Multiple Stream Dataset

Collected by:
Knowledge Discovery Lab
Department of Computer Science
Tennessee Tech University
Updated on Feb 14, 2018

This material is based upon work supported by the National Science Foundation under IIS Grant No. 1318657

Introduction

The dataset is collected from News API (https://newsapi.org/) and Twitter REST API (https://dev.twitter.com/rest/public).

The News API provides headlines from 70 worldwide sources including ABC News, BBC, Bloomberg, Business Insider, Buzzfeed, Associated Press, CNN, CNBC, ESPN, Google News etc. (A complete list of all the news sources we used to collect data from is shown in Appendix 1.) The Twitter REST API provides tweet and publicly available twitter handler information for a specified twitter handle.

The data collected in this set consists of news stories from 2/09/2017 to 6/23/2017, and associated tweets that occurred 10 days before and after the corresponding news story, based upon the twitter account (handle) mentioned in the body of the news.

How the Data was Collected?

First, we collected news data from News API. The data from News API have author name, news title, news headline, news url, published date, etc. Then, in order to get the body of the news story (which is not returned from the News API), we crawled the URL for the associated news source to get the body of the news.

Second, if the body of a news article references a twitter handle, the handle is sent to the Twitter REST API where all tweets 10 days around the published news story are collected.

The result is two separate, comma-delimited (*.csv*) files, *documents.csv* and *usertweet.csv*, corresponding to news stories and tweets respectively.

Table 1. A summary of the characteristics of Multiple Stream Dataset.

File Name	Unit of Record	Number of Record	Date Range
document.csv	news	9,917	02/09/17 - 06/23/17
usertweet.csv	tweets	9,971	02/11/17 – 06/23/17

Note: Out of the 9,917 news stories (documents), only 374 documents have corresponding tweets.

Files

document.csv

The *document.csv* file contains 7 variables. Each record is the individual news article pulled using News API. Table 2 lists the variables with a brief description.

Table 2. Variables and short description of documents.csv.

#	Variable Names	Description
1	id	Document Id (Primary Key)
2	date	Date news was published
3	author	Author who wrote the news article
4	source	News media where the news article was published on
5	domain name	The domain name where the news was published
6	url	The url of news article
6	body text	Body of the news

Figure 1 shows a snippet the actual data from the *document.csv* file.

id	date	author	source	domainName	url	bodyTEXT
1	2017-03-12T22:47:12Z	Joe Mullin	ars-technica	arstechnica.com	https://arstechnica.com/tech-po	Enlarge / Nick Denton during the
2	2017-03-12T15:00:04Z	John Timmer	ars-technica	arstechnica.com	https://arstechnica.com/science	reader comments 28 Share this
3	2017-03-12T14:00:43Z	Valentina Palladino	ars-technica	arstechnica.com	https://arstechnica.com/gadgets	Enlarge Valentina Palladino reader
4	2017-03-12T13:00:19Z	Annalee Newitz	ars-technica	arstechnica.com	https://arstechnica.com/science	reader comments 82 Share this
5	2017-03-11T19:00:21Z	Sam Machkovech	ars-technica	arstechnica.com	https://arstechnica.com/gaming	reader comments 11 Share this
6	2017-03-11T18:00:49Z	Nathan Mattise	ars-technica	arstechnica.com	https://arstechnica.com/tech-po	Enlarge / Last year at this time,
7	2017-03-11T17:00:03Z	Jonathan M. Gitlin	ars-technica	arstechnica.com	https://arstechnica.com/cars/20	reader comments 122 Share this
8	2017-03-11T15:00:24Z	David Kravets	ars-technica	arstechnica.com	https://arstechnica.com/tech-po	Otto Magus reader comments 83
9	2017-03-13T03:10:54Z	KIM TONG-HYUNG	associated-press	bigstory.ap.org	http://bigstory.ap.org/article/b6	Park's advisers offer to resign after
10	2017-03-13T02:59:31Z	MIKE FITZPATRICK	associated-press	bigstory.ap.org	http://bigstory.ap.org/article/60	Greensboro gets shot at Boeheim,
11	2017-03-13T02:58:00Z	The Associated Press	associated-press	bigstory.ap.org	http://bigstory.ap.org/article/14	NHL Capsules @import @import
12	2017-03-13T02:50:35Z	ELAINE KURTENBACH	associated-press	bigstory.ap.org	http://bigstory.ap.org/article/4e	Asian shares mostly higher on
13	2017-03-13T02:17:21Z	NOAH TRISTER	associated-press	bigstory.ap.org	http://bigstory.ap.org/article/c3	Lions reach deal with T.J. Lang
14	2017-03-13T02:15:42Z	DAVE CAMPBELL	associated-press	bigstory.ap.org	http://bigstory.ap.org/article/b3	AP Adrian Peterson takes free
15	2017-03-12T15:36:29Z	BBC News	bbc-news	bbc.co.uk	http://www.bbc.co.uk/news/uk-	Media playback is unsupported on
16	2017-03-13T02:44:07Z	BBC News	bbc-news	bbc.co.uk	http://www.bbc.co.uk/news/wo	Image copyright AFP Image
17	2017-03-13T00:00:20Z	BBC News	bbc-news	bbc.co.uk	http://www.bbc.co.uk/news/uk-	Image copyright AFP/Getty Image
18	2017-03-13T01:00:01Z	BBC News	bbc-news	bbc.co.uk	http://www.bbc.co.uk/news/uk-	Image copyright PA Image caption
19	2017-03-13T02:15:50Z	BBC News	bbc-news	bbc.co.uk	http://www.bbc.co.uk/news/uk-	Image copyright AFP The

