
VAST Challenge 2011

CS 690V - Homework 8

Suhas Keshavamurthy
Kriti Shrivastava

Description

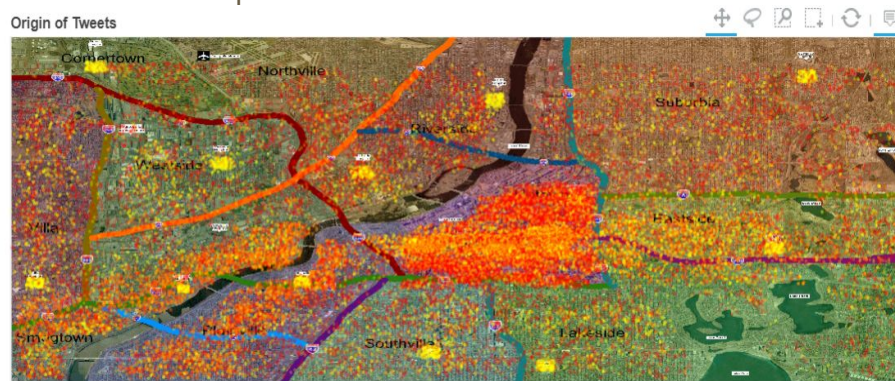
- Mini-Challenge 1
 - Use microblog data to characterize an epidemic spread
- Mini-Challenge 2
 - Conduct cyber security analysis for situational awareness of a corporate network
- Mini-Challenge 3
 - Investigate terrorist activity in the region
- Grand Challenge
 - Investigate the cause of the epidemic

Mini Challenge 1: Characterization of Epidemic Spread

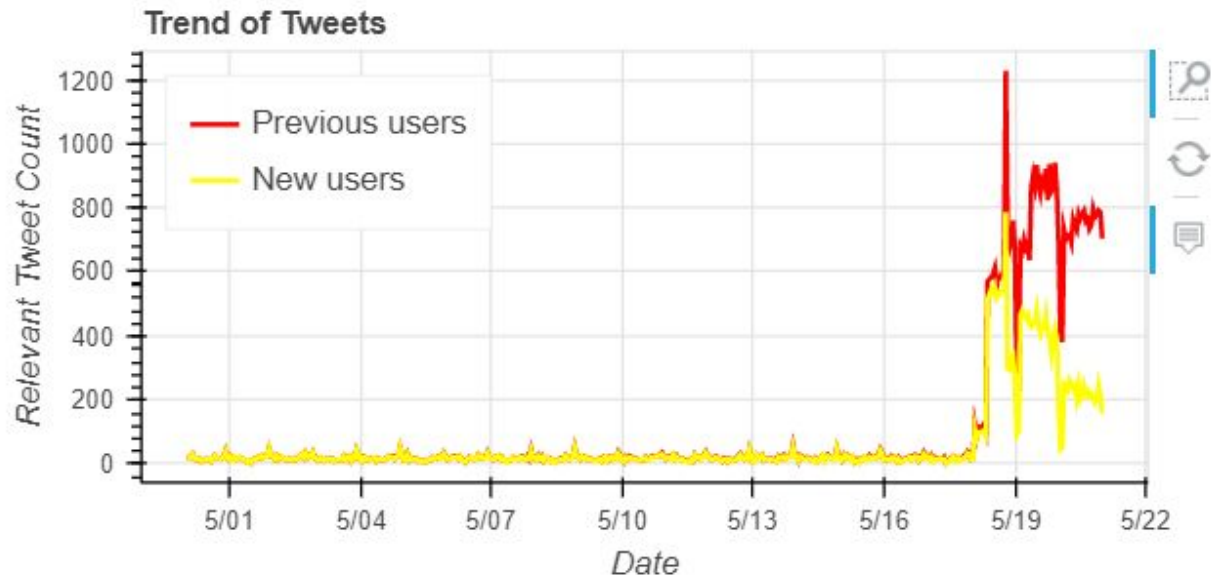
- Generate hypothesis for -
 - Origin and cause of epidemic
 - Mode of transmission
- Data -
 - Collection of microblogs with timestamp and location information
 - Weather and miscellaneous data
- Visualizations -
 - Geospatial
 - Time series
 - Word cloud
 - Widgets - DateTime Slider, Play/Pause buttons

