ANALYZING HILLARY CLINTON'S EMAILS

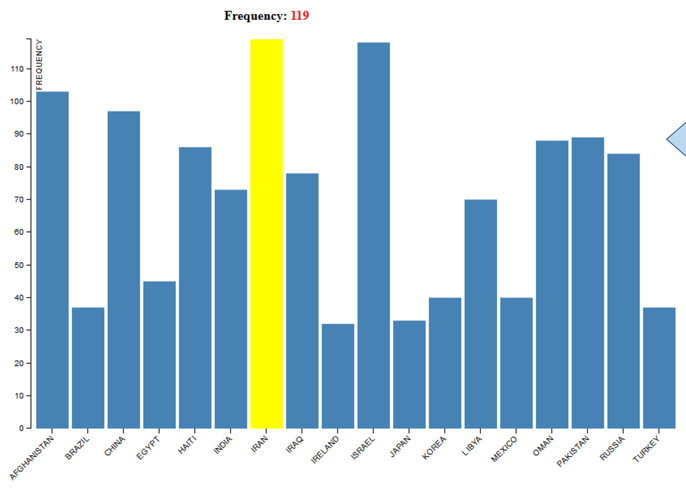
Vasundhara Dehiya and Klaus Mueller
Computer Science Department, Stony Brook University, NY, USA
SUNY Korea, Songdo, Korea

INTRODUCTION

Due to the controversy regarding the use of personal email on non-government servers by Hillary Clinton during her time as Secretary of State, her email data was made public. Yet, reading through them is impractical. In this poster, we provide a visual analysis of the content of these emails. Based on the 7,945 emails available, we identify the relation between the textual content of the emails with world policies. We unravel how the content of these emails are reflective of US emotion and behaviour with other countries around the world along with their relative importance. We also unravel correlations in the data to predict some features of content in redacted emails based on available data.

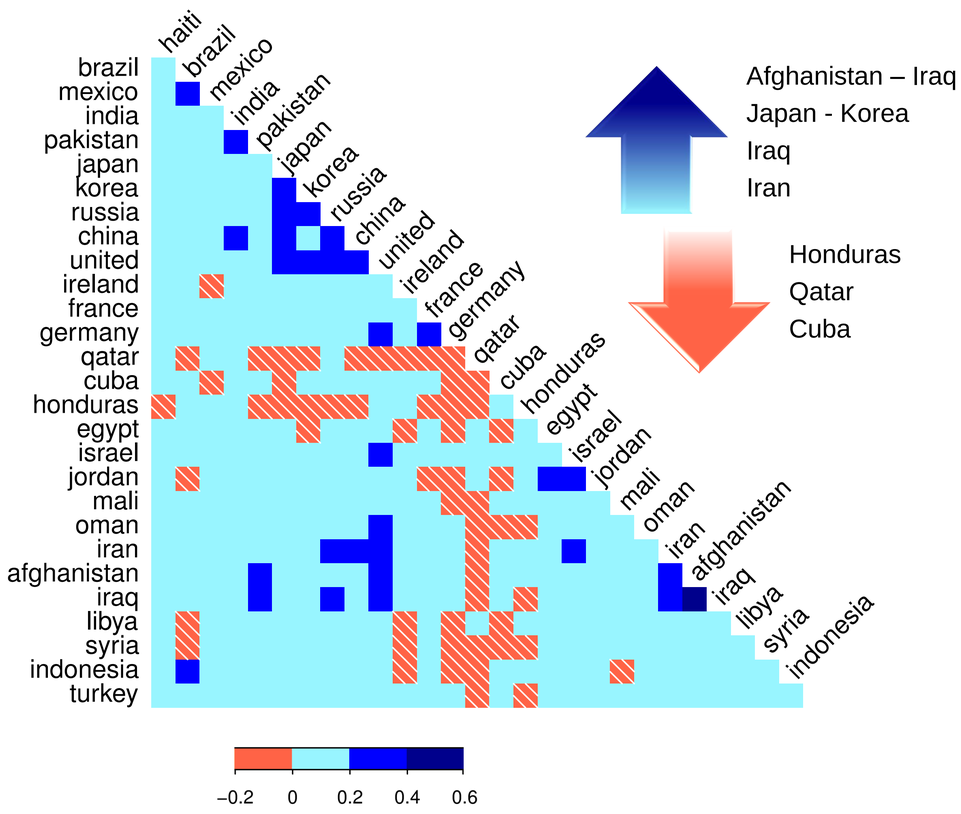
OBSERVATIONS

1. MOST FREQUENTLY DISCUSSED COUNTRIES



TOP 5
1.IRAN (119)
2.ISRAEL (118)
3.AFGHANISTAN (103)
4.CHINA (97)
5.PAKISTAN (89)

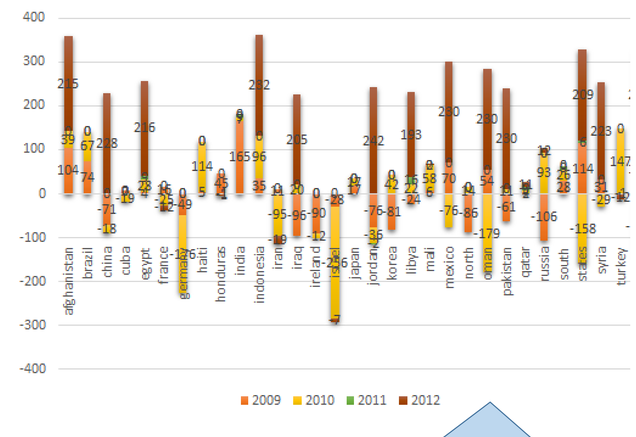
2. CORRELATION MATRIX REPRESENTING COUNTRIES DISCUSSED TOGETHER



Afghanistan – Iraq
Japan - Korea
Iraq
Iran

Honduras
Qatar
Cuba

3. NET SENTIMENT FOR EACH COUNTRY



Sentiment analysis is performed to identify US sentiment associated with each country and how it changes with time. Positive sentiment was observed to correlate with positive relations with that country and negative sentiment correlated with unfavourable US emotion towards that country (note that for 2012 only few emails were available)

4. CORRELATION OF SENTIMENT WITH EMAIL LENGTH

CONCLUSION

We identified which countries were discussed in the maximum number of emails and identified correlations between countries mentioned together. Observation 4 shows that there is a strong positive correlation for 2011 and 2012 and thus, given a retracted email, we can predict its emotion based on the length of the document. Also, Observation 2 can be used to identify which countries are being discussed in case we have partial content or names of one or more country in the text available. This information is useful in order to predict the type of content in a redacted document based on only the length of document and the names of countries present in the text.

This research was partially supported by NSF grant IIS 1527112 and the Korean Ministry of Science, ICT and Future Planning, under the IT Consilience Creative Program (ITCCP) supervised by NIPA.
Emails: { vasundhara.dehiya@stonybrook.edu, mueller@cs.stonybrook.edu }

SUNY Korea
Stony Brook University Graduate Degree Programs
The State University of New York

Stony Brook University