EPiC Series in Computing

Volume 69, 2020, Pages 362–371

Proceedings of 35th International Conference on Computers and Their Applications

Disease Outbreak Detection Using Search Keywords Patterns

Izzat Alsmadi[1,*], Zaid Almubaid[2], and Hisham Al-Mubaid[3]

[1]Texas A&M San Antonio, San Antonio, Texas, USA

[2] University of Texas, Austin, Texas, USA.

[3]University of Houston–Clear Lake, Houston, Texas, USA

*Izzat.Alsmadi@tamusa.edu

Abstract

In the recent years, people are becoming more dependent on the Internet as their main source of information about healthcare. A number of research projects in the past few decades examined and utilized the internet data for information extraction in healthcare including disease surveillance and monitoring. In this paper, we investigate and study the potential of internet data like internet search keywords and search query patterns in the healthcare domain for disease monitoring and detection. Specifically, we investigate search keyword patterns for disease outbreak detection. Accurate prediction and detection of disease outbreaks in a timely manner can have a big positive impact on the entire health care system. Our method utilizes machine learning in identifying interesting patterns related to target disease outbreak from search keyword logs. We conducted experiments on the flu disease, which is the most searched disease in the interest of this problem. We showed examples of keywords that can be good predictors of outbreaks of the flu. Our method proved that the correlation between search queries and keyword trends are truly reliable in the sense that it can be used to predict the outbreak of the disease.

Keywords: Disease outbreak detection, disease monitoring and surveillance.

1 Introduction

In recent years, the *Internet* has become highly accessible and more reliable, and people are depending on the Internet for almost all aspects of their lives including the health aspects. Furthermore, the Internet is becoming a major source of information [18]. Most people, including younger generations, and millennials in particular, are using the Internet extensively for obtaining healthcare information like disease symptoms, medication and doctor referrals among other things [2, 5, 7, 9, 11]. Most importantly more people are making healthcare decisions based on the information obtained from the *Internet*.

Researchers interested and focusing on healthcare and the *Internet* reported two highly important uses of the *internet* in healthcare: first, using the *Internet* for the detection and prediction of disease outbreaks like epidemics. Second, using the *Internet* for tracking disease geographical locations and movements.
In this paper, we investigate and analyze the usage patterns of search keywords in the *Internet* within the healthcare domain for disease outbreak detection. We want to analyze the search keyword changes related to certain diseases and a possible outbreak of those diseases over extended time periods in the USA.
Studies in healthcare informatics show that doctors and healthcare professionals rely on online search results in obtaining more information about diseases, symptoms, drugs, and other related information [5]. Also, the research showed that doctors find searching online is very helpful to get information about tracking geographical locations of disease.
The Internet search data including keywords and queries are simply the information that people are seeking and trying to find in the internet. Studying the changes in the volumes of users' popular search terms in the *Internet* can be used to study different cases of event correlation and user behavior in some cases like response to natural disasters, breaking-news, disease outbreaks, and so on.
In this work, our goal is to find out and associate popular *Internet* search terms with *Flu* outbreak or epidemic cases. We extract users popular search terms using *Google Trends* [4, 19]. Given that Google is the main search engine in the US and world, volumes and patterns of search queries extracted using Google Trends [4] can fairly represent the trend of what people are seeking in terms of healthcare information from the *Internet* [4, 11, 19]. That is, we can view the Google Trends data about the *flu* disease (or any of its related terms like *influenza*, *common cold*, etc.) as an indicator of the status of the flu and the public interest in information about flu.
The relationship, if any, between the trends/patterns of *Internet* search keywords related to certain diseases and possible outbreaks in that disease is the interest of this work. We expect to see different levels of correlations between those two points for various reasons as any spike in keywords can be correlated with event(s) related to the keywords. In this research, we focus of disease outbreaks of infectious diseases in the US and collect their historical monthly counts over several years. Then, we match data availability with the process of extracting "relevant" searched-for keywords through Google Trends.

