Spatial-temporal Analysis of International Connections Based on Textual Social Media Data

Kaidi Guo

Supervisors: Univ.Prof. Mag.rer.nat. Dr.rer.nat. Georg Gartner
Dr.-Ing. Linfang Ding

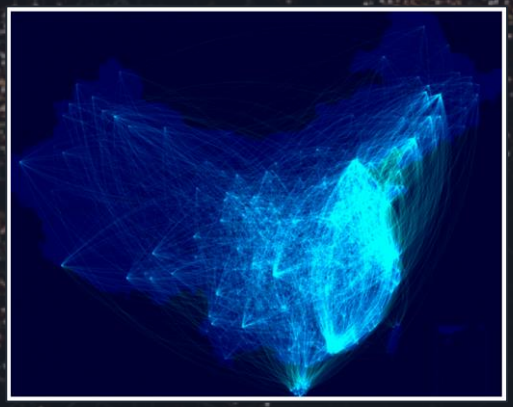
TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Background & Motivation

Social Media

WHY ? **FREE, BIG, FROM USER**

WHAT ? **Opinions, Events, Trends**



Social Media + Geoinformation

HOW ? **Geotagging**

WHAT ? **Travel behavior, local event detection**

Problems? **small percentage, bias**

Geoinformation in text

To map international connections from social media data, and reveal potential spatial and temporal patterns

Related work

Stefanidis et al. (2013): mapping virtual Syria through social media content

Problems?

How many Languages should be involved?

How to solve?

Multiple subjects, one state, one language

> In order to demonstrate how the structure of a virtual polycentric Syria can be gleaned through the analysis of social media content we use for this publication tweets that we collected over a period of one week (10 July through 17 July 2012) and included mentions to Syria or its hashtag equivalent #Syria. As per standard *Twitter*

Context & Research Questions

Sina Weibo: The biggest microblog platform in China
Almost all users are Chinese people and post in Chinese

"How famous are other countries in Chinese community?"

"What are they famous for?"

"Do different genders have different preference?"

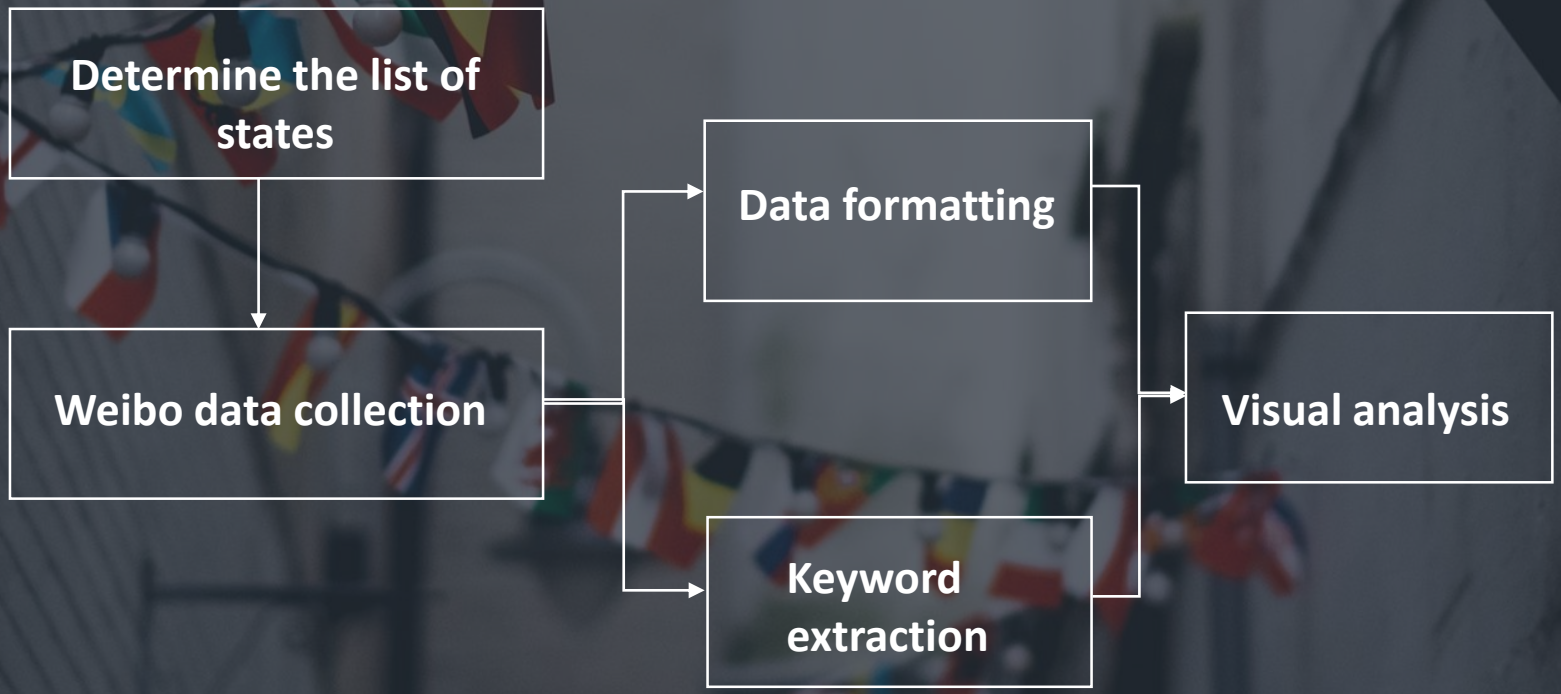
.....

