COMMIT/ infinity ADNEXT

Real-Time Storage and Processing of Geo-located Tweets in MongoDB

Ali Hürriyetoglu, Manos Tsagkias, Antal van den Bosch
29/11/2013
Center for Language Studies
Radboud University, Nijmegen

Radboud University Nijmegen

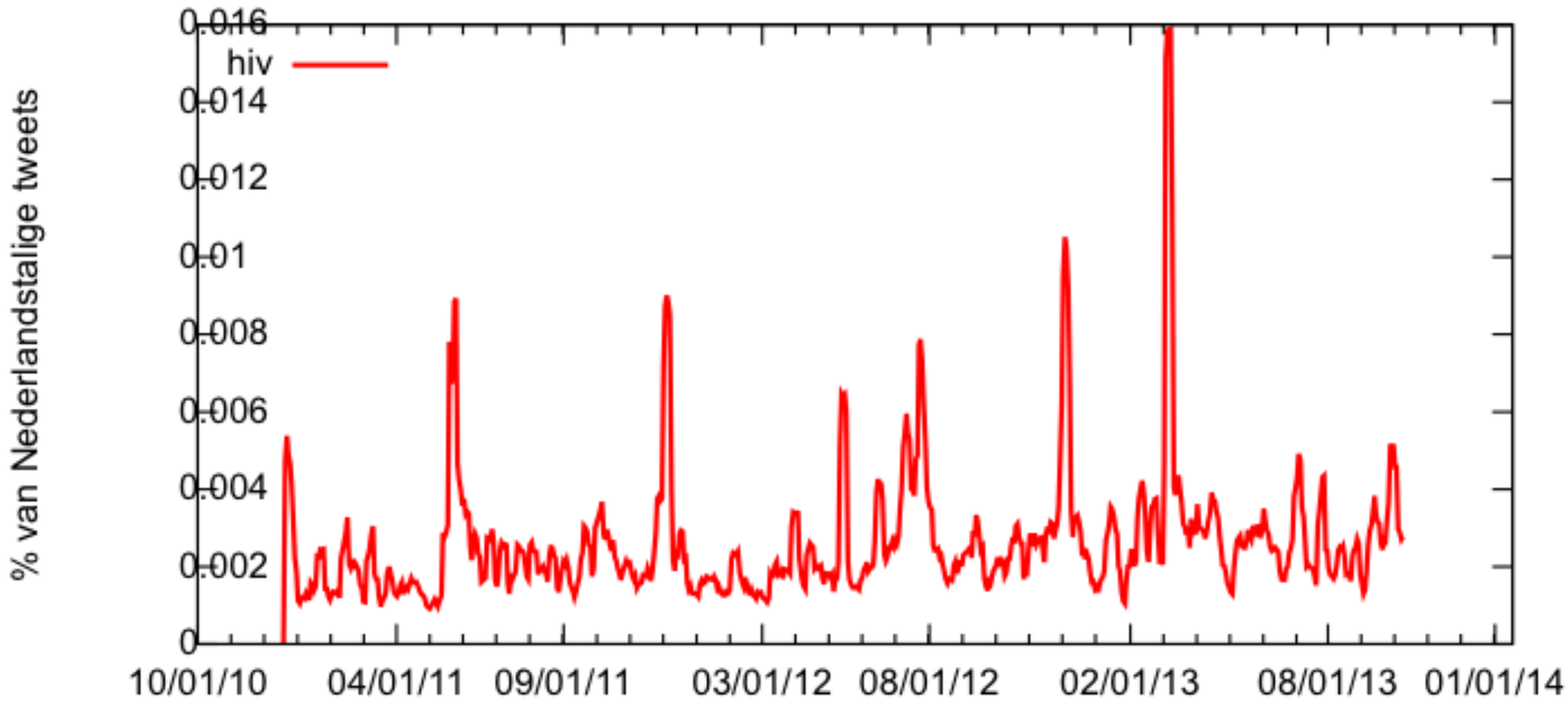
Zoek in tweets:

hiv Zoeken

18 december 2010 15 - 10 oktober 2013 01

GRAFIEK KAART WOORDEN GEBRUIKERS TWEETS HOME

Gezocht in: 2.374.168.897 tweets. Gevonden: 58.960 tweets met het trefwoord "hiv" (0.002%).



