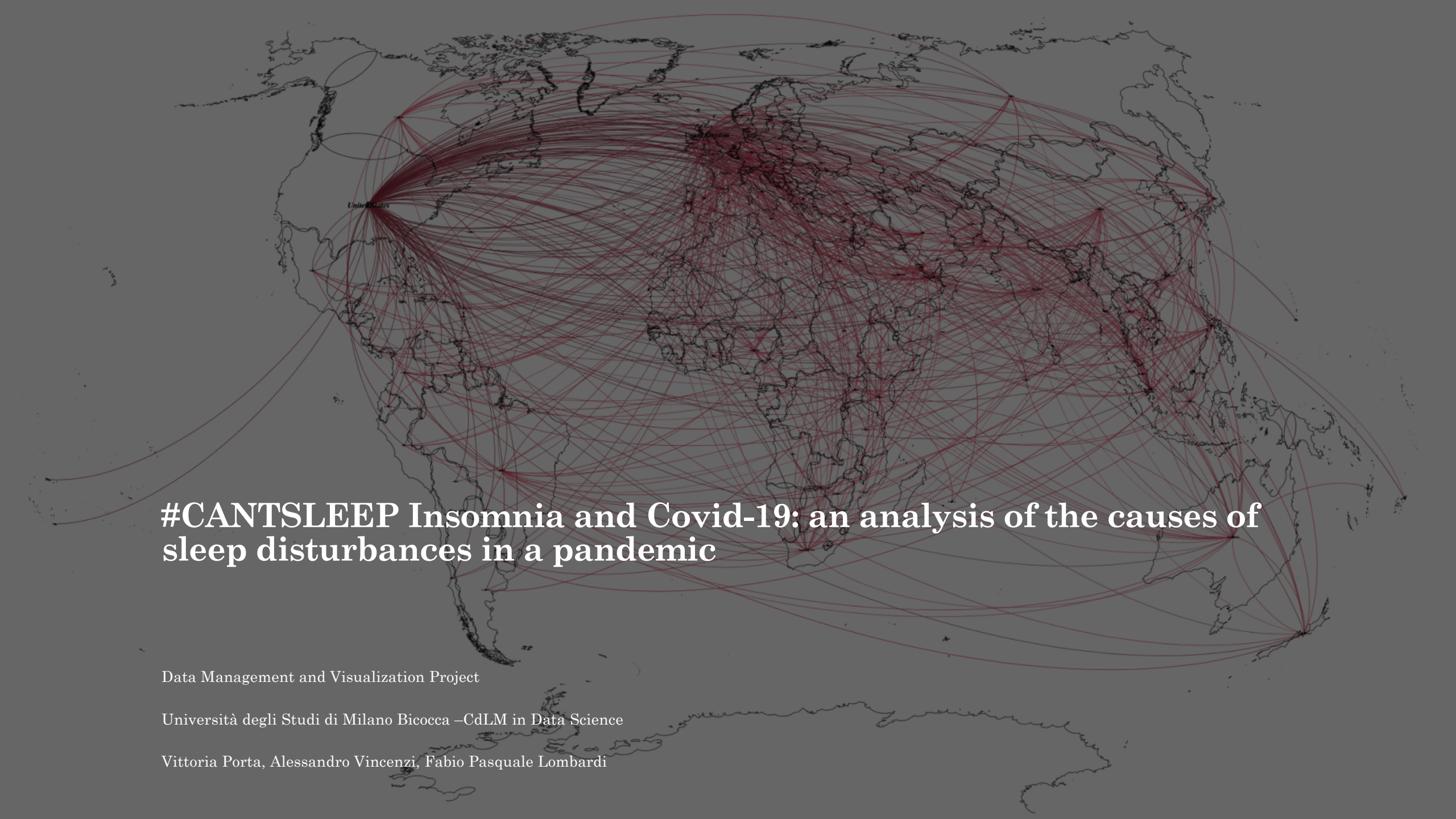


#CANTSLEEP Insomnia and Covid-19: an analysis of the causes of sleep disturbances in a pandemic

