

The Summer Undergraduate Research Fellowship (SURF) Symposium  
7 August 2014  
Purdue University, West Lafayette, Indiana, USA

## Social Media Analytics Reporting Toolkit

Yuchen Cui, Junghoon Chae and David Ebert  
Department of Electrical and Computer Engineering, Purdue University

### ABSTRACT

With the fast growth of social media services, vast amount of user-generated content with time-space stamps are produced everyday. Considerable amount of these data are publicly available online, some of which collectively convey information that are of interest to data analysts. Social media data are dynamic and unstructured by nature, which makes it very hard for analysts to efficiently and effectively retrieve useful information. Social Media Analytics Reporting Toolkit (SMART), a system developed at Purdue VACCINE lab, aims to support such analyzing. The current framework collects real-time Twitter messages and visualizes volume densities on a map. It uses Latent Dirichlet Allocation (LDA) [1] to extract regional topics and can optionally apply Seasonal-Trend decomposition using Loess (STL) [2] to detect abnormal events. While Twitter has a fair amount of active users, they account for a small portion of total active social media users. Data generated by many other social media services are not currently utilized by SMART. Therefore, my work focused on expanding data sources of SAMRT system by creating means to collect data from other sources such as Facebook and Instagram. During a test run using a collection of 88 specified keywords in search, over two million Facebook posts were collected in one week. Besides, current SMART framework utilizes only one topic model, i.e. LDA, which is considered to be slower than Non-negative Matrix Factorization (NMF) [3] model, thus I also put my effort into integrating NMF algorithm into the system. The improved SMART system can be used to fulfill a variety of analyzing tasks such as monitoring regional social media responses from different sources in disastrous events, detecting user reported crimes and so on. SMART is currently an ongoing and promising project that can be further improved by integrating new features.

### KEYWORDS

Social Media Data Analyzing, Topic Extraction using NMF, Information Search and Retrieval

### REFERENCES

- [1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.
- [2] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess (with discussion). *Journal of Official Statistics*, 6:3–73, 1990.
- [3] V. P. Pauca, F. Shahnaz, M. W. Berry, R. J. Plemmons. Text Mining using Non-Negative Matrix Factorization. From: <http://csweb.cs.wfu.edu/~pauca/publications/SIAMDM03.pdf>, January, 2004.