ENVISIONING VIRGINIA TECH

BEYOND BOUNDARIES

WHAT TWITTER REVEALS ABOUT VIRGINIA TECH'S FUTURE

AN ANLYSIS OF CONTENT RELATED TO THE BEYOND BOUNDARIES INITIATIVE

PREPARED BY:

Alice Song

Office of the Senior Fellow for Resource Development

May 10, 2015

VirginiaTech

This research analyzes social media (specifically Twitter) to gauge different aspects of life at Virginia Tech for the Beyond Boundaries initiative. Specifically, we were interested in using Twitter data to understand people's perceptions unaltered by the researcher. Other sources of data collection for Beyond Boundaries invited people to share their ideas through an idea bank, participate in an input session, comment on committee work, or submit a personal video. These data collection methods rely on people's interest in participating. Participants know that their answers are being analyzed in some way. Social media, as a data source, does not include this barrier to research.

Although we were interested in learning about particular topics related to the future of Virginia Tech (see Figure 1), we also were interested in learning from the data. What discussions were taking place that we might not have identified through our background research? Were conversations about Virginia Tech positive or negative? We were open to what the data had to offer. This provides context to our method of learning from and refining a large data set of over 1 million tweets.

Data Collection and Phase I Analysis

We used tweet collection provided by the Virginia Tech digital library research lab as our original source. This data set includes 1,277,679 tweets, which were collected using the key words "Virginia Tech," "Blacksburg," and "NRV" from October 2012 to February 2015. Because of the limitation of Twitter API, we cannot obtain all tweets with these keywords. The tweets we can collect contain around one percent of tweets on Twitter. We assume many of tweets are related to Virginia Tech. As such, we used this existing data set for continued analysis for the Beyond Boundaries project.

In addition to these data, we analyzed tweets contain the term "VTBeyond," from August 12, 2015 to April 5, 2016. These data included a total of 327 tweets.

Beyond Boundaries asks the university community to imagine Virginia Tech in a generation's time. We turned to social media to assess people's comments related to their experiences at Virginia Tech. In order to narrow the existing data set, we developed a list of keywords related to the Beyond Boundaries project (Table 1). By using these keywords, we extracted tweets from the larger data set that related to our project. We used the keywords listed in Table 1 to filter to the large tweet collection.

In addition to filtering the data set with keywords, we also cleaned the data to remove specific recurring topics that were less related to the Beyond Boundaries project. For example, in the large tweet collection, we had a number of tweets about sports. In order to get tweets more closely related to the topics concerned by Beyond Boundaries project, we cleaned the data by removing sports related tweets. For example, we removed the sports related terms: football, basketball, coach, team, sport, and so on. After the filtering and cleaning process, we narrowed the larger data set to 14,527 tweets of interest.

Table 1: Beyond Boundaries Keywords

2047	entrepreneur	MOOC	spending
academics	faculty	ncr	state legislature
access	free	nsf	strategy
afford	frontier	online education	student loans
affordable	funding	outreach	study abroad
alumni	future	partnerships	technology
campus	global	privatization	tenure
certificate	globalize	professor	trends
classroom	graduate	Prosim	tuition
collaboration	higher education	publish	undergraduate
comment	highered	quality	ut prosim
cost	innovation	rank	UTProsim
degree	international	ranking	virtual
design	laboratory	reform	voice
direction	land grant	relationships	VTBeyond
diversity	learning	research	word
education	location	scholarship	workforce
electronic	long term	service	world

After reducing the data set, we tried different methods to make sense of the content of the tweets. The best result was to perform analysis for each keyword separately. We extracted noun phrases and performed topic modeling analysis for each keyword of interest. We believe that noun phrases can give us a better understanding of content rather than a single word. We completed topic analysis using these noun phrases.

We also looked at the frequency of noun phrases in the content of the tweets. From the topic modeling analysis, we obtained a general sense of the topics discussed on tweets for each keyword. From the noun phrases ranking, we identified the most-discussed topics for each keyword.

Phase II Analysis

In a second phase of analysis, we further refined the keyword selection from those terms presented in Table 1. In this phase, we selected keywords based on topic modeling results and tweet frequency. So in this process, we eliminated keywords and tweets that did not yield meaningful information or had a small number of related tweets. Figure two is the reduced keywords and sample Tweets we get from these keywords.