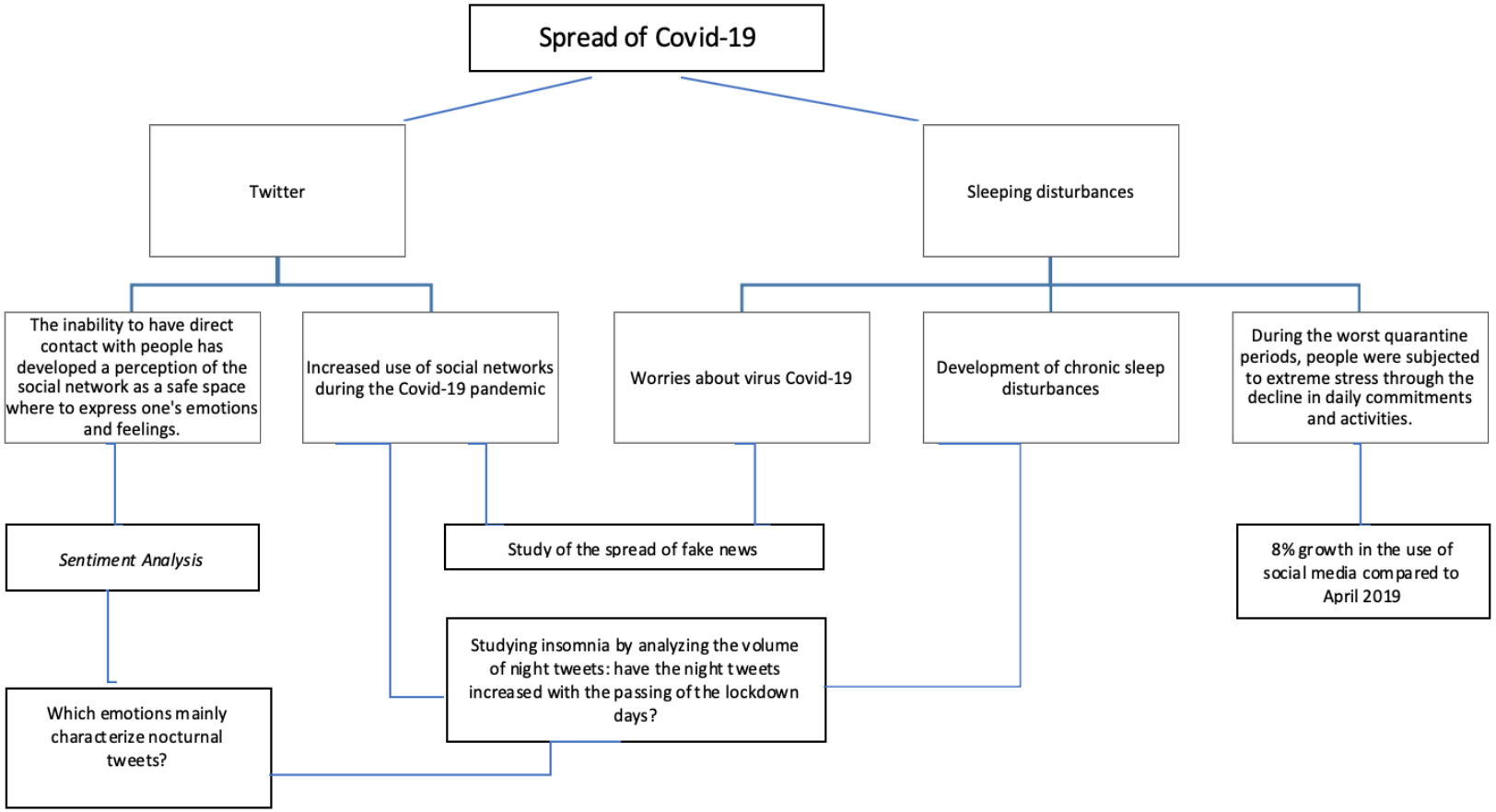
Data Management and Visualization Project

Università degli Studi di Milano Bicocca –CdLM in Data Science

Vittoria Porta, Alessandro Vincenzi, Fabio Pasquale Lombardi

Contents

The main Idea:



Introduction

Research questions:

Has the volume of night tweets increased with the passing of the days in the worst months of the pandemic?

