Web news text dataset of typhoon in China（2004-2018）

Data Documentation

I. Dataset/atlas content features

i. Abstract

This web news text dataset of typhoon in China is gathered by a web crawler from the SINA.com, which is an online news media source with the largest user group in China. The dataset includes title, time, and text. The data format is Excel. The spatial scope is China. The temporal range is 2004-2018. There are 3445 typhoon news text.

ii. Elements (content fields)

There are totally 3 fields in the dataset, whose meaning is as follows:

"rowid": News text serial number

"title": News text headlines

"date": News text release time

"body_text": News body

iii. Temporal cover

2004-2018

iv. Spatial cover

The spatial range of this dataset is China.

II. Subject/industry scope of dataset/atlas

i. Subject scope

Geography .

ii. Industry scope

Disaster Risk Reduction

iii. Other classifications (optional)

(Other categories can be applied, but should reflect the dataset/atlas characteristics.)

III. Accuracy of dataset/atlas

i. Time frequency

ii. Spatial reference, accuracy, and granularity

IV. Dataset/atlas storage management

i. Data quantity

3.79MB

ii. Type format

Excel

iii. Update management

Irregular updating

V. Quality control of the dataset/atlas

i. Production mode

News reports related to typhoon are gathered from the SINA.com website. The website provides a search page for news reports via keywords of news content and title. All names of provinces in China, and "typhoon" ("台风" in Chinese) are used as the search keywords. The search pages return a search result list that meets the search criteria. Using a web crawler, typhoon news reports can be collected, including title, time, and text.

ii. Data sources (condition selection)

The dataset is collected or downloaded from the internet. The results are classified by sorting data.

iii. Methods of the data acquisition and processing (condition selection)

Acquisition method: Gathering news reports related to typhoon from the SINA.com. The website provides a search page for news reports via keywords of news content and title. The search pages return a search result list that meets the search criteria. Using a web crawler, typhoon news reports can be collected, including title, time, and text.

Processing method: There are many repetitive and similar texts in the news reports from SINA.com According to the timing of a news release, related news reports usually emerge in large numbers within 2–3 days after a disaster. Therefore, we sort the news reports according to the release time. Then, the event location text are extracted from the news title and contrasted one by one. If the similarity of place names from two news reports is greater than 0.5, they will be considered to be duplicated, and the latest news text is retained. Finally, we obtained 5339 typhoon news reports.

VI. Sharing and usage method of the dataset/atlas

i. Sharing methods and restrictions

Fully opened sharing

ii. Contact information of the sharing service (condition selection)

Contact Information for Service：

Name: Service group of Disaster Risk Reduction Knowledge Service System of IKCEST

Address：A11 Datun Road, Chaoyang District, Beijing

Zip Code：100101

E-mail: ikcest-drr@lreis.ac.cn

iii. Conditions and methods of usage

The dataset can be read by ArcGIS software and Microsoft office.

VII. Intellectual property rights of the dataset/atlas

i. Property rights (optional)

The property of the dataset belongs to the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences.

ii. Reference method of the dataset/atlas

Web news text dataset of typhoon in China（2004-2018）. Disaster Risk Reduction Knowledge Service of International Knowledge Centre for Engineering Sciences and Technology (IKCEST) under the Auspices of UNESCO, 2018.9.28.http://drr.ikcest.org/info/9931f.

iii. Usage contacts of the datasets/atlas

Name: Service group of Disaster Risk Reduction Knowledge Service System of IKCEST

Address: A11 Datun Road, Chaoyang District, Beijing .

Postcode: 100101

Telephone: 010-64889048-8006

Email: ikcest-drr@lreis.ac.cn

VIII. Others (optional)

In addition to the above, other information must also be explained.

Data documentation author information			
Data documentation author	HanXuehua	Update time	2018-09-28
Organization	Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences.		
Contact information			

Address	A11 Datun Road, Chaoyang District, Beijing .		PostcodeS	100101
Telephone	010-64889048-8006	E-mail	hanxh@lreis.ac.cn	