Table 2: Refined Beyond Boundaries keywords with sample tweets

Keywords	Sample tweets generated by keywords
Research	AFOSR supports research in nanomaterials as embedded damage detection in composites @virginia_tech http://t.co/75ikKu4XPV #AFOSR2014 #nano
Campus	I can't even explain how in love I was when I first saw Virginia Tech's campus
Afford	My first choices were actually University of Washington and Virginia Tech but I could not afford the out of state and being 22 hours away
Tuition	First day of classes at Virginia Tech for my daughter in her final year. Am getting emotional just thinking of no more tuition payments :)
Cost	Virginia Tech on a building boom but then raises fees because of the need to keep up with rising costs. Hmm wonder why.
Scholarship	After getting back all my decisions/scholarship offers I'm proud to say that I am officially a part of Virginia Tech class of 2018 #VT18
Faculty	@virginia_tech A happy faculty make students happy? The Chron's Best Colleges to Work For. http://t.co/XqNhfRULEb http://t.co/WRlpL77i1E
Professor	@kparacha @MariaPtweets my professor from Virginia Tech has quoted stuff from my blog in his new book on Wireless Communications...
Diversity	Today's story on @virginia_tech BOV affirming commitment to inclusion & diversity: http://t.co/SQ8q1mMcHp #principlesofcommunity #highered
International	Tech leads Virginia in international students http://t.co/G4xKHYq0

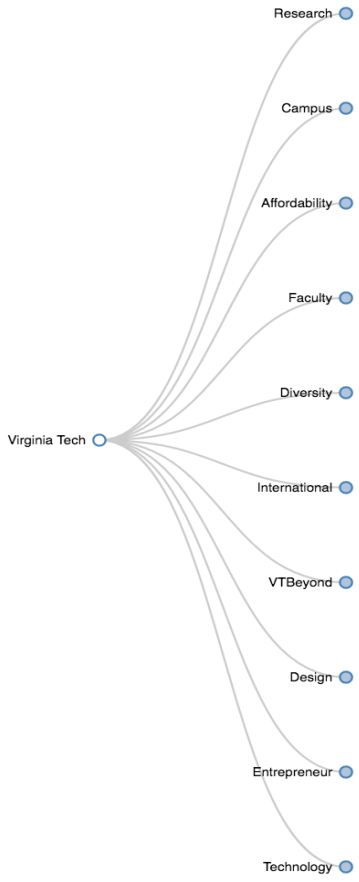
VTBeyond	RT @kpdepauw: Mayer: one example of wicked, challenging problems of the future. VT had a role to play @VTBeyond @VTSandsman https://t.co/3_Ñ_
Design	RT @vtnews: Aerospace engineering student team from Virginia Tech wins space vehicle design contest http://t.co/C6ZYHTo0
Entrepreneur	#veterans Virginia Tech hosts entrepreneurial workshop for veterans: Virginia Tech was selected because of its military com... #followme
Technology	HGTV's Doory Awards Feature $8 Million Virginia Farm http://t.co/i4LmpM8RLH #tech #technology

Collapsible tree visualization

We present the findings of the second phase analysis with collapsible tree visualizations. These are interactive graphs that show a hieratical structure of text. We used “Virginia Tech” as the main node, since all the contents are about “Virginia Tech”.

The next level of nodes contains the keywords that were of particular interest and resulted in a more significant number of tweets. Figure 1 shows these keywords as a node in the tree graph. We group some of the keywords in Table 2 together because we think they talked about similar topics. For example, we think the keywords “afford”, “cost,” ”tuition” and “scholarship” were all related to “affordability”. As such, the researchers grouped the analysis of these keywords under the original node, “affordability”.

Figure 1: First level of collapsible tree visualization



Each one of the nodes in the first level expands to a second level. The second level contains sub categories that we defined. For example, under the node “research”, according to the topic modeling results, we define five sub categories, “institutes”, “projects”, “researcher”, “resources” and “sentiment” (see Figure 2). Similar to “research”, for each original node, we used the topic modeling results to define the second-level grouping of the tree visualization.

Figure 2: Second level of collapsible tree visualization

We performed a third layer of analysis to create a third level grouping for each category in the second level. We used another round of text analysis to create this third grouping. See Figure 3.