If that were the case, people would tweet more at night since they can't sleep.

Which emotions mainly characterized night tweets?

Where the idea came from?

Generalized anxiety disorder, depressive symptoms and sleep quality during COVID-19 outbreak in China: a web-based cross-sectional survey}, Y. Huang,N. Zhao, June 2020

The technologies we used

Foto di Poeta60 / CC BY-SA 3.0

Data Sources

- Twitter as primary source

For the **data enrichment** part were used three different data sources:

- The alphabetical list of all countries and capitals of the world
- countries.csv dataset
- List of U.S. state abbreviations

For the **data transformation** part was used the list of tz database time zones.

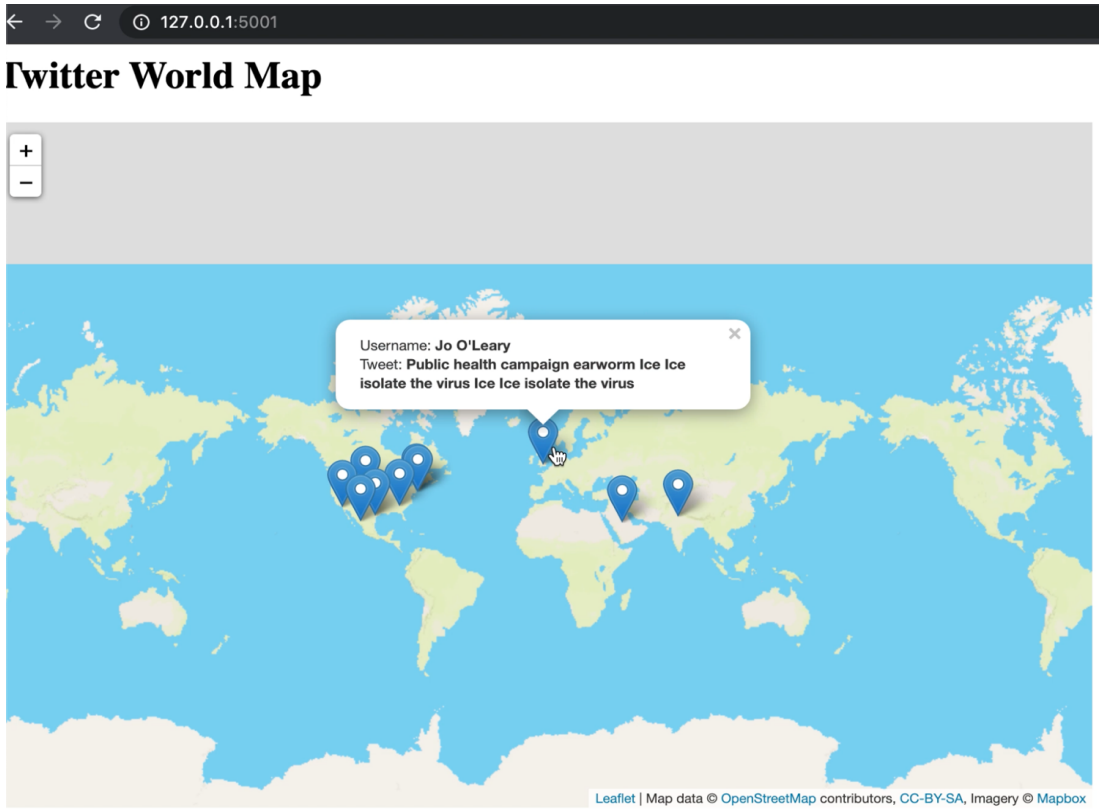
Finally, it was useful in terms of **visualization**, to focus attention on tweets from the United States of America (a place where Twitter is known to be one of the most used social networks) and for this purpose an additional collection of tweets was integrated.

