Application of Text Mining in Tweets to analyze general opinion about 'Bing It On' challenge by Microsoft

Shreya Sadhukhan, Taufique Alalm Ansari
Supervising Faculty: Dr. Goutam Chakraborty
Oklahoma State University, Stillwater, Oklahoma

Introduction

The usefulness of text mining has already been accepted worldwide by different business industries to produce effective knowledge and valuable insights of their business. In this era of digitization, vast amount of data is generated in text form every second through websites, blogs and social networking sites. If collected and analyzed efficiently, this data can be transformed into potential information to provide economic benefits as well as to predict and identify recent trends in society. Text Mining provides the tools to convert it into structured data, allowing analysis of the data in order to yield meaningful patterns and clusters for prediction and summarization purposes.

The 'Bing It On' Challenge is an online test by Microsoft that allows blind comparison of the search results by Bing and Google for five queries and the user is asked to choose a winner or declare a "draw" [1]. Microsoft research shows that people chose Bing over Google nearly 2:1 times in these blind comparison tests. Specifically, of the nearly 1,000 participants: 57.4% chose Bing more often, 30.2% chose Google and 12.4 % tests resulted in a draw [2]. Regarding this campaign there were instant positive, negative and mixed reactions from the vast user group, which was also reflected in the tweets in Twitter.com.

In this poster we have collected relevant Tweets using the directed search of the %GetTweet macro [3], and applied text mining to the data set using the SAS® Text Miner of SAS® Enterprise Miner 7.1 to summarize and portray general public opinion about this challenge and the two giant search engines. From the results, it is evident that Google is still preferred by the user community, even after they were offered goodies like Xbox as a gift to participate in the 'Bing It On' challenge.

Figure 1: Poster of Bing it On Challenge

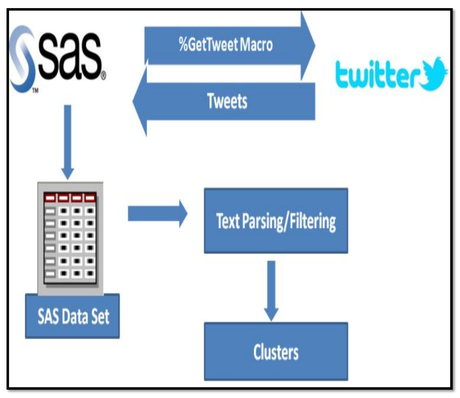

Figure 2: Overall process of collecting data from Twitter and the sequence of operations applied to the data.

Data Preparation

In order to collect data, we have used %GetTweet macro, which allows anyone to fetch data from Twitter.com after basic authentication requirement is fulfilled [3]. HTTP procedure of SAS® is used by this macro to communicate with Twitter's Search API [3]. There are several combinations of options provided to customize the search and generate desired result. Tweets that meet the provided search criteria are fetched in XML format and finally converted into SAS® data set.

For this research, we have used 'BingItOn', the name of the challenge as the keyword to gather all the tweets where the term have been specifically mentioned. Because the collected data was very directed, we had only to remove the duplicate tweets to clean the data. This task was performed in Microsoft Excel. Then the data is converted into SAS® data set, on which Text Mining is applied. Figure 3 shows the nodes of text mining and the sequence of operations applied to the data.

