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

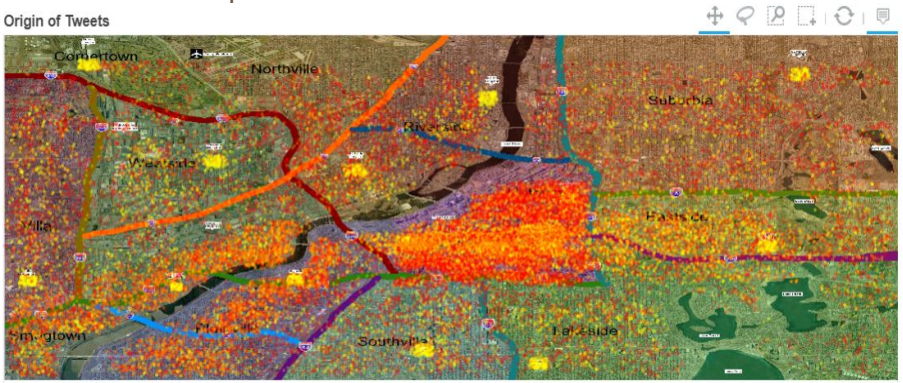
- Generate hypothesis for -
  - Origin and cause of epidemic
  - Mode of transmission

- Data -
  - Collection of microblogs with timestamp and location information
  - Weather and miscellaneous data

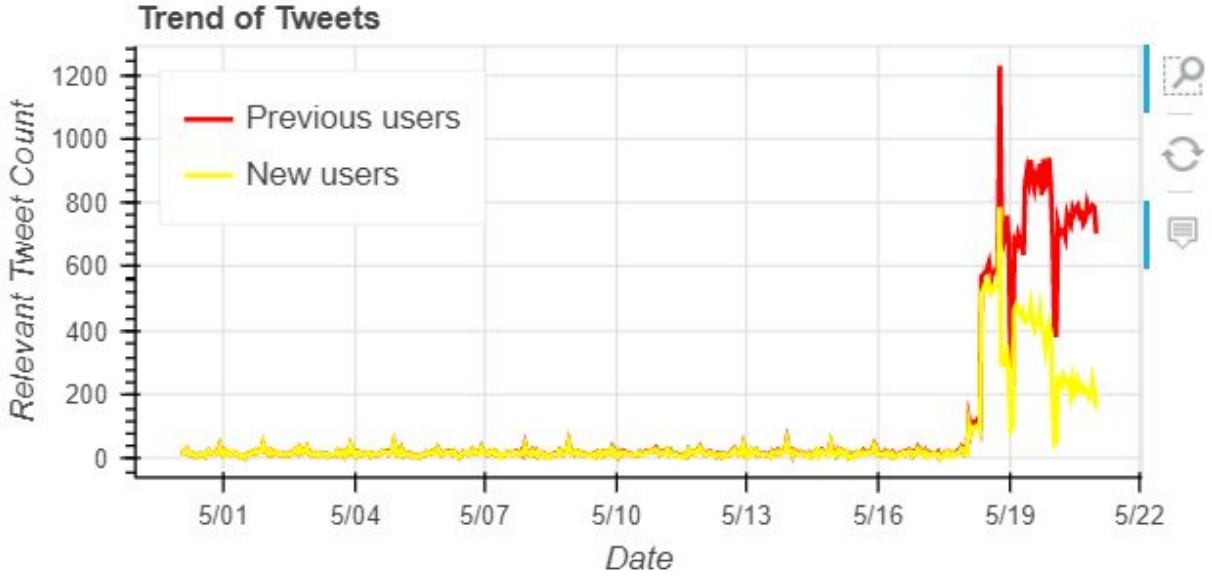
- Visualizations -
  - Geospatial
  - Time series
  - Word cloud
  - Widgets - DateTime Slider, Play/Pause buttons

