Chapter 1
Introduction

Twitter®[1] is a massive social networking site tuned towards fast communication. More than 140 million active users publish over 400 million 140-character "Tweets" every day.[2] Twitter's speed and ease of publication have made it an important communication medium for people from all walks of life. Twitter has played a prominent role in socio-political events, such as the Arab Spring[3] and the Occupy Wall Street movement.[4] Twitter has also been used to post damage reports and disaster preparedness information during large natural disasters, such as the Hurricane Sandy.

This book is for the reader who is interested in understanding the basics of collecting, storing, and analyzing Twitter data. The first half of this book discusses collection and storage of data. It starts by discussing how to collect Twitter data, looking at the free APIs provided by Twitter. We then goes on to discuss how to store this data for use in real-time applications. The second half is focused on analysis. Here, we focus on common measures and algorithms that are used to analyze social media data. We finish the analysis by discussing visual analytics, an approach which helps humans inspect the data through intuitive visualizations.

1.1 Main Takeaways from This Book

This book provides hands-on introduction to the collection and analysis of Twitter data. No knowledge of data analysis, or social network analysis is presumed. For all the concepts discussed in this book, we will provide in-depth description of the underlying assumptions and explain via construction of examples. The reader will

[1] http://twitter.com

[2] https://blog.twitter.com/2012/twitter-turns-six

[3] http://bit.ly/N6illb

[4] http://nyti.ms/SwZKVD

S. Kumar et al., *Twitter Data Analytics*, SpringerBriefs in Computer Science,
DOI 10.1007/978-1-4614-9372-3_1, © The Author(s) 2014

gain knowledge of the concepts in this book by building a crawler that collects Twitter data in real time. The reader will then learn how to analyze this data to find important time periods, users, and topics in their dataset. Finally, the reader will see how all of these concepts can be brought together to perform visual analysis and create meaningful software that uses Twitter data.

The code examples in this book are written in Java®, and JavaScript®. Familiarity with these languages will be useful in understanding the code, however the examples should be straightforward enough for anyone with basic programming experience. This book does assume that you know the programming concepts behind a high level language.

1.2 Learning Through Examples

Every concept discussed in this book is accompanied by illustrative examples. The examples in Chap. 4 use an open source network analysis library, JUNG™,[5] to perform network computations. The algorithms provided in this library are often highly optimized, and we recommend them for the development of production applications. However, because they are optimized, this code can be difficult to interpret for someone viewing these topics for the first time. In these cases, we present code that focuses more on readability than optimization to communicate the concepts using the examples. To build the visualizations in Chap. 5, we use the data visualization library D3™.[6] D3 is a versatile visualization toolkit, which supports various types of visualizations. We recommend the readers to browse through the examples to find other interesting ways to visualize Twitter data.

All of the examples read directly from a text file, where each line is a JSON document as returned by the Twitter APIs (the format of which is covered in Chap. 2). These examples can easily be manipulated to read from MongoDB®, but we leave this as an exercise for the reader.

Whenever "…" appears in a code example, code has been omitted from the example. This is done to remove code that is not pertinent to understanding the concepts. To obtain the full source code used in the examples, refer to the book's website, *http://tweettracker.fulton.asu.edu/tda*.

The dataset used for the examples in this book comes from the Occupy Wall Street movement, a protest centered around the wealth disparity in the US. This movement attracted significant focus on Twitter. We focus on a single day of this event to give a picture of what these measures look like with the same data. The dataset has been anonymized to remove any personally identifiable information. This dataset is also made available on the book's website for the reader to use when executing the examples.

[5] http://jung.sourceforge.net/

[6] http://d3js.org

To stay in agreement with Twitter's data sharing policies, some fields have been removed from this dataset, and others have been modified. When collecting data from the Twitter APIs in Chap. 2, you will get raw data with unaltered values for all of the fields.

1.3 Applying Twitter Data

Twitter's popularity as an information source has led to the development of applications and research in various domains. Humanitarian Assistance and Disaster Relief is one domain where information from Twitter is used to provide situational awareness to a crisis situation. Researchers have used Twitter to predict the occurrence of earthquakes [5] and identify relevant users to follow to obtain disaster related information [1]. Studies of Twitter's use in disasters include regions such as China [4], and Chile [2].

While a sampled view of Twitter is easily obtained through the APIs discussed in this book, the full view is difficult to obtain. The APIs only grant us access to a 1 % sample of the Twitter data, and concerns about the sampling strategy and the quality of Twitter data obtained via the API have been raised recently in [3]. This study indicates that care must be taken while constructing the queries used to collect data from the Streaming API.

References

1. S. Kumar, F. Morstatter, R. Zafarani, and H. Liu. Whom Should I Follow? Identifying Relevant Users During Crises. In *Proceedings of the 24th ACM conference on Hypertext and social media*. ACM, 2013.
2. M. Mendoza, B. Poblete, and C. Castillo. Twitter Under Crisis: Can we Trust What We RT? In *Proceedings of the First Workshop on Social Media Analytics*, 2010.
3. F. Morstatter, J. Pfeffer, H. Liu, and K. Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *International AAAI Conference on Weblogs and Social Media*, 2013.
4. Y. Qu, C. Huang, P. Zhang, and J. Zhang. Microblogging After a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake. In *Computer Supported Cooperative Work and Social Computing*, pages 25–34, 2011.
5. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.