Social Media Monitoring and Surveillance for Vector Borne Diseases

This is a research project for developing a strategy and infrastructure to use social media to monitor topics for the purpose of public health surveillance in Pima County, Arizona.

These include influenza, vector borne disease, and heat related illness.

The research will in collaboration with the Pima County Health Department and it will develop and implement strategies to accomplish the following:

1. Track and monitor social media for key public health topics
2. Elicit public feedback and reporting on topics of public health relevance
3. Measure the impact and reach of social media campaigns

1. Data Collection

Twitter offers streaming APIs to give developers and researchers low latency access to its global stream of data. Public streams, which can provide access to the public data flowing through Twitter, were used in this study. Studies have estimated that using Twitter's Streaming API, researchers can expect to receive 1% of the tweets in near real-time. Tweepy -- a Python library for the Twitter API, was used to access tweet information from the Twitter Streaming API.

Keywords (vector borne diseases symptoms)

The **vector borne diseases** symptoms-related stream is to collect only tweets containing any of 64 related keywords that were suggested by clinical collaborators. Figure 1 shows the keywords used in the data collection process.

Table 1. Keywords Table

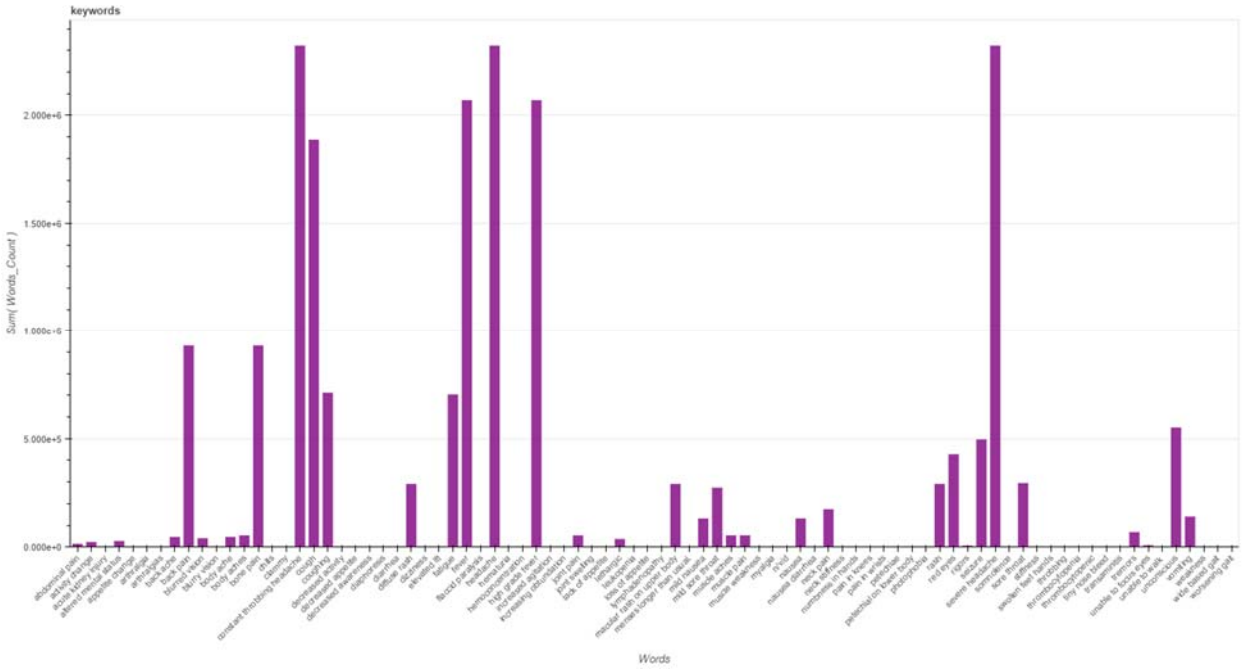
	abdominal pain	activity change	acute kidney injury	altered mental status	appetite change	arthralgia (arthralgias)	back ache (back pain)	blurred vision (blurry vision)
	bone pain	chills	clammy	constant throbbing headache	cough (coughing)	decreased activity	decreased appetite (lack of appetite)	decreased awareness
	diarrhea	diffuse rash (rash)	dizziness	elevated lft	fatigue	fever	flaccid paralysis	headache

	abdominal pain	activity change	acute kidney injury	altered mental status	appetite change	arthralgia (arthralgias)	back ache (back pain)	blurred vision (blurry vision)
	hemoconcentration	high grade fever	increased agitation	increasing obtundation	joint pain	joint swelling	lethargic	leukopenia
	lymphadenopathy	macular rash on upper body	menses longer than usual.	mild nausea	mild sore throat	muscle aches (muscle pain, muscle weakness)	myalgia	n/v/d
	nausea diarrhea	neck pain (neck stiffness)	numbness in hands	pain in knees	pain in wrists	petechiae	photophobia	red eyes
	seizure	severe headache	somnolence	sore throat	stiffness	swollen feet hands	throbbing	thrombocytopenia
	transaminitis	tremors	unable to focus eyes	unable to walk	unconscious	vomiting	weakness	wide based gait

Figure 1 shows the number of tweets in our Twitter stream for all of the keywords used in data collection.

Open Figure 1

Figure 1. keywords and Frequency of Tweets



Data summary

Our Twitter dataset was collected from Apr 04, 2016, through Sep 20, 2016, and contains **26,422,695** tweets. On average, about **170,000** tweets were generated from all over the world per day. This demonstrates that Twitter is a promising data source for social media monitoring and surveillance and should be explored further.

Table 2. Data summary

Dataset	Number of tweets	collection period	Geographical area	Number of keywords	Dataset size
Vector borne dataset	26,422,695	2016-4-15 ~ 2016-9-20	All over the world	64	38.819+ GB

