Paper 1066-2017

A Big Data Challenge: Visualizing Social Media Trends about Cancer using SAS® Text Miner

Scott Koval, Yijie Li, and Mia Lyst, Pinnacle Solutions, Inc.

ABSTRACT

Analyzing big data and visualizing trends in social media is a challenge that many companies face as large sources of publically available data become accessible. While the sheer size of usable data can be staggering, knowing how to find trends in unstructured textual data is just as important of an issue. At a Big Data conference, data scientists from several companies were invited to participate in tackling this challenge by identifying trends in cancer using unstructured data from Twitter users and presenting their results. This paper explains how our approach using SAS analytical methods was superior to other Big Data approaches in investigating these trends.

INTRODUCTION

In recent years, public interest and participation have become the heart of big data. In particular, data from social media has increased exponentially. At the 2016 Indy Big Data Conference, a Visualization Challenge offered a way for companies to share their methodologies for handling large and complex datasets. The Visualization Challenge required participants to mine a defined set of Twitter tweets and produce visualizations that offer trends and insights to cancer. The raw data contained more than 143,845,720 individual tweets with three attributes identifying a user ID, date and time, and text content of each tweet. The data was not allowed to be augmented in any way.

Given the size and complexity of the data, only four companies ultimately participated in this challenge with only two weeks to generate results and a presentation.

The traditional way of analyzing this data is hypothesis-based where data is examined based on a particular question of interest. While other companies followed this approach using modules of the Hadoop framework or other Visualization tools (e.g. Hadoop, Apache Spark, Apache Solr, Datameer), we offered a data-driven, analytical solution by combining SAS Enterprise Guide, SAS Enterprise Miner and SAS Visual Analytics along with other tools (Hadoop, Python, and Spark). This allowed the data to tell the story rather than restricting the outcomes based on our limited knowledge of current cancer trends.

METHODOLOGY

Our solution included a blend of different technologies in order to apply the best features from each to the appropriate function. As a result, Python, Spark and Scala were used to process the data in a timely manner and SAS was used for text analytics and visualizations (Figure 1) to identify trends.

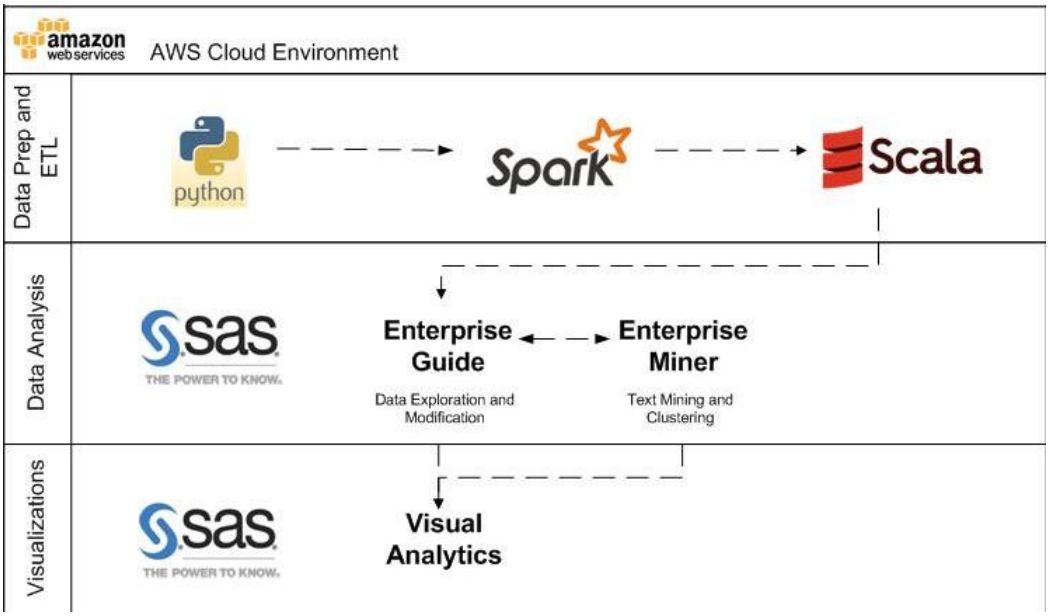

Figure 1. Indy Big Data Challenge Solution

The raw data provided for this challenge existed in ~555,000 separate CSV files (18GB). We used Python, Spark, and Scala to merge these files together and then imported the data into SAS Enterprise Guide for additional preprocessing. A query was created to determine whether or not a tweet contained a reference to the word 'cancer' or any related terms. This was used to help filter down the data to help investigate the topic at hand. Retweets, or messages that are simply shared, were also removed from the data in order to prevent a bias in the results. Overall, the filtered data included 1.9 million cancer related tweets to analyze.