MC1 - Solution Approach

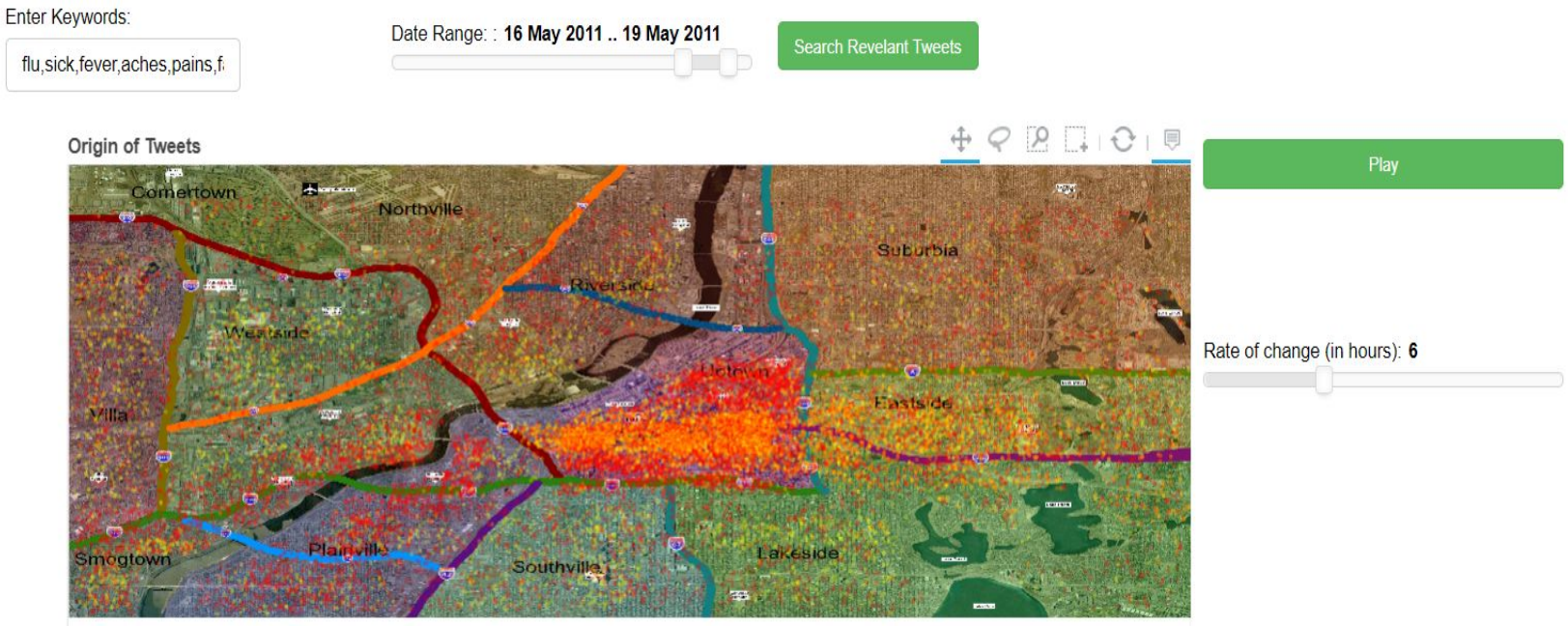
- MC 1.1 Origin and Epidemic spread
  - Preprocess data
  - Extract relevant tweets
    - Stem the keywords (symptoms from challenge description)
    - Add similar words to the keyword search list (using Word2Vec)
    - Filter tweets (from around a million to 55k)
  - Plot relevant tweets on map