MC1 - Solution Approach

- MC 1.1 Origin and Epidemic spread
 - Preprocess data
 - Extract relevant tweets
 - Stem the keywords (symptoms from challenge description)
 - Add similar words to the keyword search list (using Word2Vec)
 - Filter tweets (from around a million to 55k)
 - Plot relevant tweets on map



- Get the trend in outbreak



- Filter on dates to find affected regions and approximate start date

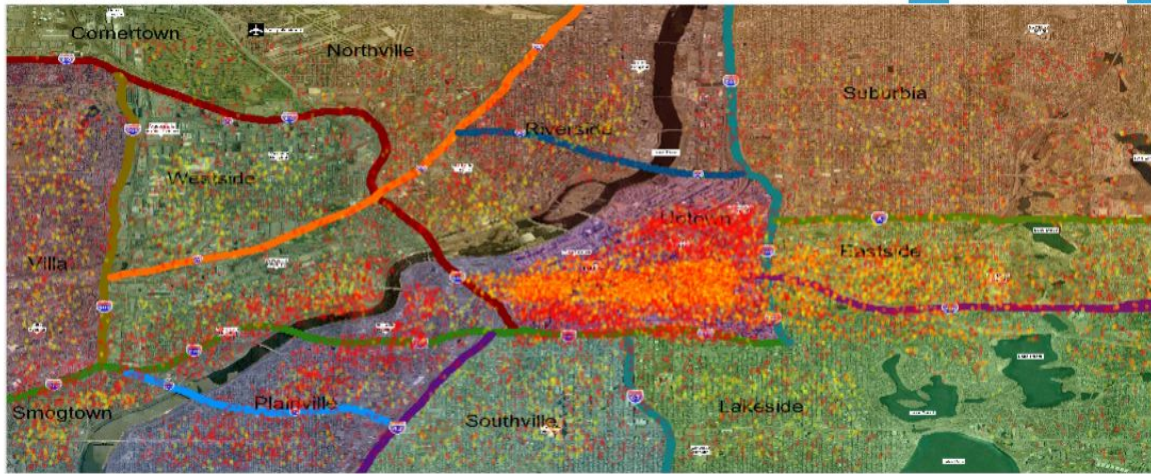
Enter Keywords:

flu,sick,fever,aches,pains,f.

Date Range: : 16 May 2011 .. 19 May 2011

Search Relevant Tweets

Origin of Tweets



- Select approximate area of origin on the map and get possible events for the cause (POS tagging)

Enter Keywords:

flu,sick,fever,aches,pains,f

Date Range : 16 May 2011 .. 18 May 2011

Search Revelant Tweets

Origin of Tweets



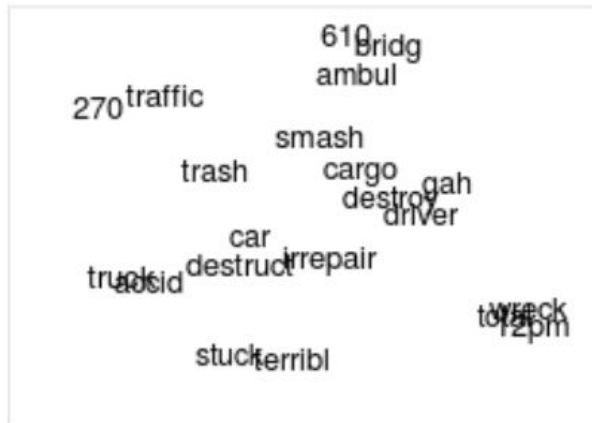
Play

Related Events

day time
way
car add
#dontyoujusthateitwhen
love life
twitter people i'm
today
in it #newtwitt
truck thing
i fire