While only three columns were provided in the raw data, we were able to create additional fields to help investigate the data. This included breaking up the date field into year, month, day, day of week, and time of day variables. An additional field was created to form a cleaned up version of the tweet, which retained alphanumeric values. Binary variables were also created to flag the message as containing a mention or hashtag. The number of mentions and hashtags each message had used was also calculated.

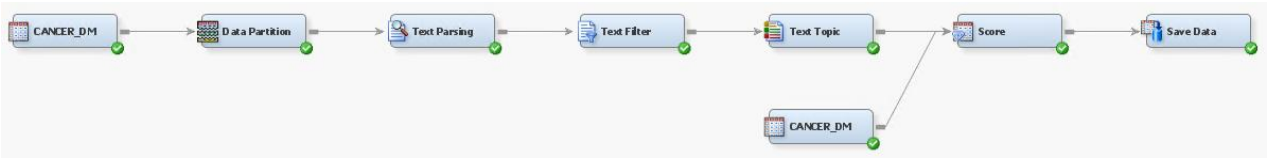

Figure 2. SAS Enterprise Miner process flow for creating text topics on cancer

These processed data were imported into SAS Enterprise Miner for analysis (Figure 2). This program contains an add-on called SAS Text Miner, a useful tool for analyzing unstructured text data in order to identify underlying topics and segments of words. The concept of cancer is a very broad topic, and in order to explore trends, we used this software to determine a list of text topics present in the data in order to detect any underlying themes.

After the data were randomly sampled, the first step in the analysis was to parse it using the Text Parsing node in SAS Text Miner. This contains a series of tasks used to tokenize, stem, and restructure the data. A stop list was also used in this step in order to drop frequently used English words from the analysis. Examples of these common terms include, "a", "the", "of", "at", etc. A spell check was also used during

this step to help correct commonly misspelled words and standardize the data and reduce noise during analysis.

Next, the data were processed through the Text Filter node. The default parameter was used to determine both the frequency (Log) and term (Entropy) weighting. After running the node, we used the Interactive Filter Viewer to further refine the results. Unnecessary words were manually dropped from analysis, and synonyms were created to group like-worded terms. An example of a synonym term being created would be to combine the terms “SKIN CANCER” and “MELANOMA”.

Now the data were ready to explore using the Text Topic node. We created up to 50 single-term topics and 25 multi-term topics (Fig 3). Upon exploring the results, it appeared that 6 of the 50 single-term topics contained specific cancers. These included Breast Cancer (n = 321,745), Colon Cancer (n = 39,016), Lung Cancer (n = 62,014), Ovarian Cancer (n = 60,226), Prostate Cancer (n = 34583), and Skin Cancer (n = 35,853). Six new cancer specific datasets were then created based on the tweets that were flagged for each of these types of cancer.

Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Single	1	0.001	0.001	+breast cancer	1	128646
Single	2	0.001	0.001	+help	1	55096
Single	3	0.001	0.001	+support	1	36231
Single	4	0.001	0.001	+today	1	36202
Single	5	0.001	0.001	+awareness	1	33645
Single	6	0.001	0.001	+patient	1	33575
Single	7	0.001	0.001	+treatment	1	32422
Single	8	0.001	0.001	+fight	1	31899
Single	9	0.001	0.001	+research	1	30567
Single	10	0.001	0.001	+risk	1	29693
Single	11	0.001	0.001	+lung cancer	1	24843
Single	12	0.001	0.001	+prostate cancer	1	24004

Figure 3. Table containing cancer text topic results

Each of these six new datasets was repeated through the same Text Parsing and Text Filtering techniques mentioned above. After, the Text Cluster node was used for each of them to create a hierarchical cluster and segment the tweets based on frequently occurring terms (Fig 4). For each of these cancer types, we reviewed the clusters produced and categorized them with appropriately named clusters.

