data analytics, and online reviews.

Ramesh Sharda is the Vice Dean for Research and the Watson Graduate School of Management, Watson/ConocoPhillips Chair and a Regents Professor of management science and information systems in the Spears School of Business at Oklahoma State University. He has co-authored two textbooks (*Analytics, Data Science, and Artificial Intelligence: Systems for Decision Support, 11th edition*, Pearson and *Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition*, Pearson). His research has been published in major journals in management science and information systems including *Management Science*, *Operations Research*, *Information Systems Research*, *Decision Support Systems*, *Interfaces*, *INFORMS Journal on Computing*, and many others. He is a member of the editorial boards of journals, such as *Decision Support Systems*, *Decision Sciences*, *ACM Database*, and *Information Systems Frontiers*. He served as the Executive Director of Teradata University Network through 2020 and was inducted into the Oklahoma Higher Education Hall of Fame in 2016. He is a Fellow of INFORMS and AIS.

APPENDIX

BigInsights, BigSheets, Eclipse, Jaql, BigR, R, and Aql Exercise

This week's lab exercise is analyzing social media data on Hadoop with InfoSphere BigInsights tools.

Learning Objectives:

After completing this exercise, you will have learned how to complete the following tasks:

1. Create an InfoSphere BigInsights project
2. Build your crawling Application to access Twitter using Eclipse InfoSphere BigInsights tool
3. Create and populate a Jaql program with application logic to deal with JSON files using Eclipse Jaql InfoSphere BigInsights tool
4. Publish your application to the InfoSphere BigInsights catalog and deploy and run it on the Hadoop cluster
5. Create Visualizations of the Twitter data in BigSheets
6. Use Big R functions to analyze data located on the InfoSphere BigInsights server with an R environment
7. Perform text analytics with R packages
8. Create visualizations in R
9. Use Text Analytics module in Eclipse to write AQL statements to extract the information that you want to extract.

In this assignment, you will gather flu related tweets for the entire world using the Twitter & Flutrack APIs. The words used as monitoring tags are common flu symptoms: Influenza, flu, fever, cough, fever, chills, headache, sore throat, runny nose, dry cough and sneezing. You are going to process, analyze and visualize all the influenza related Twitter messages! Follow the procedures in the following steps to do this assignment:

I. Twitter Crawling

Start BigInsights Hadoop Cluster and follow the instructions in the following link to develop and execute a Jaql application to the BigInsights Hadoop cluster. All you need to do is create a BigInsights project within Eclipse, then create a Jaql program, test it and finally publish it in the BigInsights applications catalog.

Reference:
https://www.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.analyze.doc/doc/t_analyze_bd_jaql.html

Here is the script that needs to be copied and pasted into your jaql program. The following code writes the results of the Jaql query as both a delimited file and a seq file to the /user/biadmin/Twitter_flu/ directory in your distributed file system.

```
// TODO: Add JAQL content here
term = "InfluenzaORfluORfeverORcoughORchillsORheadacheORsore throatORrunny noseORsneezingORdry cough";
// Get the current system time
time=now();
results = read(http("http://api.flutrack.org/?s="+term));
tweets=jaqlGet("http://api.flutrack.org/?s="+term);
tweets -> transform {
                $.aggravation,
                $.latitude,
                $.longitude,
                $.tweet_date,
                $.tweet_text,
                $.user_name,
                };
// Write the results of the Jaql query as a delimited file in HDFS.
tweets -> write(del("/user/biadmin/Twitter_flu/tweets.del", schema = schema
{aggravation,latitude,longitude,tweet_date,tweet_text,user_name}));

// Write the results of the Jaql query as a Hadoop sequence file in HDFS.
tweets_text = tweets -> transform $.tweet_text;
tweets_text -> write(seq("/user/biadmin/Twitter_flu/tweets_text.seq"));
```

```
tweets_text -> write(lines(location='/user/biadmin/Twitter_flu/tweets_text.txt'));
```

Here is a very brief explanation what does the script do:

- jaqlGet() loads JSON text file from a URL.
- tweets -> transform {$.aggravation,$.latitude,...} take original JSON structure, and transform it to the simpler one.
- tweets -> write(del("/user/biadmin/Twitter_flu/tweets.del", schema = schema {…} writes the new structure to a comma-delimited file to a file in HDFS.
- tweets -> transform $.tweet_text; extracts just the text part of tweets.
- tweets_text -> write(seq("….")); stores the extracted text content of tweets into a hadoop sequence file (again in this example to HDFS file system).

