Readme for Social Honeypot Dataset

Updated on July 15, 2011

1 Data Description

We provide social honeypot dataset collected from December 30, 2009 to August 2, 2010 on Twitter. The dataset was used in the paper *Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter* in ICWSM 2011.

The dataset contains 22,223 content polluters, their number of followings over time, 2,353,473 tweets, and 19,276 legitimate users, their number of followings over time and 3,259,693 tweets.

2 Paper Citation

Please cite the following paper when using the dataset.

K. Lee, B. Eoff, and J. Caverlee. *Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter*. In Proceeding of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM), Barcelona, July 2011. (Bibtex)

3 Data Format

Six text files are in the package:

- “content_polluters.txt” contains content polluters’ profile information in the form of “UserID\tCreatedAt\tCollectedAt\tNumberOfFollowings \tNumberOfFollowers \tNumberOfTweets\tLengthOfScreenName\tLengthOfDescriptionInUserProfile”.

1
• “content_polluters_followings.txt” contains user information in the test set in the form of “UserID\tSeriesOfNumberOfFollowings (each number of following is separated by ,)”.

• “content_polluters_tweets.txt” contains tweets in the form of “UserID \tTweetID\tTweet\tCreatedAt”.

• “legitimate_users.txt” contains legitimate users’ profile information in the form of “UserID\tCreatedAt\tCollectedAt\tNumberOfFollowings \tNumberOfFollowers \tNumberOfTweets\tLengthOfScreenName\tLengthOfDescriptionInUserProfile”.

• “legitimate_users_followings.txt” contains user information in the test set in the form of “UserID\tSeriesOfNumberOfFollowings (each number of following is separated by ,)”.

• “legitimate_users_tweets.txt” contains tweets in the form of “UserID \tTweetID\tTweet\tCreatedAt”.

4 Contact

Please feel free to contact Kyumin Lee if you have any question about the dataset.

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

5 Agreement

Social Datasets by [Infolab] is licensed under a [Creative Commons Attribution-Noncommercial 3.0 United States License]