RQ1: What are the possible methods to harvest weibos that contain textual geoinformation at a state level.

RQ2: How should we process the textual content of weibo?

RQ3: What information visualization techniques should be used to present the different information we harvested from actual weibos and their metadata?

Methodology



Workflow

- Determine the list of states

  193 member states of UN, Vatican, Palestine, Taiwan.
  Transform into commonly used Chinese name

- Weibo data collection based on the state list

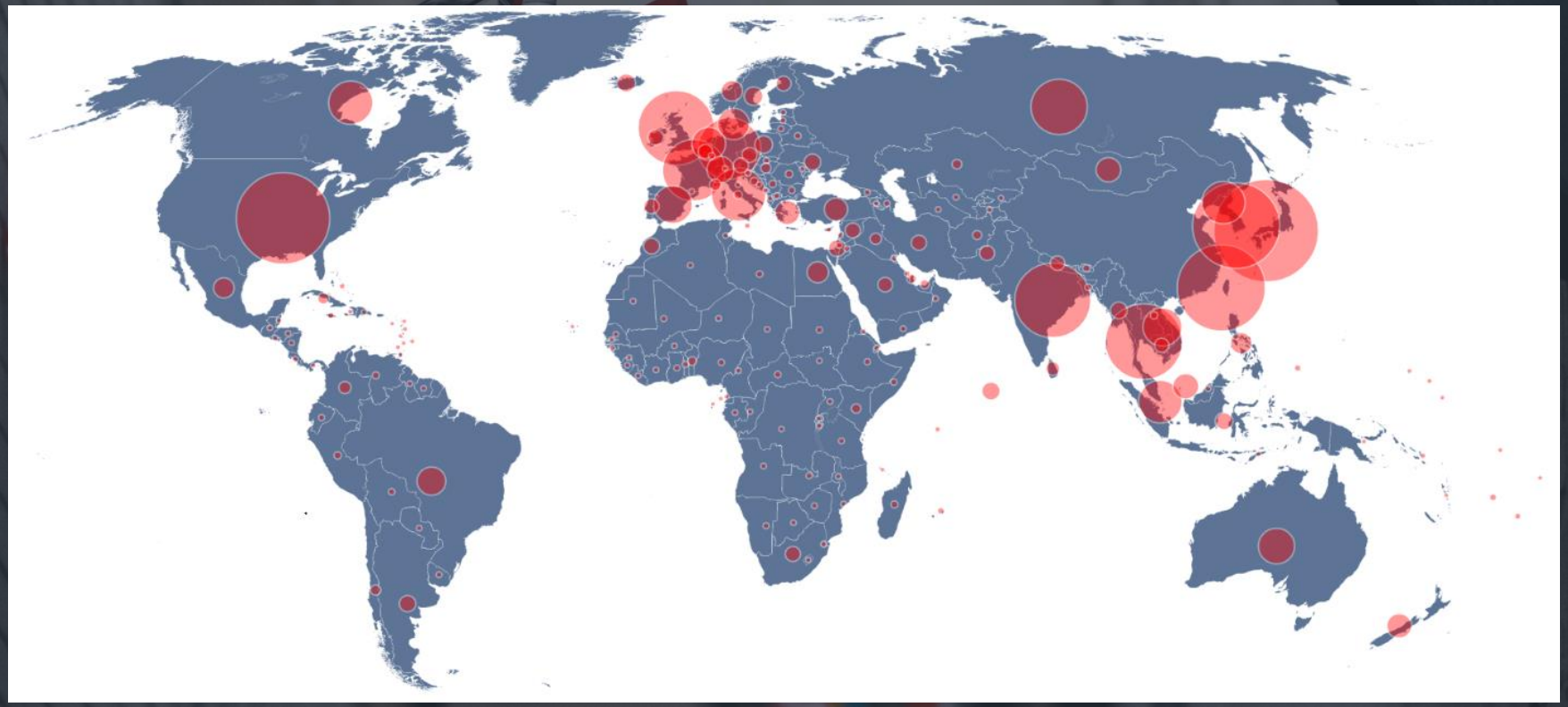
  No available Sina API. A web crawler was created to harvest JSON data

