after implementation, we found that the recall of the algorithm was relatively low, such that the set of tweets that discuss a particular event was often split across several clusters. In CrisisTracker, all new clusters are therefore compared with the current clusters to check for overlap. We refer to such a cluster of clusters as a story, and as the next section explains, this method also enables human intervention in the clustering process. Initial informal evaluation suggests that our approach greatly improves recall without substantially reducing precision, but accurate measurement of cluster recall for Twitter-scale corpuses is a research problem in itself. We, therefore, leave the specific cluster evaluation for future work and instead focus in this paper on evaluating the general ability of CrisisTracker to improve situation awareness and support decision-making.

A limitation of any bag-of-words-based event detection technique is that clusters do not necessarily correspond to events, as tweets can potentially have high textual similarity and be grouped together without discussing the same event. We have observed this with sensor-based feeds that publish regular updates about weather and earthquakes. In addition, because the system would quickly run out of storage space if all content was retained, increasingly larger stories and all their content are deleted with increasing age, unless they have been tagged by a human. Stories consisting of a single tweet are kept for approximately one day.

**Crowd curation and metadata creation**

The purpose of clustering the tweet stream into stories is to facilitate crowd curation. De-duplication (ideally) eliminates redundant work, directly reduces the number of items to process per time unit, enables size-based ranking of stories, and groups together reports that mention the same event but contain different details necessary for piecing together a complete narrative.

Search and filtering requires metadata for stories. Some of this metadata is extracted automatically, such as time of the event (timestamp of first tweet), keywords, popular versions of the report, and number of unique users who mention the story (referred to as story size). Story size enables CrisisTracker to estimate how important the message is to the community that has shared it [6, 9]. Users of the system can rank stories by their size among all Twitter users or among the 5,000 users most frequently tweeting about the disaster. In our experience, the top-5,000 option better brings out stories with detailed incremental updates to the situation, whereas the full rank more frequently includes summary articles, jokes, and opinions. Because metadata is assigned per-story, it also covers future tweets in the same story.

Curators are directed toward recent and extensively shared stories but can self-select which stories to work on. The first curation step is to further improve the clustering, by optionally merging the story with possible duplicate stories that are textually similar but fall below the threshold for automated merging. Misclassified content can also be removed from stories, which are then annotated (Figure 2) with geographic location, deployment-specific report categories (e.g., infrastructure damage or violence), and named entities. Stories deemed irrelevant (e.g., a cooking