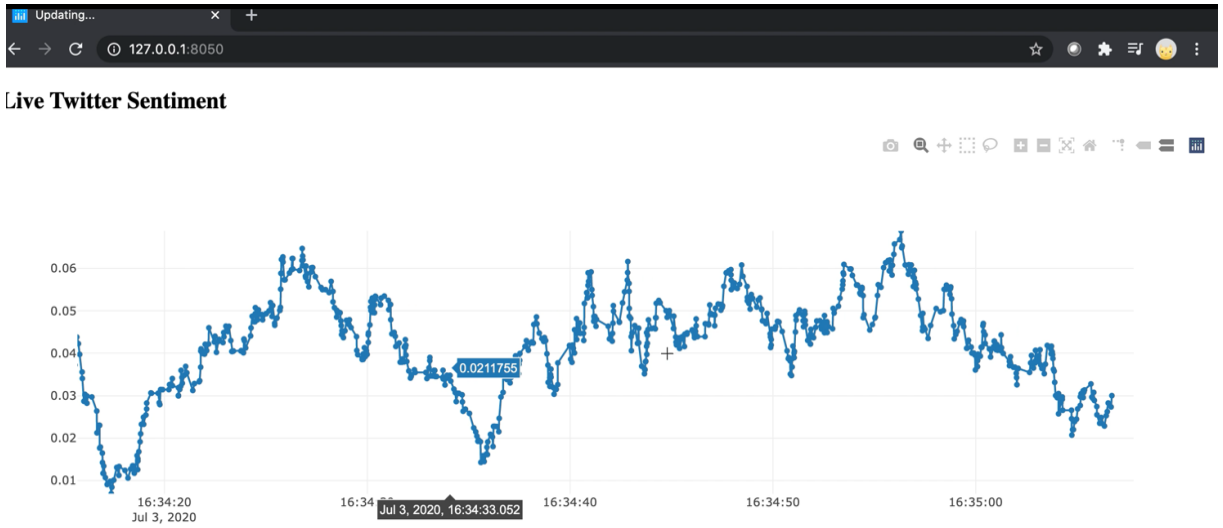
Data extraction and ingestion

Data extraction: Tweepy

Data ingestion: Apache Kafka connection to Python through PyKafka Library

Connection to MongoDB and use of PyMongo





Real Time Sentiment Line Chart

technologies used: Python, Flask. vaderSentiment Python library (Valence Aware Dictionary and sEntiment Reasoner)

Preliminary Analysis: Real Time Interactive Dashboards

Since the data extraction process would take a long time, we wanted to have a tool that could visualize the data we were streaming in real time, allowing us to constantly check the quality of the data visually.