(download data van grafiek)

(smoothfactor: 10264; verlagen; verhogen)

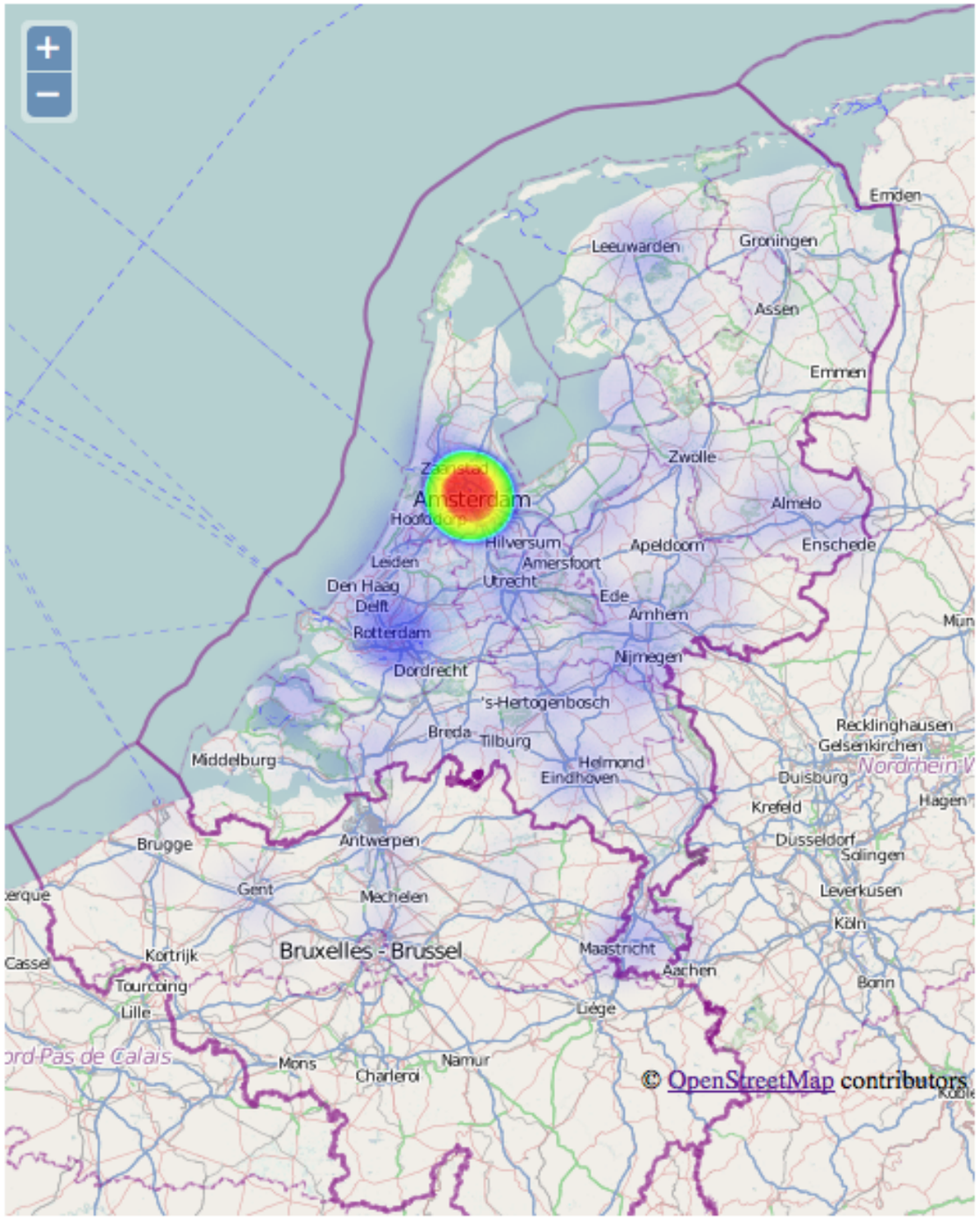
Created at: 19 November 2013 16:46:56. erikt(at)xs4all.nl (clusterstatus)

Zoek in tweets: twinl-geo+hiv Zoek

18 december 2010 15 - 15 oktober 2013

GRAFIEK KAART WOORDEN GEBRUIKERS

Gezocht in: 2.378.670.058 tweets. Gevonden: 674 tweets van de categorie "twi



(download data van kaart) (kaart met positiemarkeringen)

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Created at: 13 November 2013 10:30:16. erikt(at)xs4all.nl (clusterstatus)