2. Geographic Analysis

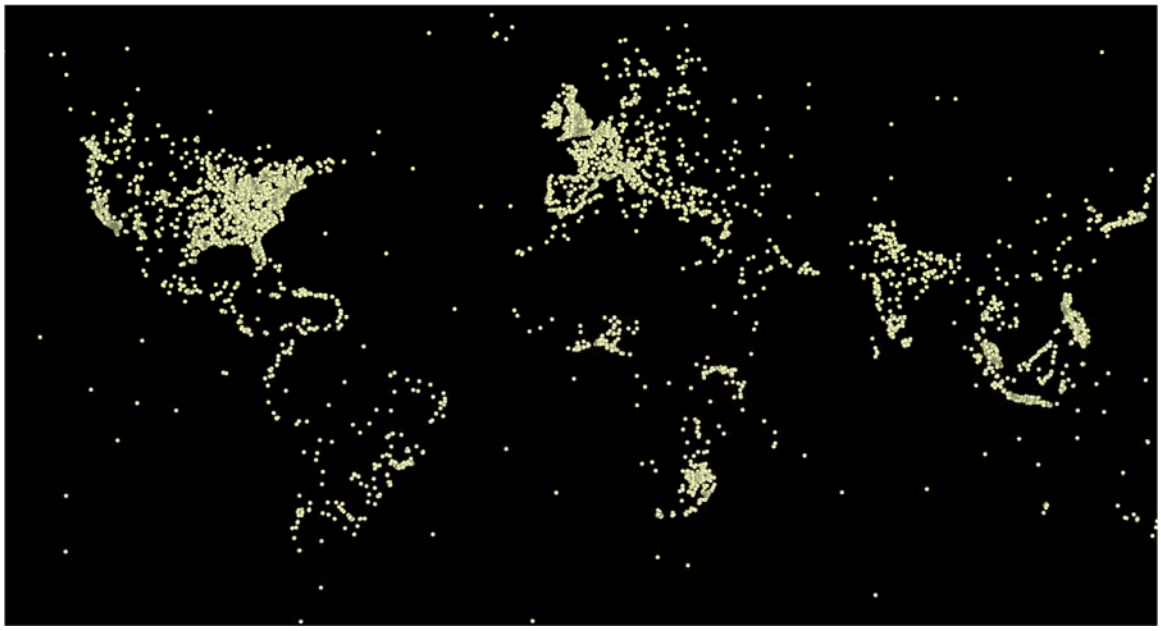
The geographic location for each tweet was identified using: **location coordinates** (latitude/longitude) and/or **user-specified location**. Read More

Among all the tweets we collected, **0.14% (37,964 / 26,422,695)** of the tweets revealed their coordinates. **0.07% (18,287 / 26,422,695)** of the tweets are from United States (US) Northern most latitude: +48.987386; Southern most latitude: +18.005611; West most longitude: -124.626080; East most longitude: -62.361014 .

There were **61.31% (16,200,147 / 26,422,695)** tweets containing location information. However, there is considerable ambiguity in the location information, e.g., “I am somewhere on the earth” or “Moon”.

Geo-location extraction was performed by using (1) latitude and longitude mapping for the coordinates field and (2) regular expressions analysis for the location field, i.e., matching state names or postal abbreviations on the US states (e.g., Texas or TX), followed by matching city names (e.g. Dallas). We were able to identify location for approximately **9.09% (2,403,950 / 26,422,695)** tweets as one of the US states name.

Figure 2. Geographic Visualization
(2016-4-15 ~ 2016-9-20, All over the world, 0.14% 37,964 Tweets)

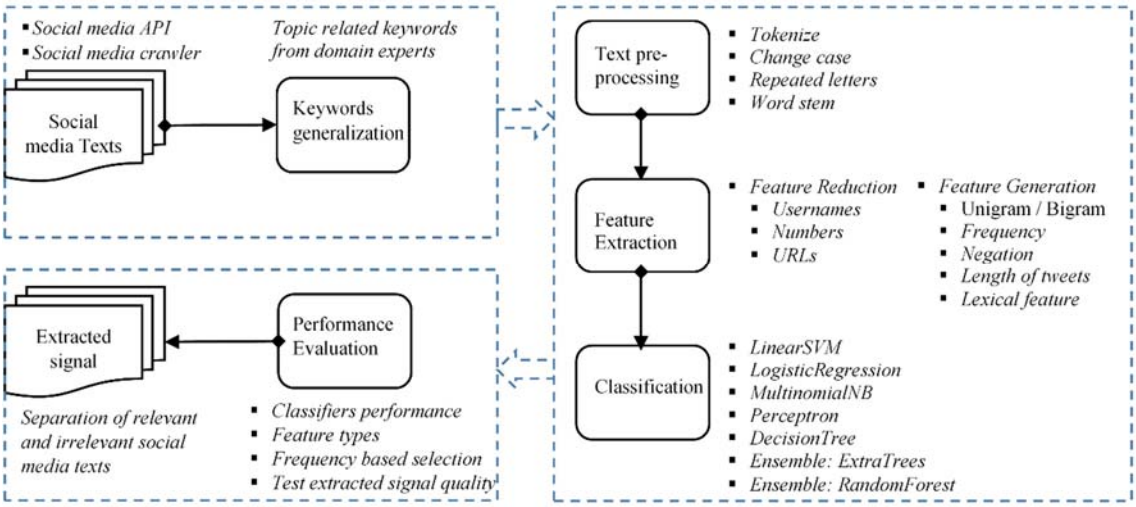


3. Natural Language Analysis

Natural Language Analysis involves natural language understanding and enabling computers to derive meaning from human or natural language input (i.e., tweets textual information).

Our approach adapts and proposes a combination of different Natural Language Processing feature generation methods along with machine learningclassfiers to identify and remove various types of noise from social media text as shown in Figure 2.

Figure 3. Methodology for Signal Extraction from Social Media Text

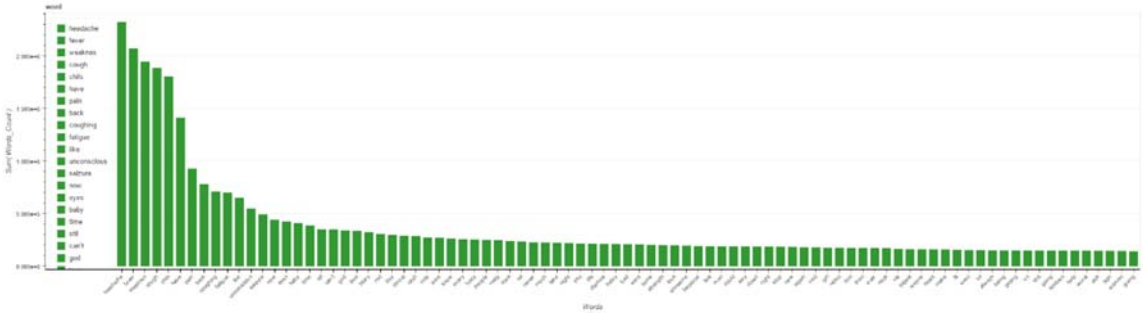


Top words used by Twitter users

We summarize the top words used by Twitter users in the dataset. "headache", "fever", "weakness", "cough (coughing)", "chils", "pain", "fatigue", "unconscious" are meaningful top words used by Twitter users.

Figure 4. Top Words Used by Twitter Users

Open Figure 4



We further analyze the lexical feature of words used by Twiiter users. Part-of-speech (POS) tagging (also called grammatical tagging) was applided. POS tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its ***definition*** and its ***context***.
Based on the result of POS tagging, we summarize the top nouns and top verbs used by Twitter users in the dataset.