2 Related Work

An epidemic is a quick spread of an infectious disease in a specific area and over a short period of time [1, 17]. We could not verify the fact that a disease outbreak is called *epidemic* when it goes out of control. A number of research projects and studies in the literature examined and confirmed the correlation between disease outbreaks and online search keywords and trends, but still various issues need to be further investigated. In [5], Nawaz et al. (2017) showed that people can get vital first-hand healthcare information from *Google* and *Twitter* [5]. They investigated whether it is possible to use tweets to track, monitor and predict diseases, especially Influenza epidemics. Their results show that healthcare institutes and professionals' use social media to provide up-to-date health related information and interact with the public [5].
In general, using the *Internet* for obtaining healthcare information is limited to search engines and social media [9, 11, 13]. By and large, social media can be represented by *Facebook* and *Twitter*, while search engines can be represented by *Google* [11].
Carneiro and Mylonakis (2009) reported that *Google Trends* for Flu has strong correlations with retrospective surveillance data from the CDC and accurately estimated influenza levels 1–2 weeks earlier than published CDC reports [11].

Aitken et al. (2014) found that there is a clear direct relationship between drug sales and Wikipedia traffic for a selection of approximately 5,000 health-related articles [17].
In [15], Ye et al. (2016) reported clear statistical correlations between online posts and diseases [15]. Specifically, in their research, the Pearson correlation coefficient in the ANOVA (Analysis of Variance) was 0.954, suggesting that the number of new Weibo posts was significantly correlated (at the 2-tailed 0.01 level) with the number of new dengue fever [15]. Aramaki et al. (2011), in [9], mention a telephone triage service, which is a public service, to give advice to users via telephone. They investigated the number of telephone calls and reported a significant correlation with influenza epidemics [9]. Nuti et al. (2014) present a comprehensive study on the research using Google Trends in the health care domain [19].

3 Approaches for Disease Outbreak Detection

One of the goals of this paper is to examine if we can use/utilize search queries and keywords in Google Trends for disease outbreak prediction. In other words, are the search queries and keyword trends truly reliable to be used for the prediction of disease outbreaks? To investigate this problem, we propose a method based on *Google Trends* and the historical disease data from official sources for selected infectious diseases. Our overall process steps of this work are as follows.

(1) Select a target disease (disease of interest *D*). (2) Let *N(D)* be the set of all names, acronyms, synonyms, symptoms (major ones) and diagnostics of *D* {*note: this is the set of the keywords used by the population to search for disease D*}. (3) Let *P(d)* be the search pattern (user keyword volume pattern for searching the terms/keywords in the set *N(D)* in step 2 above. (4) Analyze the pattern (*volume changes in search*) from step 3 above against official disease outbreak data from *CDC*. (5) Build the model to identify the most significant keywords that can accurately predict disease outbreaks based on real historical official disease statistics (*CDC*) and verify the significance of the results with $p < 0.05$.

Initially, for a specific target disease, we would like to compose an accurate set of keywords for the disease. This step is conducted by researching the literature on that target disease and its related terms, e.g. *flu*. We also use *Google Trends* (*GT*) in this step:

- In GT, we analyze the "*related queries*" section in *GT* and extract the *top* and *rising* search terms as shown in Table 1. *GT* provides, along with the trend data for the topic of interest, a set of other related queries and label them as '*top*' or '*rising*'. As shown in Table 1, *GT* distinguishes terms breakout and rising keywords in terms of how quick and how long such terms have been on the search rise. As an example, for the *Malaria* disease, the search trend data as obtained from *GT* are illustrated in Figure 1.
- For a search term to be selected by our technique in the evaluation dataset, it should have two requirements: (1) It should be extracted only from the "*related queries*" in GT results in the *top* section or listed with more than 90% *rising* terms. (2) It should be repeated for more than 4 times in the "*related queries*" different results, i.e. from different initial search terms.

Keyword	Subject	Related topic	Value
Malaria	Top	Symptom	100
Malaria	Top	Disease	80
Malaria	Top	Pharmaceutical drug	53
Malaria	Top	Vaccine	51
Malaria	Top	Therapy	45
Malaria	Rising	Zika virus	Breakout
Malaria	Rising	Zika fever	Breakout
Malaria	Rising	Imperialism	850%
Malaria	Rising	Ebola virus disease	750%
Malaria	Rising	Chikungunya virus infection	600%

Table 1: The first few top and rising topics related to keyword *Malaria*. The Value column indicates whether it is breakout or percentage value only for rising topic terms.

