Weibo Sentiment Index DataSet related to the Flood in Shouguang, China (2018)

Data Documentation

I. Dataset/atlas content features
   i. Abstract
      This dataset is Weibo Sentiment Index Data related to the Flood in Shouguang, China. Using web crawlers and Weibo API, original Weibo messages were collected with “Shouguang” as the keyword. The following information was extracted: user ID, timestamp (i.e., the time when the message was posted), text (i.e., the text message posted by a user), and location information. Then sentiment index was calculated by ‘Tencent’ natural language processing (NLP) API. The data format is Excel. The spatial scope is China. The temporal range is 00:00 on August 19 2018 and 00:00 on August 28 2018. There are 77,581 Weibo texts.
   ii. Elements (content fields)
      There are totally 5 fields in the dataset, whose meaning is as follows:
      “ID”: Weibo text serial number
      “UserLocation”: the user location of Weibo text
      “CreateTime”: Weibo text release time
      “Text”: Weibo text
      “Sentiment_index”: sentiment, range 0-1, the closer to 1, the more positive, the closer to 0, the more negative.
   iii. Temporal cover
      00:00 on August 19 2018 to 00:00 on August 28 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
      9.31MB
   ii. Type format
      Excel
   iii. Update management
      Irregular updating

V. Quality control of the dataset/atlas
   i. Production mode
      Data acquisition and preprocessing are carried out by means of network crawler, regular expression and
natural language processing. The ‘Tencent’ natural language processing (NLP) platform is used to calculate sentiment index

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: Using web crawlers and Weibo API, 28,608 original Weibo messages were collected with “Shouguang” as the keyword with timestamps between 00:00 on August 19 and 00:00 on August 28. The following information was extracted: user ID, timestamp (i.e., the time when the message was posted), text (i.e., the text message posted by a user), and location information.

Processing method: The original Weibo texts contain interference information, for instance, http hyperlink, spaces, punctuation marks, hashtag, and @user. Text filtering is necessary to eliminate noise and improve the efficiency of word segmentation. These types of interference information were removed by the regular expression operations (‘re’ module) in Python. Very short Weibo texts (less than four words) and duplicated Weibo texts were deleted. The ‘Tencent’ natural language processing (NLP) platform is used to calculate sentiment index

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**

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.

<table>
<thead>
<tr>
<th>Data documentation author information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data documentation author</td>
</tr>
<tr>
<td>Organization</td>
</tr>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>Contact information</td>
</tr>
<tr>
<td>Address</td>
</tr>
<tr>
<td>Telephone</td>
</tr>
</tbody>
</table>