Figure 1. Example data from document.csv file.

The detail description of each field in *document.csv* is given below:

1. **Variable Name:** id
   Type: Num
   Summary Statistics
   Range: 1 - 10,240
   Missing: 323
   Note: This is the primary key or unique identifier for all the news articles/documents. This document ID carries no information about the documents, and is provided solely for reference and data processing purposes.

2. **Variable Name:** date

Type: char

Format: YYYY-MM-DDTHH:MM:SSZ

Summary Statistics

Range: 02/09/17 – 06/23/17

Missing: 0

Note: The date when the news was published on the website.

3. **Variable Name:** author

Type: char

Summary Statistics

Range: N/A

Missing: 0

Unique Value: 2,412

Note: The name of the author who wrote the news. There are total 2,412 unique authors. In addition, 612 documents do not have an associated author.

4. **Variable Name:** source

Type: char

Summary Statistics

Range: N/A

Missing: 0

Unique Value: 38

Note: The name of the news media who published the news. This corresponds to the news source in Appendix 1.

5. **Variable Name:** domain name

Type: char

Summary Statistics

Range: N/A

Missing: 0

Unique Value: 49

Note: The domain name of the associated news website.

6. **Variable Name:** url

Type: char

Summary Statistics

Range: N/A

Missing: 0

Note: The URL of the news article, where it was pulled from.

7. **Variable Name:** body text

Type: char

Summary Statistics

Range: N/A

Missing: 0

Note: The body of the news article. The News API doesn't provide the body text of the news article. We crawled the URL given by the News API and grabbed the body text from that URL, which could potentially include a mentioned twitter handle. If a twitter handle is mentioned, we pulled corresponding tweets using the Twitter REST API. These tweets are stored in usertweet.csv.

usertweet.csv

The *usertweet.csv* file contains 6 variables. Each record is the individual tweet.

Table 3. Variables and short description of usertweet.csv.

#	Variable Names	Description
1	id	Tweet Id (Primary Key)
2	docid	Document Id that correspond to each news document in document.csv
3	news date	Date news was published
4	tweet date	Date tweet was done
5	screen name	Tweeter screen name
6	tweet	JSON dump of the tweet

Figure 2 shows a snippet of the actual real data from *usertweet.csv* file.

