ANALYTICS PROJECT PRESENTATION - SUMMER 2015

APPLICATION OF MACHINE LEARNING OF FOR IMPROVED CRISIS FORECASTING

YARK Team : Ariel Dexler, Michael Rawson, Kania Azrina, Yixue Wang

This paper aims to leverage data mining for improved crisis forecasting. Using sequential patterning and neural networks to analyze news archives and financial data, we aim to create a prediction model for domestic and international crises.

MOTIVATION

Who are the users of this analytic?

- UNICEF Operation Center (OPSCEN)

Who will benefit from this analytic?

- OPSCEN Officer

Why is this analytic important?

- Forecasting event sequences
- Preparing humanitarian aid (money, water, and sanitation)
- Targeting humanitarian aid location
- Avoiding human bias

DATA SOURCES

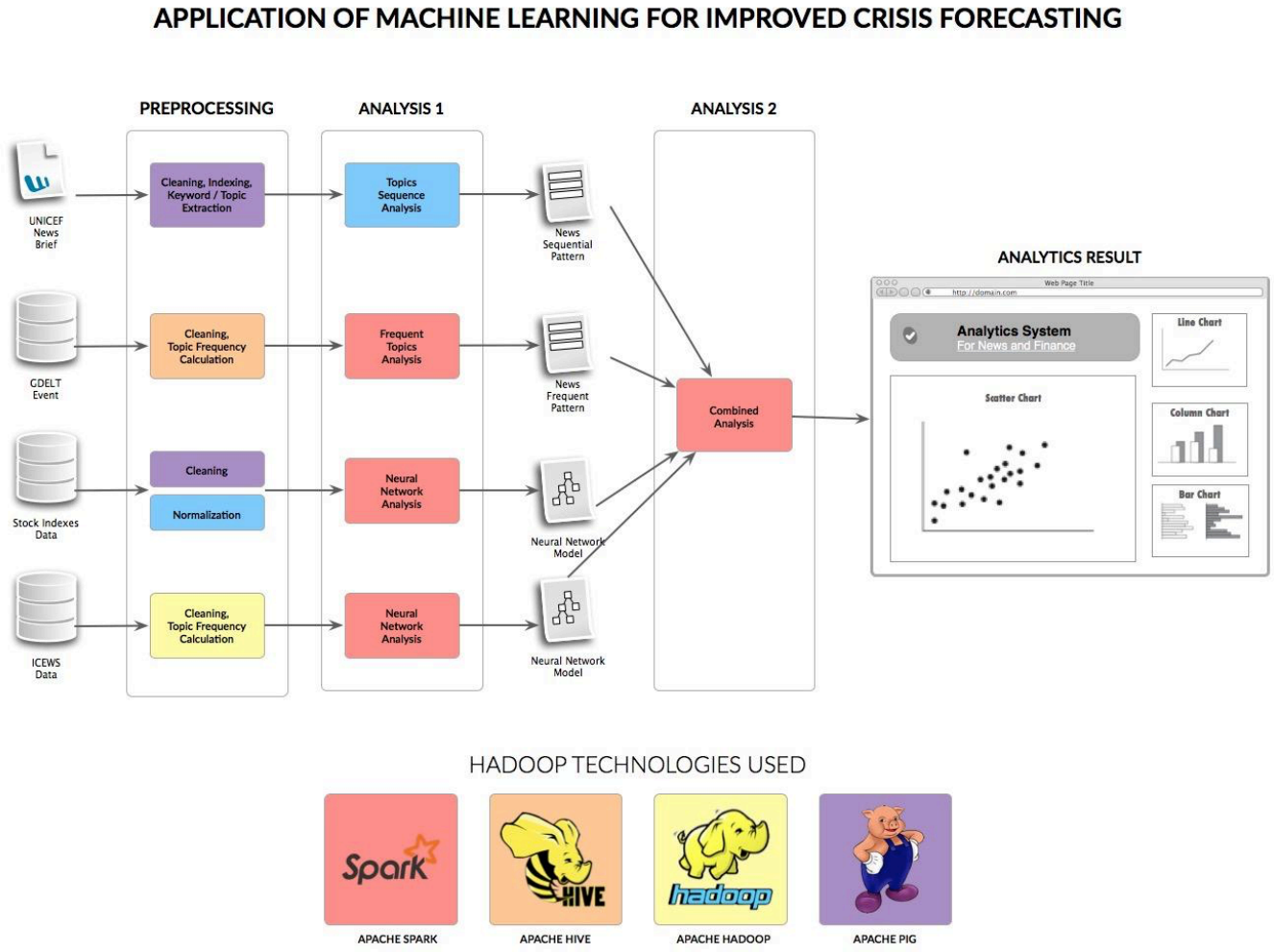
Name: UNICEF OPSCEN's News Brief 2004-2015
Description: Headline news from all over the world

Name: ICEWS Data 1995-2014
Description: Daily events coded ranked by scale of hostility vs. cooperation. Monthly indicators for events of interest per country.