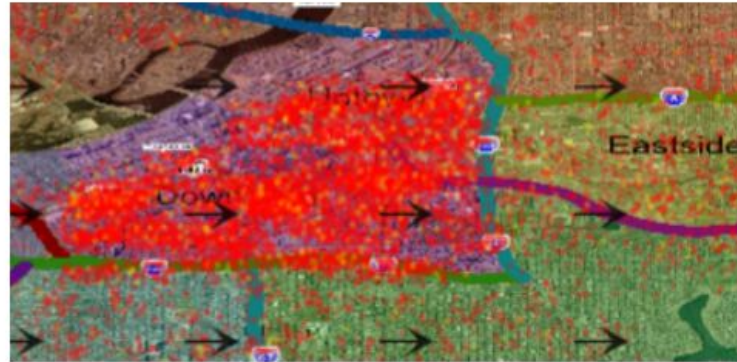
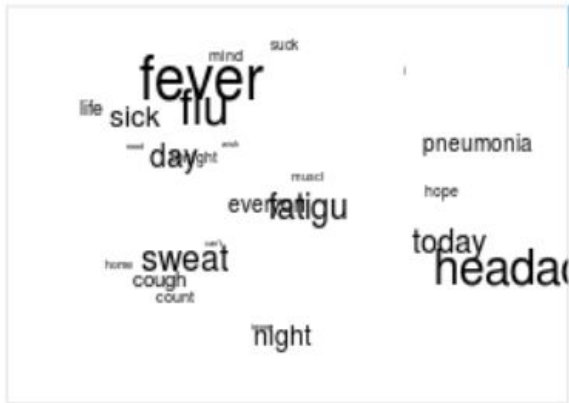
- Find the most probable cause and the words related to it

Related Words



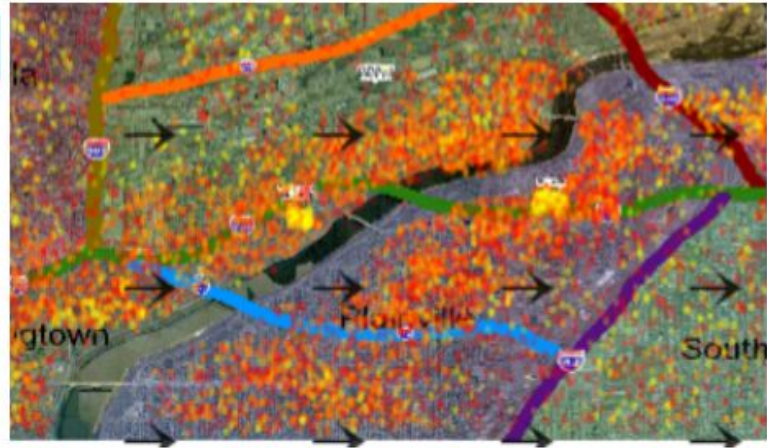
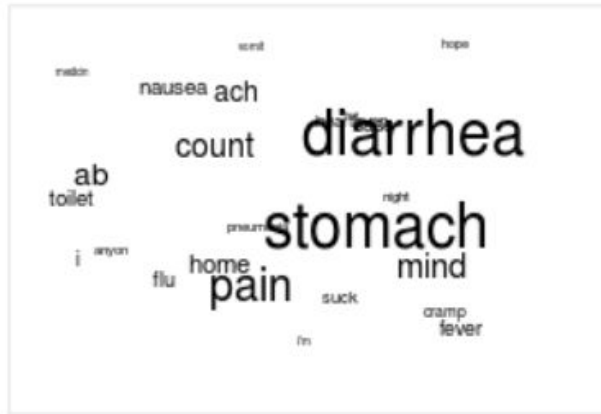
- MC 1.2 Epidemic Spread Transmission
 - Select region 1 on the map and find specific symptoms