Run the Jaql program and provide the screenshot of the JSON data that you have supplied to the HDFS from BigInsights along with the first three tweets from the JSON file.

Reference: More information about reading and writing JSON files is available here.
Notice that Jaql can be also accessed from the command-line interface: **Jaql shell**. You can open a terminal and type: $BIGINSIGHTS_HOME/jaql/bin/jaqlshell and then execute the above Jaql script from the Jaql shell.

II. Visualization with BigSheets

In the next step, import the Tweets data to BigSheets. Use Comma Separated Value (CSV) Line Reader to edit your data. Also, rename the column names in order of aggravation, latitude, longitude, tweet_date, tweet_text, user_name.
Now, let's get out of the virtual machine into the main machine and log in to the InfoSphere BigInsights Console in the Chrome web browser. Use the IP address of your virtual machine along with the port number of BigInsights web Console (8080) to get access to the InfoSphere BigInsights Console from the Chrome web browser.
Use geolocation's coordinates (longitude and latitude), for mapping the tweets on both World map and the US map. Use ***Density Map*** Chart to do so! Please discuss the charts and provide screen shots of the charts.
In the next step, click *add sheets* and select *Group*. Group tweets based on longitude and latitude data. Use ***Value Map*** for mapping the tweets! Please discuss the maps and provide screen shots of the charts. Also, provide a screenshot of the tweet-location occurrence sheet (matrix).
Reference: Learn more about charts and maps type in BigInsights here. Learn more about BigSheets in BigInsights here.

III. Twitter Data Analysis with Big R

In this step, you will use a combination of R and Big R functions to analyze your *Tweets* data located in the BigInsights/Hadoop cluster.
Let's get back to the virtual machine and install BigInsights BigR. You will have a big R icon on your desktop. Click on it and let it install.
Click on Big R and let it install. If you get errors in the installation process, follow the instructions to fix it and complete the process. More information about installing Big R is available here.
After you install Big R, open a terminal from the desktop and type R followed by enter to dive into the R environment! Next, load the Big R package and connect to the InfoSphere BigInsights server:

```
install.packages("bigr")
install.packages("rJava")
library(bigr)
library(rJava)
bigr.connect(host="bivm",
  port=7052, database="default",
  user="biadmin", password="biadmin")
```

Verify that the connection was successful.

```
is.bigr.connected()
```

If you ever lose your connection during this exercise, run the following line to reconnect.

```
bigr.reconnect()
```

Once you connected, you will be able to browse the HDFS file systems. Examine the *tweets* dataset that has already been loaded onto the cluster. Answer the following questions:

1: What is your R working directory?

2: List all the functions that the current version of Big R has.

3: List files on the BigInsights file system.

```
bigr.listfs()  #list files under root"/"
```

Locate your *tweets* file in the hdfs and provide a screenshot.

```
bigr.listfs("/user/biadmin/") # List files under /user/biadmin
```

Recall that you stored the tweets data as a comma-delimited file in the hdfs. Now let's connect to it and explore it a bit. This is done by creating data set as a bigr.frame over the dataset.

```
tweets <- bigr.frame(dataSource="DEL",dataPath="/user/biadmin/Twitter_flu/tweets.del")
```

4: Check the class of "tweets" to see if it is an object of "bigr.frame". Provide a screenshot.

5: What is the difference between a bigr.frame object and a regular data.frame when it comes to load that data into memory?

6: Explore the structure of the dataset and discuss it.

7: Examine the dimensions of the dataset. Find how many rows and columns are in the dataset.

Next clean the dataset by dropping the first column and relabeling the other columns. Run *str* command to check the new structure of the data.

```
library ("plyr")
tweets$V1=NULL
tweetsRevised<-rename (tweets,
c("V2"="Latitude","V3"="Longitude","V4"="Tweet_time","V5"="Tweet_text","V6"="User_name"))
```

8: Use bigr.sample function in Big R to draw a 15% random sample from the tweets data. Examine the class and the dimension of the random sample.

```
#Generate a 15% random sample of data
tweetsRevisedSample<-bigr.sample(tweetsRevised, 0.15)
```

9: Using random sampling support in Big R split the *tweets* data into training set (~70) and test set (~30%).

```
splitstweets <- bigr.sample(tweets, c(0.7, 0.3))
train <- splitstweets[[1]]
test <- splitstweets[[2]]
```

What is the class of the splitstweets object?