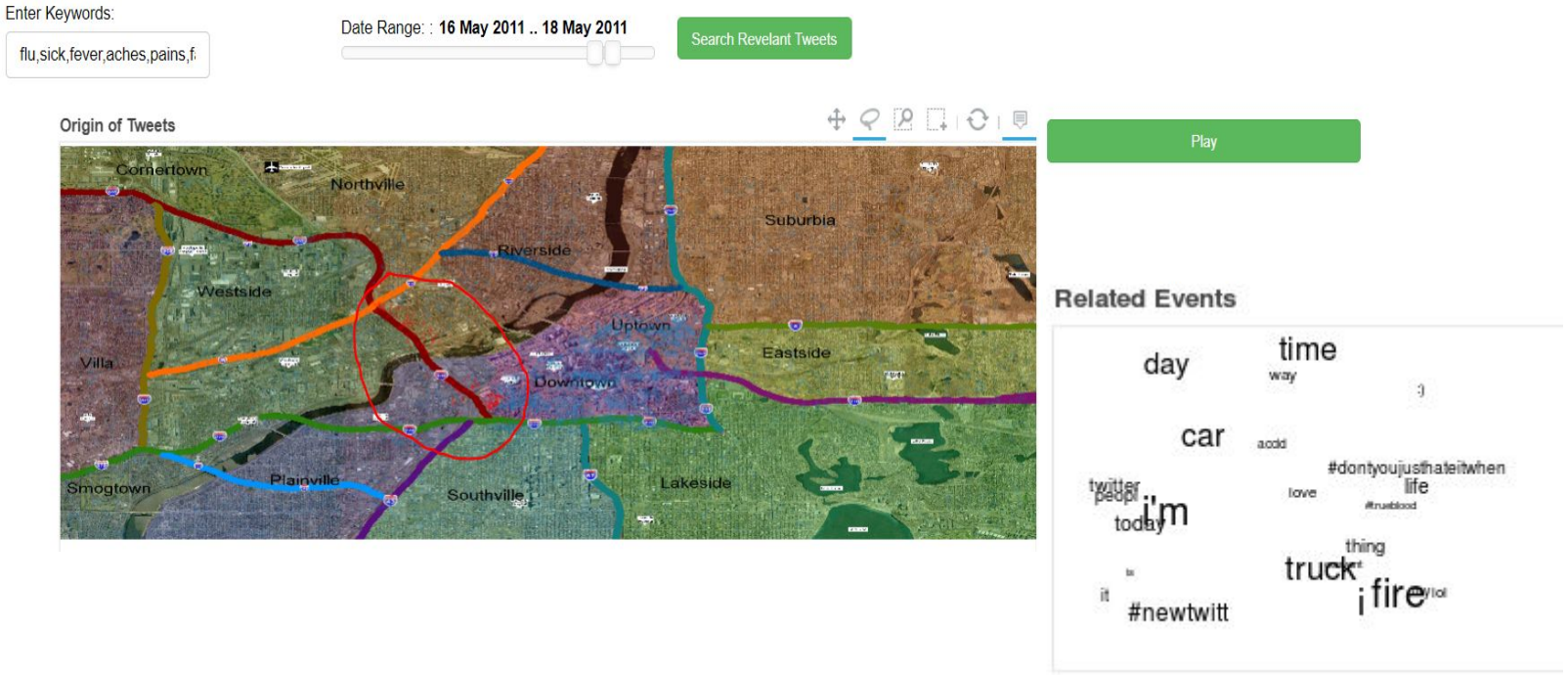
- Get the trend in outbreak



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

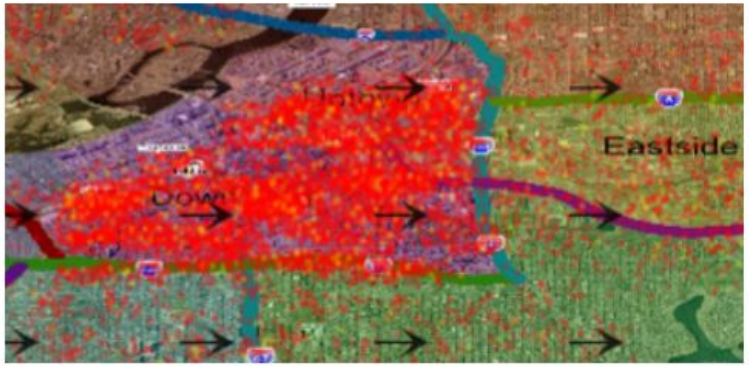


- Select approximate area of origin on the map and get possible events for the cause (POS tagging)