Figure 5. Top Nouns Used by Twitter Users

Open Figure 5

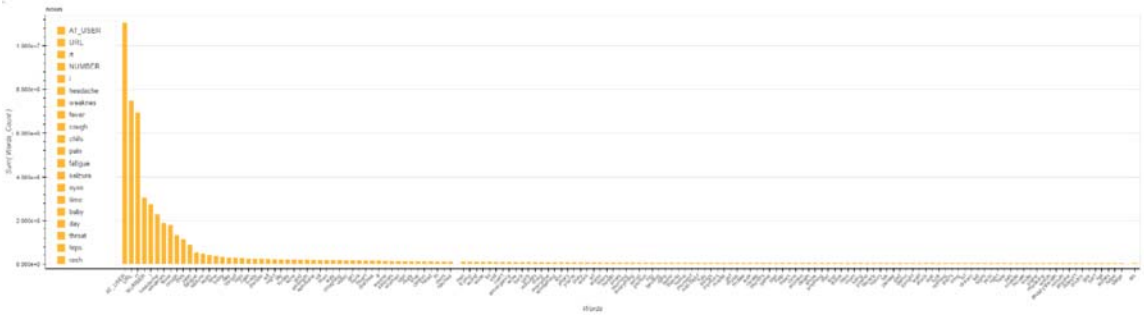
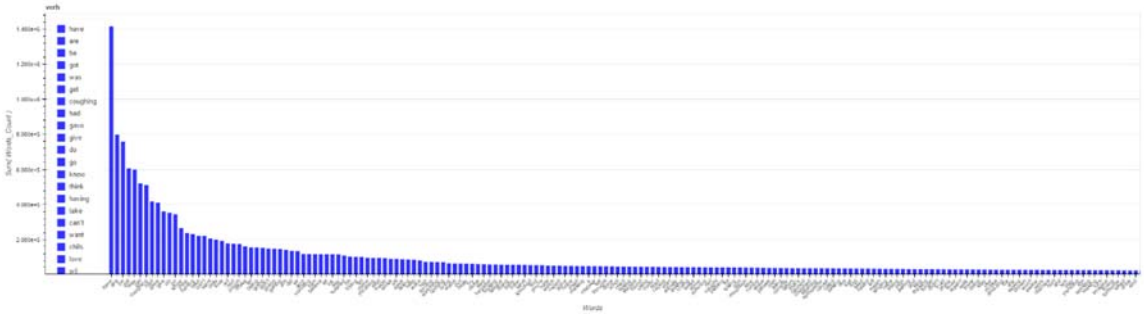


Figure 6. Top Verbs Used by Twitter Users

Open Figure 6

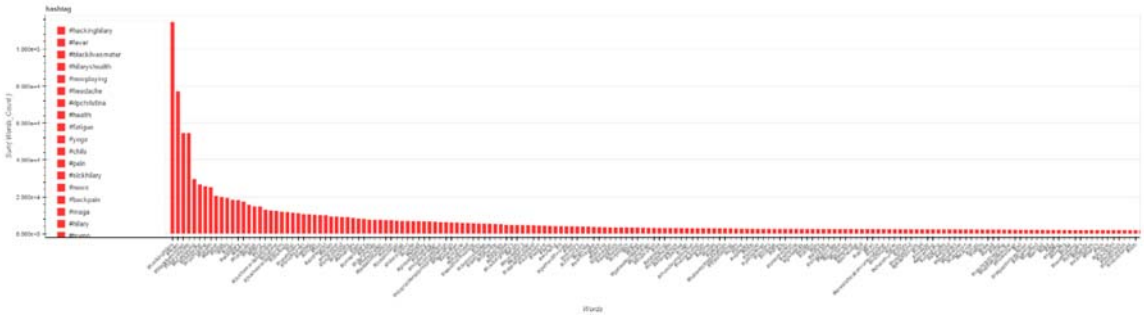


Categorize Tweets by Keyword

A hashtag—written with a # symbol—is used to index keywords or topics on Twitter. This function allows people to easily follow topics they are interested in. Hashtagged words that become very popular are often Trending Topics.
We summarize the top-used hashtags in the dataset. Most of the topics are health related, for example, #fever, #headache, #health, #fatigue, #pain; also political related topics are not uncommon, for example, #hackinghilary, #blacklivesmater, #hilaryshealth, #sickhilary.

Figure 7. Top #hashtags Used by Twitter Users

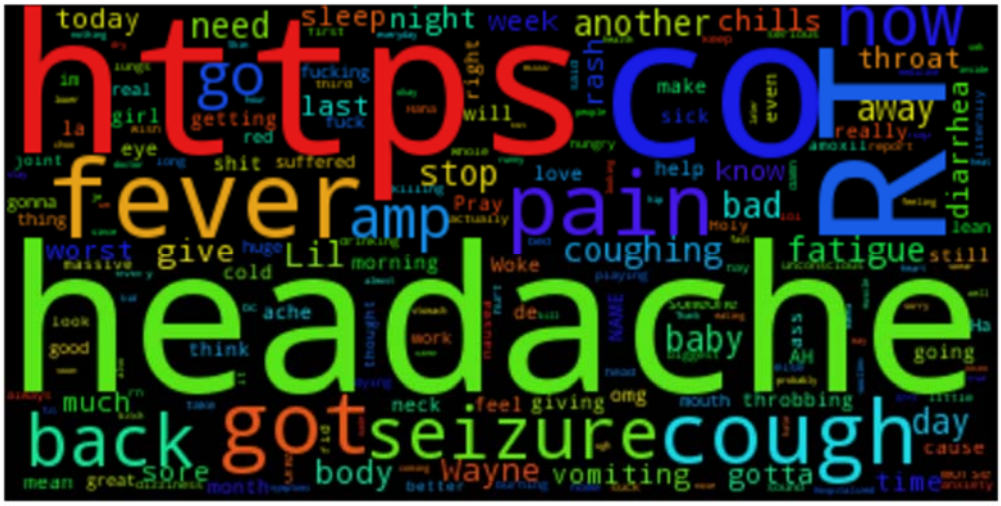
Open Figure 7



Key conversations (topics) and trends

Topic modeling is a form of text mining, a way of identifying patterns in a dataset. Latent Dirichlet Allocation (LDA) algorithms was applied in order to discover the latent topics.
LDA generates topics based on word frequency from a set of documents. LDA is particularly useful for finding reasonably accurate mixtures of topics within a given document set.
Below are some example topics.

Figure 8. Topic A: Symptoms



Example tweets from this topic:

- I have the worst headache☹

- All day long with headache and when i finally get some pills i t hrow up.Great! :)

- RT @Cheaats: Eating 3 or 4 Marshmallows helps get rid of a sore throat.

- When I stay inside all day and watch TV my body start to ache. I 'm an active person sitting around isn't for me.

