Packaging Twitter Data for the Repository

WADL – JCDL 2016

Mark Phillips
UNT Libraries
June 23, 2016
After we collect Twitter data, what do we do with it?
For me the answer typically is...

“Put it in the repository”
Featured Collections

United States Experiment Station Record

Published by the United States Department of Agriculture, the *Experiment Station Record* provides abstracts of agricultural experiments conducted during the year, along with statistics, convention reports, bulletins, and bibliographies. The collection contains volumes 1-95 covering 1889 to 1946. These materials were donated by Vanderbilt University and digitized by the UNT Libraries.

Featured Partner

UNT Institute of Applied Science

The mission of the Institute of Applied Science (IAS) is to foster, facilitate and conduct science-based interdisciplinary environmental research that seeks to understand how human actions impact...
"Stand With Wendy" Twitter Dataset

**Date:** 2013-06-25/2013-07-03  
**Creator:** Phillips, Mark Edward  
**Description:** This dataset contains Twitter JSON data for several Twitter search queries collected the week following the filibuster by Wendy Davis in the Texas Senate related to Senate Bill 5, using the twarc (https://github.com/edsu/twarc) package that makes use of Twitter’s search API. A total of 560,954 Tweets make up the combined dataset.

**Contributing Partner:** UNT Libraries

---

"Nelson Mandela" Twitter Dataset

**Date:** December 15, 2013  
**Creator:** Phillips, Mark Edward  
**Description:** This dataset contains Twitter JSON data for several Twitter search queries that were collected the week following the death of Nelson Mandela on December 5, 2013 using the twarc (https://github.com/edsu/twarc) package that makes use of Twitter’s search API. A total of 10,678,479 Tweets make up the combined dataset.

**Contributing Partner:** UNT Libraries
"Stand With Wendy" Twitter Dataset

Description: This dataset contains Twitter JSON data for several Twitter search queries collected the week following the filibuster by Wendy Davis in the Texas Senate related to Senate Bill 5, using the twarc (https://github.com/edsu/twarc) package that makes use of Twitter's search API. A total of 560,954 Tweets make up the combined dataset.

Creator(s): Phillips, Mark Edward
Location(s): United States - Texas
Creation Date: 2013-06-25/2013-07-03
Partner(s): UNT Libraries
Collection(s): UNT Data Repository
Usage: Total Uses: 200
Past 30 days: 5
Yesterday: 0

Feedback:
If you are having problems, need to report errors, or have questions or comments for the staff, please use our Feedback Form.

Support the Digital Library
The UNT Digital Library operated by the UNT Libraries provides resources to the UNT Community and users around the world. Please consider supporting the UNT Digital Library today. Donate Today.
## "Stand With Wendy" Twitter Dataset

<table>
<thead>
<tr>
<th>Filename</th>
<th>Size</th>
<th>Format</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>#IStandWithWendy_.json.gz</td>
<td>10.2 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>#PissedAtPerry.json.gz</td>
<td>7.5 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>#SB5.json.gz</td>
<td>22.1 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>#istandwithwendy.json.gz</td>
<td>10.1 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>#sittdownwendy.json.gz</td>
<td>10.9 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>#standwithwendy.json.gz</td>
<td>114.1 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>#txlege.json.gz</td>
<td>40.4 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>@GovernorPerry.json.gz</td>
<td>9.0 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>@WendyDavisTexas.json.gz</td>
<td>23.8 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>README.txt</td>
<td>2.3 KB</td>
<td>text/plain</td>
<td>Download</td>
</tr>
<tr>
<td>Rick_Perry.json.gz</td>
<td>13.5 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>Wendy_Davis.json.gz</td>
<td>30.7 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
</tbody>
</table>
Readme for StandWithWendy Related Tweets (JSON) Dataset:

This dataset is being made available using a Creative Commons CC0 Public Domain Dedication. See http://creativecommons.org/publicdomain/zero/1.0/

Data contained within each tweet sample is copyright by the tweet's authors and is subject to all copyright law. The creator of the dataset makes no assertion of rights related to the content of each tweet.

The dataset is provided 'as is' without warranty or any representation of accuracy, timeliness or completeness.

The dataset format is line-oriented JSON in utf8 encoding.