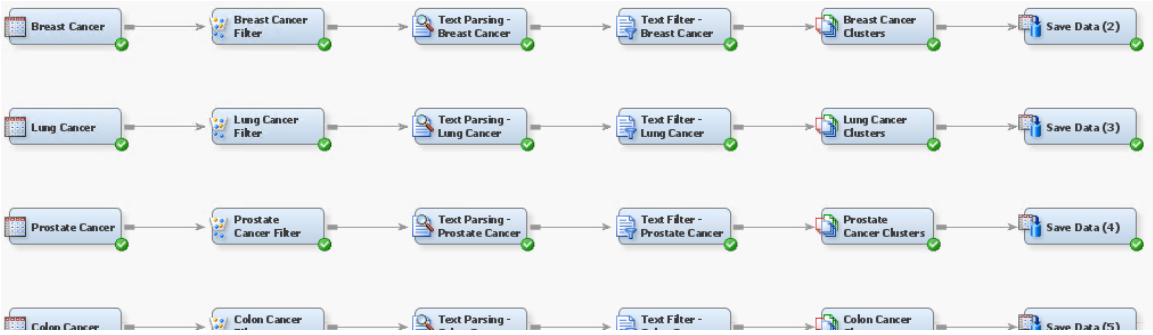

Figure 4. Diagram featuring flows to create text clusters for each type of cancer

SAS Visual Analytics was used to help display the results by creating several reports and explorations. These visualizations helped analysts further explore the findings and infer trends.

RESULTS

The results of the topic analysis corresponded to the top 6 Most Common Cancers in 2016, except Ovarian Cancer (Tables 1 and 2). We suspect Ovarian Cancer may have surfaced in the analysis since it is the 5th leading cancer-related cause of death in women.

Cancer Type	Estimated New Cases	Estimated Deaths
Breast (Female – Male)	*246,660 – 2,600*	*40,450 – 440*
Lung (Including Bronchus)	*224,390*	*158,080*
Prostate	*180,890*	*26,120*
Colon and Rectal (Combined)	*134,490*	*49,190*
Bladder	76,960	16,390
Melanoma	*76,380*	*10,130*
Non-Hodgkin Lymphoma	72,580	20,150
Thyroid	64,300	1,980
Kidney (Renal Cell and Renal Pelvis) Cancer	62,700	14,240
Leukemia (All Types)	60,140	24,400
Endometrial	60,050	10,470
Pancreatic	53,070	41,780

Table 1. 2016 Cancer facts and figures

	Topic Modeling	Awareness	Prevention & Screening	Fundraising/ Campaigns	Research & Studies	Treatments	Risks
Cancer Type	% of Tweets	% of Topic	% of Topic	% of Topic	% of Topic	% of Topic	% of Topic
Breast	*58.6%*	*54.2%*	*24.0%*	*6.7%*	*7.4%*		*7.7%*
Lung	*11.2%*	*25.6%*			*40.3%*	*15.4%*	*18.7%*
Prostate	*10.8%*	*32.6%*	*34.7%*		*12.6%*	*15.4%*	*4.6%*
Colon and Rectal	*6.9%*	*23.0%*	*14.0%*		*52.2%*	*10.7%*	
Melanoma	*6.4%*	*5.3%*	*68.1%*		*4.0%*	*13.6%*	*9.1%*
Ovarian	*6.1%*	*59.7%*	*14.2%*	*12.6%*	*13.5%*		

Table 2. Cancer cluster topic results

SAS Visual Analytics allowed us to easily see seasonal trends in tweets by plotting the frequency of tweets over time for each of the main cancer types. Instances of tweets about breast cancer spiked every year during the month of October for the annual awareness month. The same is true for colon cancer in March, lung cancer in November, ovarian and prostate cancer in September, and skin cancer in May (Fig 6). This would indicate that the awareness campaigns are effective in raising discussions of the associated diseases during specific times of the year.

