Extracting a Topic Specific Dataset from a Twitter Archive

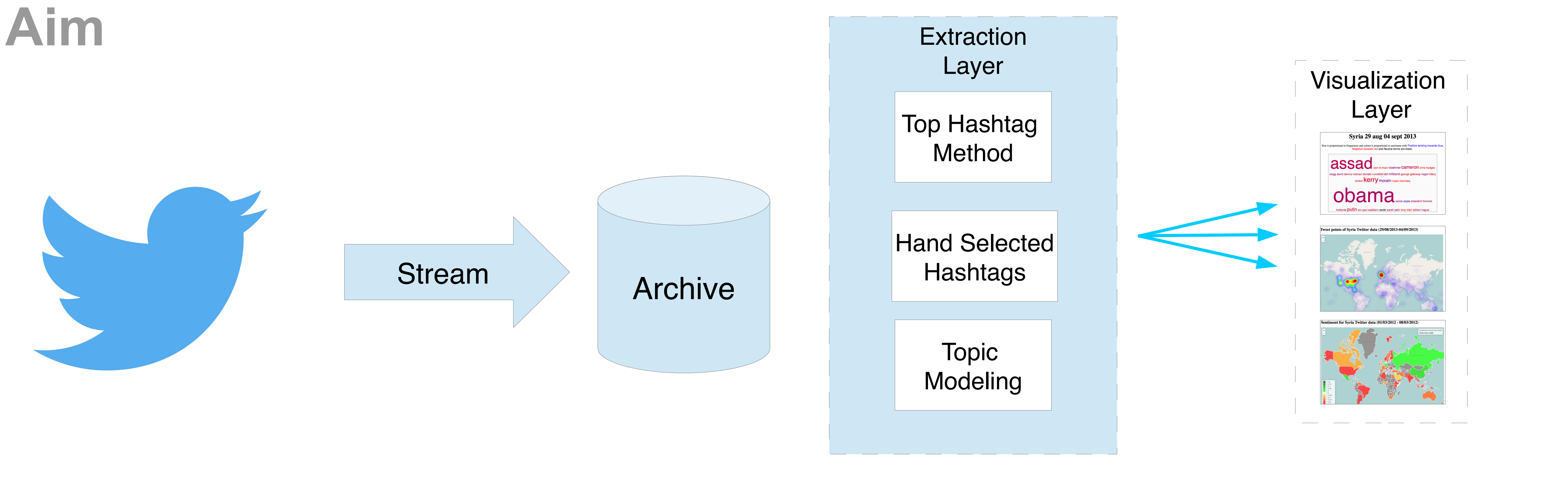
Summary

Here we describe and compare three methods for extracting tweets to form a topic-specific dataset from a Twitter archive. The data was streamed from the Twitter API and limited to English tweets. The aim was to select tweets that discussed Syria and Syrian specific events.

To create a base set from which we could expand we initially gathered all tweets from the archive that contained the term ‘syria’ in either upper or lower case. The three methods used to expand the dataset were 1) using common hashtags in the set as search terms, 2) using hand selected hashtags from the set as search terms and 3) using LDA topic modeling to cluster data and selecting Syrian related clusters. The data sets extracted were across two time frames: a week in March 2012, and a week in August and September 2013. During the 2012 week the UK embassy in Damascus was closed. During the 2013 week the UK Parliament voted not to authorize military action over chemical weapons use in Syria.

We found that the relevance scores for the top hashtags method were low. The relevance scores for the hand selected hashtags method was high but the set was small. The topic modeling approach gave a larger but slightly less relevant set.

Aim



Methods

The base set of data was extracted using the term 'syria'. This data is augmented using three methods:

- Top Hashtags

From the base set the top 40 hashtags from both time periods were selected and normalized The hashtag terms (hashtags with the # removed) were used as search terms to gather more data. Not all tweets in the set were about the Syrian conflict, for example, the hashtag #UK collected tweets about various activities that were happening in the UK in the selected weeks.

