Siddhartha Chatterjee
Michal Krystyanczuk

Python Social Media Analytics

Analyze and visualize data from Twitter, YouTube, GitHub, and more

Packt>

Python Social Media Analytics

Analyze and visualize data from Twitter, YouTube, GitHub, and more

Siddhartha Chatterjee
Michal Krystyanczuk

Packt>

BIRMINGHAM - MUMBAI

Python Social Media Analytics

First published: July 2017

Production reference: 1260717

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham
B3 2PB, UK.
ISBN 978-1-78712-148-5

www.packtpub.com

About the Authors

Siddhartha Chatterjee is an experienced data scientist with a strong focus in the area of machine learning and big data applied to digital (e-commerce and CRM) and social media analytics.

He worked between 2007 to 2012 with companies such as IBM, Cognizant Technologies, and Technicolor Research and Innovation. He completed a Pan-European Masters in Data Mining and Knowledge Management at Ecole Polytechnique of the University of Nantes and University of Eastern Piedmont, Italy.

Since 2012, he has worked at OgilvyOne Worldwide, a leading global customer engagement agency in Paris, as a lead data scientist and set up the social media analytics and predictive analytics offering. From 2014 to 2016, he was a senior data scientist and head of semantic data of Publicis, France. During his time at Ogilvy and Publicis, he worked on international projects for brands such as Nestle, AXA, BNP Paribas, McDonald's, Orange, Netflix, and others. Currently, Siddhartha is serving as head of data and analytics of Groupe Aéroport des Paris.

Michal Krystyanczuk is the co-founder of The Data Strategy, a start-up company based in Paris that builds artificial intelligence technologies to provide consumer insights from unstructured data. Previously, he worked as a data scientist in the financial sector using machine learning and big data techniques for tasks such as pattern recognition on financial markets, credit scoring, and hedging strategies optimization.

He specializes in social media analysis for brands using advanced natural language processing and machine learning algorithms. He has managed semantic data projects for global brands, such as Mulberry, BNP Paribas, Groupe SEB, Publicis, Chipotle, and others.

He is an enthusiast of cognitive computing and information retrieval from different types of data, such as text, image, and video.

Credits

Authors
Siddhartha Chatterjee
Michal Krystyanczuk

Reviewer
Rubén Oliva Ramos

Commissioning Editor
Amey Varangaonkar

Acquisition Editor
Divya Poojari

Content Development Editor
Cheryl Dsa

Technical Editor
Vivek Arora

Copy Editor
Safis Editing

Project Coordinator
Nidhi Joshi

Proofreader
Safis Editing

Indexer
Tejal Daruwale Soni

Graphics
Tania Dutta

Production Coordinator
Arvindkumar Gupta

Acknowledgments

This book is a result of our experience with data science and working with huge amounts of unstructured data from the web. Our intention was to provide a practical book on social media analytics with strong storytelling. In the whole process of analytics, the scripting of a story around the results is as important as the technicalities involved. It's been a long journey, chapter to chapter, and it would not have been possible without our support team that has helped us all through. We would like to deeply thank our mentors, Air commodore TK Chatterjee (retired) and Mr. Wojciech Krystyanczuk, who have motivated and helped us with their feedback, edits, and reviews throughout the journey.

We would also like to thank our co-author, Mr. Arjun Chatterjee, for sharing his brilliant technical knowledge and writing the chapter on *Social Media Analytics at Scale*. Above all, we would also like to thank the Packt editorial team for their encouragement and patience with us. We sincerely hope that the readers will find this book useful in their efforts to explore social media for creative purposes.

About the Reviewer

Rubén Oliva Ramos is a computer systems engineer with a master's degree in computer and electronic systems engineering, teleinformatics, and networking specialization from University of Salle Bajio in Leon, Guanajuato, Mexico. He has more than five years of experience in developing web applications to control and monitor devices connected with Arduino and Raspberry Pi using web frameworks and cloud services to build Internet of Things applications.

He is a mechatronics teacher at University of Salle Bajio and teaches students studying the master's degree in Design and Engineering of Mechatronics Systems. He also works at Centro de Bachillerato Tecnologico Industrial 225 in Leon, Guanajuato, Mexico, teaching electronics, robotics and control, automation, and microcontrollers at Mechatronics Technician Career. He has worked on consultant and developer projects in areas such as monitoring systems and datalogger data using technologies such as Android, iOS, Windows Phone, Visual Studio .NET, HTML5, PHP, CSS, Ajax, JavaScript, Angular, ASP .NET databases (SQlite, MongoDB, and MySQL), and web servers (Node.js and IIS). Ruben has done hardware programming on Arduino, Raspberry Pi, Ethernet Shield, GPS, and GSM/GPRS, ESP8266, and control and monitor systems for data acquisition and programming.