Study Outline

- Gathering data & management
- Content analysis of the data

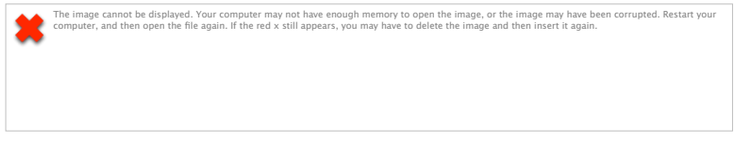
Twitter API

- Twitter API provide JSON format of the tweets, %1 of the stream
- It is possible to query it by using key terms, geo-boxes: [-180,-90,180,90]
- It provides a random sample which is not clear what covers

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

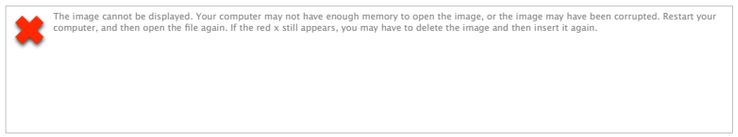
MongoDB

- Flexible storage possibility allows compact storage:
  - Default valued and duplicate attributes of a tweet are ignored, 2/3 of a document
  - Store just geo-time series which has some activity in them
- Geo-coordinate indexes and query operators: $near, $geoWithin, $geoIntersects
- GeoJSON compatibility, 2dsphere index



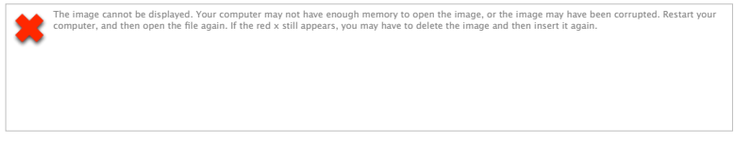
MongoDB

- Month numbers start from 0 in MongoDB
- Tweet creation time should be compared with datetime.utcnow()
- Incremental map-reduce: update every 5 minutes



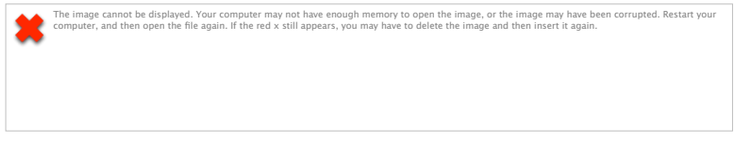
MongoDB - Map

```
function geotimeMap(){

        var mydatetime = this.created_at;

        var mylon = Math.floor(this.coordinates.coordinates[0]/10);
        var mylat = Math.floor(this.coordinates.coordinates[1]/10);

        var created_at_minute = new Date(mydatetime.getFullYear(),
                                         mydatetime.getMonth(),
                                         mydatetime.getDate(),
                                         mydatetime.getHours(),
                                         mydatetime.getMinutes());

    emit({created_at_min:created_at_minute, longitude:mylon, latitude:mylat}, {count: 1});
    }
```



MongoDB - Reduce

```
function geotimeReduce(key,values){

        var total = 0;
        for (var i = 0; i < values.length; i++){
                total += values[i].count;
        }
        return {count: total};
}
```



Tweets since page loac

- - 8 2 0 3 9

Legend

tweet cluster

lastest tweet

Filters

cluster view heatmap view

keywords filter

hashtags filter

5 most popular hashtags

HürTeşebbüseDARBE (6456)

MentionSomeoneYouAreThankfulFor (3

HappyThanksgiving (2312)