Let's bring the data from the Hadoop server (currently as bigr.frame) into a data.frame. Notice that R data.frames are held in memory and it may exhaust memory if the data is very large!

```
tweetsdf<-as.data.frame(tweetsRevisedSample)
```

To learn more about the Big R concepts and functions, view the IBM tutorial here link.

In the next step, we want to build a term-Document Matrix (tdm) and Document-term Matrix (dtm) for the Twitter dataset that we have. Basically, we want to cluster words and tweets to find groups of words. Package *tm* provides functions for text mining:

http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Run the following commands to install and load required packages in your R section:

```
install.packages ("tm")
install.packages ("ggplot2")
install.packages("NLP")
install.packages ("wordcloud")
library ("tm")
```

Before, we use *tm* functions for text mining and building up the tdm and dtm matrices, we should clean up our data a bit from some "exotic" characters within the tweets. Otherwise, the *tm* functions will not be able to handle it and will throw an error during the tdm and dtm matrices creation process. One way to find out and drop those tweets with weird characters from our tweets collection is convert the tweet texts explicitly into Unicode (utf-8) using the following command:

```
tweetsdf<-as.data.frame(tweetsRevisedSample)
convTweets <- iconv(tweetsdf$Tweet_text, to = "utf-8")
```

This leaves some vector entries "NA" associated with those tweets that cannot be handled. Then, we remove the "NA" entries from the entire tweet dataset using the following command:

```
tweets <- (convTweets[!is.na(convTweets)])
```

Now next, build a corpus of Tweets and store in myCorpus using the following command:

```
myCorpus <- Corpus(VectorSource(tweets))
```

10: Type myCorpus to see how many tweets are in this corpus. Discuss the result!

11: Using *inspect* function to display detailed information on the corpus.

```
inspect (myCorpus [1:10])
```

12: In the next step, create term-document matrix from the above corpus with function TermDocumentMatrix().

```
stopwords("english")
tdm<-TermDocumentMatrix(myCorpus,control=list(removePunctuation=TRUE,removeNumbers=TRUE, stopwords=TRUE))
```

The control argument is a list of transformations that need to be performed upon the Corpus, including changing letters to lower case, and removing punctuations, numbers and stop words.

12-1: type *tdm* to see how sparse or dense your matrix is.

12-2: Next, examine the Term-Document Matrix using function inspect and provide a snapshot of the result and discuss it:

```
inspect (tdm)
inspect(tdm[1:10,1:20])
tdm
```

13: Repeat the steps mentioned in 12 but to create the Document-Term matrix: **(**0.25M)

```
dtm<-DocumentTermMatrix(myCorpus,control=list(removePunctuation=TRUE,removeNumbers=TRUE, stopwords=TRUE))
```

15: Let's look at the popular words and the association between words. Using function findFreqTerms() finds frequent terms with frequency no less than 50. Provide a snapshot of the result and discuss the result.

```
findFreqTerms(tdm, lowfreq=50)
```

findFreqTerms() finds frequent terms with frequency no less than 50.

16: Alternatively, you can develop a bar plot to visualize Term Document Matrix by running the following codes:

```
termFrequency<-rowSums(as.matrix(tdm))
termFrequency<-subset(termFrequency,termFrequency>=50)
library("ggplot2")
qplot(names(termFrequency),termFrequency,geom="bar",xlab="Terms")+coord_flip()+geom_bar(position="stack",stat="identity
")
```

Provide a snapshot of your bar chart.

17: World Cloud: In the next step, we can show the importance of words with a word cloud (also known as a tag cloud), which can be easily produced with package 'wordcloud'. Run the following commands to

```
library ("wordcloud")
m<-as.matrix(tdm)
wordFreq<-sort(rowSums(m),decreasing=TRUE)
set.seed(375)
grayLevels<-gray((wordFreq+5)/(max(wordFreq)+5))
wordcloud(words=names(wordFreq),freq=wordFreq,min.freq=10,random.order=F,colors=grayLevels)
```

Provide a snapshot of your ***WordCloud*** and discuss it.