- Extract keywords from Weibo content with jieba NLP package

  7 other attributes: State, user name, user gender, platform, follower, repost count, post time. (key: weibo ID)

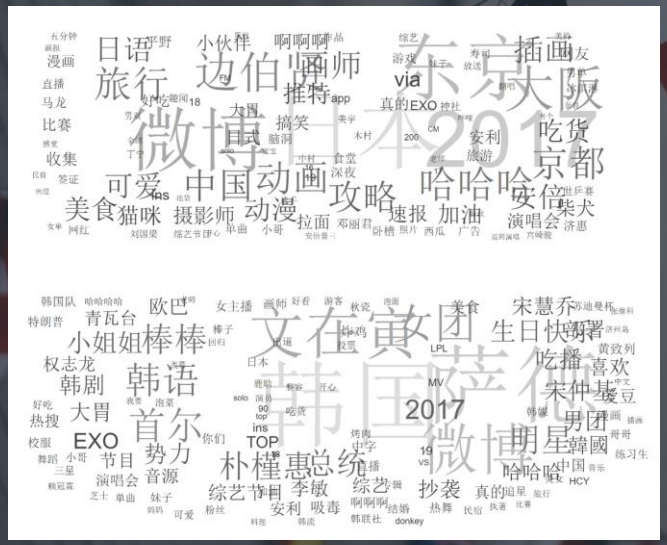
- Visual analysis of the data

Results & Discussion



1. *Closer states intend to have higher level of concern*
2. *The space-time distance has less influence (Saint Vincent And The Grenadines has 5 mentions)*

Results & Discussion



Higher culture similarity states, weibo user discussed more about popular culture