Name: Stock Indexes/Economic Data 1990-2015
Description: Daily closing prices back to 1990

Name: GDELT Event Data 1979-2014
Description: Daily records including date, country, event code, latitude and longitude, etc.

DESIGN DIAGRAM



Platform(s) on which the analytic ran:

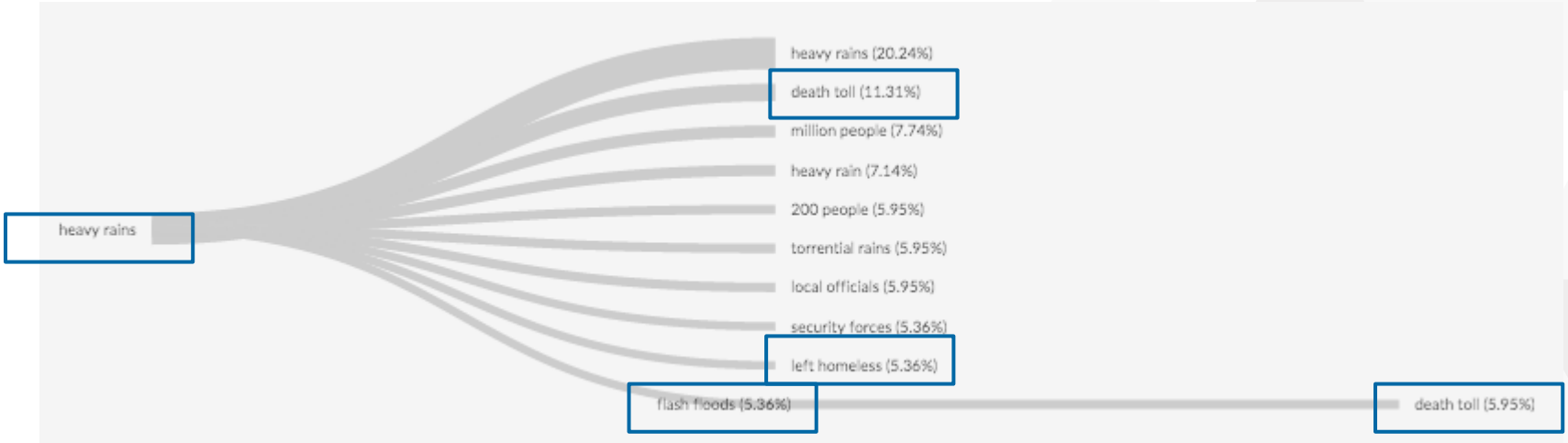
Quickstart VM, NYU HPC cluster, NYU CUSP cluster

VISIT OUR WEBSITE

http://bit.ly/crisisforecasting

RESULTS

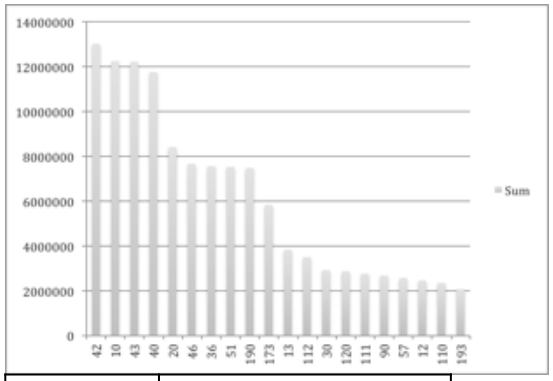
1. UNICEF News Sequence from 'Heavy Rain' Topic