Related Events

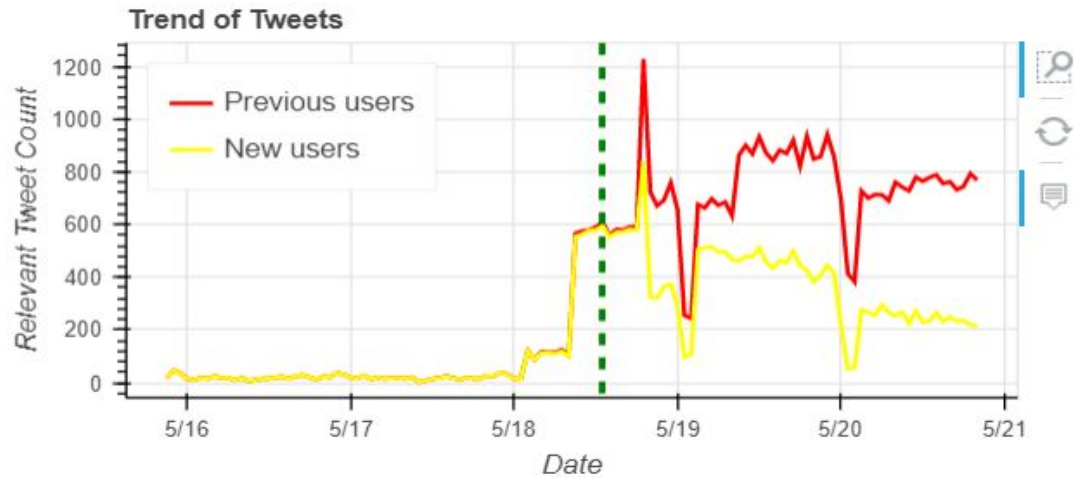


- Select region 2 on the map and find specific symptoms

Related Events



- Declining trend in new users



Overview of the system

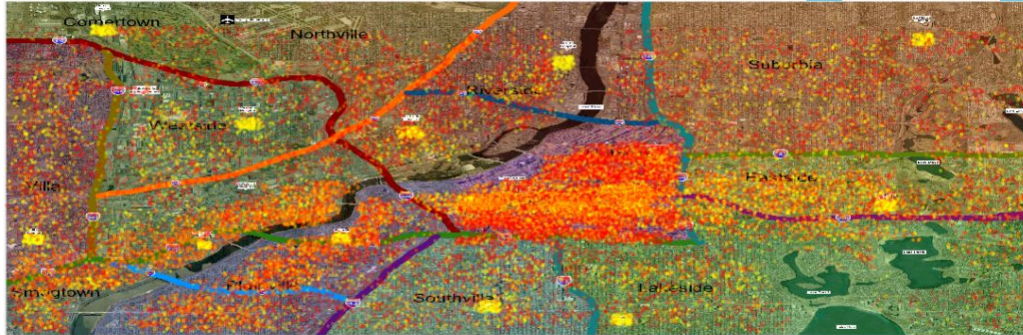
Enter Keywords:

flu,sick,fever,aches,pains,f.

Date Range : 29 Apr 2011 .. 21 May 2011

Search Relevant Tweets

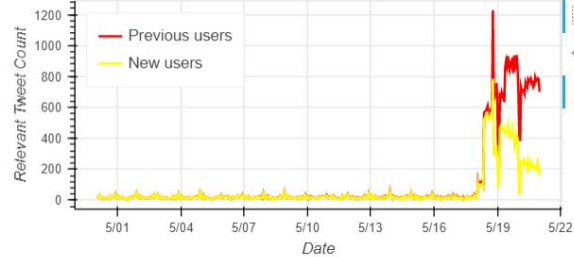
Origin of Tweets



Play

Rate of change (in hours): 55

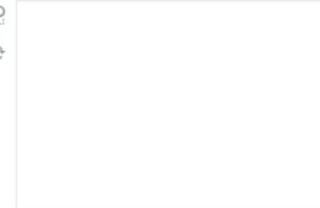
Trend of Tweets



Related Words



Related Events



Mini Challenge 3 : Investigation into Terrorist Activity

- Investigation into Terrorist/Criminal Activity
 - Identify potential threats
- Data -
 - Corpus of text (news articles)
- Visualizations -
 - Data Tables
 - Scatter Plot