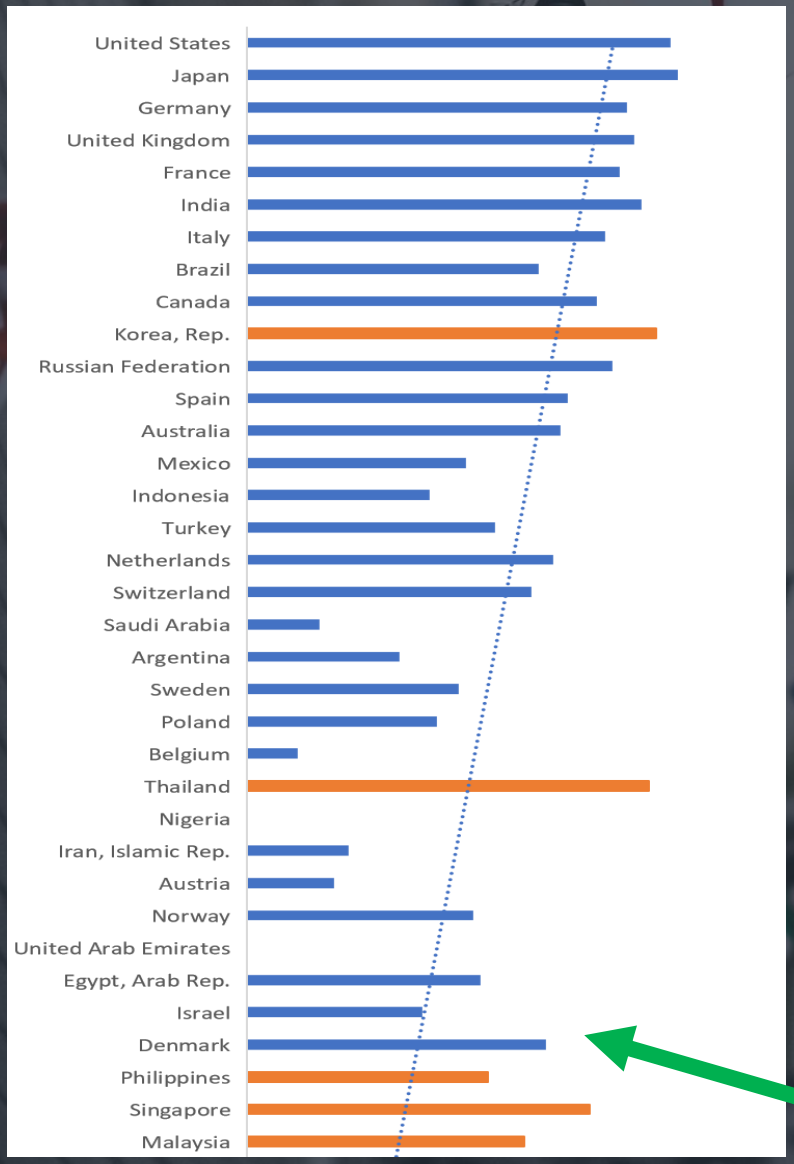
- *Japanese animation*
- *South Korea pop music*

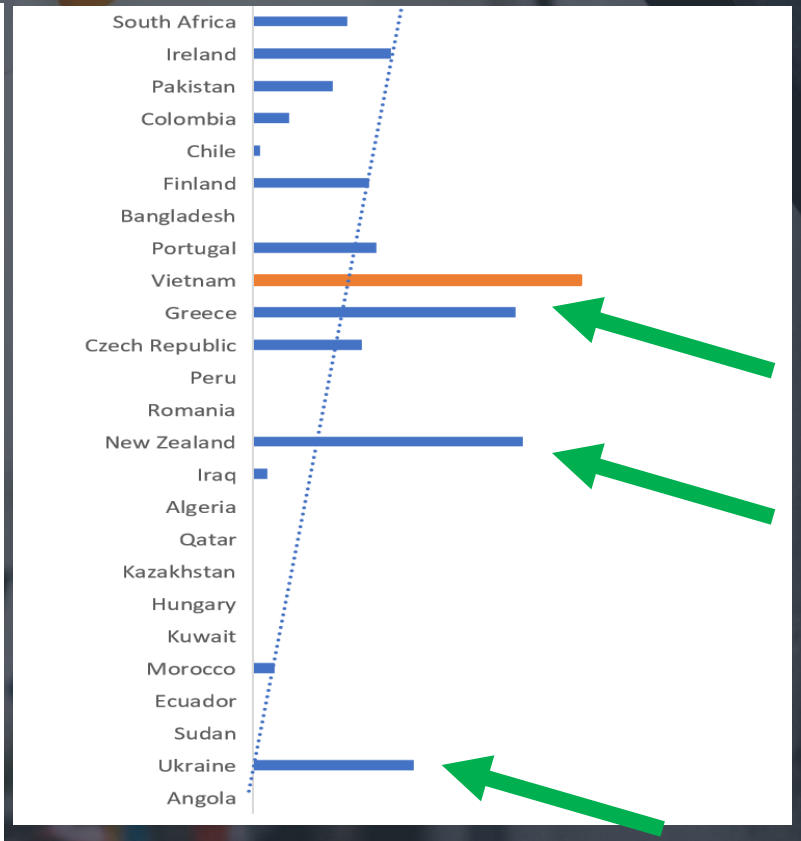
lower culture similarity states, the discussions lean to general news or specialized interests