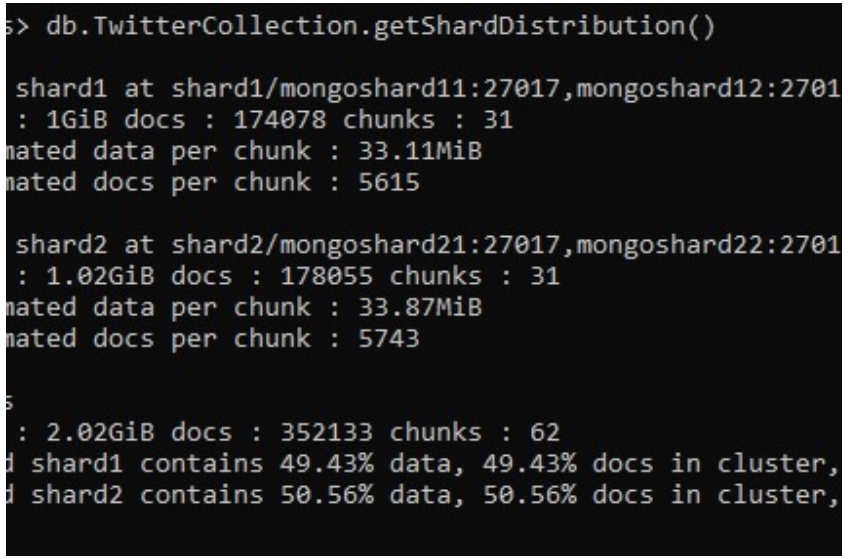
Live Twitter Map

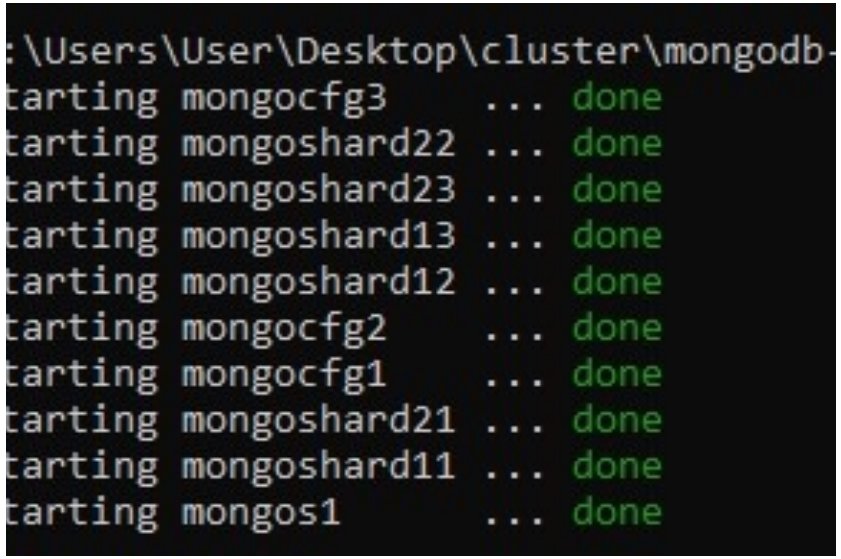
technologies used Python,Flask, Leaflet

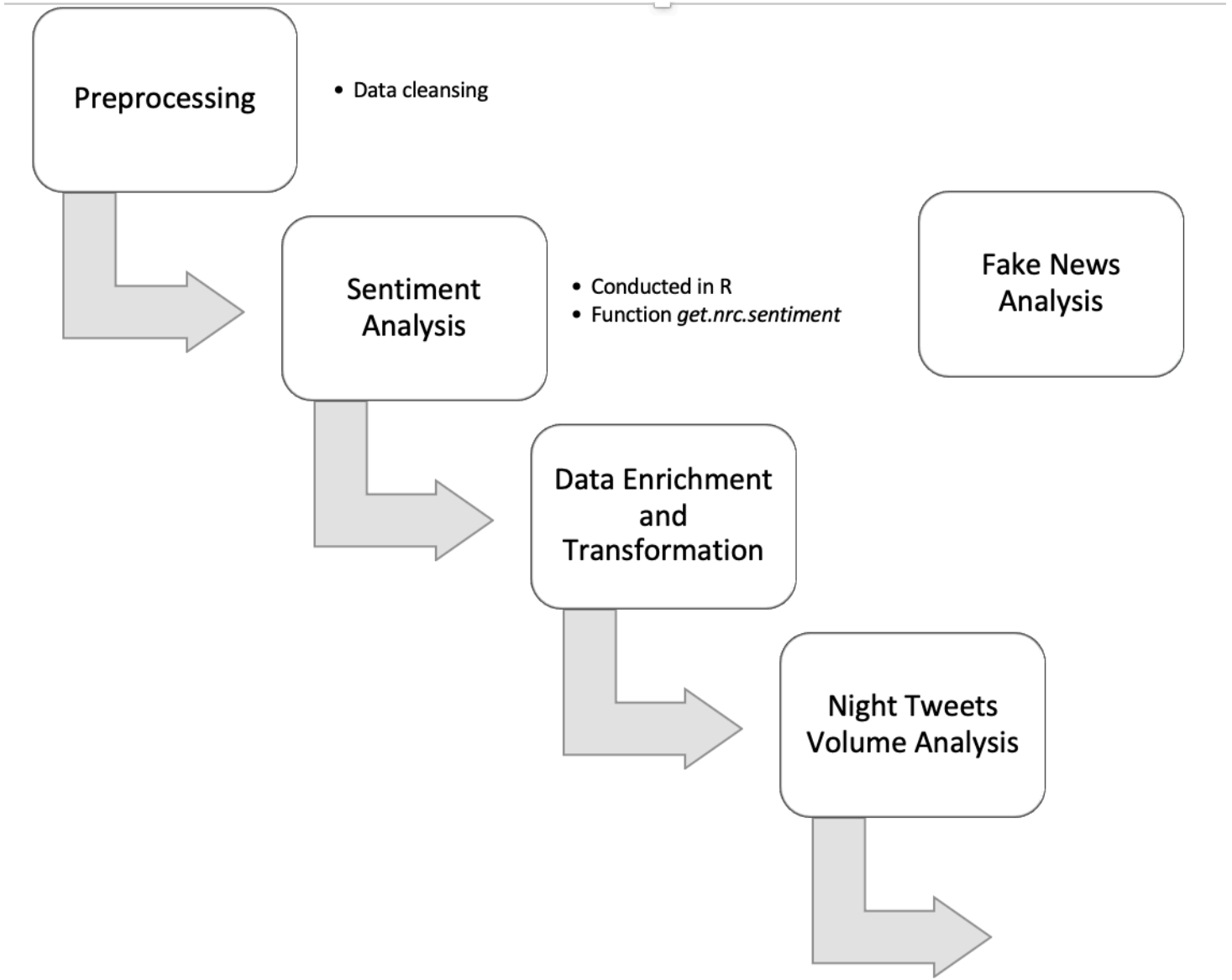
Demo available at https://drive.google.com/drive/folders/1T8rJVkNA05zm8P_yzRKrGoVWSA0IR4nG?usp=sharing

Sharded Architecture

- The architecture that has been deployed has:
- 2 Shard clusters in replica set (3 nodes for each cluster)
  - mongoshard11, mongoshard12, mongoshard13, mongoshard21, mongoshard22, mongoshard23
- Config cluster in replica set (3 nodes)
  - mongocfg1, mongocfg2, mongocfg3
- 1 Router mongos instance
  - mongos1
- First, the collection was saved in MongoDB, after which, using the --mongorestore command, it has been reinserted into a new sharded collection.







Analysis

- **Conducted in R**
- **Pre-processing and missing values treatment and data filtering**
- **Data cleaning**
- Data Enrichment and Transformation on field ‘user.location’ and ‘created_at’

Data Quality

To analyse the locations, we assessed the presence of **homonyms** (same name for different concepts) that could give rise to any conflicts, both between states, between cities and between state-code.

Refer to secondary sources.

Constantly monitor the stream through real time dashes.