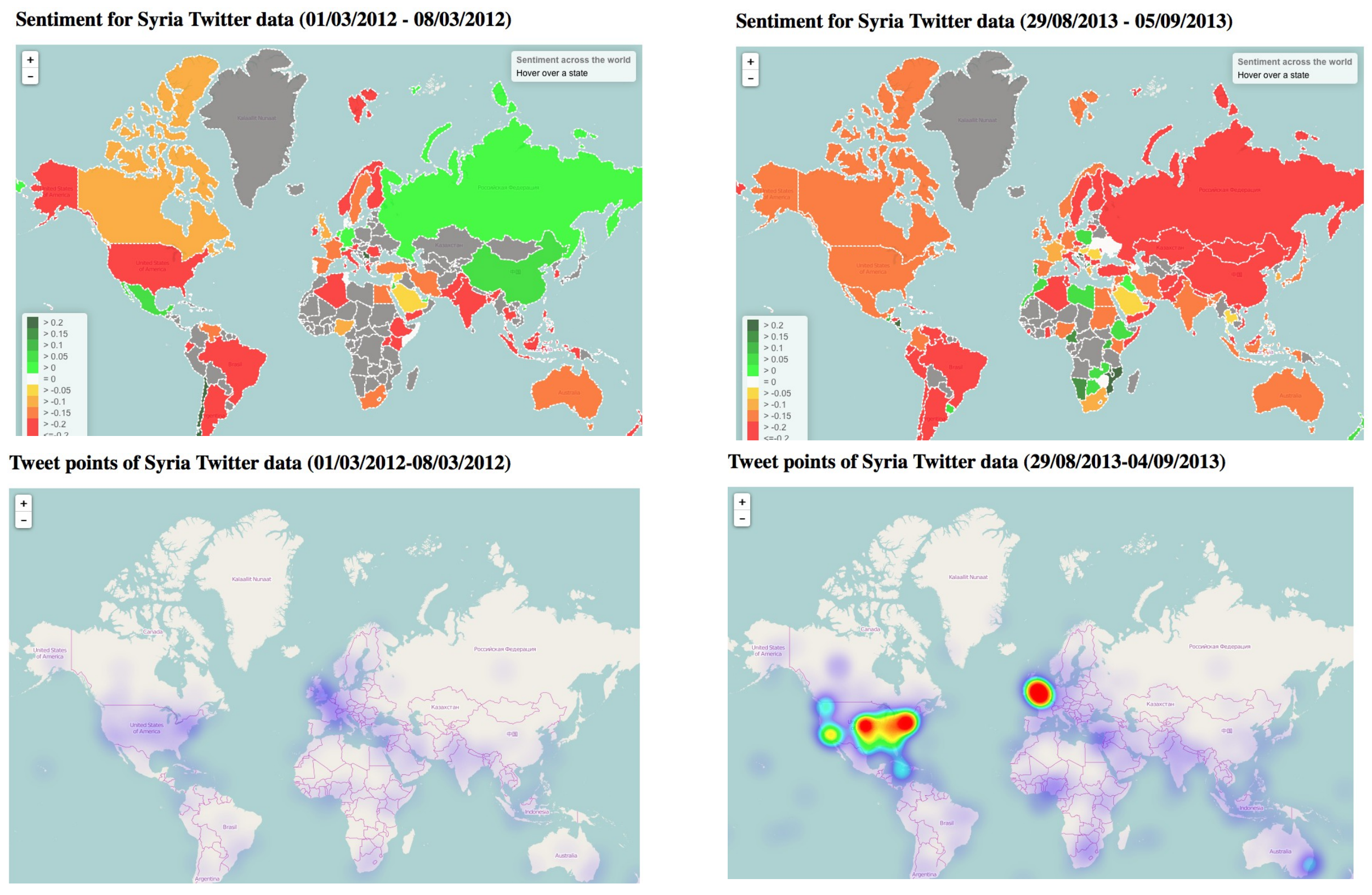
- Hand Selected Hashtags

From the base set all hashtags that had a higher frequency than 10 (2012 set) or 20 (2013 set) were selected. Each hashtag was annotated by two human coders as either directly relating to the Syrian conflict or not. This included all locations, people and institutions from Syria or formed to deal with Syria or anything with any of those items incorporated into a compound term, for example ‘norway4syria’. The human coders were in perfect agreement on which tags were related giving 32 which were used as search terms to gather more data from the archive.