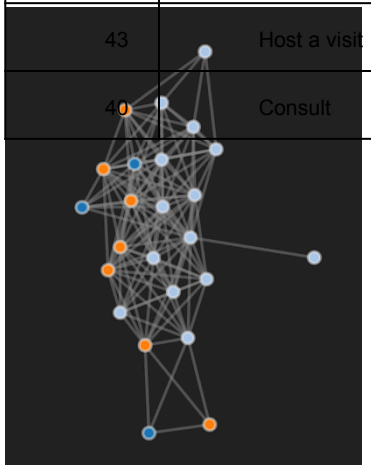
5.36% = 8/169 countries

RESULTS

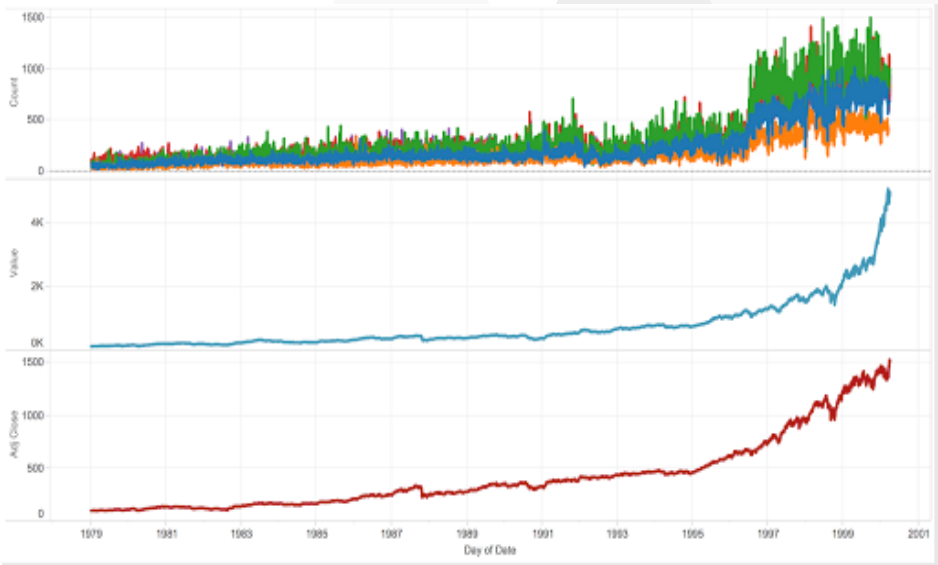
2. GDELT Frequent Item Analysis

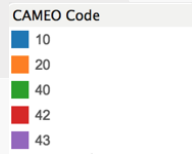


42	Make a visit
10	Make statement
43	Host a visit
40	Consult



Frequent Event vs. Stock Data

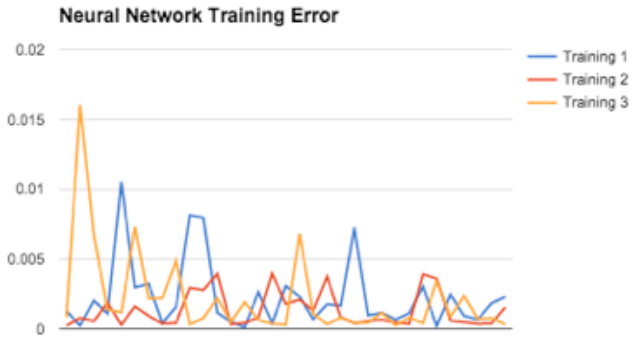


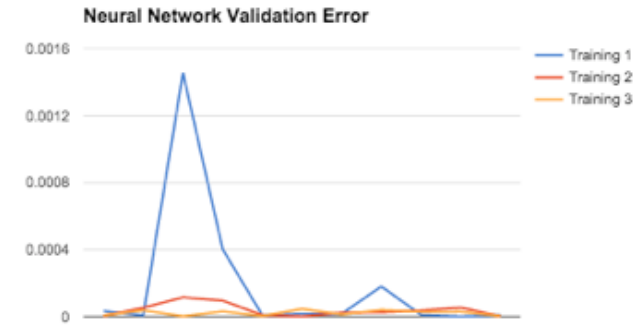


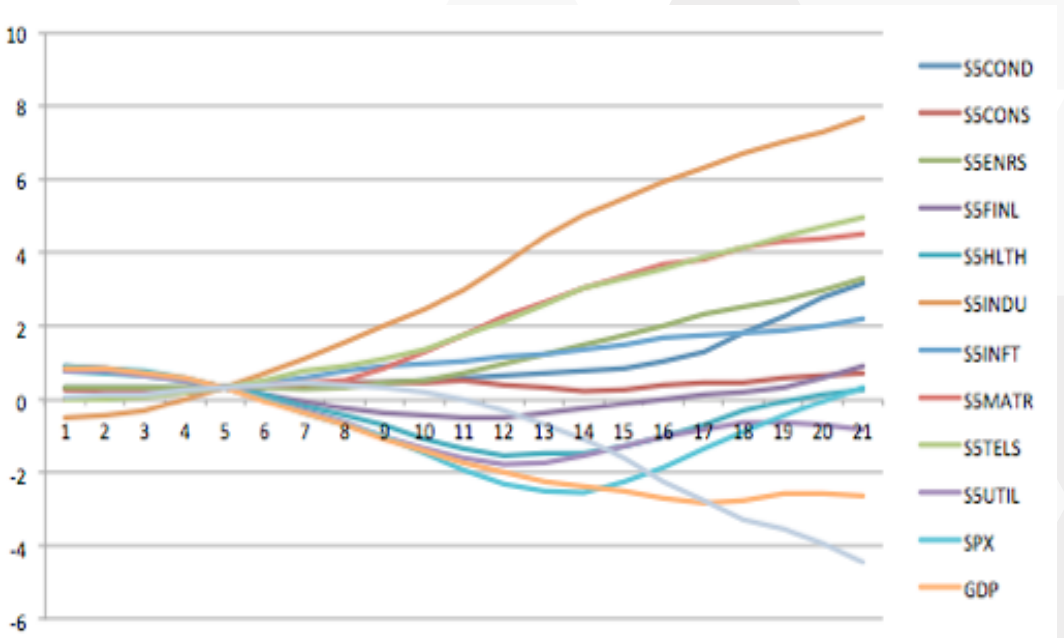
0.819

RESULTS

3. Neural Network Analysis of Stock Indexes/Economic Data







- Neural network shows cyclic nature of financial crises
- Show importance of energy and consumer staples sectors for future growth

RESULTS

4. Decision Tree Analysis of ICEWS Data

Index	Variable Name
5427	INStALLeventsct
5018	GOVtUAFhosttotals
6443	NOTGOVtALLhosscaleav
1059	ALLtINSeventsct
941	ALLtALLhosscaleav
5438	INStALLhosttotals
5433	INStALLhighhostilityct
2867	DIStGOVeventstotals
1395	BUDtALLeventsct

OBSTACLES

1. Keyword extraction algorithm effectiveness

- Produces meaningless keywords ex: ‘200 people’ or ‘Recent years’

2. Handling a file with thousands of columns

- Had to scale with MapReduce job

3. Low availability on HPC

- Hive Databases intermittently accessible

4. Neural Network

- Computational Power for Models with Thousands of Dimensions
- Infinite Number of Samples for Continuous Time Series

SUMMARY

- Frequent events related to communication and cooperation among countries
- Some events have a high correlation with following events
- Prominence of buddhist insurrection