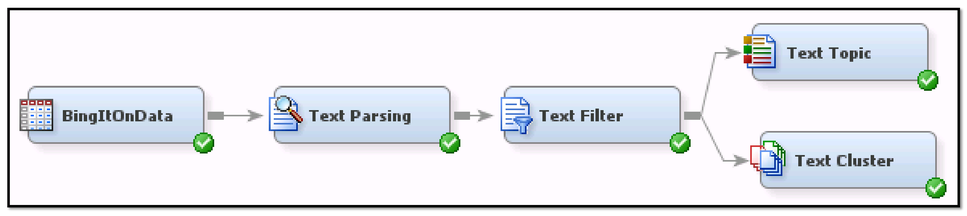

Figure 3: Text mining nodes applied to the dataset

• First, Text Parsing is applied to identify the unique terms and phrases eventually used to create a term-by-document matrix where each entry represents the number of the times a term appears in a document [4].
• A customized stop list was prepared and employed by identifying the low frequency as well as the irrelevant terms present in the parsed terms list.
• The next step was Text Filtering, where total number of parsed terms is reduced to most relevant and valuable terms.
• Standard English Dictionary has been used to spell check the terms.
• In the interactive filter viewer the relevant terms and their respective concept link diagrams can be displayed, as shown in figures 4, 5 and 6.
• Finally text topics have been generated using the Text Topic node. Topics are collection of terms that characterizes a main theme or idea [4], as shown in figure 7. These topics have been further used in this research to recognize distinct public opinions.
• Clustering technique has also been applied, which classifies the dataset into several disjoint groups using relative weights of the significant terms. The clusters generated in this case are shown in figure 8 and the descriptive terms are used to understand the theme of the respective cluster.

Results

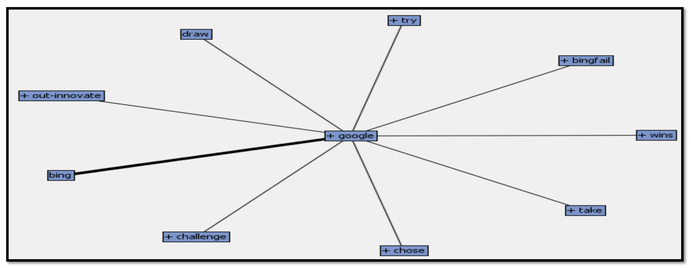

Figure 4: Concept link diagram of 'Google'

The concept link diagram of 'Google' shows strong association with terms 'wins', 'chose', 'out-innovate', 'challenge', which implies that Google has mostly won the challenge of 'Bing It On'.

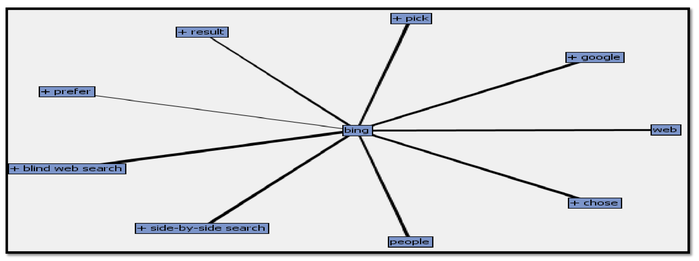

Figure 5: Concept link diagram of 'Bing'

The concept link diagram of 'Bing' shows strong association with only terms like 'side-by-side search', 'people', 'blind web search', 'chose', which can only be interpreted in the way that Bing was part of the side by side blind search competition.

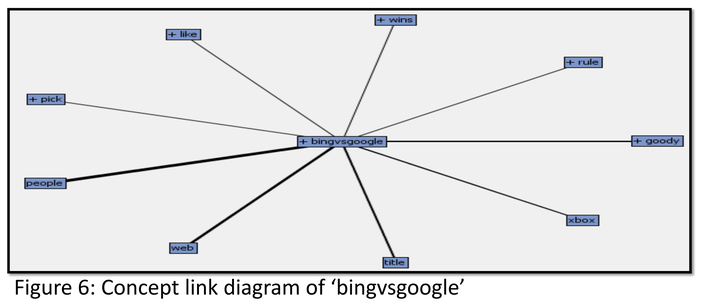

Figure 6: Concept link diagram of 'bingvsgoogle'

The concept link of 'BingvsGoogle' challenge shows association with terms 'xbox', 'goodies' along with the other terms, which depicts the fact that people were offered gifts to participate in the challenge.