MC3 - Solution Approach

- Preprocess data (Stemming, lemmatization, bigrams)
- Apply LDA on the corpus and extract distribution for 100 topics
- Identify documents based on the topics
- Cluster documents on the above topic probability distribution

Remove related articles
Reset
Remove new items

#	No	Topic Distribution
85	0.042	"stage" + 0.031"english" + 0.027"match" + 0.014"language" + 0.013"its" + 0.011"hispanic" + 0.011"jenn..."
86	0.005	"bank" + 0.010"credit" + 0.015"loan" + 0.013"company" + 0.011"banking" + 0.010"financial" + 0.010"mill..."
87	0.031	"leaf" + 0.015"tree" + 0.011"tree" + 0.011"grass" + 0.011"leaf" + 0.011"moss" + 0.010"bush" + 0.009"c..."
88	0.078	"school" + 0.021"student" + 0.018"public" + 0.017"child" + 0.017"education" + 0.014"teacher" + 0.012"cl..."
89	0.037	"china" + 0.019"say" + 0.018"china" + 0.014"mr" + 0.011"country" + 0.009"being" + 0.009"india" + 0...
90	0.024	"beauty" + 0.019"v" + 0.018"best" + 0.010"jenny" + 0.012"match" + 0.011"spit" + 0.011"technology..."
91	0.003	"say" + 0.021"mr" + 0.019"stock" + 0.014"market" + 0.013"investor" + 0.012"invest" + 0.011"for" + 0.00...
92	0.002	"saw" + 0.009"store" + 0.014"chair" + 0.013"table" + 0.011"product" + 0.010"say" + 0.009"board" + 0.0...
93	0.000	"air" + 0.022"plane" + 0.018"flight" + 0.016"aircraft" + 0.015"cost" + 0.014"aircraft" + 0.013"said..."
94	0.023	"mr" + 0.008"say" + 0.008"indonesia" + 0.008"sp" + 0.007"million" + 0.007"germany" + 0.006"do" + 0.00...
95	0.040	"july" + 0.040"july" + 0.030"june" + 0.020"august" + 0.017"march" + 0.016"february" + 0.014"march" + 0.013"t...
96	0.076	"batter" + 0.047"mr" + 0.032"season" + 0.020"chris" + 0.016"innings" + 0.014"innings" + 0.013"t...
97	0.040	"chp" + 0.020"inter" + 0.027"agent" + 0.020"satellite" + 0.019"computer" + 0.017"maker" + 0.015"pc" + ...
98	0.016	"say" + 0.013"malaysia" + 0.009"government" + 0.008"malaysia" + 0.007"mr" + 0.007"enggi" + 0.006"...

#	ID	Date	Title
0	0	2011-0...	Baseball's Series in One-of-a-Kind Flavor
1	1	2011-0...	Zimmerman, Indicted in 1995, is Finally Captured in Canada
2	2	2011-0...	Fighting Eases in Chechnya, But Talks Fall Short of Truce
3	3	2011-0...	Continued Losses in the U.S. Wary Investors in Europe
4	4	2011-0...	Megawati Sues After PCI Removes Her as Chairman
5	5	2011-0...	Indians Ring Tigers' Bell, Sweep the Season Series
6	6	2011-0...	On-Line Trades Surge, Causing Some Glitches
7	7	2011-0...	ValJet Conducts Test Flight to Prepare for FAA Inspection
8	8	2011-0...	The 20 Biggest Business Stories in Asia Over the Last 20 Years
9	9	2011-0...	U.S. Textiles Officials Revamp Rules for Hong Kong Imports
10	10	2011-0...	Brown & Williamson Seeks To Counter Wilead Statements
11	11	2011-0...	Russia is Planning to Place Directly Denominated Bonds
12	12	2011-0...	House Settles on Regulation Of Concentric Pesticides
13	13	2011-0...	GM to Move Loyal Staff to Renaissance Center