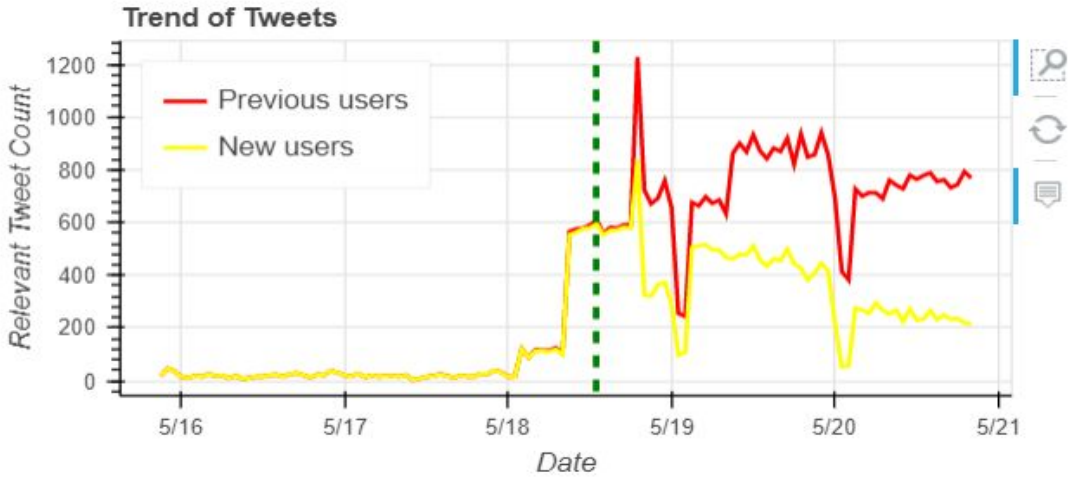
- Find the most probable cause and the words related to it