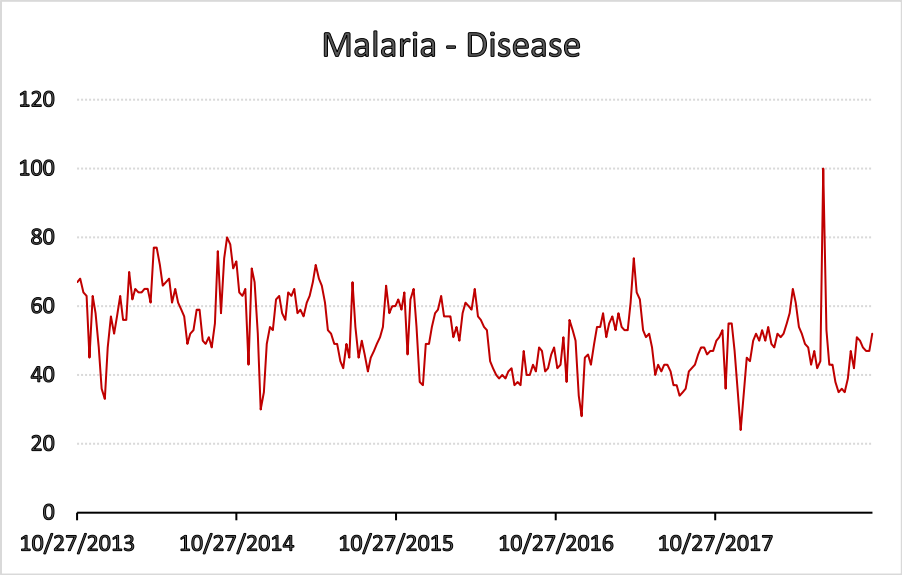

Figure 1: Results from GT about the trends of Internet search queries for the keyword *Malaria* as a disease.

Protease inhibitors HIV	Seronegative	Prep
HIV	Antiretroviral Therapy	Prevalence
HIV1	Celibacy	Concordant
HIV2	Antiretroviral	Noncompliance
Enzyme inhibitor	Serosorting	Regimen
AIDS	HIV Neutral	Viral Load
Protease	Longevity	Asymptomatic
Integrase	Seropositive	Western Blot Test

Table 2: HIV Selected Search Keywords

In three different experiments, we evaluated correlations between the *Trends* and *Disease* arrays. We examined both *Pearson* and *Spearman* and found no significant differences; and therefore, we report here with only *Pearson* correlation. As an example, Table 2 shows the selected keywords for the HIV disease.

Google Trends [4] is a comprehensive information source and database for user search queries and keywords over a long period with analysis and visualization features [4, 11, 19]. This resource offers a tremendous amount of information about the trends and patterns of searching the *internet* via *Google* search engine which is the most popular search engine in the *Internet* in the US. From GT, one can find and analyze the popularity of top search queries in Google across various regions and languages and over various timeframes like within days, months, and up to ten years. Moreover, GT offers graphs (see Figure 1) to compare the search volume of different queries over time [4, 11, 19].

If the search keywords related to a target disease *D* are highly correlated with the disease outbreak, then can we determine the small set of such keywords related to *D* that can best predict the outbreak of *D*. If disease *D* has *n* related keywords including disease names, acronyms, symptoms, etc. then we induce a set S_D of related keywords that can best represent the outbreak of disease *D*:

$R_D = \{R_{D1} \dots R_{Dn}\}$: the set of all related keywords to disease *D*.

$S_D = \{S_{D1} \dots S_{Dm}\}$: the set of significant keywords related to disease *D*.

In terms of correlation, no significant positive or negative correlation is shown in the volume of those terms and cases volumes. The highest keywords in terms of correlation (negative or positive) were: Seronegative, Viral Load, and HIV.

4 Experiments and Analysis

4.1 Disease Breakout prediction using SVM

We developed a learning model based on *support vector machines* (SVM) for disease breakout prediction using on collected features and with different experiments. In our learning model, and for a given disease, we use keywords selected from the *GT* where the records represent monthly data of the disease of interest. This makes all features the same, with similar values range, except the class. In the class, we relied on the official disease data from USA *Center for Disease Control and Prevention* (CDC) [22]. In USA, the CDC center is a well-known and highly regarded as the most reliable source of all health information and disease data in official and accurate form; therefore, it can be a good source of official and accurate disease data in the US including disease outbreak statistics.