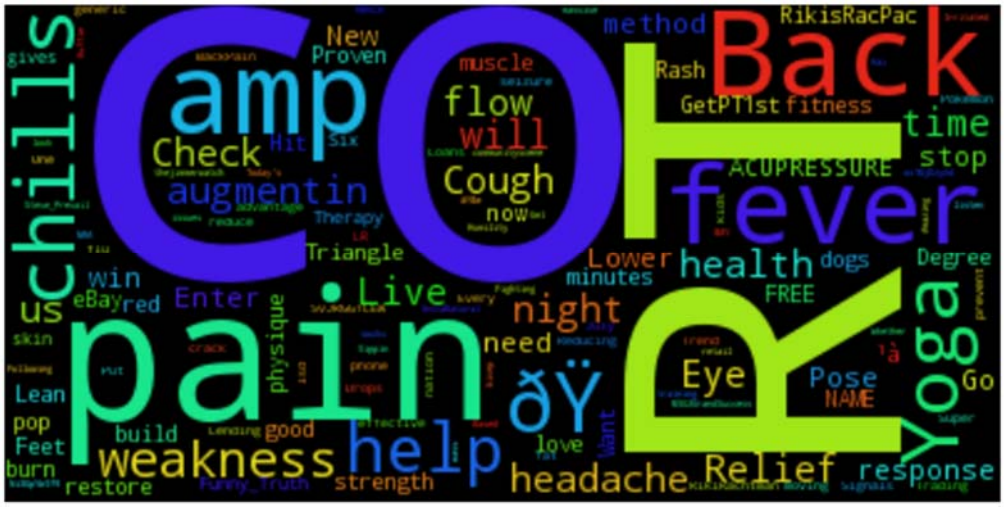
- hi. i have a fever. what? i don't remember the last time i actua lly had a fever. help. what is illness.

Figure 9. Topic B: News

Example tweets from this topic:

```
- Lil Wayne suffers third seizure in under a month https://t.co/yf
qdqaPV23 https://t.co/ol9xTp6m1Q

- Heath Cast: Back Surgery at Any Age - Almost 25 percent of all d
octors' visits for low back pain are patients o... https://t.co/9h
4QVYj5oQ

- RT https://t.co/Jds90xA0Eu Have back pain? We'll have help from
noon to 5:30 p.m. tomorrow with @Intermountain. #â€¦ https://t.co/
HHOoZz9pGk

- Researchers say that back pain affects about two-thirds (68%) of
 women either during or following a pregnancy https://t.co/1kylj5X
Wh8

- RT @Mixmag: The science behind why we get chills when listening
to music - explained. https://t.co/6I5rSrmmcA https://t.co/FH8Uf3X
fz4
```

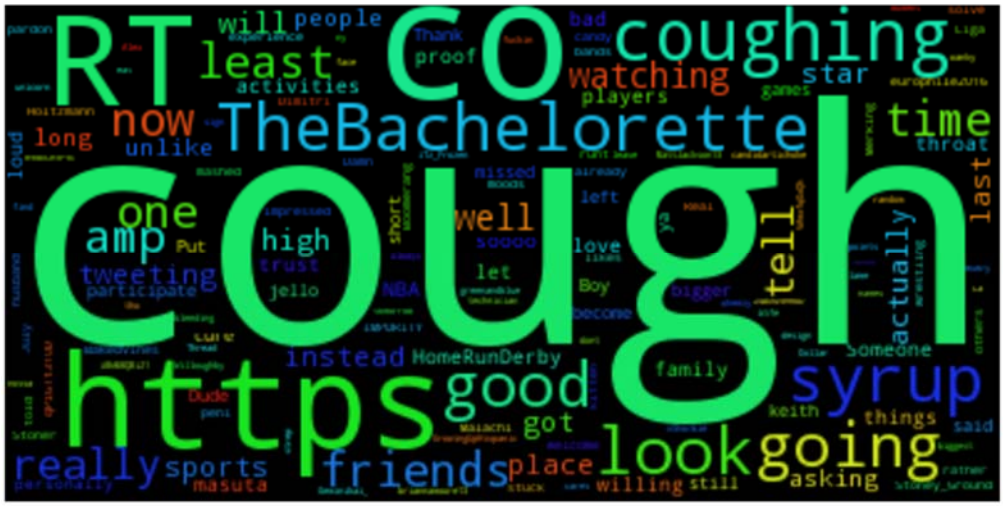
Figure 10. Topic C: Advertisement



Example tweets from this topic:

```
- Do you have a retail business? Have you experienced the FEVER of
 Pokemon-Go? Take advantage of the foot traffic... ://t.co/SE8DXpU
5pG

- for more far out, headache material check out my latest film on
lsd ://t.co/CAc1Dehimk

- Amazing Stretches To Alleviate Lower Back Pain ¦ : ://t.co/o26X5
xRYhF ,,, ://t.co/R4qtN6abPK

- Cough up less for Lenovo Yoga 3 Pro - 80HE011XUS ! Bid on it! Go
 to eBay ://t.co/XGPjiJKWPd
```

Figure 11. Topic D: Irrelevant: Cough



Example tweets from this topic:

- RT @emotionlessvibe: Stop asking me to trust you when i'm still coughing up water from the last time you let me drown.

- RT @zef_aliens: Thread of vines that have made me cough up blood

- At least most of the bigger players in the MLB are willing to participate in all star activities unlike some sports *cough cough* NBA.

- I caught have strep throat. ATTN NYERS: Can you please cover your mouth before you cough on me in the future. Thank you.

Figure 12. Topic E: Irrelevant: Fever

Example tweets from this topic:

- @LilKim !!! Give them fever !!! https://t.co/ycSYt02UOd

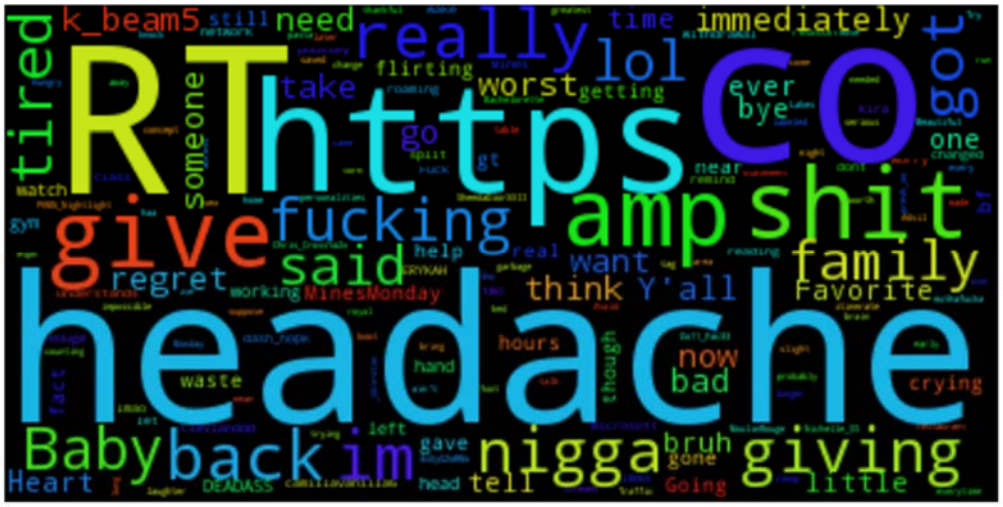
- The nurse is hot....i think I'm developing a fever...ITS TIME TO TEST OUT THIS ACTING CAREER EARLY

- The latest Libra's Gospel Music Fever Digest!