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

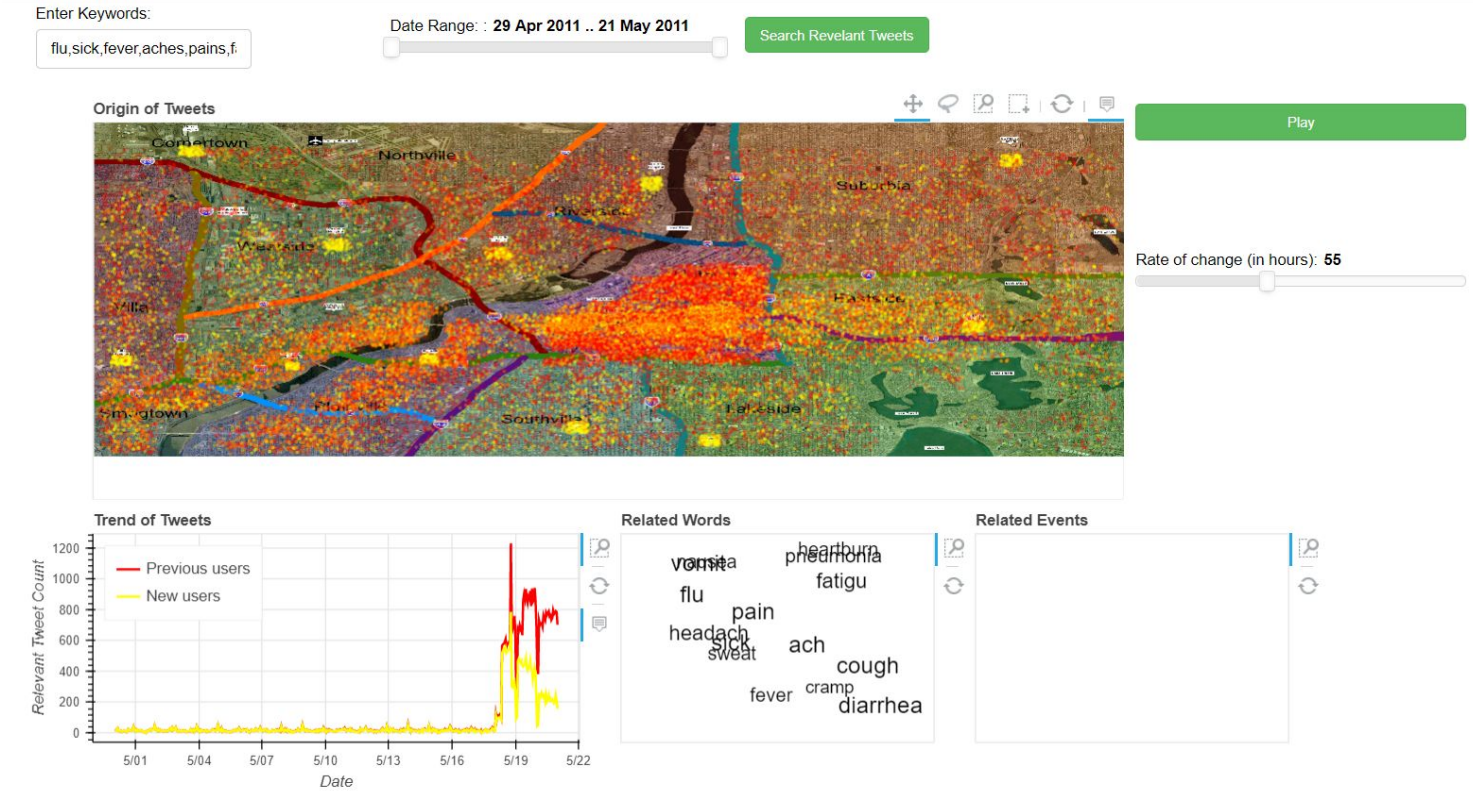


- Select region 2 on the map and find specific symptoms

- Declining trend in new users



Overview of the system

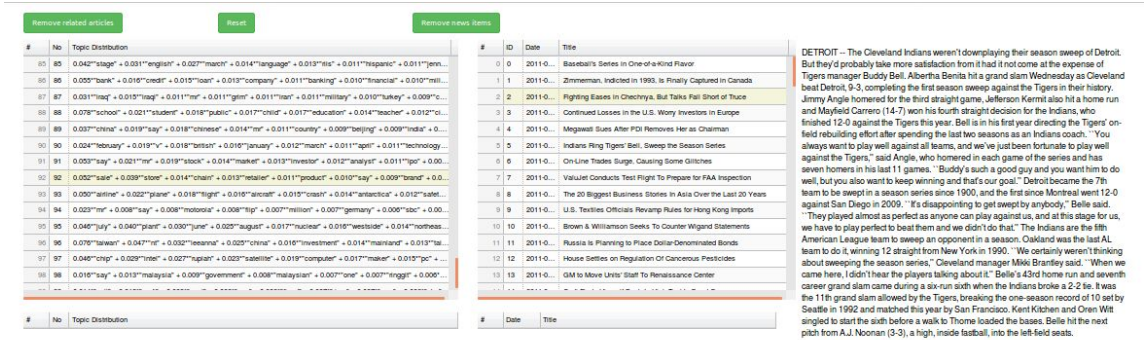


Mini Challenge 3 : Investigation into Terrorist Activity

- 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



0.031*"run" + 0.020*"hit" + 0.020*"game" + 0.016*"inning" + 0.016*"two" + 0.012*"first" + 0.011*"three" + 0.009*"homer" + 0.009*"season" + 0.008*"yard" 0.032*"team" + 0.021*"player" + 0.016*"season" + 0.012*"last" + 0.011*"cup" + 0.010*"m" + 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

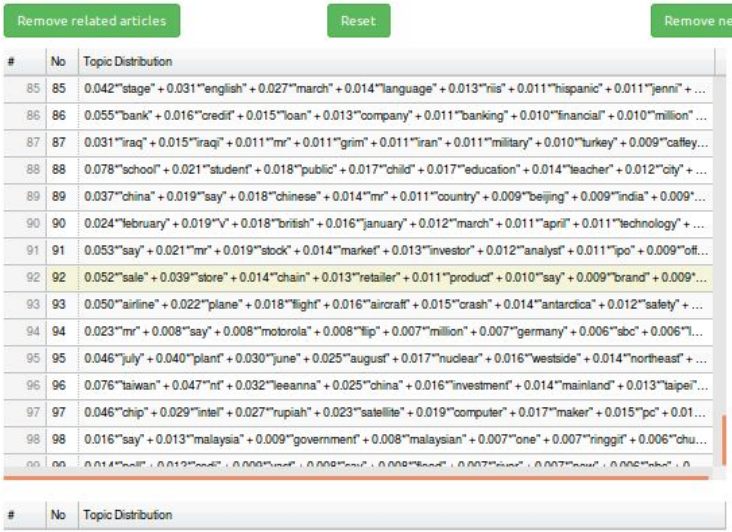
- LSA -> uses SVD, and as a result the topics are assumed to be orthogonal.
- pLSA -> Treats topics as word distributions, uses probabilistic methods, and topics are allowed to be non-orthogonal.
- **LDA -> similar to pLSA, but with dirichlet priors for the document-topic and topic-word distributions. 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*"tci" + 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"')
```



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 sw
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 alr
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 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
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