- *Terrorism attack in UK and Brexit*
- *Football and political talks under Germany*

Word clouds could indicate the “stereotypes”

Results & Discussion

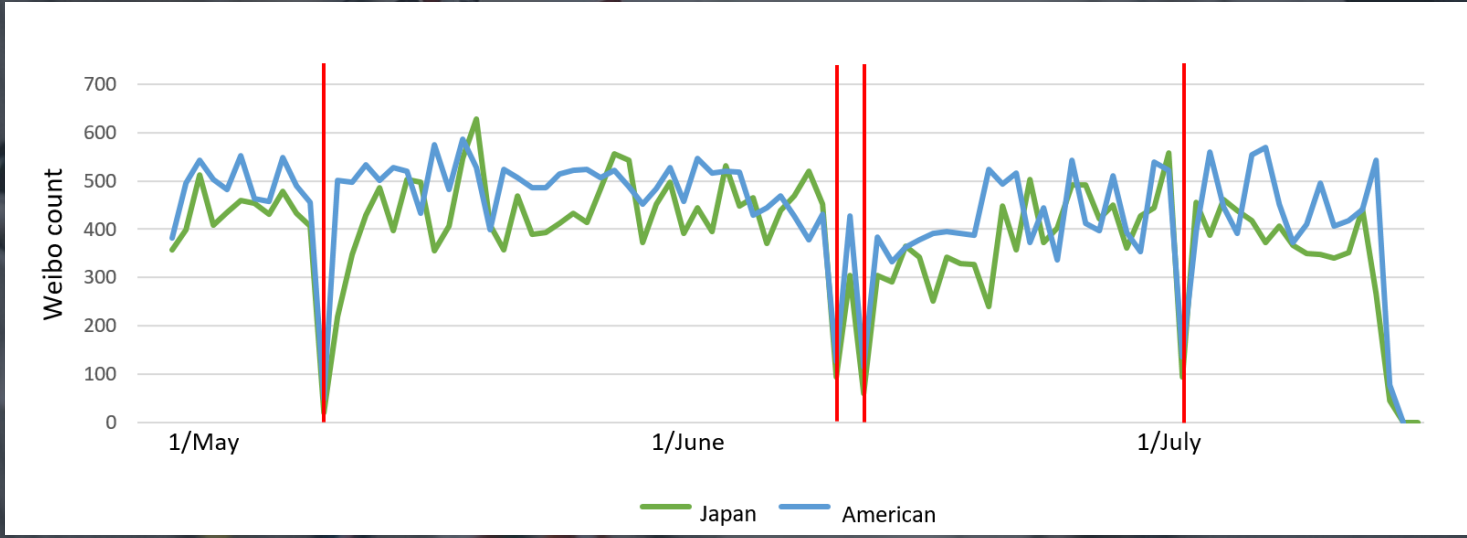


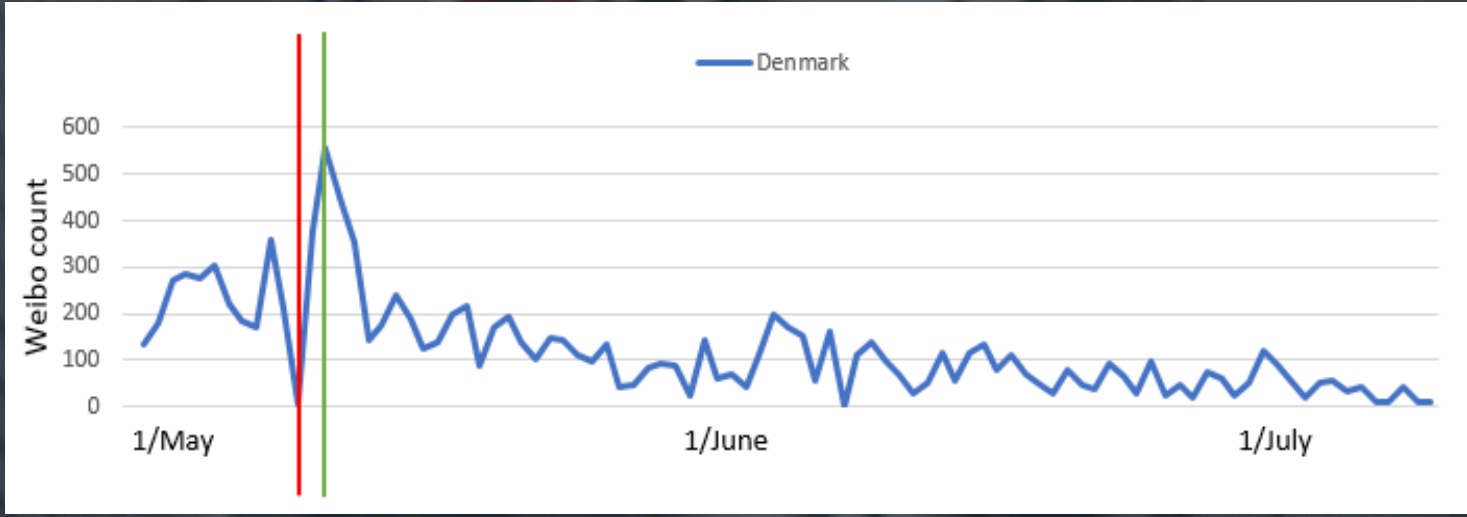


2016 global GDP ranking VS data ranking

- the data amount rank generally consistent with GDP– or, “soft power”

Results & Discussion





Results & Discussion

Denmark

spokesperson

oyster

qianxi

little prince

qianxi

- Search for the weibos with highest repost count in May

Danish time on May 11 at 9 am, TFBOYS members of the members of the customs of the Danish tourism image spokesperson awarded the ceremony was held in Odense City, Denmark. Andersen hometown of the city of Odense and the Danish National Tourism Administration co-awarded the Commissioner Xi Danish tourism spokesperson certificate.

Conclusion

- Textual geoinformation contained in social media could be harvested to depict international connections.
- The statistic features could be visualized. The spatial and temporal patterns could reveal interesting political and cultural events.
- With the help of nature language processing, visual result combined with human thought could lead to meaningful conclusions.