- Baby fever 😷🤰

- Sun is out!! 😊☀ ... Fever is in huhu

Figure 13. Topic F: Irrelevant: Headache



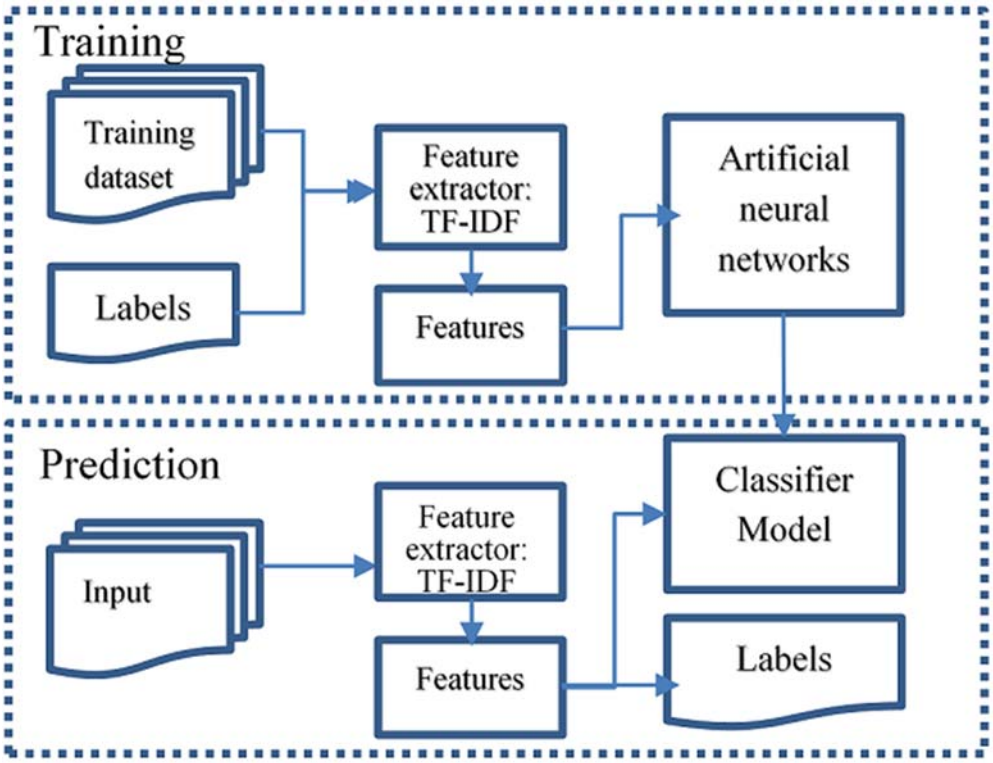
Example tweets from this topic:

```
- If I'm tired and have a headache she expects me to change my att
itude☺☹

- My greatest headache and favorite stress reliever is my family
💞

- Hungry and happy with a headache☺😉

- Being in a relationship is a fucking headache...
```

Theme Analysis

One of the challenges we needed to address is to extract signal from the noisy Twitter dataset i.e., to distinguish tweets that are relevant to symptoms from tweets that mentioned keywords in an irrelevant context. Figure 13 shows the process used for categorizing Twitter dataset.

Figure 14. Theme Analysis



Training data

According to the result of LDA topic modeling, the **vector borne diseases** symptoms-related tweets can be classified as 4 categories:

Table 3. Tweets Categories

Label	Theme
I	Irrelevant: not a symptom
R	Relevant: disease symptoms, e.g., fever, diarrhea, ……
N	News: news, policy and government themes
A	Advertisement: marketing messages

Classification

Classification is applied to identify to which of a set of categories a single tweet belongs to. To this purpose, a small amount of data needs to be labeled by human operators. 15 junior and senior students from University of Arizona were invited to label **1,537 tweets** (randomly sampled from the dataset) as *"Relevant"*, *"Irrelevant"*, *"News"* and *"Advertisement"*.

The classifier Logistic Regression was trained on this dataset. A 10 fold-cross validation was executed to evaluate the performance.

Table 4. Classification Performance

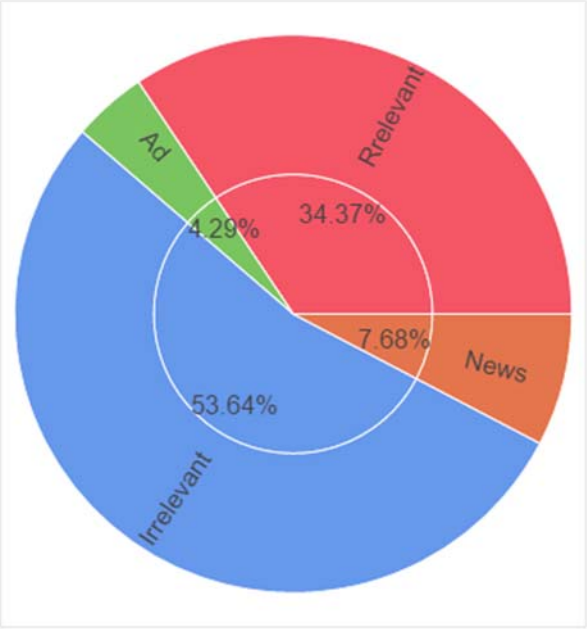
LogisticRegression	Irrelevant	Relevant	News	Advertisement
precision	0.690	0.653	0.677	0.267
recall	0.820	0.597	0.316	0.197
f1-score	0.746	0.616	0.413	0.136
accuracy	0.670			

Percentage of theme categories

The pie chart below shows the percentage of each category .

Figure 15. Theme Composition (9.09% tweets from US)

Open Figure 15



4. Time Series Analysis

Time series analysis is an illustration of the number of Tweets at successive time intervals. We could extract meaningful characteristics of the data and predict future values based on previously observed patterns.

Figure 16. Number of Tweets per Day (US)