Year	Epis	Month	Epi-Class	Year	Epis	Month	Epi-Class
2003	499	1	1	2004	223	4	0
2003	427	2	1	2004	17	5	0
2003	89	3	0	2004	66	10	0
2003	39	4	0	2004	154	11	0
2003	52	10	0	2004	358	12	0
2003	244	11	0	2005	447	1	1
2003	847	12	1	2005	578	2	1
2004	503	1	1	2005	741	3	1
2004	694	2	1	2005	327	4	0
2004	645	3	1	2005	61	5	0

Table 3: A sample of extracting Flu Epi-Classes

Determining the Flu Epidemic Models (Epis): We extracted the *flu* (aka. *influenza*) details from CDC and based on the CDC initial categories (for the state-by-state model of weekly activities) six categories are found: *No Activity*, *Sporadic*, *Local Activity*, *Regional*, *Widespread*, and *No Report*. To estimate epidemic class we used the following model where we distinguish two epidemic classes *1:Yes*, *0:No*; based on CDC estimates as follows:

(1) We collected GT data at the US national level; and in order to match this with CDC data, we aggregated data from all states.

(2) Since GT reports data on monthly basis and in order to match that with CDC data, we aggregated CDC weekly data into monthly data.

(3) Our model includes data from 2004 to 2017.

(4) We transformed CDC disease activity categories into five (we merged *No Activity* and *No Report* into one category here) as follows:

0: No Activity & No Report	3: Regional Activity
1: Sporadic	4: Widespread
2: Local Activity	

And the final total values range 0 to 1000.

(5) We computationally calculated the best *cutoff* to be at 450 for the epidemic class as shown in Table 3.

Date	Influenza	Malaise	Myalgia	Nasal con	Nausea	Neuramin	Orthomyx	Vomiting	Zanamivir	Class1
1/1/2004	11	21	76	33	21	0	0	27	0	1
2/1/2004	21	23	57	36	23	0	0	28	0	1
3/1/2004	27	29	81	14	23	0	0	27	0	1
4/1/2004	10	31	92	24	24	0	0	29	3	0
5/1/2004	0	25	82	15	23	59	0	31	0	0
6/1/2004	0	42	70	22	23	0	0	34	0	0
7/1/2004	0	43	88	12	26	0	0	26	0	0
8/1/2004	27	23	74	22	26	0	0	28	0	0
9/1/2004	9	22	76	27	27	0	0	25	3	0
10/1/2004	17	30	100	40	24	0	65	31	5	0
11/1/2004	27	29	96	24	26	0	0	29	6	0
12/1/2004	8	33	74	25	29	0	0	36	0	0
1/1/2005	7	32	87	46	29	0	0	37	4	1
2/1/2005	59	27	76	39	25	0	0	37	2	1
3/1/2005	38	30	98	28	27	0	0	32	8	1

Table 4: Flu SVM model using Google Trend Popular Terms

Keyword	Corr.	Keyword	Corr.
Viral pneumonia	0.621308	Influenza	0.389872
Common cold	0.473166	Fever	0.359736
Influenza A virus	0.419551	Flu season	0.342614
Nasal congestion	0.418517	Cough	0.335666
Gastroenteritis	0.389948	Cat flu	0.301257

Table 5: Popular Search Terms Correlation with Epis Classification Model, Popular terms from: Generous et al. 2014

For the selection of Flu Google Trends keywords, we started with the set of keywords from *Generous et al.* (2014) [1] {*note: in [1] Generous et al. studied 54 related keywords to Flu*}. An *SVM* model is then built for the *Flu* with *Google Trends* keywords as features and Epi-Class, shown in Table 4. In terms of keywords correlation with Epis monthly total, following keywords, Table 5, showed the highest (more than 30%).

Our extended dataset showed more keywords with correlation higher than 30%, Table 6. This indicates that our selected list of popular terms can better predict *flu* disease outbreaks.

We will then analyze prediction based on *Decision Trees* using the developed *SVM* model. We evaluated different feature extraction methods in Decision Tree classification (Figure 2) to extract features that can best predict class change from None-Epi to Epi and vice versa {*the classification, feature extractions, .etc. are implemented in Java*}. Results can be summarized as follows:

In the first *Flu* dataset, we used conditional feature selection (*CfsSubsetEval*) method for feature selection along with *GreedyStepwise* Search algorithm, and the best selected features are as follows:

Amantadine	*Common cold*	*Gastroenteritis*
Influenza A virus subtype H1N1	*Influenza-like illness*	*Viral pneumonia*

And the overall prediction performance as follows:

Accuracy:	97.6
Precision:	97.6
Recall:	97.6
F1-score:	97.6

The model above shows detailed results where 45 (or 93.8%) of the 48 instances of the positive class (*disease-outbreak* class) were predicted correctly based on the best selected attributes (i.e., only 10 attributes). For 120 records of the negative class, *i.e.*, the *No-Flu Outbreak* class, 117 (or 97.5%) where predicted successfully. This is interpreted as that the 10 attributed listed (Amantadine, Influenza-like illness, Viral pneumonia, ….) can predict the existence of an epidemic with ~97.6% accuracy (more detailed results are also in Figure 3).