- Some mined topic sequences have no semantic meaning
- Neural network shows cyclic nature of financial crises
- Show importance of energy and consumer staples sectors for future growth
- Further tuning and combination of the models could produce useful leads for investigation

ACKNOWLEDGEMENTS

- HPC team at NYU, including Akhil Kundh and Steve Leak
- Dr. Huy Vo, NYU CUSP
- Eva Kaplan, Tara Kinch, Dipti Jain, PPU and OPSCEN, UNICEF
- Professor McIntosh, NYU

REFERENCES

- M. D. Ward et al. Stepping into the Future: The Next Generation of Crisis Forecasting Models. International Studies Review, 2013.
- J. Dean et al. Large Scale Distributed Neural Networks. NIPS 2012: Neural Information Processing Systems, 2012.
- D.C. Anastasiu et al. Big Data Frequent Pattern Mining. Frequent Pattern Mining, Springer, 2014.
- R. Gwadera et al. Mining and Ranking Streams of News Stories using Cross-Stream Sequential Patterns. Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM, 20
- Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. In M. W. Berry & J. Kogan (Eds.), Text Mining: Theory and Applications: John Wiley & Sons.
- Zaki, Mohammed J. "SPADE: An efficient algorithm for mining frequent sequences." Machine learning 42.1-2 (2001): 31-60.
- Jiawei Han , Jian Pei , Yiwen Yin, Mining frequent patterns without candidate generation, Proceedings of the 2000 ACM SIGMOD international conference on Management of data, p. 1-12, May 15-18, 2000, Dallas, Texas, USA

THANK YOU!