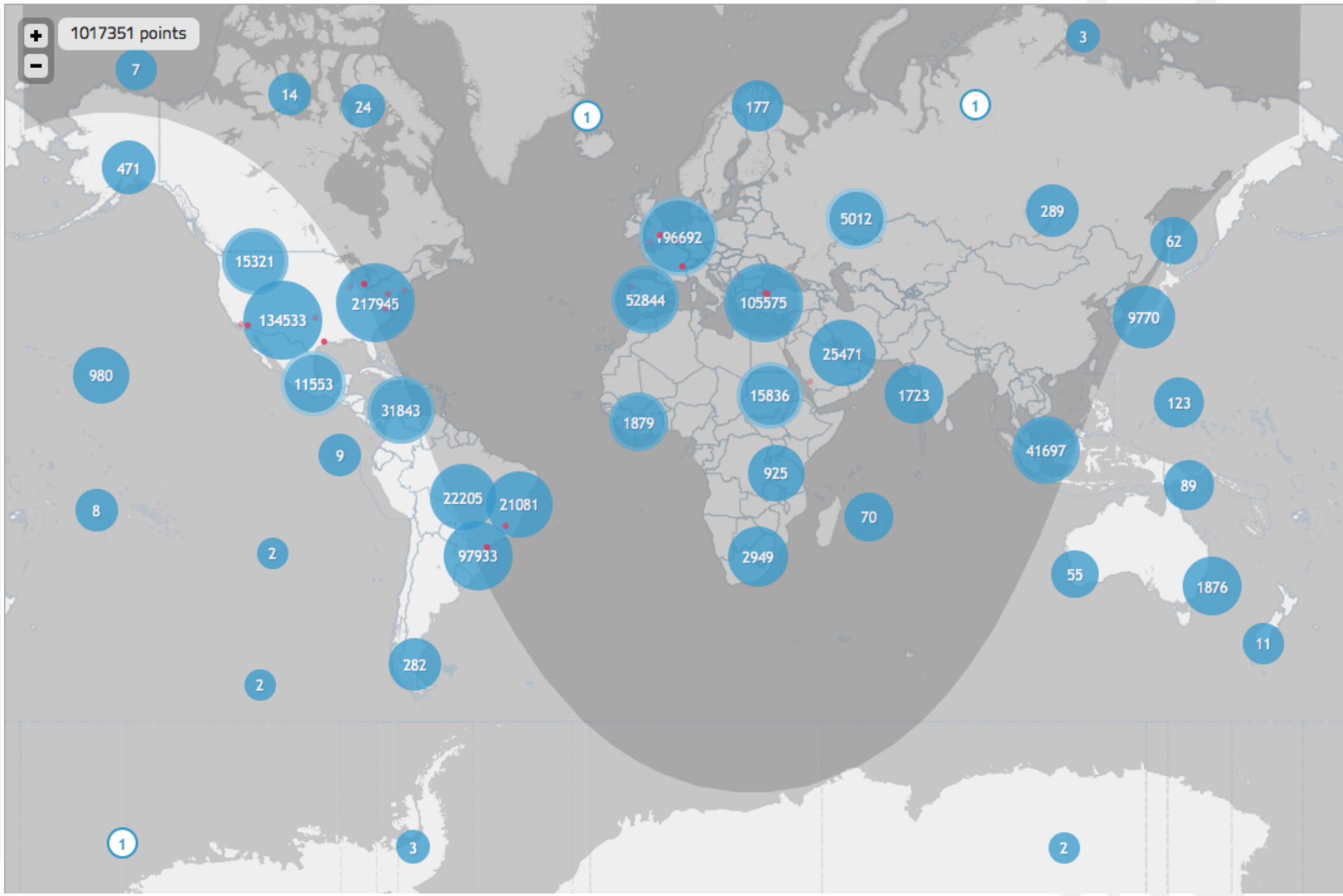
Thanksgiving (2058) PeoplesChoice (20

No Time Filtering

If you want to start with a blank map, clic reset map button (reload the page to see tweets again).