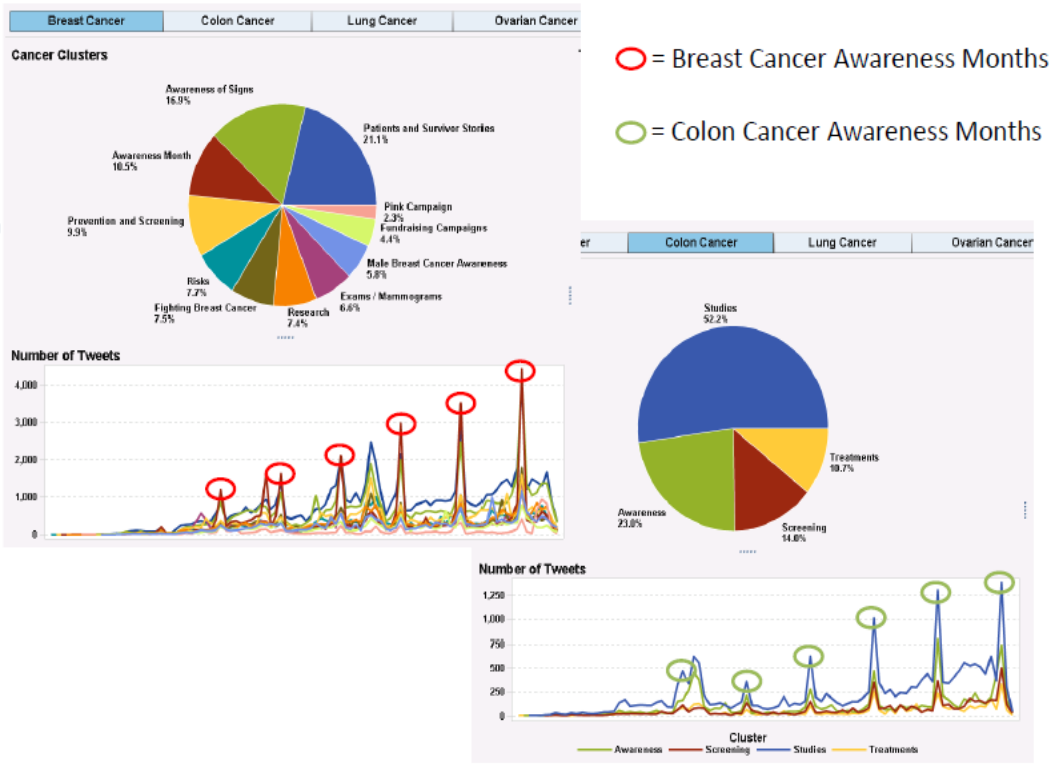

Figure 6. Cancer awareness months work!

Of the 6 diseases, breast cancer had the highest frequency of tweets which speaks to the very salient and established campaigns put out by organizations like Pink and Susan G. Komen (Fig 7). In addition, Breast and Ovarian cancer were the only two cancer topics to actually have clusters formed around fundraising campaigns.