Keywords-based prediction using linear regression

As predictor class (i.e. historical disease volumes) and all dependent variables (i.e. historical Google Trends' Keywords) are continuous values, we decided to use linear regression to study whether and how much Google Trends' variables can reflect disease volumes. The results from Regression coefficients' summary (*not reported*) proved that Flue historical disease volumes represent the first predictor class in the table with the 40 *Google* keywords as dependent variables (with $p<0.05$).

Moreover, the results indicated that 10 keywords (Intercept is related to Flu disease volumes and not a keyword) from our dataset (i.e. Equine.influenza, Influenza.A.virus, Influenza.A.virus.subtype.H10N7, Influenza.like.illness, Malaise, Oseltamivir, Shivering, Viral.pneumonia, Vomiting, Zanamivir) are good predictors for Flu disease volumes (with $p < 0.05$). Their level of impact (i.e. Estimate column) varies where the keywords (Vomiting, Influenza.A. virus. subtype.H10N7, Viral.pneumonia, Shivering, Influenza.like.illness, Influenza.A.virus) are the highest with positive impact, and (Zanamivir, Oseltamivir, Malaise, and Equine.influenza) are the highest with negative impact or correlation. All those picked keywords have a *p-value* of less than 5%. In other words, with all those keywords, we can reject the null hypothesis and say that a relationship exists between those keywords and Flu disease volumes.

In order to extract only results with enough accuracy credibility, we used Excel VB scripts to pick only keywords with $p\text{-value} < 0.05$. The intercept record was included always regardless of their *p-value*. The estimate value in those selected records shows the following information:

1. Which keyword can show a significant impact (i.e. all those that were selected with a *p-value* of less than 5%)?
2. The value of that impact (i.e. the higher the value, the higher the impact).

3. The value of those variables represent a correlation coefficient between -1 to 1, where if the value is positive and high, then it is a high positive correlation, and vice versa.
4. Second, we want to look at the "weight of the influence" measures by the "Estimate" values.

The top positive estimate values from the experiments (*not shown*) proved how much such keywords can influence the prediction accuracy.

Keyword	Corr.	Keyword	Corr.
Viral pneumonia	0.621308	Influenza	0.387644
the common cold	0.515474	Cold	0.379964
viral gastroenteritis	0.496596	Fever	0.368357
Common cold	0.473166	flu symptoms	0.349797
is influenza a virus	0.451952	flu influenza	0.344205
the cold	0.433808	the flu season	0.343357
Influenza A virus	0.419551	Flu season	0.342614
Nasal congestion	0.418517	Cough	0.335666
Incubation period	0.390718	cough syrup	0.331482
Gastroenteritis	0.389948	Bacteria	0.321042
Cat flu	0.301257	how to get rid of a cough	0.312222

Table 6: Popular Search Terms Correlation with Epis Classification Model, Our own extracted popular terms

```
Viral pneumonia <= 38
|   Shivering <= 24: 0 (72.0/1.0)
|   Shivering > 24
|   |   Amantadine <= 25: 0 (10.0)
|   |   Amantadine > 25
|   |   |   Human flu <= 8: 1 (4.0)
|   |   |   Human flu > 8: 0 (2.0)
Viral pneumonia > 38
|   Nasal congestion <= 27: 0 (10.0)
|   Nasal congestion > 27
|   |   Influenza-like illness <= 24
|   |   |   Influenza vaccine <= 17
|   |   |   |   Malaise <= 89
|   |   |   |   |   Flu season <= 11
|   |   |   |   |   |   Human flu <= 6: 0 (4.0)
|   |   |   |   |   |   Human flu > 6
|   |   |   |   |   |   |   Zanamivir <= 2: 1 (13.0)
|   |   |   |   |   |   |   Zanamivir > 2
|   |   |   |   |   |   |   |   Malaise <= 41: 1 (6.0/1.0)
|   |   |   |   |   |   |   |   Malaise > 41: 0 (6.0)
|   |   |   |   |   Flu season > 11: 1 (15.0)
|   |   |   |   Malaise > 89: 0 (4.0)
|   |   |   Influenza vaccine > 17: 0 (14.0/2.0)
|   |   Influenza-like illness > 24: 1 (8.0)
```

Figure 2: Decision tree prediction, best-selected features

Accuracy	0.9762
Precision	0.976
Recall	0.976
F1	0.976
MCC	0.941
TP rate	0.976
FP rate	0.047
AUC	0.989

Figure 3: Decision tree prediction, overall performance metrics.