Open Figure 16

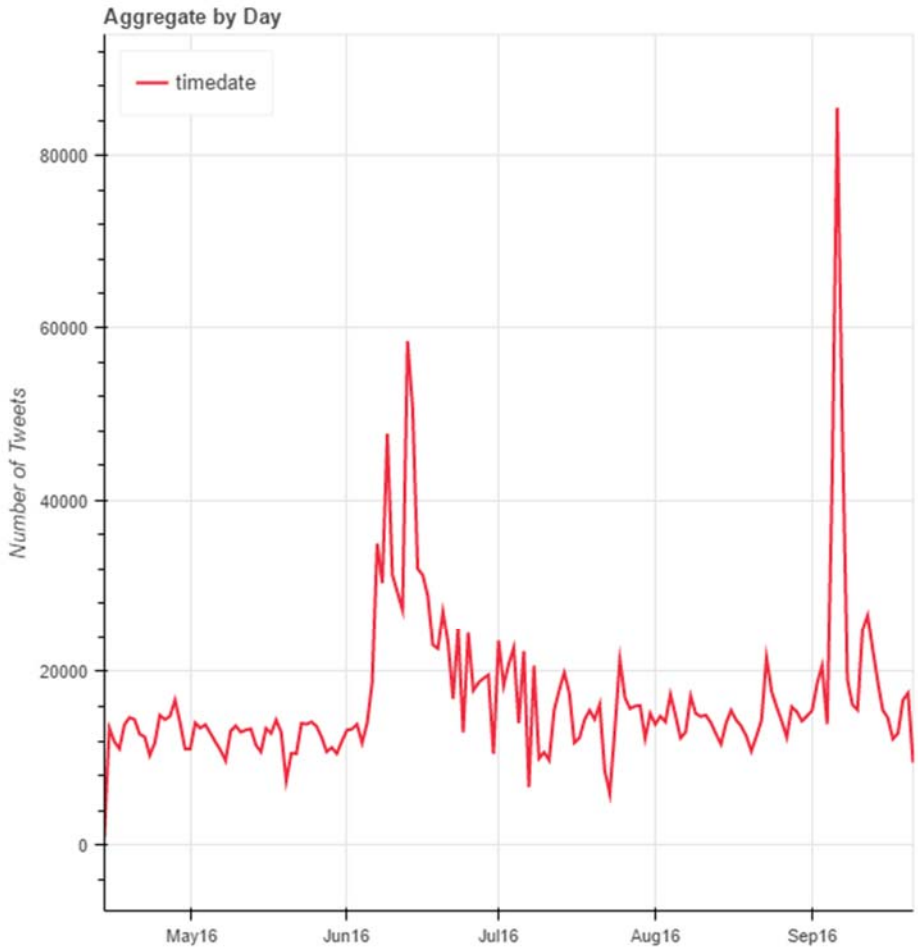


Figure 17. Number of Tweets per Hour (US)

Open Figure 17

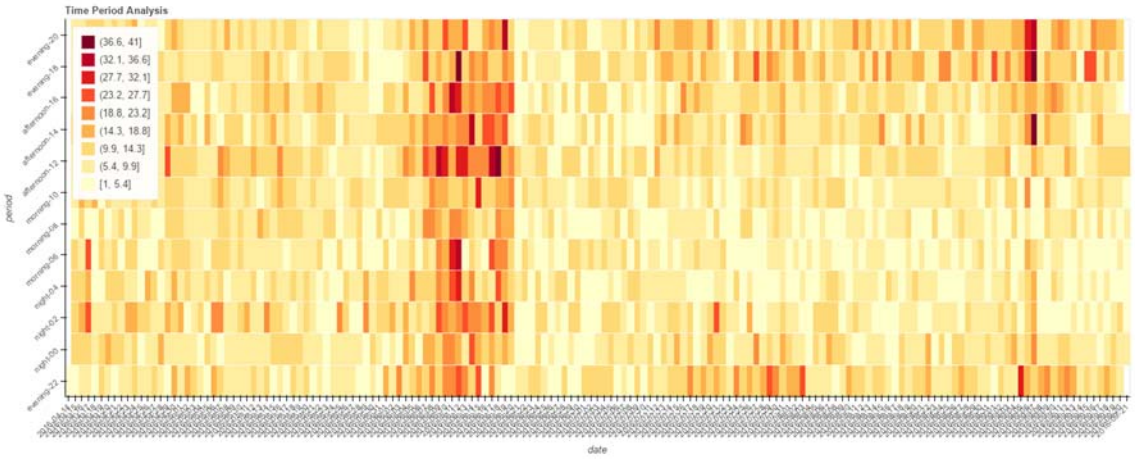
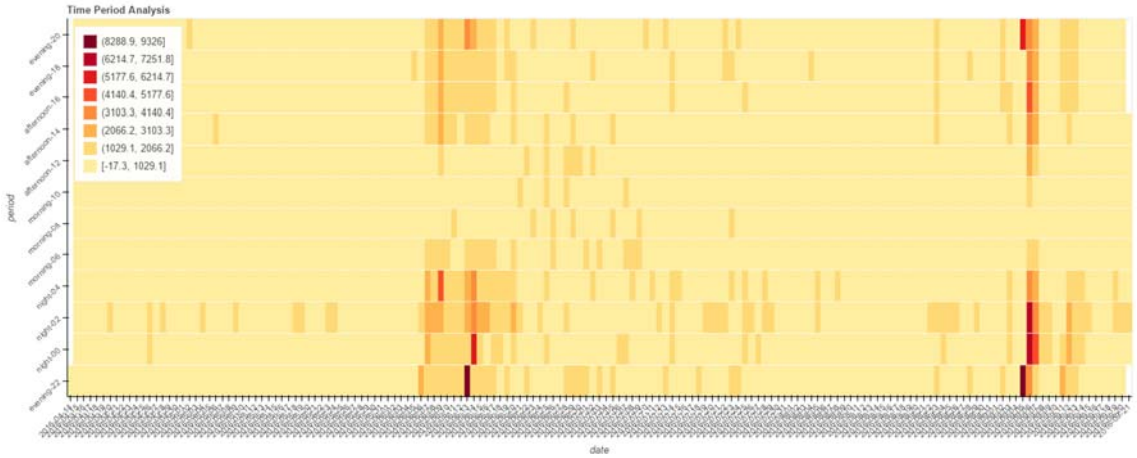


Figure 18. Number of Tweets per Hour (World)

