Crash Tweets and Crash Reports from Ma3Route

Overview

Using the Twitter API, tweets were scrapped from the twitter handle @Ma3Route. Ma3Route is a mobile/web/SMS platform that crowdsources transport data and provides users with information on traffic, road traffic crash (RTC), matatu directions and driving reports. Users post RTC or traffic information to Ma3Route, where Ma3Route then publishes the post on Twitter. Tweets were obtained in order to identify tweets that reported RTC.

This documentation describes a "truth dataset" of tweets that were manually coded to determine which reported a crash and the location of any reported crashes.

Dataset

We create a "truth dataset" to inform using tweets from Ma3Route to identify crashes. The truth dataset consists of all tweets from July 1, 2017 – July 31, 2018 that contain RTC related keywords. Specifically, these are tweets that contain one of the following keywords:

> *accident, accidents, ajali, collision, crash, crashes, crashs, crush, crushed, damage, disaster, emergency, fatal, fatality, fender-bender, fender bender, hazard, hit, hit-and-run, incident, incidents, injuries, injury, magari zmegongana, mishap, overturn, pileup, rol, rold, roled, roll, rolld, rolled, smash, smashed, wreck, wreckage, zilicrash, zimecrash*

and misspelled versions of accident, incident, crash and crashed. We consider a word misspelled version of "accident" or "incident" if the word has one or two character differences from accident or incident and consider a word a misspelled version of "crash" or "crashed" if the word has a one character difference from crash or crashed.

Six team members were trained to manually label tweets. Each tweet was manually coded by two different team members. The process followed the following steps:

Crash Information

First, tweets were labeled as to whether they referenced a crash. If the tweet did reference a crash and contained location information, the coordinates of the crash were recorded and any landmarks and roads mentioned in the tweet used to determine the coordinates were recorded. After two team members labeled the crash, and additional team member checked for and resolved any discrepancies.

For instructions provided to team members to complete the task, see:
ma3map_coding_instructions.pdf

Clustering Crashes to Crash Reports

Multiple tweets may refer to the same crash. After all tweets were labelled, team members grouped crash reports by applying a crash ID. Team members referenced the time, location and content of tweets to determine if multiple tweets referenced the same crash. Discrepancies between team members were not resolved, as this process can be somewhat subjective; rather, we provide two sets of crash IDs.

For instructions provided to team members to complete the task, see:
crash_clustering_protocol.pdf

Dataset

The **tweets_truth.dta** dataset contains the following variables:

- **uid:** Unique ID
- **tweet_id:** Tweet ID, which can be hydrated to obtain the tweet and metadata. From January – March 2018, only tweets were obtained—not tweet IDs; consequently, tweet IDs are missing in those month.
- **created_at:** Date/time of tweet (in East Africa Time).
- **crash_report:** Whether the tweet reports or references a crash
- **latitude:** latitude of the reported crash
- **longitude:** longitude of the reported crash
- **crash_id_c1:** Crash ID, from one round of coding
- **crash_id_c2:** Crash ID, from a separate round of coding
- **crash_landmark:** Landmark used to geocode the crash