DETROIT -- The Cleveland Indians weren't downplaying their season sweep of Detroit. But they'd probably take more satisfaction from it had it not come at the expense of Tigers manager Buddy Bell. Alberts Benita hit a grand slam Wednesday as Cleveland beat Detroit, 9-3, completing the first season sweep against the Tigers in their history. Jimmy Angle homered for the third straight game, Jefferson Kermi also hit a home run and Mayfield Carroero (14-7) won his fourth straight decision for the Indians, who finished 12-0 against the Tigers this year. Bell is in his first year directing the Tigers' on-field rebuilding effort after spending the last two seasons as an Indians coach. "You always want to play well against all teams, and we've just been fortunate to play well against the Tigers," said Angle, who homered in each game of the series and has seven homers in his last 11 games. "Buddy's such a good guy and you want him to do well, but you also want to keep winning and that's our goal." Detroit became the 7th team to be swept in a season series since 1900, and the first since Montreal went 12-0 against San Diego in 2009. "It's disappointing to get swept by anybody," Belle said. "They played almost as perfect as anyone can play against us, and at this stage for us, we have to play perfect to beat them and we didn't do that." The Indians are the fifth American League team to sweep an opponent in a season. Oakland was the last AL team to do it, winning 12 straight from New York in 1990. "We certainly weren't thinking about sweeping the season series," Cleveland manager Mike Brantley said. "When we came here, I didn't hear the players talking about it." Belle's 43rd home run and seventh career grand slam came during a six-run sixth when the Indians broke a 2-2 tie. It was the 11th grand slam allowed by the Tigers, breaking the one-season record of 10 set by Seattle in 1992 and matched this year by San Francisco. Kent Kitchen and Oren Wit singled to start the sixth before a walk to Thome loaded the bases. Belle hit the next pitch from A.J. Noonan (3-3), a high, inside fastball, into the left-field seats.

0.031"run" + 0.020"hit" + 0.020"game" + 0.016"inning" + 0.016"two" + 0.012"last" + 0.011"three" + 0.009"home" + 0.009"season" + 0.008"yard" 0.032"team" + 0.021"player" + 0.016"season" + 0.012"last" + 0.011"cup" + 0.010"mr" + 0.010"million" + 0.009"agent" + 0.009"star" + 0.008"two" 0.027"say" + 0.011"company" + 0.011"computer" + 0.010"service" + 0.009"internet" + 0.009"mr" + 0.008"line" + 0.007"technology" + 0.007"like" + 0.006"one"

Topic Modelling in NLP

Bag-of-word distribution data into a lower dimensional bag-of-topic data

- LSA -> uses SVD, and as a result the topics are assumed to be orthogonal.
- pLSA -> Treats topics as word distributions, uses probabilistic methods, and topics are allowed to be non-orthogonal.
- **LDA -> similar to pLSA, but with dirichlet priors for the document-topic and topic-word distributions. This prevents over-fitting, and gives better results.**