Figure 3: Third level of collapsible tree visualization

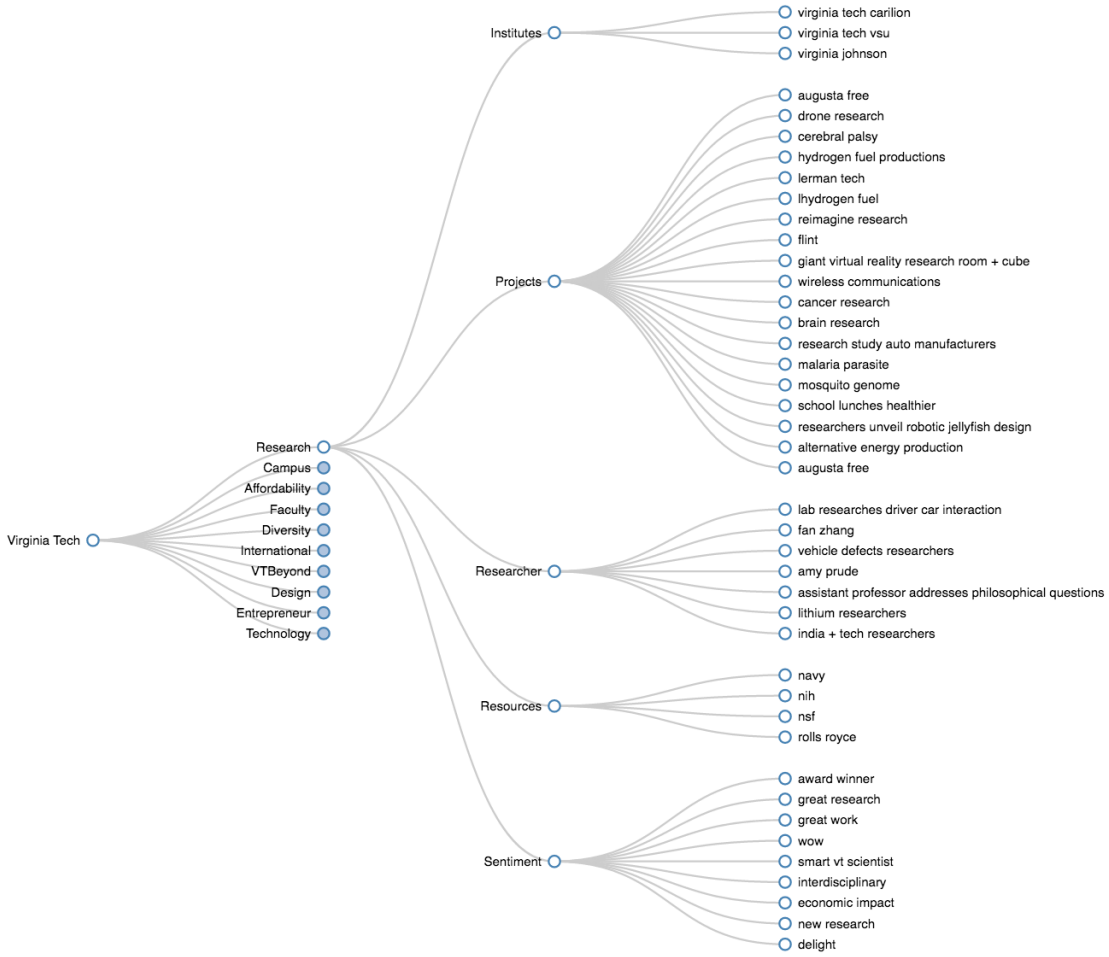


Figure 3 shows one example of the expanded third level. For example, for the word “research,” the source data are tweets containing keywords “research.” The second level shows the main content in our tweet collection related to research—“institutes,” “projects,” “researcher,” “resources,” and “sentiment.” We used the topic modeling results and frequency to create both the second and then a third level groupings. Under the “projects” node, we show the research projects related to Virginia Tech that were discussed most on Twitter. Similarly, for “researcher,” we show the identities of researchers that were most discussed in the tweets related to “research.”