5 Conclusion

In this paper, our main goal was to produce a dataset of user search queries that can best predict epidemic diseases. We focused on *Flu* disease and used our model to classify *Flu* volumes as breakout or not based on an initial classification proposed by the US Center for Disease Control and Prevention (CDC). We proposed a model to evaluate the monthly trigger in Flu volume of cases from non-Epidemic to Epidemic. Such a model can be used for example, for future disaster management and planning. We compared the prediction performance of our created dataset with an earlier one. Using the same settings, our selected dataset of popular user search terms indicates better prediction performance with the Flu Epidemic prediction model.

References

[1] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, Reid Priedhorsky, Global Disease Monitoring and Forecasting with Wikipedia, Published: November 13, 2014, https://doi.org/10.1371/journal.pcbi.1003892

[2] Wiemken TL, Furmanek SP, Mattingly WA, Wright MO, Persaud AK, Guinn BE, Carrico RM, Arnold FW, Ramirez JA. Methods for computational disease surveillance in infection prevention and control. Am J Infect Control. 2018;46 (2):124-132. doi: 10.1016/j.ajic.2017.08.005.

[4] Google Trends: https://trends.google.com/trends/

[5] M. Saqib Nawaz, Raza Ul Mustafa, M. Ikram Ullah Lali. Role of Online Data from Search Engine and Social Media in Healthcare Informatics. 2018, IGI Global. DOI: 10.4018/978-1-5225-2607-0.ch011

[7] Grishman, R., Huttunen, S., & Yangarber, R. (2002). Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35(4), 236–246.

[9] Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter catches the flu: Detecting Influenza epidemics using. Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1568-1576).

[11] Carneiro, H. A., & Mylonakis, E. (2009). Google Trends. A web-based tool for real-time surveillance of disease outbreaks. Clinical Infectious Diseases, 49(10), 15557–15564. doi:10.1086/630200 PMID:19845471

[13] Finding Healthcare Issues with Search Engine Queries and Social Network Data International journal on Semantic Web and information systems (2017). Vol.13, No.1, 2017 DOI: 10.4018/IJSWIS.2017010104.
[15] Xinyue Ye, Shengwen Li, Xining Yang and Chenglin Qin. Use of Social Media for the Detection and Analysis of Infectious Diseases in China. ISPRS Int. J. Geo-Inf. 2016, 5, 156; doi:10.3390/ijgi5090156
[17] M. Aitken, T Altmann, and D. Rosen (2014) Engaging patients through social media. IMS Institute for healthcare informatics. 2014.
[18] Wang L, Wang J, Wang M, Li Y, Liang Y, Xu D. Using Internet search engines to obtain medical information: a comparative study. J Med Internet Res. 2012; 14 (3):e74. doi:10.2196/jmir.1943
[19] S.V. Nuti, B. Wayda, I. Ranasinghe, S. Wang, R.P. Dreyer, et al. (2014). The Use of Google Trends in Health Care Research: A Systematic Review. PLOS ONE 9(10):e109583.
[20] Hassid, B.G., Day, L.W., Awad, M.A. et al. (2017). Using Search Engine Query Data to Explore the Epidemiology of Common Gastrointestinal Symptoms. Digestive Diseases and Sciences journal; vol. 62, issue 3, 2017, pp.588-592. https://doi.org/10.1007/s10620-016-4384-y
[21] M. Radin, and S Sciascia. Infodemiology of Systemic Lupus Erythematous Using Google Trends. Lupus 26, no. 8. 2017, pp.886–89. doi:10.1177/0961203317691372.
[22] Centers for Disease Control and Prevention (CDC), USA, *retrieved August 2018.* https://www.cdc.gov/flu/weekly/usmap.htm
[23] Wang J, Zhang T, Lu Y, Zhou G, Chen Q, Niu B (2018). Vesicular stomatitis forecasting based on Google Trends. PLoS ONE 13(1): e0192141. https://doi.org/10.1371/journal.pone.0192141