# Specifications of RSS crawlers

rsscrawler.py is unique main code for crawling all RSS sources from one news.

Command format as follows:

rsscrawler.py -n <news name> -i <sources file> [-s <stopwords file>]

For example, news name is '*cnn*'. This name is used to indicate the subdirectory for storing the crawling results. In addition, the name is also used as config mark of special processing.

"sources file" format as follows:

*http://rss.cnn.com/rss/edition.rss*
*http://rss.cnn.com/rss/edition_world.rss*
*http://rss.cnn.com/rss/edition_africa.rss*
*http://rss.cnn.com/rss/edition_americas.rss*
...

"stopwords file" format as follows:

*a*
*the*
*who*
…

The following explanation gives the purpose of some the files that are generated in the process:

1. RSSName_wordsfreq.db stores words frequency file generated from fetched web pages in order to visualize words trends of News API.
2. fileName.db (fileName means one RSS source link address) stores all links of web pages fetched in terms of one RSS link source.
3. MERGE.TXT stores updated links, title, source, and date of web pages fetched from last commit in terms of one News source.
4. filename.html stores the whole webpage fetched, where filename is gotten by news' title.

Algorithm as follows:

```
1.      while  a new source is still  not fetched:
2.              load all fetched Links from a filename .db file;
3.              load all words frequency statistic from a  RSSName_wordsfreq.db file
4.              get a XML file from a RSS source;
5               while new items are still not read in the XML:
6.                      if there is still new link to not fetched:
7.                              store the link, title, date, and name in MERGE.TXT;
8.                              store a whole webpage in a HTML file in terms of the link;
9.                              calculate words frequency;
10.             if there is new links to fetched in the process:
11.                     update all new links into a filename .db file;
12.                     update words frequency record file ,  RSSName_wordsfreq.dbt;
13.                     generate words frequency record file as format
                         "date \t source \t words \t frequency" to visualize news words trends;
```

The code is shared at Github , that is, https://github.com/fangnster/RSScrawler .