Outlook

Limitations

Limited choice of states

Unstable data acquisition

Based only on Chinese

Rudimentary NLP function

No user side spatial analysis

Future work

Involve more entities. e.g. Crimea

Perform future study with API

Study other countries through Twitter

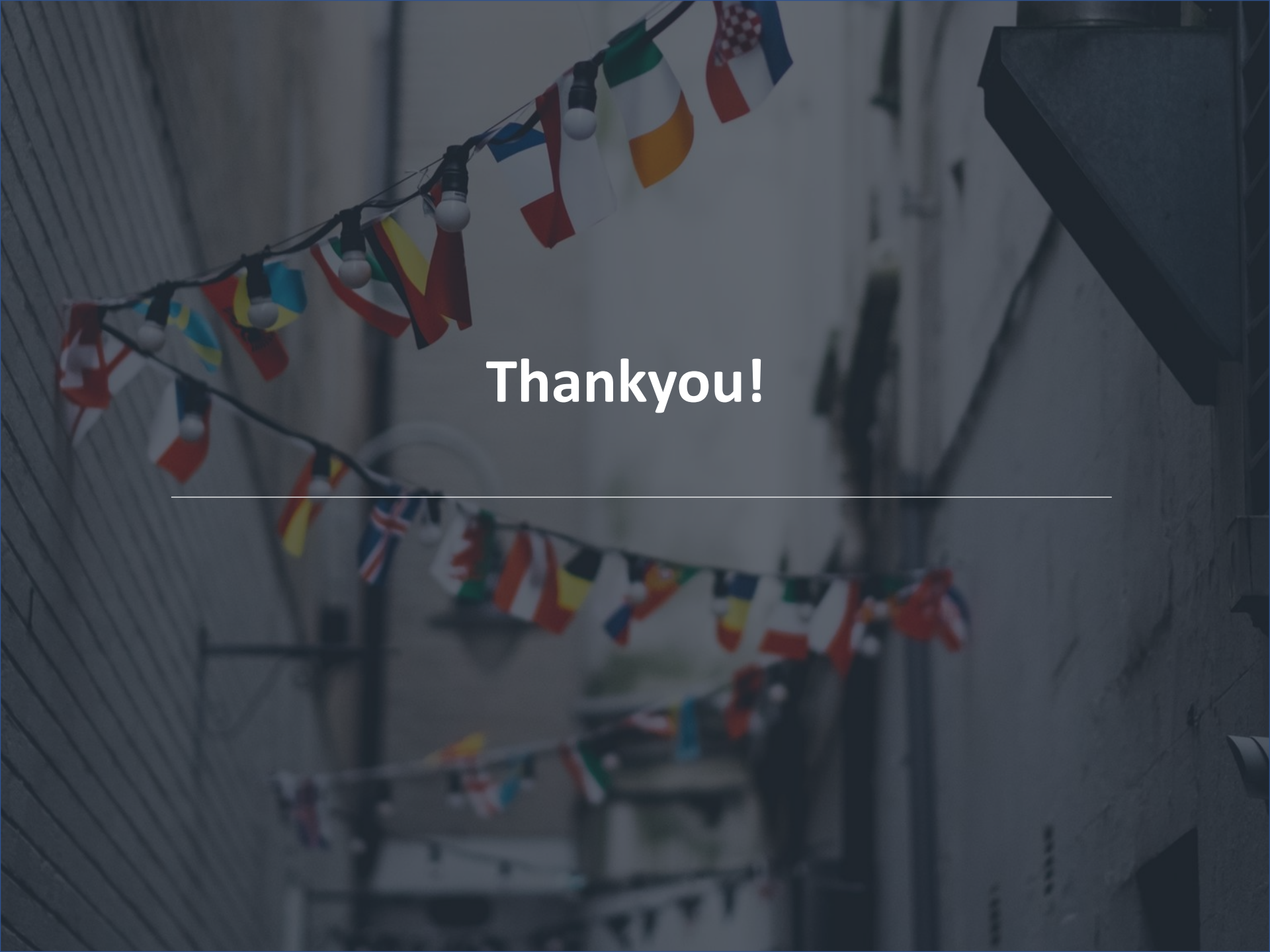
Include sentiment analysis

Consider user location study

Outlook

"Our object was to highlight these emerging geographic opportunities as a result of the proliferation of social media, rather than to create a definitive map of virtual Syria."

- Stefanidis et al. (2013)


Thankyou!

TF-IDF

frequency–inverse document frequency (TF-IDF)

- Important words tend to appear in the article more.
- Some commonly used words are frequently present in more than one articles

The importance of a word is proportional to the number of times it appears in one document, but it is inversely proportional to the frequency it appears in the corpus

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Visual Data Exploration

Initially, it is easy for the user to collect a huge amount of data from any other automated process, however

- If the data is displayed textually, the amount of data able to be presented is very limited.
- Without human participation, the value of the data cannot be interpreted.

Therefore, Visual data exploration (Visual data mining) aims at integrating human in data exploration process.