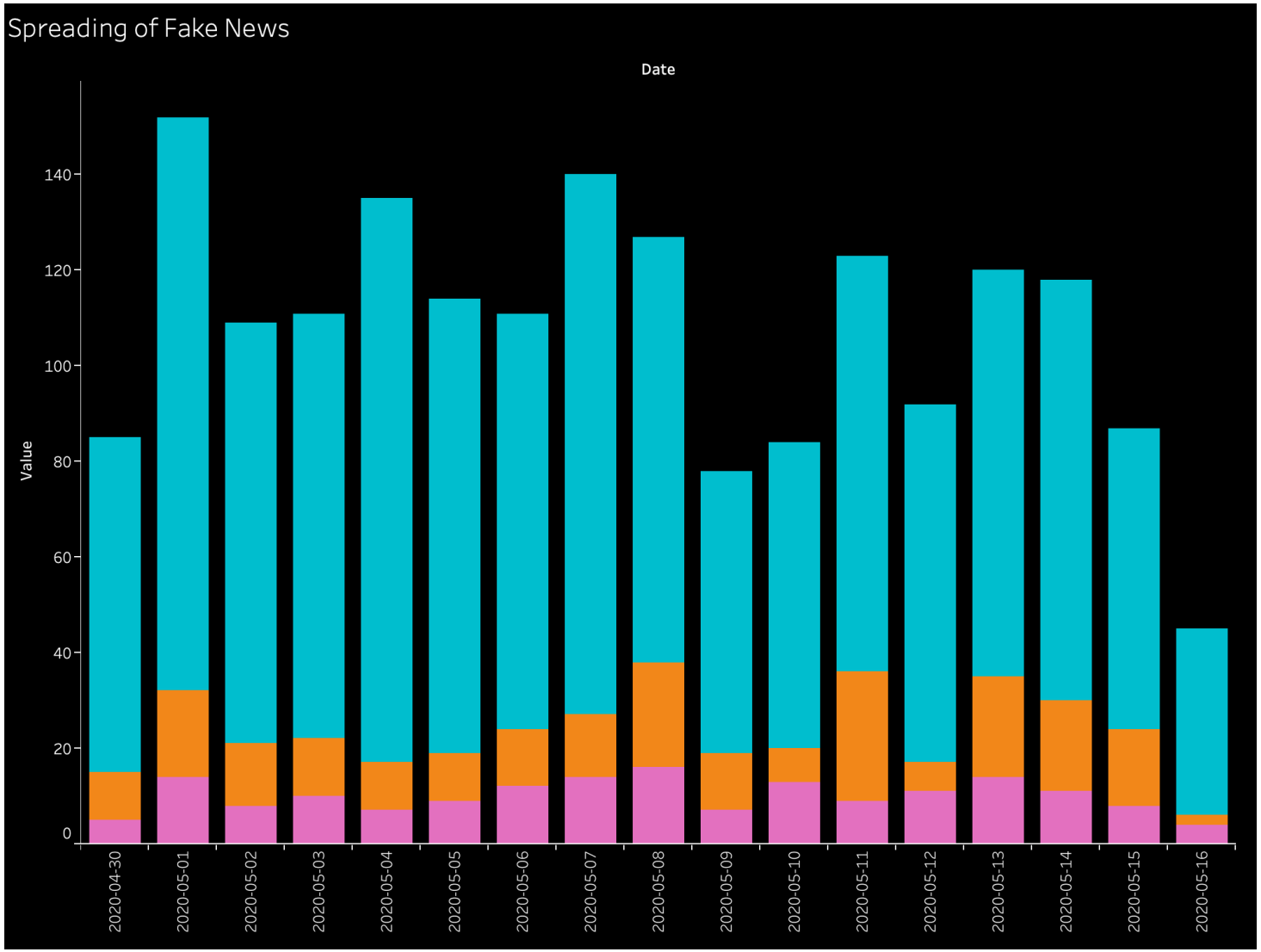
Sentiment Analysis

Python: vaderSentiment Python library (Valence Aware Dictionary and sEntiment Reasoner)