id	docid	news date	tweet date	screen name	tweet
1	8	3/11/17	3/9/17	dmkravets	{u'contributors': None, u'truncated': False, u'text': u'RT @jbrodkin: @dmkravets CELTICS', u'i
2	8	3/11/17	3/2/17	dmkravets	{u'contributors': None, u'truncated': False, u'text': u'@Driveguy2000 @arstechnica Our story
3	44	3/12/17	3/3/17	usattybharara	{u'contributors': None, u'truncated': False, u'text': u'Everyone deserves to be free from fear
4	58	3/12/17	3/22/17	morningmika	{u'contributors': None, u'truncated': False, u'text': u'Very informative new report on the Arct
5	58	3/12/17	3/21/17	morningmika	{u'contributors': None, u'truncated': False, u'text': u'Humility. How refreshing. https://t.co/4
6	58	3/12/17	3/21/17	morningmika	{u'contributors': None, u'truncated': False, u'text': u'"I really love this country and I believe ir
7	58	3/12/17	3/21/17	morningmika	{u'contributors': None, u'truncated': False, u'text': u"I'll disagree with some of his major rulin
8	58	3/12/17	3/21/17	morningmika	{u'contributors': None, u'truncated': True, u'text': u'It may sound dry but Meacham strategic
9	58	3/12/17	3/21/17	morningmika	{u'contributors': None, u'truncated': False, u'text': u'@jpaceDC asked important questions th
10	58	3/12/17	3/21/17	morningmika	{u'contributors': None, u'truncated': True, u'text': u'Why does Kellyanne Conway speak as if t
11	58	3/12/17	3/21/17	morningmika	{u'contributors': None, u'truncated': False, u'text': u'RT @JoeNBC: RYAN: You\u2019re annou
12	58	3/12/17	3/21/17	morningmika	{u'contributors': None, u'truncated': False, u'text': u'Mark on Squawk https://t.co/R1PcjHQD
13	58	3/12/17	3/22/17	cowspiracy	{u'contributors': None, u'truncated': False, u'text': u'#Repost @truthordrought\n\u30fb\u30
14	58	3/12/17	3/22/17	cowspiracy	{u'contributors': None, u'truncated': False, u'text': u'\u201cOne of the most shocking things t
15	58	3/12/17	3/22/17	cowspiracy	{u'contributors': None, u'truncated': False, u'text': u'RT @TryVeg: DYK animal ag is a leading
16	58	3/12/17	3/21/17	cowspiracy	{u'contributors': None, u'truncated': False, u'text': u'\U0001f331\u270c\U0001f3fc\U0001f3
17	58	3/12/17	3/20/17	cowspiracy	{u'contributors': None, u'truncated': False, u'text': u'"Just as Cowspiracy caused people to re
18	58	3/12/17	3/19/17	cowspiracy	{u'contributors': None, u'truncated': False, u'text': u'"With the arctic melting at an alarming r
19	58	3/12/17	3/18/17	cowspiracy	{u'contributors': None, u'truncated': False, u'text': u"#Repost @onegreenplanet\n\u30fb\u3
20	58	3/12/17	3/17/17	cowspiracy	{u'contributors': None, u'truncated': False, u'text': u'RT @JVM: Conspiracy & WhattheH

Figure 2. Example data from usertweet.csv file.

The detail description of each field in *usertweet.csv* is given below:

1. **Variable Name:** id
   Type: Num
   Summary Statistics
      Range: 1-10,239
      Missing: 268
      Note: This is the primary key or unique identifier for all the tweets. This tweet ID carries no information about the tweet or document, and is provided solely for reference and data processing purposes.

2. **Variable Name:** docid
   Type: Num
   Summary Statistics
      Range: 1-10,240
      Unique Value: 374 (Only 374 news documents have corresponding tweet)
      Note: This is the foreign key to documents. This tweet refers to or is present in the document with the associated *docid*. This document ID carries no information about the documents, and is provided solely for reference and data processing purposes.

3. **Variable Name:** news date
   Type: char
   Format: M/DD/YY
   Summary Statistics

Range: 2/28/17 – 6/23/17

Missing: 0

Note: The date when the news was published on the website.

4. **Variable Name:** tweet date

   Type: char

   Format: M/DD/YY

   Summary Statistics

   Range: 10 days before and after the document (news) was published

   Missing: 0

   Note: The date when the tweet was done by the user. We have pulled all the tweets associated with the corresponding Twitter handle mentioned on the news body text before and after 10 days of the news published.

5. **Variable Name:** screen name

   Type: char

   Summary Statistics

   Range: N/A

   Missing: 0

   Unique Value: 401

   Note: Twitter screen name/user account. These are pulled from the news body text. Every twitter handle mentioned in that news are included for that document.

6. **Variable Name:** tweet

   Type: char

   Summary Statistics

   Range: N/A

   Missing: 0

   Note: JSON dump of the corresponding tweet. It can be parsed using any JSON parser to extract the necessary information.

Appendix 1

We used various English language news source from the News API. The list of all news sources used in this dataset from the News API are as follows:

SN.	News Source	Category
1.	Ars Technica	Technology
2.	Associated Press	General
3.	BBC News	General
4.	BBC Sports	Sport
5.	Business Insider	Business
6.	Business Insider (UK)	Business
7.	CNBC	Business
8.	CNN	General
9.	Entertainment Weekly	Entertainment
10.	Fortune	Business
11.	FourFourTwo	Sports
12.	Google News	General
13.	Hacker News	Technology
14.	Independent	General
15.	Mashable	Entertainment
16.	Mirror	General
17.	MTV News	Music
18.	MTV News (UK)	Music
19.	New Scientist	Science & Nature
20.	Newsweek	General
21.	New York Magazine	General
22.	NFL News	Sports
23.	Recode	Technology
24.	Reddit /r/all	General
25.	The Guardian (AU)	General
26.	The Guardian (UK)	General
27.	The Huffington Post	General
28.	The Lad Bible	Entertainment
29.	The Next Web	Technology
30.	The Telegraph	General
31.	The Verge	Technology
32.	The Wall Street Journal	Business
33.	The Washington Post	General
34.	USA Today	General
35.	The Times of India	General