Dataset Overview:

This dataset contains Twitter JSON data for several Twitter search queries that were collected the week following the filibuster by Wendy Davis in the Texas Senate related to Senate Bill 5, using the twarc (https://github.com/edsu/twarc) package that makes use of Twitter’s search API. The following searches are included in this dataset.

#IStandWithWendy
#PissedAtPerry
#SB5
#istandwithwendy
#sitdownwendy
#standwithwendy
#txlege
@GovernorPerry
@WendyDavisTexas
"Rick Perry"
"Wendy Davis"

Twitter data making up this dataset was captured between 2013-06-25 and 2013-07-03.

All Twitter data is stored as line-oriented JSON (each line is a complete JSON document), and is exactly what was received from the Twitter API.

Files:

IStandWithWendy.json.gz : 21,975 Tweets for the search IStandWithWendy
#PissedAtPerry.json.gz : 16,199 Tweets for the search #PissedAtPerry
#SB5.json.gz : 40,037 Tweets for the search #SB5
#istandwithwendy.json.gz : 21,520 Tweets for the search #istandwithwendy
#sitdownwendy.json.gz : 20,745 Tweets for the search #sitdownwendy
**Description:** This dataset contains Twitter JSON data for several Twitter search queries that were collected the week following the death of Nelson Mandela on December 5, 2013 using the twarc (https://github.com/edsu/twarc) package that makes use of Twitter’s search API. A total of 10,678,479 Tweets make up the combined dataset.

**Creator(s):** Phillips, Mark Edward

**Creation Date:** December 15, 2013

**Partner(s):** UNT Libraries
About | Browse this Partner

**Collection(s):** UNT Data Repository
About | Browse this Collection

**Usage:**
- Total Uses: 73
- Past 30 days: 1
- Yesterday: 0
"Nelson Mandela" Twitter Dataset

<table>
<thead>
<tr>
<th>Filename</th>
<th>Size</th>
<th>Format</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>#NelsonMandela.json.gz</td>
<td>454.4 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>#RIPNelsonMandela.json.gz</td>
<td>771.7 MB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>Nelson_Mandela.json.gz</td>
<td>4.5 GB</td>
<td>application/x-gzip</td>
<td>Download</td>
</tr>
<tr>
<td>README.txt</td>
<td>1.6 KB</td>
<td>text/plain</td>
<td>Download</td>
</tr>
</tbody>
</table>

Feedback:
If you are having problems, need to report errors, or have questions or comments for the staff, please use our Feedback Form.

Support the Digital Library
The UNT Digital Library operated by the UNT Libraries provides resources to the UNT Community and users around the world. Please consider supporting the UNT Digital Library today. Donate Today.
Readme for Nelson Mandela Related Tweets (JSON) Dataset :

This dataset is being made available using a Creative Commons CC0 Public Domain Dedication. See http://creativecommons.org/publicdomain/zero/1.0/

Data contained within each tweet sample is copyright by the tweet's authors and is subject to all copyright law. The creator of the dataset makes no assertion of rights related to the content of each tweet.

The dataset is provided 'as is' without warranty or any representation of accuracy, timeliness or completeness.

The dataset format is line-oriented JSON in utf8 encoding.

Dataset Overview :

This dataset contains Twitter JSON data for several Twitter search queries that were collected the week following the death of Nelson Mandela on December 5, 2013 using the twarc (https://github.com/edsu/twarc) package that makes use of Twitter's search API. The following searches are included in this dataset.

#NelsonMandela
#RIPNelsonMandela
"Nelson Mandela"

Twitter data making up this dataset was captured between 2013-06-25 and 2013-07-03.

All Twitter data is stored as line-oriented JSON (each line is a complete JSON document), and is exactly what was received from the Twitter API.

Files :

#NelsonMandela.json.gz : 833,273 Tweets for the search #NelsonMandela
#RIPNelsonMandela.json.gz : 1,475,992 Tweets for the search #RIPNelsonMandela
Nelson_Mandela.json.gz : 8,369,214 Tweets for the search "Nelson Mandela"

If you find errors or have questions, contact Mark Phillips at mark.phillips@unt.edu
Adding social media collections to an existing repository infrastructure seems to be a logical first step in sharing this content.
Attention should be paid to agreements between the researcher and the social media platform that take the form of an End User License Agreement.
A simple txt readme document is an important component to provide context and provenance to the social metadata content being added to the repository.
Questions?

mark.phillips@unt.edu