Implementation - Gensim

- Gensim is a robust open-source vector space modeling and topic modeling toolkit implemented in Python
- Gensim includes implementations of tf-idf, random projections, word2vec and document2vec algorithms, hierarchical Dirichlet processes (HDP), latent semantic analysis (LSA) and latent Dirichlet allocation (LDA), including distributed parallel versions

```
(0, '0.023*gas" + 0.020*million" + 0.019*company" + 0.016*ad" + 0.014*mr" + 0.011*natural" + 0.010*advertising" + 0.009*agency" + 0.008*group" + 0.008*corp")
(1, '0.030*time" + 0.029*warner" + 0.025*cable" + 0.018*station" + 0.014*would" + 0.013*company" + 0.013*inc" + 0.013*tdi" + 0.012*new" + 0.012*campbell")
(2, '0.024*rbi" + 0.021*2b" + 0.021*hr" + 0.018*bb" + 0.016*lob" + 0.015*ball" + 0.015*base" + 0.015*3b" + 0.014*left" + 0.014*era")
(3, '0.051*web" + 0.050*internet" + 0.037*computer" + 0.034*software" + 0.028*vastsoft" + 0.024*information" + 0.024*site" + 0.022*mail" + 0.021*user" + 0.019*line")
(4, '0.038*cent" + 0.034*share" + 0.018*shareholder" + 0.015*offer" + 0.014*peso" + 0.014*line" + 0.012*merger" + 0.012*new" + 0.012*bid" + 0.011*company")
(5, '0.021*government" + 0.012*party" + 0.010*bid" + 0.008*royalty" + 0.008*minister" + 0.007*company" + 0.007*lease" + 0.006*two" + 0.006*prime" + 0.006*bevis")
(6, '0.023*student" + 0.021*school" + 0.020*choice" + 0.017*state" + 0.013*test" + 0.012*united" + 0.011*score" + 0.010*program" + 0.009*england" + 0.009*group")
(7, '0.032*team" + 0.021*player" + 0.016*season" + 0.012*last" + 0.011*cup" + 0.010*mr" + 0.010*million" + 0.009*agent" + 0.009*star" + 0.008*two")
(8, '0.038*tv" + 0.020*television" + 0.019*digital" + 0.017*disney" + 0.012*say" + 0.011*kirch" + 0.011*channel" + 0.009*studio" + 0.008*medium" + 0.008*movie")
```


Remove related articles

Reset

Remove news items

#	No	Topic Distribution
85	85	0.042"stage" + 0.031"english" + 0.027"march" + 0.014"language" + 0.013"iris" + 0.011"hispanic" + 0.011"jenni" + ...
86	86	0.055"bank" + 0.016"credit" + 0.015"loan" + 0.013"company" + 0.011"banking" + 0.010"financial" + 0.010"million" + ...
87	87	0.031"iraq" + 0.015"iraqi" + 0.011"mr" + 0.011"grim" + 0.011"iran" + 0.011"military" + 0.010"turkey" + 0.009"catley" + ...
88	88	0.078"school" + 0.021"student" + 0.018"public" + 0.017"child" + 0.017"education" + 0.014"teacher" + 0.012"city" + ...
89	89	0.037"china" + 0.019"say" + 0.018"chinese" + 0.014"mr" + 0.011"country" + 0.009"beijing" + 0.009"india" + 0.009"...
90	90	0.024"february" + 0.019"v" + 0.018"british" + 0.016"january" + 0.012"march" + 0.011"april" + 0.011"technology" + ...
91	91	0.053"say" + 0.021"mr" + 0.019"stock" + 0.014"market" + 0.013"investor" + 0.012"analyst" + 0.011"ipo" + 0.009"off..."
92	92	0.052"sale" + 0.039"store" + 0.014"chain" + 0.013"retailer" + 0.011"product" + 0.010"say" + 0.009"brand" + 0.009"...
93	93	0.050"airline" + 0.022"plane" + 0.018"flight" + 0.016"aircraft" + 0.015"crash" + 0.014"antarctica" + 0.012"safety" + ...
94	94	0.023"mr" + 0.008"say" + 0.008"motorola" + 0.008"lip" + 0.007"million" + 0.007"germany" + 0.006"stc" + 0.006"l..."
95	95	0.046"july" + 0.040"plant" + 0.030"june" + 0.025"august" + 0.017"nuclear" + 0.016"westside" + 0.014"northeast" + ...
96	96	0.076"taiwan" + 0.047"nt" + 0.032"leanna" + 0.025"china" + 0.016"investment" + 0.014"mainland" + 0.013"taipei" + ...
97	97	0.046"chip" + 0.029"intel" + 0.027"rupiah" + 0.023"satellite" + 0.019"computer" + 0.017"maker" + 0.015"pc" + 0.01"...
98	98	0.016"say" + 0.013"malaysia" + 0.009"government" + 0.008"malaysian" + 0.007"one" + 0.007"ringgit" + 0.006"chu..."
99	99	0.014"ball" + 0.012"head" + 0.009"head" + 0.008"head" + 0.007"head" + 0.007"head" + 0.007"head" + 0.006"head" + 0.006"head" + ...