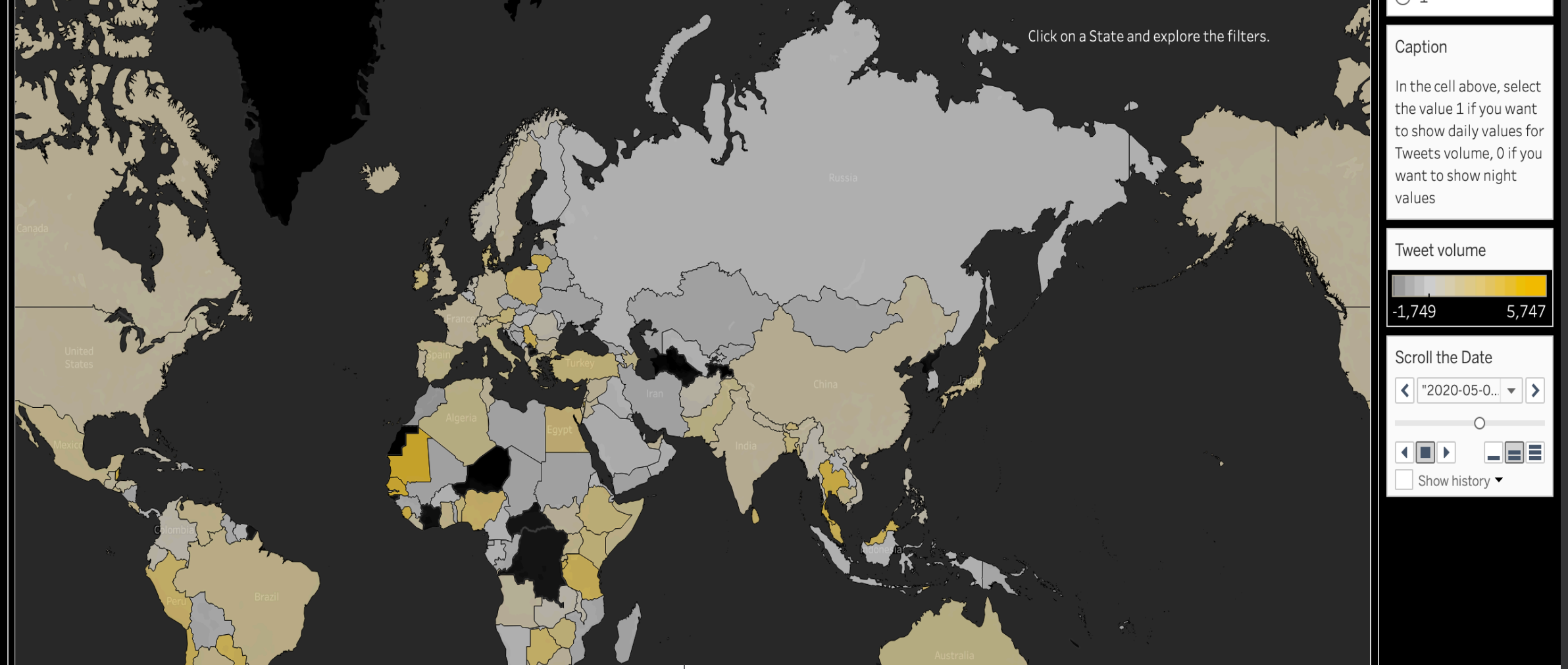
R :get_nrc_sentiment' which calls the NRC sentiment dictionary to calculate the presence of eight different emotions

- anger
- anticipation
- disgust,
- fear
- joy
- sadness
- surprise
- trust

Retweet Analysis and Spreading of the Fake News analysis



- **Retweet Analysis**: the importance of creating a dataset *ad hoc*. Identify who has been retweeted and who retweeted.
- **Fake News Analysis:** The fake news concerning the Covid-19 virus can refer to three main categories: the conspiracy (or the belief that the virus was somehow the result of the failure of diplomacy between states); false homemade treatments and transmission. The analysis was conducted in order to search the tweet texts of the keywords that referred to one or more categories



Data Visualization

- Tableau
- Gephy
- Real Time Dashboards