Reset map

Maptimize

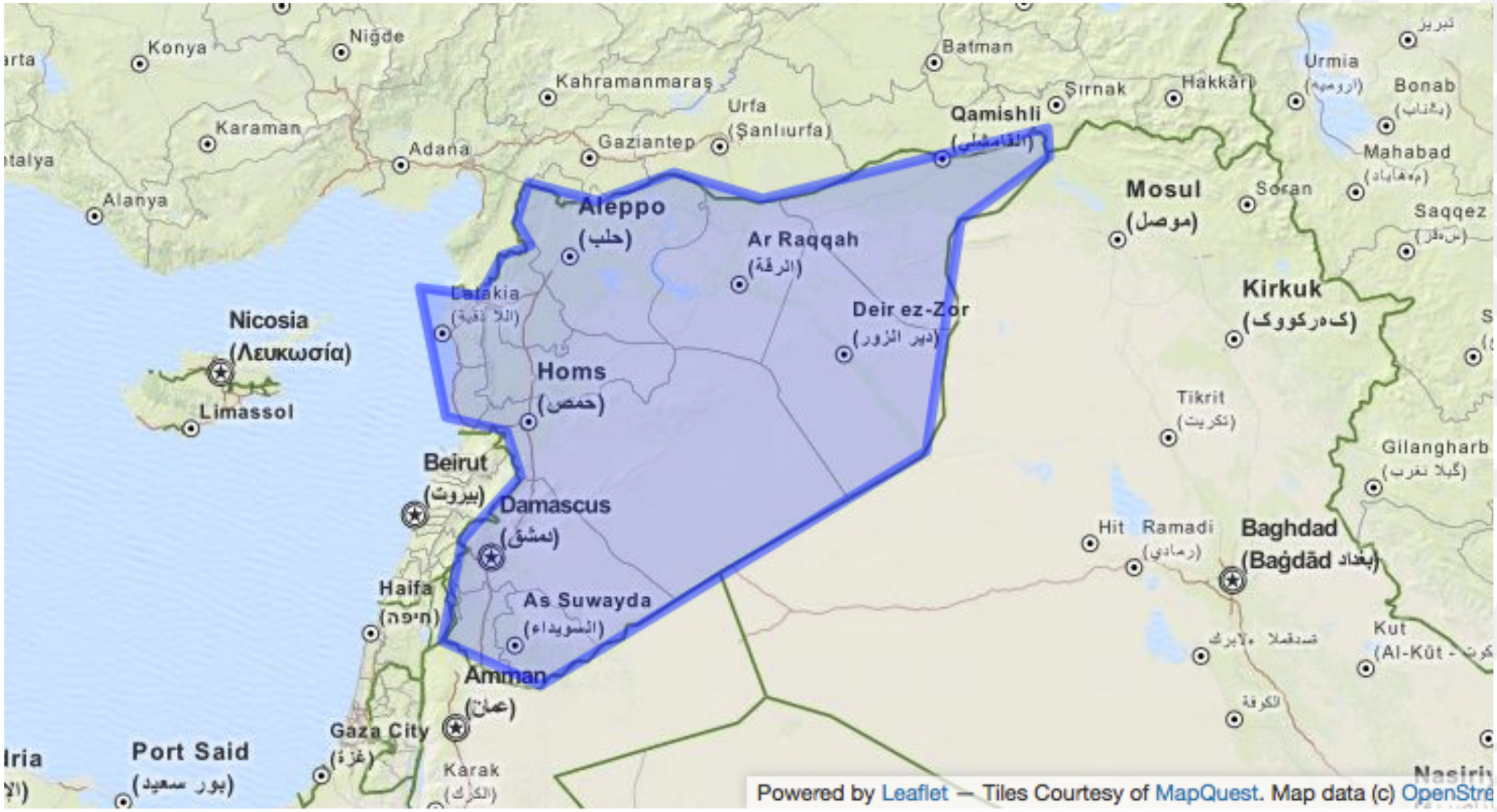


The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Study Outline

- Content analysis of the data

Content Analysis - Syria



The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Content Analysis - Syria

- Platform diversity
  - Fousquare, Path, Here, Instagram, etc.
- Keyword independence
  - Although tweets are sparse, their content is diverse

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

foursquare

I'm looking for...

Rotterdam, NL

KrSd B. checked in at **KarKamış**

Turkey | September 6 via foursquare for iPhone

LIKE

sınırda bir hareketlilik var tanklar Suriye tarafına saf tutmuşlar ayrıca siyasi bir hareketlilikte var

Log in to foursquare to leave a comment!

About this check-in

Holding down the mayorship!

KarKamış

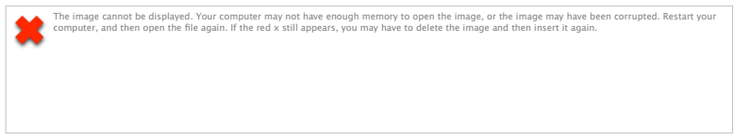
General Travel

SAVE

Switch Language: English

Conclusion

- Geo-tagged tweets can provide geographical region based information
- We can gather as diverse/rich as possible information
- We can gather textual/meta data independent of language and domain
- By using geographical region relevant hashtags and links we can search for them in the non-geotagged tweets to enrich the data
- We can track the content over time and geo-regions
- MongoDB enable us to store and process this information effectively



Future Research

- Extend the study to other regions
- Use geographical and user based information
- Spatio-temporal analysis of hashtags, links
- Compare the random sample of Twitter API

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

References

- http://geojsonlint.com/
- http://twiqs.nl/
- http://onemilliontweetmap.com/

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Thanks for your time !

Any Questions or Comments ?