Topic ID	Topic
1	+blind test,+chose,+pick,+side-by-side search,bing
2	+side-by-side search,+prefer,+result,bing,title
3	+take,+wins,+challenge,+google,+bingfail
4	title,+challenge,+bingfail,+good,+google
5	xbox,+wins,+goody,+rule,+challenge

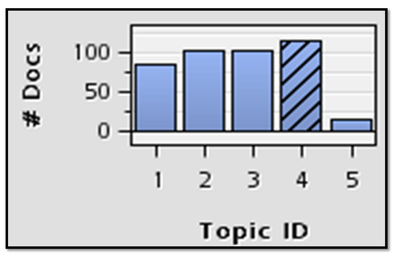

Figure 7: Topics and Documents by Topic

Cluster ID	Descriptive Terms
1	out-innovating title bingiton +bingvsgoogle +make best course +challenge
2	'blind web search' 'people chose' +'blind test' +chose +pick blind people search
3	draw fans field took won +challenge +take +google
4	goodies rules xbox win +challenge bingiton bing

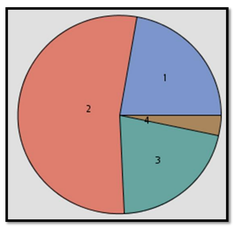

Figure 8: Clusters and Clusters by frequency

Discussion

- SAS® Text Miner is a very powerful tool to summarize text and identify meaningful patterns which is importance to a company that wishes to evaluate their customers opinions and sentiments.
- In this research, we show how text mining can be applied to detect general opinion about the Bing It On challenge.
- Cluster 4, Topic 5 and the concept link diagram of 'bingvsgoogle' (figure 6) shows that there are gifts associated with the challenge.
- Topic 3 and 4 in figure 7 clearly depicts that people are talking positively about Google, which does not hold for the case of Bing. In particular, people are declaring Google as the winner, not Bing.
- Cluster 3 in figure 8 represents the same fact as in topic 3 and 4, that people took the challenge and Google won, which is also evident in figure 4, concept link diagram of Google.
- Therefore, in this case SAS® Text Miner shows that the general opinion about the challenge is not similar to what Microsoft claims. As a matter of fact, the text mining of tweets result seems to show that "Google' wins the challenge more frequently and is still preferred by the users. This is of course subject to the self-selection bias imposed by the use of specific tweets that contain the keyword 'BingItOn' used n this study.

Reference

[1] "Bing Challenges nation to 'Bing It On'" from News Center of Microsoft.
http://www.microsoft.com/en-us/news/Press/2012/Sep12/09-06BingChallengePR.aspx
[2] "People Prefer Bing Web Search Results Over Google Nearly 2:1 – Bring on the Bing It On Challenge!" blog from Bing
http://www.bing.com/community/site_blogs/b/thedetails/archive/2012/08/06/bingchallenge.aspx?form=MFESOC&publ=BGITON&crea=TEXT_MFESOC_Challenge_Desktop-Landing_prblog-desktop-upper_1x1
[3] %GetTweet: SAS® Macro to Fetch and Summarize Tweets- Satish Garla and Goutam Chakraborty, Oklahoma State University, Stillwater, OK
[4] "Introduction to Text Miner" In SAS® Enterprise Miner Help.SAS® Enterprise Miner 6.2. SAS® Institute Inc., Cary, NC.

Acknowledgement

We wish to express our sincere thanks to Dr. Goutam Chakraborty for his valuable guidance.

Contact Information

Shreya Sadhukhan and Taufique Alam Ansari are pursuing masters degree in MIS along with SAS and OSU Data Mining Certificate in Oklahoma State University, Stillwater, OK 74078. E Mail address: shreya.sadhukhan@okstate.edu, taufique.ansari@okstate.edu.
Dr. Goutam Chakraborty is Professor(Marketing) and Founder of SAS and OSU Business Analytics Program, Oklahoma State University, Stillwater, OK 74078