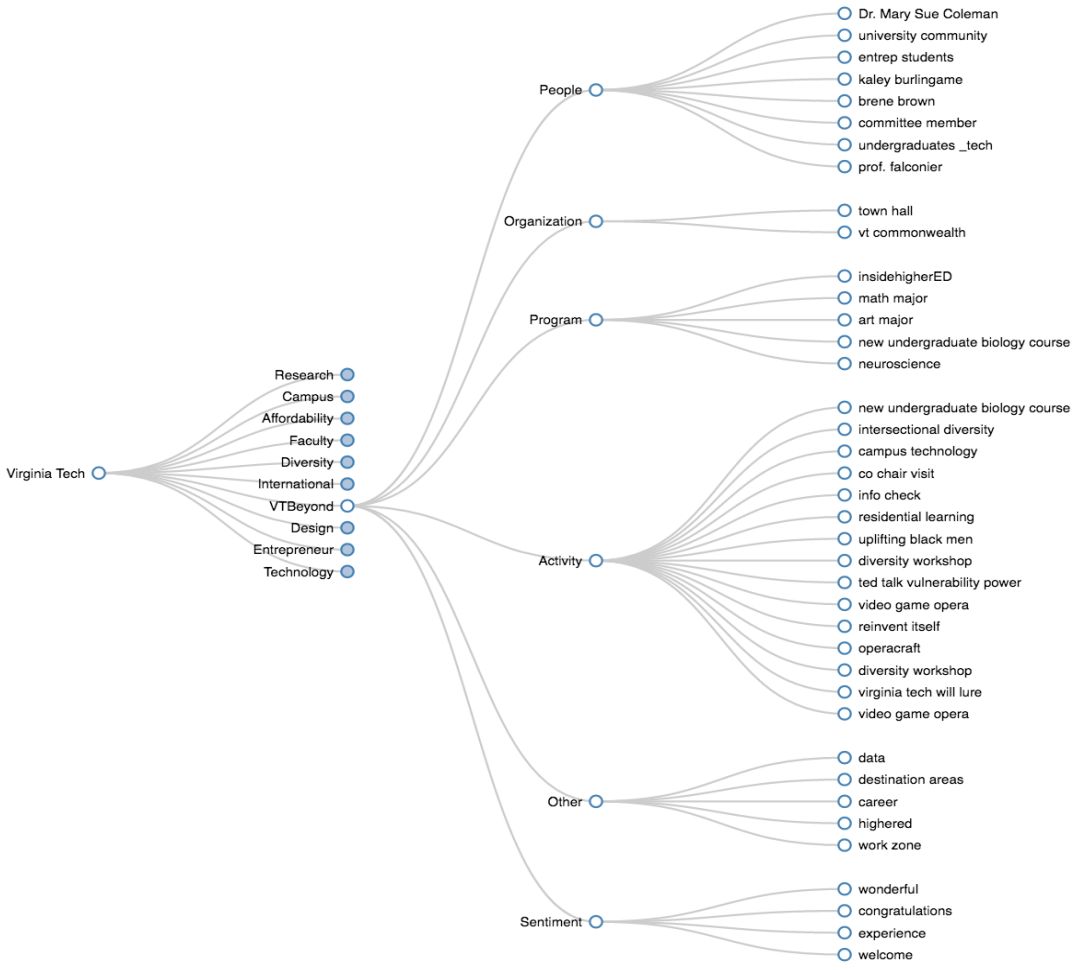
By using collapse tree visualization, we built an interactive way to view the contents of tweets in a hieratical structure. From each node in the first level, see aspects related to Beyond Boundaries as were discussed on Twitter. From the second level, you may find the categories of discussion. On third level, we reveal the main content of each discussion thread.

Keyword: VTBeyond

This paper offers an exploratory look and organization of topics related to Beyond Boundaries. We were particularly interested in the keyword, “VTBeyond” as it is highly relevant to the project. In this section, we take an in-depth look at this Twitter content.

The keyword “VTBeyond” expands from the first level into six different categories in the second level. Within these categories, you may get a sense of the main contents of the 327 tweets containing the term or hashtag “VTBeyond”. (See Figure 4)

Figure 4: Expanded node “VTBeyond”



Under the node “VTBeyond”, the first category is “people”, and contains names of people that are mentioned most in tweets related to “VTBeyond” during the timespan of data collection. For example,

“Dr. Mary Sue Coleman,” and “prof. falconier” were mentioned several times in tweets related to “VTBeyond.” Dr. Mary Sue Coleman launched the Beyond Boundaries project with an inspiring lecture to the university community. Professor Falconier has won a $7.2M award to study the well being of couples. University communities and students also fall under this category. The category “activity” contains activities that related to the project. Under “program”, we can find math major, art major, neuroscience and so on. It suggests that they are actively involved in the discussion on Twitter.

We believed that most of the conversation on Twitter relates to beyond boundaries will be collected by the keyword “VTBeyond”. By analyzing the content, we can know the dominant topics, events and people related to “VTBeyond” discussion on Twitter. We also think that it is important to build visualization tools that can make sense of these analyses. We believed that collapse tree visualization could assist readers to better understanding the content of conversations. From the tree graph, we can easily find the related people that were frequently discussed, the activities mentioned most, the sentiment of people’s discussions and so on. Although this involves some of manual inspection, it is still valuable for readers to look at it and get a general idea about what people are saying about “Beyond Boundaries”.

Conclusion

We believe that discussions on social media, especial Twitter, are valuable to us. In this research, we explored the topics that mattered most about “Virginia Tech” and “Beyond Boundaries.” We defined keywords of interest and extracted a set of twitter data that best satisfied our needs. After analysis, we built visualization tools to show the content analysis results to the reader, involving manually categorizing the text analysis results. This research is the beginning of further exploration of topics related to the Beyond Boundaries initiative. In the future, we can work on developing algorithms to generate categories in the tree graph automatically and perform visualization for dynamic Twitter data.