18. Clustering Words: Next try to find clusters of words in the dataset with a hieratical clustering method, so called *Ward*. First, sparse terms are removed, so that the clustering plot would not be crowded with too many less important words! Then, using dist the distance between terms is calculated and after that, the terms are clustered with hclust. Also, the method of clustering is set to *ward* and the number of clusters is set to 10.

```
# clustering words
tdm2<-removeSparseTerms(tdm,sparse=0.95)
dtm2<-removeSparseTerms(dtm,sparse=0.95)
m2<-as.matrix(tdm2)
distMatrix<-dist(scale(m2))
fit<-hclust(distMatrix,method="ward")
plot(fit)
rect.hclust(fit,k=10)
```

Provide a snapshot of your ***Cluster Dendrogram*** and discuss it.

19. In this part of exercise, we put a social network analysis into perspective. Using *igraph* package, we will build a network of terms based on their co-occurrence in tweets. First, we transform the term-document matrix into term-term adjacency matrix and based on that, a network graph of terms is built.

```
# A network of Terms
install.packages ("igraph")
library ("igraph")
# term-term matrix
tdmMatrix<-as.matrix(tdm2)
dtmMatrix<-as.matrix(dtm2)
termMatrix<- tdmMatrix %*% dtmMatrix
g<-graph.adjacency(termMatrix,weighted=T,mode="undirected")
#remove loops and multiple edges
g<-simplify(g)
V(g)$lable<-V(g)$name
V(g)$degree<-degree(g)
set.seed(3952)
layout1<-layout.fruchterman.reingold(g)
plot(g,layout=layout1)
```

Provide a snapshot of your ***network of terms*** and discuss the result.
Try different layouts:

```
# Other layouts
layout1<-layout.kamada.kawai(g)
plot(g,layout=layout1)
layout1<-layout.spring(g)
plot(g,layout=layout1)
layout1<-layout.sphere(g)
plot(g,layout=layout1)
layout1<-layout.fruchterman.reingold.grid(g)
```

```
plot(g,layout=layout1)
# Regraph to set the lable size of verticies based on their degrees
V(g)$label.cex <- 2.2 * V(g)$degree / max(V(g)$degree)+ .2
V(g)$label.color <- rgb(0, 0, .2, .8)
V(g)$frame.color <- NA
egam <- (log(E(g)$weight)+.4) / max(log(E(g)$weight)+.4)
E(g)$color <- rgb(.5, .5, 0, egam)
E(g)$width <- egam
# plot the graph in layout1
plot(g, layout=layout1)
```

20. Finally, convert tweets timestamps to actual time and create some histograms to see for example which days of the week and which month of the year people are tweeting a lot during their flu epidemics. Discuss the histogram!

```
//convert timestamp to actual time
tweetstime<-as.numeric(tweetsdf$Tweet_time)
actual_tweet_time<-as.POSIXct(tweetstime, origin = "1970-01-01", tz = "GMT")
actual_tweet_time
weekdays(actual_tweet_time)
hist(actual_tweet_time, "day", freq = TRUE,main="Histogram of Number of Tweets per day")
hist(actual_tweet_time, "month", freq = TRUE, main="Histogram of Number of Tweets per week"))
counts <- table(weekdays(actual_tweet_time))
barplot(counts, main="bar chart of Number of Tweets per weekday")
```

21. Run the following code to scrape the flu data from the CDC website.

```
## CDC Data
library(cdcfluview)
library(magrittr)
library(dplyr)

flu <- get_flu_data(region="hhs",
                    sub_region=1:10,
                    data_source="ilinet",
                    years=2013:2015)
 get_state_data<- get_state_data(2013:2015)
get_weekly_flu <-get_weekly_flu_report()
flu %<>%
mutate(season_week=ifelse(WEEK>=40, WEEK-40, WEEK),
       season=ifelse(WEEK<40,
                     sprintf("%d-%d", YEAR-1, YEAR),
                     sprintf("%d-%d", YEAR, YEAR+1)))

prev_years <- flu %>% filter(season != "2013-2015")
curr_year <- flu %>% filter(season == "2013-2015")
 curr_week <- tail(flu, 1)$season_week
```

22. Perform monthly/weekly/daily cross-correlation between CDC data and Twitter data.

```
##Weekly Cross Correlation
cor(curr_week$count[8:13],Tweetflu$count[7:12])
ccf(curr_week$count[17:47],Tweetflu$count[17:47], ylab = "cross-correlation",main = "Weekly Flu Activity & Weekly Flu
Twitter Activity")
```