- Topic Modeling

Data from the full set was clustered using LDA topic modeling. A score for each tweet for each of the various topic is given. We assign each tweet to the topic for which it has the highest score. A list of the top 20 words in each topic was compiled. The topics that were classed as relevant for this task were those which have ‘syria’ as one of these terms. Tweets that were allocated one of these topics were classed as relevant and assigned to the dataset. This approach was implemented using the Mallet tool-kit.

Visualization


Sentiment for Syria Twitter data (01/03/2012 - 08/03/2012)

Sentiment for Syria Twitter data (29/08/2013 - 05/09/2013)

Tweet points of Syria Twitter data (01/03/2012-08/03/2012)

Tweet points of Syria Twitter data (29/08/2013-04/09/2013)

Results

The size of the data sets is shown to the right. The aim of the process is to get a large relevant set. The percentage that are relevant give an overview of the likely pollution of the dataset and the F-score gives an indication of accuracy for each method.

Size	Size of Dataset (No. Tweets)	
Data set	2012	2013
Full Set	9,988,193	112,722,991
Top Hashtags	25,753	231,724
Hand Selected Hastags	2,555	23,838
Topic Modelling	2,292	60,013

Relevance

The percentage of tweets that are relevant was calculated through a manual evaluation. For each of the 6 datasets (one for each of the 3 methods for each time period) 100 tweets were randomly selected for manual examination. Each tweet was coded as relevant or irrelevant to the conflict in Syrian by two annotators.

- The top hashtags approach gives very low relevance results (large data set not relevant to the topic)
- The hand selected hashtags method gives high relevance scores (small data set very relevant to the topic)
- The topic modeling approach provides a high level of relevance for the smaller 2012 set but lower for the larger 2013 set

Percent Relevant	2012			2013		
	Annotator 1 (%)	Annotator 2 (%)	Kappa	Annotator 1 (%)	Annotator 2 (%)	Kappa
Top Hashtags	9	8	0.936	14	17	0.886
Hand Selected Hashtags	95	91	0.695	100	100	1.00
Topic Modeling	92	89	0.826	61	57	0.876

Accuracy

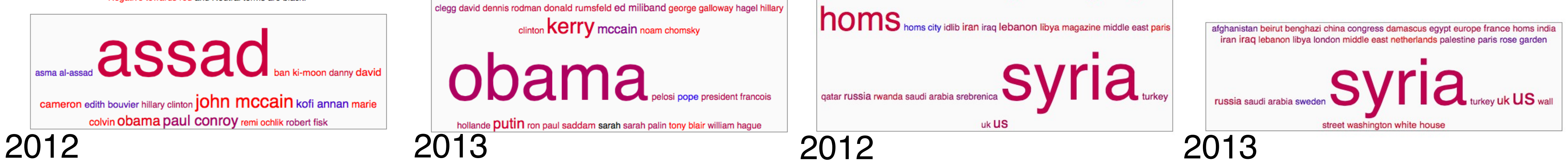
An F-score was calculated by comparing the automatically generated results against a gold standard set. The tweets used to create the gold standard were 1000 randomly chosen tweets from each time period extracted from the top hashtag set (this gave a set with a higher number of relevant tweets and, therefore, made the accuracy evaluation task difficult and the results more robust). Each tweet was annotated as relevant or not.

- The highest F-score was for the hand selected approach for 2012 set.
- Both approaches showed a drop in F-score for the larger 2013 set (smaller for the topic modeling)
- While the precision score of the hand selected approach increased for the 2013 set the recall score decreased.
- The hand selected approach did select appropriate tweets but it also missed many, providing a relevant but small set. The opposite happens for the topic modeling approach.
- Overall, the F-scores for both datasets are lower but there was a lower drop in accuracy between the two sets. As a drop precision is balanced by a rise in the recall.

Accuracy	2012			2013		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Hand Selected Hastags	0.92	0.98	0.95	0.95	0.66	0.78
Topic Modelling	0.76	0.80	0.78	0.60	0.90	0.72

Visualization

Size is proportional to frequencey and colour is proportional to sentiment with Positive tending towards blue, Negative towards red and Neutral terms are black.



THE UNIVERSITY of EDINBURGH
informatics

Clare Llewellyn (C.A.Llewellyn@sms.ed.ac.uk),
Claire Grover, Bea Alex, Jon Oberlander and Richard Tobin
School of Informatics
University of Edinburgh