Open Figure 18



5. Mapping and Visualization

Figure 19. Mapping Timeline

Open Figure 19

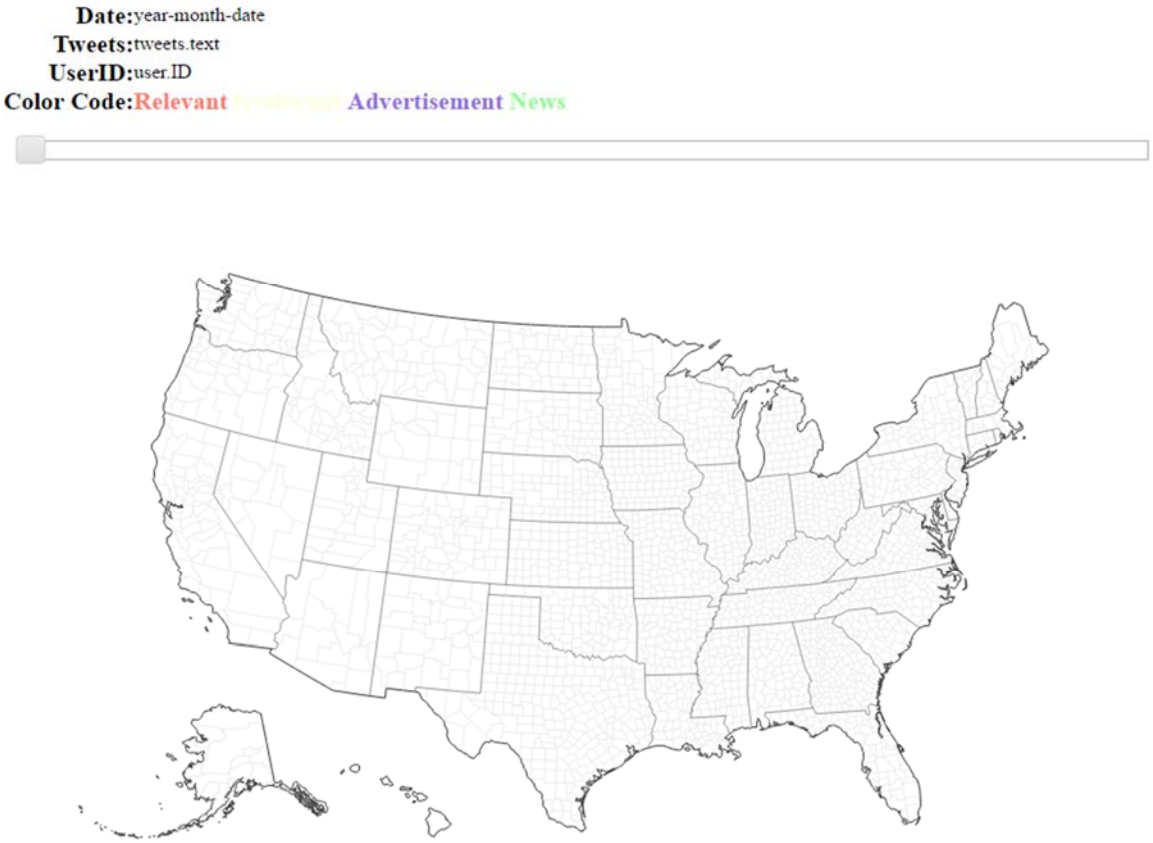
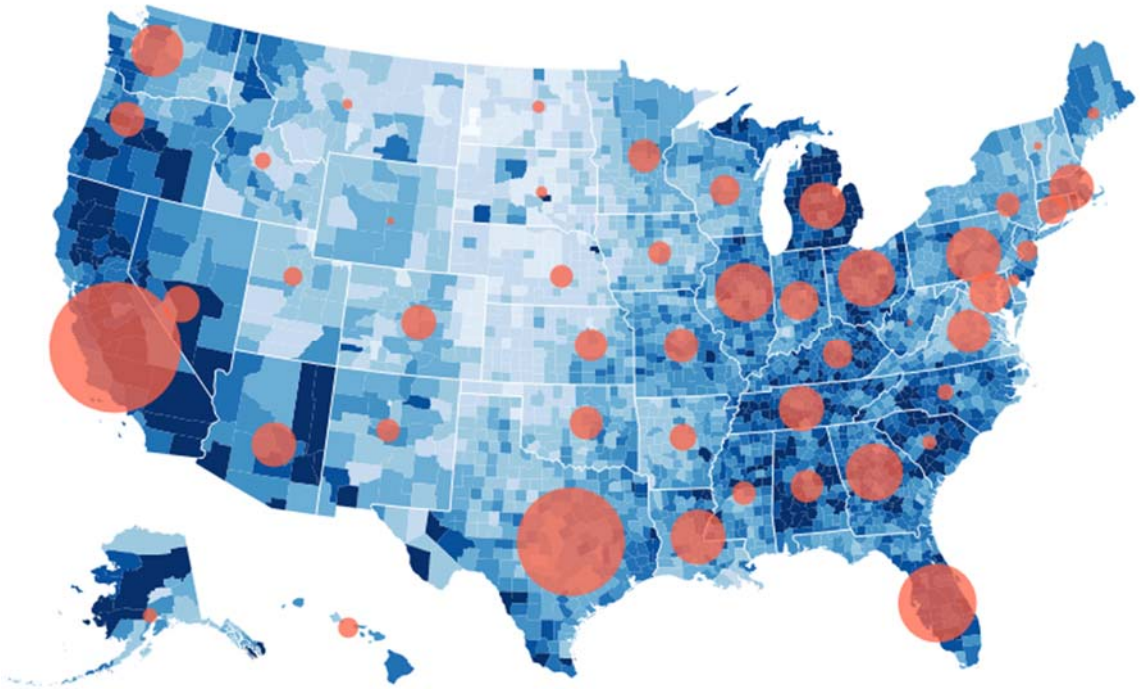


Figure 20. Mapping Number of Tweets aggregated by US States

Open Figure 20



6. Flights and Travel Analysis

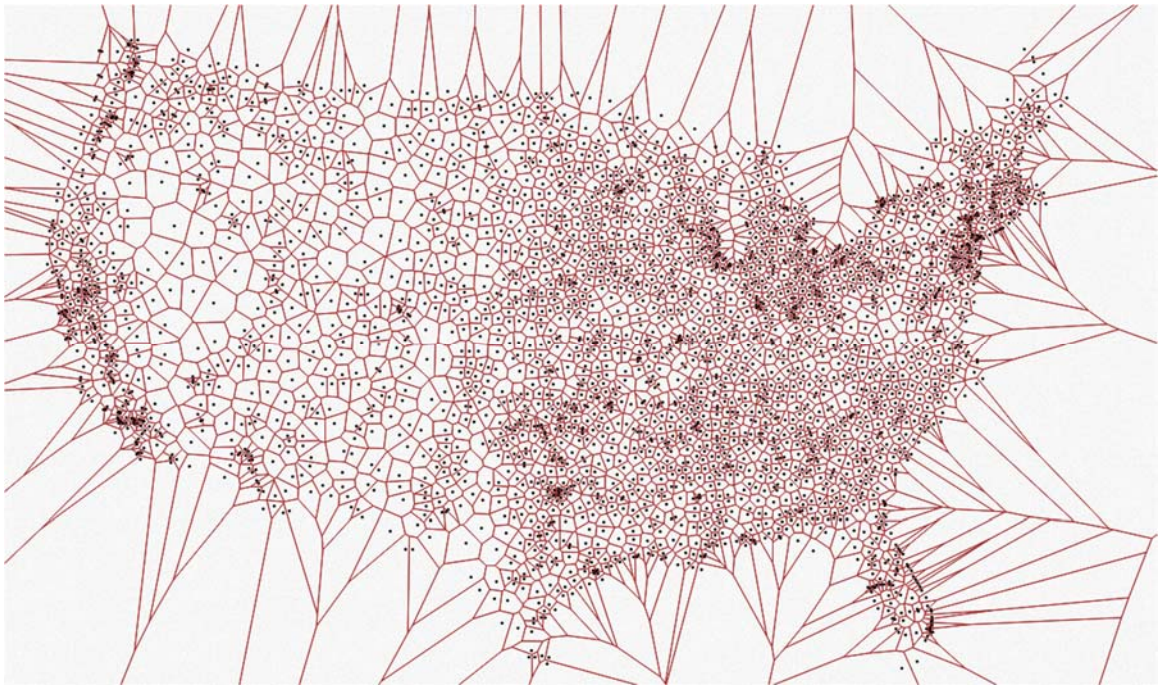
Airports and Flights

According to the literature review, with the outbreaks of vector borne diseases, the number of cases among travelers visiting or returning to the US from affected areas will likely increase. These imported cases could result in local spread of the virus in other parts of the US.

We visualized all US airposts and Voronoi algorithm was applied. Voronoi diagram is a partitioning of a plane into regions based on distance to points in a specific subset of the plane. For each seed there is a corresponding region consisting of all points closer to that seed than to any other. These regions are called Voronoi cells.