I would like to thank my savior and lord, Jesus Christ, for giving me strength and courage to pursue this project, to my dearest wife, Mayte, our two lovely sons, Ruben and Dario. To my father, Ruben, my dearest mom, Rosalia, my brother, Juan Tomas, and my sister, Rosalia, whom I love, for all their support while reviewing this book, for allowing me to pursue my dream, and tolerating not being with them after my busy day job.

www.PacktPub.com

For support files and downloads related to your book, please visit `www.PacktPub.com`. Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at `www.PacktPub.com` and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at `service@packtpub.com` for more details. At `www.PacktPub.com`, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

`https://www.packtpub.com/mapt`

Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at `https://www.amazon.com/dp/1787121488`.

If you'd like to join our team of regular reviewers, you can email us at `customerreviews@packtpub.com`. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products!

Table of Contents

Preface

Social media in the last decade has taken the world by storm. Billions of interactions take place around the world among the different users of Facebook, Twitter, YouTube, online forums, Pinterest, GitHub, and others. All these interactions, either captured through the data provided by the APIs of these platforms or through custom crawlers, have become a hotbed of information and insights for organizations and scientists around the world. *Python Social Media Analytics* has been written to show the most practical means of capturing this data, cleaning it, and making it relevant for advanced analytics and insight hunting. The book will cover basic to advanced concepts for dealing with highly unstructured data, followed by extensive analysis and conclusions to give sense to all of the processing.

What this book covers

`Chapter 1`, *Introduction to the Latest Social Media Landscape and Importance*, covers the updated social media landscape and key figures. We also cover the technical environment around Python, algorithms, and social networks, which we later explain in detail.

`Chapter 2`, *Harnessing Social Data - Connecting, Capturing, and Cleaning*, introduces methods to connect to the most popular social networks. It involves the creation of developer applications on chosen social media and then using Python libraries to make connections to those applications and querying the data. We take you through the advantages and limitations of each social media platform, basic techniques to clean, structure, and normalize the data using text mining and data pre-processing. Finally, you are introduced to MongoDB and essential administration methods.

`Chapter 3`, *Uncovering Brand Activity, Emotions, and Popularity on Facebook*, introduces the role of Facebook for brand activity and reputation. We will also introduce you to the Facebook API ecosystem and the methodology to extract data. You will learn the concepts of feature extraction and content analysis using keywords, hashtags, noun phrases, and verbatim extraction to derive insights from a Facebook brand page. Trend analysis on time-series data, and emotion analysis via the AlchemyAPI from IBM, are also introduced.

`Chapter 4`, *Analyzing Twitter Using Sentiment Analysis and Entity Recognition*, introduces you to Twitter, its uses, and the methodology to extract data using its REST and Streaming APIs using Python. You will learn to perform text mining techniques, such as stopword removal, stemming using NLTK, and more customized cleaning such as device detection. We will also introduce the concept and application of sentiment analysis using a popular Python library, VADER. This chapter will demonstrate the classification technique of machine learning to build a custom sentiment analysis algorithm.

`Chapter 5`, *Campaigns and Consumer Reaction Analytics on YouTube - Structured and Unstructured*, demonstrates the analysis of both structured and unstructured data, combining the concepts we learned earlier with newer ones. We will explain the characteristics of YouTube and how campaigns and channel popularity are measured using a combination of traffic and sentiment data from user comments. This will also serve as an introduction to the Google developer platform needed to access and extract the data.

`Chapter 6`, *The Next Great Technology - Trends Mining on GitHub*, introduces you to GitHub, its API, and characteristics. This chapter will demonstrate how to analyze trends on GitHub to discover projects and technologies that gather the most interest from users. We use GitHub data around repositories such as watchers, forks, and open issues to while making interesting analysis to infer the most emerging projects and technologies.

Chapter 7, Scraping and Extracting Conversational Topics on Internet Forums, introduces public consumer forums with real-world examples and explains the importance of forum conversations for extracting insights about people and topics. You will learn the methodology to extract forum data using Scrapy and BeautifulSoup in Python. We'll apply the preceding techniques on a popular car forum and use Topic Models to analyze all the conversations around cars.

`Chapter 8`, *Demystifying Pinterest through Network Analysis of Users Interests*, introduces an emerging and important social network, Pinterest, along with the advanced social network analysis concept of Graph Mining. Along with the Pinterest API, we will introduce the technique of advanced scraping using Selenium. You will learn to extract data from Pinterest to build a graph of pins and boards. The concepts will help you analyze and visualize the data to find the most influential topics and users on Pinterest. You will also be introduced to the concept of community detection using Python modules.

`Chapter 9`, *Social Data Analytics at Scale - Spark and Amazon Web Services*, takes the reader on a tour of distributed and parallel computing. This chapter will be an introduction to implementing Spark, a popular open source cluster-computing framework. You will learn to get Python scripts ready to run at scale and execute Spark jobs on the Amazon Web Services cloud.