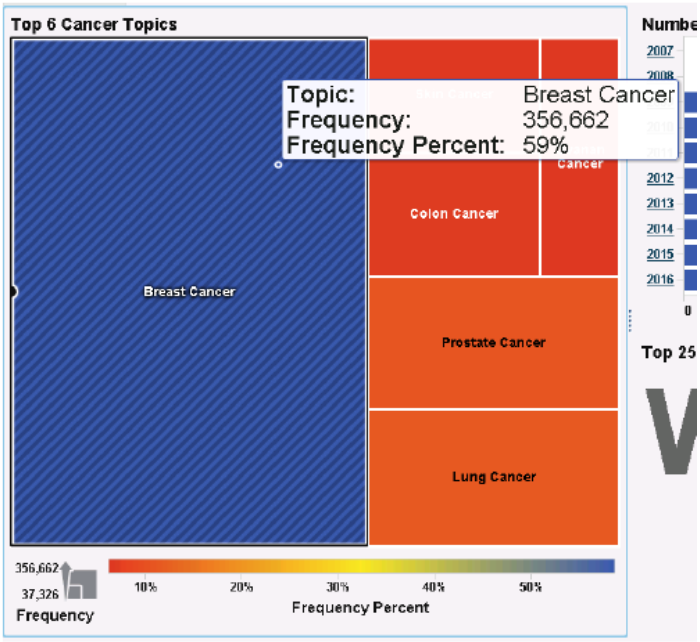

Figure 7. Breast cancer has the highest percentage of tweets

When focusing more on the types of clusters to emerge, lung and colon cancer had the highest frequencies of tweets categorized into research and studies segments. This could be due to high mortality rates of these specific diseases and amount of funding spent on research.

Word clouds of the hashtags and mentions for each cancer categories displayed some meaningful top mentioned words. At first, we examined result from breast cancer data. The top mentions in breast cancer include Taylor Swift, The Ellen Show, Kylie Minogue, Joan Lunden, Carolina Herrera, Christina Applegate and Robin Roberts, indicating a clear celebrity effect.

In addition, from the word cloud of the Male Breast Cancer Awareness cluster, we find that the top mentioned word in this cluster is NFL (Fig 8).