Figure 21. All Airports in US

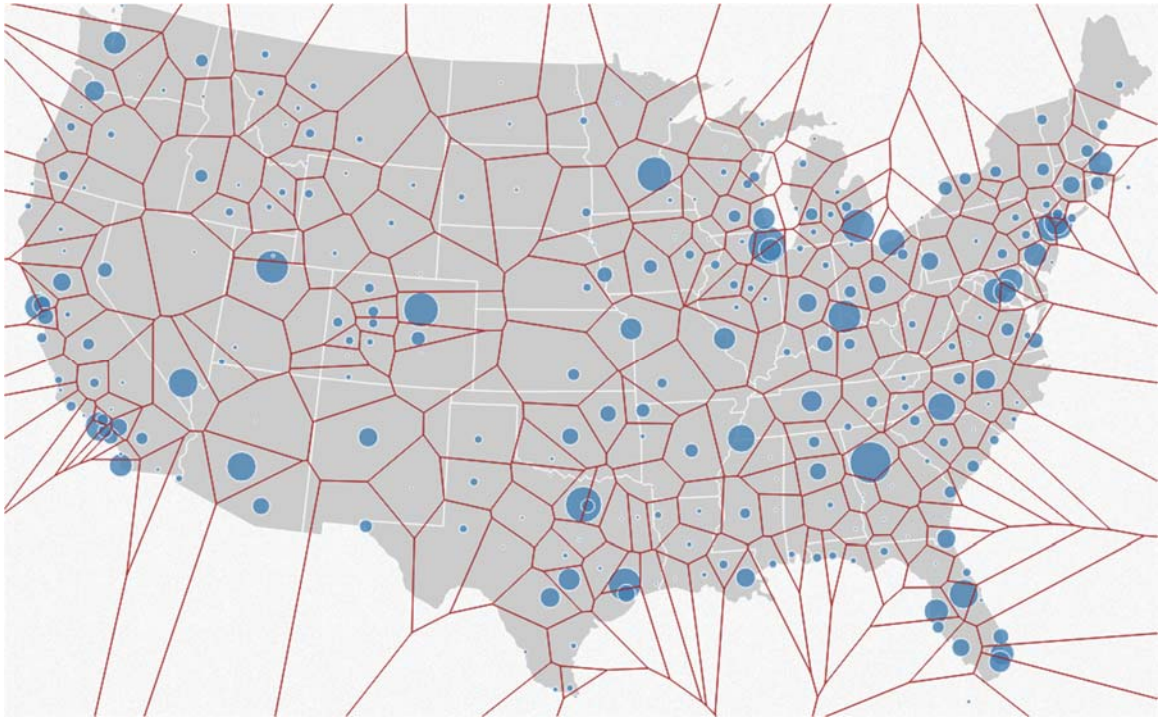
Open Figure 21



Then airposts with only one flight were filtered out. We can hover to see flight routes associated with each airport.

Figure 22. Mapping Airports and flights

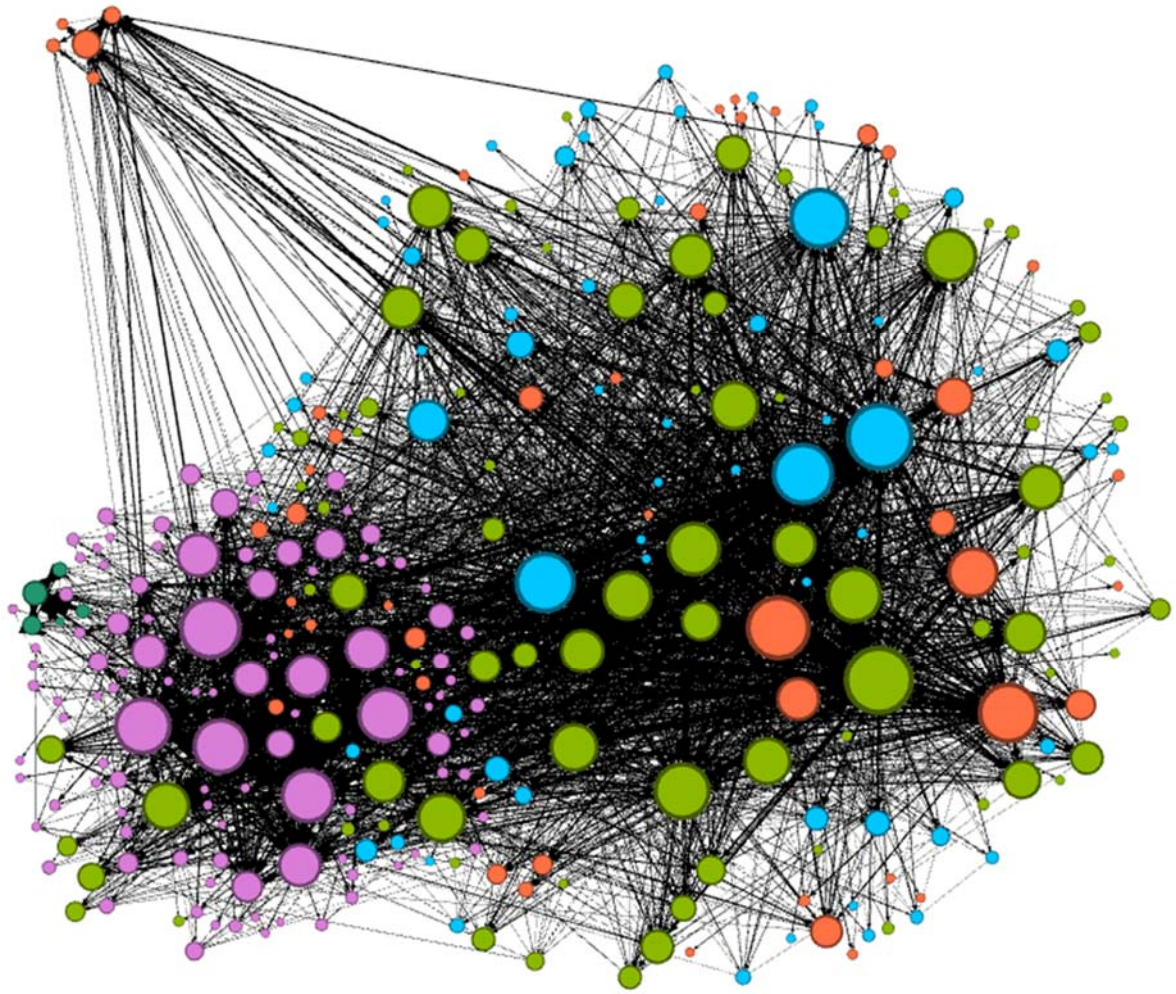
Open Figure 22



Flights Network Analysis

Figure 23. Flights Network Analysiss

Open Figure 23



7. Twitter Dataset Visualization for Specific Vector (mosquito)

We represented "mosquito" related tweets as a full representation with nodes for Users, Tweets, Hashtags, Urls and Media.