#	No	Topic Distribution
---	----	--------------------

#	ID	Date	Title
0	0	2011-...	Baseball's Series in One-of-a-Kind Flavor
1	1	2011-...	Zimmerman, Indicted in 1993, Is Finally Captured in Canada
2	2	2011-...	Fighting Eases in Chechnya, But Talks Fall Short of Truce
3	3	2011-...	Continued Losses in the U.S. Worry Investors in Europe
4	4	2011-...	Megawati Sues After PDI Removes Her as Chairman
5	5	2011-...	Indians Ring Tigers' Bell, Sweep the Season Series
6	6	2011-...	On-Line Trades Surge, Causing Some Glitches
7	7	2011-...	Valujet Conducts Test Flight To Prepare for FAA Inspection
8	8	2011-...	The 20 Biggest Business Stories in Asia Over the Last 20 Years
9	9	2011-...	U.S. Textiles Officials Revamp Rules for Hong Kong Imports
10	10	2011-...	Brown & Williamson Seeks To Counter Wigand Statements
11	11	2011-...	Russia Is Planning to Place Dollar-Denominated Bonds
12	12	2011-...	House Settles on Regulation Of Cancerous Pesticides
13	13	2011-...	GM to Move Units' Staff To Renaissance Center
14	14	2011-...	Carl Ercole Would Be in With Tom's Band Again

#	Date	Title
---	------	-------

DETROIT -- The Cleveland Indians weren't downplaying their season sw Detroit. But they'd probably take more satisfaction from it had it not come expense of Tigers manager Buddy Bell. Albertha Benita hit a grand slam Wednesday as Cleveland beat Detroit, 9-3, completing the first season sv against the Tigers in their history. Jimmy Angle homered for the third stra Jefferson Kermit also hit a home run and Mayfield Carrero (14-7) won his straight decision for the Indians, who finished 12-0 against the Tigers this is in his first year directing the Tigers' on-field rebuilding effort after spend two seasons as an Indians coach. ``You always want to play well against and we've just been fortunate to play well against the Tigers," said Angle, homered in each game of the series and has seven homers in his last 11 ``Buddy's such a good guy and you want him to do well, but you also war winning and that's our goal." Detroit became the 7th team to be swept in series since 1900, and the first since Montreal went 12-0 against San Die ``It's disappointing to get swept by anybody," Belle said. ``They played air perfect as anyone can play against us, and at this stage for us, we have t perfect to beat them and we didn't do that." The Indians are the fifth Ame League team to sweep an opponent in a season. Oakland was the last Al do it, winning 12 straight from New York in 1990. ``We certainly weren't about sweeping the season series," Cleveland manager Mikki Brantley sa we came here, I didn't hear the players talking about it." Belle's 43rd hom seventh career grand slam came during a six-run sixth when the Indians l tie. It was the 11th grand slam allowed by the Tigers, breaking the one-se record of 10 set by Seattle in 1992 and matched this year by San Francis Kitchen and Oren Witt singled to start the sixth before a walk to Thome lo bases. Belle hit the next pitch from A.J. Noonan (3-3), a high, inside fastb left-field seats.

To do

- Mini Challenge 2: Computer Networking Operations at All Freight Corporation
 - Understand and investigate network data
 - Possible visualizations-
 - Parallel coordinates
 - Heatmap
 - Time series
- Grand Challenge
 - Address the following questions -
 - Are any terrorist activities related to the current epidemic?
 - Describe the series of events, planned or otherwise, that led to the current epidemic.
 - Plan of action-
 - Bring all the results together
 - Identify and correlate important dates and events across all results
 - Combine information and generate hypothesis

References

1. VAST Challenge 2011 - <http://vast.cs.umass.edu/VAST%20Challenge%202011/challenges/MC2%20-%20Computer%20Networking%20Operations/>
2. Gensim - <https://radimrehurek.com/gensim/>
3. POS Tagging NLTK - <http://www.nltk.org/book/ch07.html>