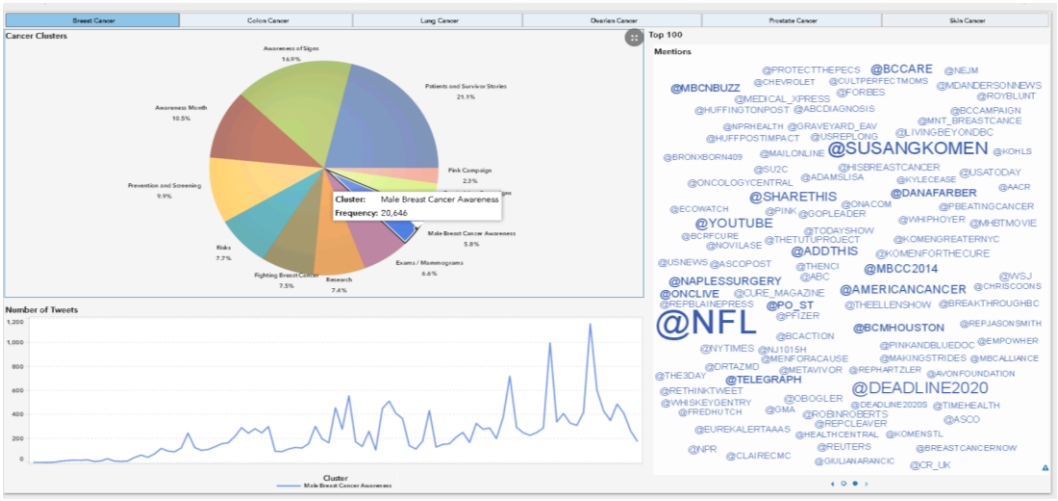

Figure 8. Breakdown of tweets in the Male Breast Cancer Awareness cluster

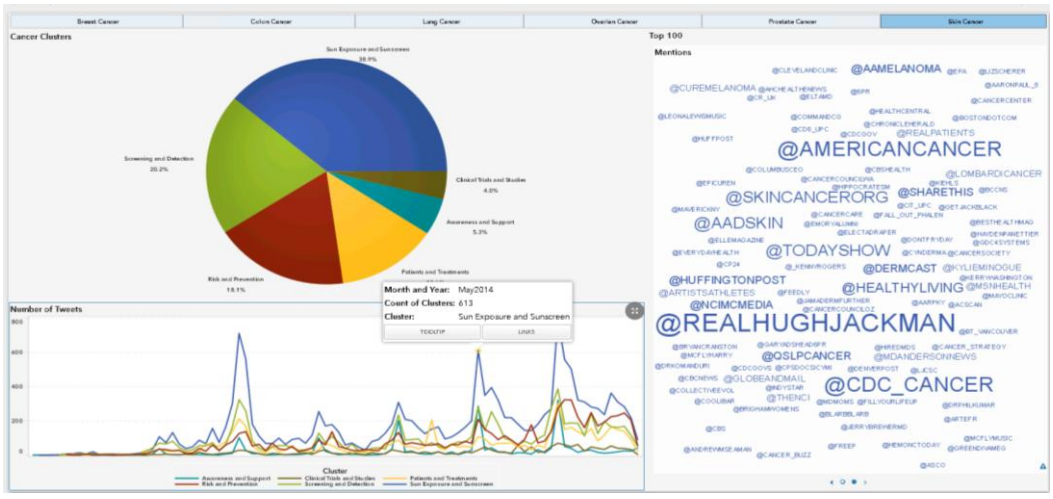

Figure 9. Celebrities influence spikes in cancer tweets

Such kind of celebrity effect can also be found in other subcategories of cancer.

In May 2014, Hugh Jackman posted his skin cancer scare on Instagram and reminded people to wear sunscreen. This can be correlated to the spike in tweets about Sun Exposure and Sunscreen and the top mention for that month, @REALHUGHJACKMAN (Fig 9).

This further confirms that anyone with high celebrity profile can help to increase public awareness in cancer prevention and treatment.

A table of all the trends and insights obtained in this analysis is shown below (Table 3):

	Description
General Trends	Top 6 Cancer types obtained from Topic modeling correspond to the 2016 Most Common Canter Types, except Ovarian Cancer.
	Ovarian Cancer tweets may have surfaced in Topic modeling due to high mortality rates.
	Breast Cancer had highest number of tweets.
	Cancer Awareness months show significant increase of tweets of the particular cancer for that month.
	Breast and Ovarian cancer were the only two topic areas that surfaced fundraising/campaign tweets.
	Colon and Lung Cancer formed the largest clusters for Research & Studies.
	People tweet about Prevention & Screening for cancers where early prevention screening affects survival rates.
Breast Cancer Insights	The top 100 mentions contain several celebrities such as, Taylor Swift (mom), Ellen Show (mom), Kylie Minogue, Joan Lunden, Carolina Herrera (designer), Christina Applegate, Robin Roberts (Good Morning America), and Oprah
	@NFL is the top mention for the Male Breast Cancer cluster.
Ovarian Cancer Insights	Spike tweets for Clinical Studies cluster in May 2015 shows @THEROCATEST as the top mention. This correlates to the results of ROCA test which were shown to be twice as effective for early detection as other screenings.
	Visible UK mentions (@OVARIANCANCERUK).
Prostate Cancer Insights	Spike in tweets for Screening cluster in May 2011 shows Coffee as a Top 25 word. This correlates to a study that showed that coffee reduces the risk of prostate cancer.
	Visible UK mentions (@PROSTATEUK).
Skin Cancer Insights	Spike in tweets for Sun Exposure and Sunscreen cluster in May 2014. This correlates to an Instagram that Hugh Jackman posted where he reveals his skin cancer scare and reminds people to wear sunscreen.

Table 3. Trends and insights found in cancer tweet clusters

CONCLUSION

Although we are not subject matter experts in current cancer trends, the use of SAS software in big data analytics allowed us to simplify a fairly complex problem and identify several trends and insights using only three columns of twitter data. By combining these tools and techniques together, we were able to let the data speak for itself rather than relying on ad-hoc analysis. While other participants used the power

of the Hadoop framework and big data visualization tools to process the all of the data, they did not perform analytical techniques in order to uncover hidden trends the data had to offer.

REFERENCES

American Cancer Society (ACS). 2016. “Cancer Facts & Figures 2016.” http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-047079.pdf.

École Polytechnique Fédérale de Lausanne (EPFL). 2016. “Scala logo.png”. By Source, Fair use, https://en.wikipedia.org/w/index.php?curid=21286998.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Scott Koval
Pinnacle Solutions, Inc.
(317) 423-9143
scott.koval@thepinnaclesolutions.com
www.thepinnaclesolutions.com

Yijie Li
Pinnacle Solutions, Inc.
(317) 423-9143
yijie.li@thepinnaclesolutions.com
www.thepinnaclesolutions.com

Mia Lyst
Pinnacle Solutions, Inc.
(317) 423-9143
mia.lyst@thepinnaclesolutions.com
www.thepinnaclesolutions.com