Readme for Twitter Dataset

Updated on Nov 7, 2010

1 Data Description

We provide both training set and test set (collected from September 2009 to January 2010) in the paper You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users in CIKM 2010.

The training set contains 115,886 Twitter users and 3,844,612 updates from the users. All the locations of the users are self-labeled in United States in city-level granularity.

The test set contains 5,136 Twitter users and 5,156,047 tweets from the users. All the locations of users are uploaded from their smart phones with the form of "UT: Latitude,Longitude".

2 Paper Citation

Please cite the following paper when using the dataset.

Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In Proceeding of the 19th ACM Conference on Information and Knowledge Management (CIKM), Tonronto, Oct 2010. (Bibtex)

3 Data Format

Four text files are in the package:

- "training_set_users.txt" contains user information in the training set in the form of "UserID\tUserLocation".

- "training_set_tweets.txt" contains tweets in the training set in the form of "UserID\tTweetID\tTweet\tCreatedAt".
- "test_set_users.txt" contains user information in the test set in the form of "UserID\tUserLocation".
- "test_set_tweets.txt" contains tweets in the test set in the form of "UserID\tTweetID\tTweet\tCreatedAt".

4 Contact

Please feel free to contact **Zhiyuan Cheng** if you have any question about the dataset.

Email: zcheng AT cse DOT tamu DOT edu
Homepage: http://students.cse.tamu.edu/zcheng/

5 Agreement