23. Visualize Twitter data on a US map. Identify top 10 locations from which a majority of the flu related tweets originated.

```
## Heat map
# Subsetting the data to Only US region
Tweets_flu_US <- subset(Tweets_flu, (tweetsdf$longitude < -60  &Tweets_flu$longitude > -135)&( Tweets_flu$latitude > 23
&Tweets_flu$latitude <50) )

Tweets_flu_US <- sqldf('select Latitude, Longitude ,Count(Tweet_text) as count1  from tweets group by Latitude, Longitude')
```

```
# mapping all the points in the map
map <- get_map("US", zoom=3, source = "google")
p <- ggmap(map)
p <- p + geom_point(data = Tweets_flu_US , aes(x = Tweets_flu_US$Longitude, y = Tweets_flu_US$Latitude ))
p

# Subsetting the US data to Only users who have tweeted more than 1000 tweets
Tweets_flu_US_1000 <- subset(Tweets_flu_US, Tweets_flu_US$count >=1000)
Tweets_flu_US_1000
map <- get_map("US", zoom=4, source = "google")
p <- ggmap(map)
p <- p + geom_point(data = Tweets_flu_US_1000 , aes(x = Tweets_flu_US_1000$Longitude, y = Tweets_flu_US_100$Latitude,
col = "red", size = 10 ))
p
```

Hint: Run the programs line-by-line to understand what is happening!

IV. Text Analytics with Aql (Basic)

Now you are going to use BigInsights Text Analytics tool and write Annotation Query Language (AQL) statements to extract some features that you want to extract from tweets. Follow the instructions in the following link to guild you through the process.

Reference:

http://www-01.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.tut.doc/doc/tut_Mod_TxtAna.html

First, from the HDFS file browser, download the tweets_text.txt file and put it in the /home/biadmin/ directory. Go to Eclipse tool and click on biginsights text analytics workflow. Follow step-by-step wizard to create a text extractor. Extract one or two medication features from the tweets.

This is a sample AQL code that defines Antibiotic as an extractor and finds tweets with Antibiotic content.

```
module Antibiotic_BasicFeatures;

create view Antibiotic as
extract regex /antibiotics/
   on R.text as match
from Document R;
output view Antibiotic;
```

Provide a screenshot of the "text analytics result" window.

Twitter Crawling in JAVA: The twitter data can be downloaded in various ways. The intension is to provide few resources to do the same. The IBM Infosphere BigInsights already has flume installed in it. All you want to do is change the flume config file and start the flume service.

Flume location: /opt/ibm/biginsights/flume

Go to bin and run the flume-ng to start the agent.

In this section I am also providing the instructions to download twitter data using Twitter4j.

For those who are new to java you can go through http://www.tutorialspoint.com/java/ to learn basics

Note that according to the rate limits and application we should modify java code. Please go through the link

http://twitter4j.org/en/code-examples.html

https://github.com/yusuke/twitter4j/tree/master/twitter4j-examples/src/main/java/twitter4j/examples

Which gives us various snippets of the code that can be used at various scenarios. Also, we should understand that the API supplies us with tweets as java objects to make our life easier when we submit a request.

Download the zip file. Extract the project. Open eclipse and import the project. Now right click on the project -> Build Path ->configure build path -> add jars-> browse to lib folder and click on the temboo_java_sdk_2.8.0.jar and click ok once it is added to the list. Make sure after this step there are no errors in the project. Change the query in the program. Now right click on the project and run. There will be several json files that are generated, and each file will have close to 2000 tweets.

Note if you quit the program abruptly the last json file generated may not be in proper format. Make sure that the file ends with ‘]’.

1. Now that you have several json files on your local. Mention at least three possible ways that you can use process these files to use in your analysis in IBM big insights Hadoop environment. Explain how? (You are expected to research)
2. Using at least 4000 tweets by any of the ways to get the tweets. How many tweets are repeating? provide the tweets that are retweets and replies explain how you achieved this using screenshot (Do not use Hive)
3. Run the word count application from the web console for only the texts of the tweets and display the results.

For those who are interested in Programming, you can deploy the above java program as java application as application in BigInsights. Try doing this by embedding this into a Map reduce program or JAQL.

Apart from this there are several API on various programming languages like python, PHP, .NET, GO etc. to download the twitter data.