Socially-Informed Timeline Generation Corpus (first released on March 2015)
URL: http://www.cs.cornell.edu/~luwang/data.html

This corpus is distributed together with:

Socially-Informed Timeline Generation for Complex Events
Lu Wang, Claire Cardie, and Galen Marchetti.
Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2015

==== Content ====

I. Description of the datasets
II. Contact

==== I. Description of the datasets ====

There are two main datasets under directory /PATH/TO/socialtimeline:

1) NYT2013: New York Times news articles and user comments.
It is under /PATH/TO/socialtimeline/nyt2013.

2) TIMELINE2014: New York Times, CNN, and BBC news articles and user comments on four major events happened in 2014.
It is under /PATH/TO/socialtimeline/events (articles and comments) and /PATH/TO/socialtimeline/goldstandard (gold standard timelines).

======== 1) NYT2013 ========

We collected all articles with comments from NYT in 2013 to form a training set for learning importance scoring functions on articles sentences and comments (see details in Section 3). NYT2013 contains 3,863 articles and 833,032 comments.

Articles are stored as /PATH/TO/socialtimeline/nyt2013/articles/[articleID], and the corresponding comments are stored as /PATH/TO/socialtimeline/nyt2013/comments/[articleID].

======== 2) TIMELINE2013 ========

>> ARTICLES AND COMMENTS
We crawled news articles and comments (if there is any) from New York Times (NYT), CNN, and BBC on four trending events in 2014: the missing Malaysia Airlines Flight MH370 (MH370), the political unrest in Ukraine (Ukraine), the Israel-Gaza conflict

(Gaza), and the NSA surveillance leaks (NSA). More descriptions can be found in the paper (Section 2).

The corresponding articles and comments from different sources for each event can be found under /PATH/TO/socialtimeline/events/[mh370|ukraine|gaza|nsa]/[nyt/cnn/bbc]/articles/[articleID] and /PATH/TO/socialtimeline/events/[mh370|ukraine|gaza|nsa]/[nyt/cnn/bbc]/comments/[articleID].

Notice that we didn't collect any comment for BBC articles, and for NSA dataset, we only collected data from NYT and CNN.

>> GOLD-STANDARD TIMELINES
We constructed gold-standard timelines for each of the four events from the corresponding Wikipedia page(s), NYT topic page, or BBC news page. The timelines can be found under /PATH/TO/socialtimeline/goldstandard/[mh370|ukraine|gaza|nsa].

Here are the links to the web pages from which the gold-standard timelines are constructed. Please note that Wikipedia editors constantly make changes for the article pages. You may find the information in the URLs below different from the gold-standard used in our experiments.

>>>> MH370:
http://en.wikipedia.org/wiki/Timeline_of_Malaysia_Airlines_Flight_370

>>>> Ukraine:
http://en.wikipedia.org/wiki/Timeline_of_the_2014_pro-Russian_unrest_in_Ukraine
http://en.wikipedia.org/wiki/Timeline_of_the_annexation_of_Crimea_by_the_Russian_Federation
http://www.bbc.com/news/world-middle-east-26248275

>>>> Gaza:
http://en.wikipedia.org/wiki/Timeline_of_the_2014_Israel%E2%80%93Gaza_conflict

>>>> NSA:
http://topics.nytimes.com/top/reference/timestopics/organizations/n/national_security_agency/index.html

==== II. Contact ====

Should you have any questions, please contact luwang@cs.cornell.edu (Lu Wang).