SOMECODE Documentation

Release latest

November 11, 2016

Contents

SOMECODE is a research platform for serious observation and analysis of Twitter data. SOMECODE brings together 9 years of unbroken continuity in developing social media research tools. Previous tools and processes developed by the contributor team are in daily use by many FORTUNE100 companies and major advertising agencies. SOMECODE is the solution we always wanted to build, but due to the kinds of restraints commercial entities have, never got to.

```
pip install somecode
```

All you need to have is Python 2.7 and the somecode installation will take care of all the dependencies.

CHAPTER 1

GREAT FOR

Somecode is great for researching a variety of topics, for example:

- Public figures
- Brands and organizations
- Special events (e.g. election)
- Emerging or ongoing crisis
- Ideas (e.g. radicalization)
- Websites and communities

Somecode comes with built-in scoring system and various signals that are specifically targeted at identifying bot, spam and troll accounts in Twitter, to help researchers better understand malicious techniques used in computer-aided propaganda.

CHAPTER 2

BENEFITS

Somecode takes you from an idea to understanding with one command in 20 seconds and allows any researher to start with serious social media research in minutes from being a total novice.

- Equal or better capabilities vs. best industrial solution
- Up to 10 million tweets per day with single API key
- Optmized for minimizing out-of-scope time use (configuration, data wrangling, etc.)
- Supports both streaming and rest API endpoints for both status (tweet) and user objects
- Provides an ideal environment for academic research and publication

SOMECODE is built by researchers for researchers and is very easy to extend / customize to suit specific needs. For most research scopes SOMECODE will work “out of the box”. It has very few depedencies (below) and takes minutes to deploy for your first research project.

CHAPTER 3

TYPICAL WORKFLOW

Depending on the need of the researcher at the time of use, a typical scenario involves no more than 2 or 3 function calls and depending on the size of the sample and the function used for collecting the data, no more than 1 minute. Such a scenario may involve:

1. Use one of the ‘data collector’ functions to get the data you need
2. Use one or more of the ‘report’ functions to visualize the data
3. Based on the findings, use one of the ‘data collectors’ to get drill-down data
4. Export report and dataset for reference / later use

CHAPTER 4

GETTING STARTED

The easiest case is:

```
pip install somecode
```

The hardest case is to first run this in shell:

```
sudo apt-get update -y; sudo apt-get install python-pip -y; sudo pip install --upgrade pip; sudo pip
```

And then run this in Jupyter / python shell:

```
import nltk; nltk.download('vader_lexicon'); nltk.download('vader_lexicon');
```

It will take less than a minute and because we're only using common packages, the installation should be painless in any case.

4.1 DESKTOP

SOMECODE runs on any regular low-end laptop with a common web browser. Tested on Linux and Mac OS X. Jupyter is highly recommended.

4.2 SERVER/CLOUD

SOMECODE runs on Amazon Nano (or similar) for $5 per month. Tested on Ubuntu 14.04LTS

For both options:

```
pip install somecode
```

SOMECODE is very easy to customize / extend if you would feel the need to do it. Even if you are a beginner python learner.

100% fun, 0% mindless wrangling.

CHAPTER 5

FEATURES

Collecting, processing and analyzing Twitter data comes with many caveats and obstacles, a factor that has kept most researchers oblivious to the potential Twitter data has. Many of Somecode's features are related with making all of that completely dissapear.

Some of the things we've figured out for you include:

- System performance
- Twitter API rate-limit mangagement
- JSON parsing
- Signal selection
- Plot configuration

The 'data collector' functions all return the same format pandas DataFrame, which means that you can use any plots, models, etc. to go where you want to go with the data.

CHAPTER 6

DATAFRAME COLUMNS

All of the 'data collection' methods ('search','stream','timeline','flatfile') return a pandas dataframe with direct signals from Twitter, together with SOMECODE scores and other inferred metrics.

VARIABLE NAME	ABOUT	DTYPE
days_since_creation	user	int64
influence_score	scores	int64
reach_score	scores	int64
quality_score	scores	int64
retweet_count	tweet	int64
text	tweet	object
user_tweets	user	int64
user_favourites	user	int64
user_followers	user	int64
user_following	user	int64
user_listed	user	int64
handle	user	object
created_at	user	datetime64
default_profile	user	bool
egg_account	user	bool
description	user	object
location	user	object
timezone	user	object
compound	sentiment	float
neu	sentiment	any
neg	sentiment	any
pos	sentiment	any

CHAPTER 7

FUNCTIONS

There are four categories of functions in SOMECODE:

- Data Collection
- Data Processing
- Reporting
- Export

7.1 DATA COLLECTION

There are four ways to get data in to SOMECODE.

FUNCTION	REQUIRED PARAMETERS	DATA SIZE
search()	keyword	max 3200 tweets
stream()	keyword or userid	up to 100 per second
timeline()	handle	up to 3200 tweets
flatfile()	filename	any size

To get 1000 tweets for keyword “election”:

```
some.search("election",1000)
```

To pen a stream with keyword “election”:

```
some.stream("election")
```

To get maximum number of tweets from @realdonaldtrump timeline:

```
some.timeline("realdonaldtrump")
```

To get tweets from a file:

```
some.flatfile("some_stream.json")
```

7.2 DATA PROCESSING

While it is possible to call any of the 20 or so modules included in Somecode as standalone functions, the ‘data processing’ modules are a little different in the sense that they are not called directly with the exception of keywords() which makes sense also to call directly.

FUNCTION	REQUIRED PARAMETERS	COMMENTS
data_frame()	Somecode dataframe	max 3200 tweets
data_prep()	Somecode dataframe	Just for flatfile()
keywords()	Any series with text	basic keyword stats

To compute entropy and other singals for textual data:

```
some.keywords(df)
```

Various additional semantic analysis is possible as part of freq_plot() and cooc_plot() reporting function.

7.3 REPORTING

There are two kinds of reporting capabilities; plots and tables. The tables come from the pretty.py library and plots are heavily customized Seaborn and Matplotlib plots.

FUNCTION	KIND OF REPORT
age_plot()	Bubble chart
bars()	Bar chart
cooc_plot()	Bubble chart
freq_plot()	Side-by-side bar
hist_plot()	Histogram
neg_plot()	Bubble chart
neg2_plot()	Bar chart
retweet_plot()	Bubble chart

For the Pretty descriptive tables:

FUNCTION	KIND OF REPORT
pretty.header()	Produces pretty header
pretty.table()	Produces pretty table
pretty.data()	Prepares data for table
pretty.toggle()	Hide code cells
pretty.warnings()	Turns of warnings

7.4 PLOTS

When you are on Jupyter, on the first line you must declare:

```
%matplotlib inline
```

Otherwise you will not see the plots.

7.5 PERFORMANCE

During the 2016 election, SOMECODE topical, sentiment, scoring and other computations have been tested in up to 200,000 tweets per hour volume using a single $50 per month server (8gb RAM) where the computations required for every 10 minute cycle were generally completed in 20 seconds.

7.6 UPGRADE

At anytime, you can update the current version of SOMECODE:

```
sudo pip install somecode --upgrade
```

7.7 BUILT ON

Frankly speaking, SOMECODE would not be possible without all the amazing technology solutions it's based on. What SOMECODE does, is put a few key technologies together, with "business logic" that came from working on over a thousand social media research projects since 2005. Somecode uses pandas, numpy, seaborn and matplotlib libraries heavily.

Other than that, dependent on the system, you should have minimal dependencies to worry about. Also if you're not using it already, I highly recommend Jupyter (http://jupyter.org/). It helps make programming much more about fun, and less about frustration.