Assignment 4, Data Processing & Visualization QMSS G4063

Adnan Hajizada ah3326@columbia.edu

While solving this problem set I have consulted my group members: Saad Khalid, Alessandra Plassaras and Surubhi Bajpai.

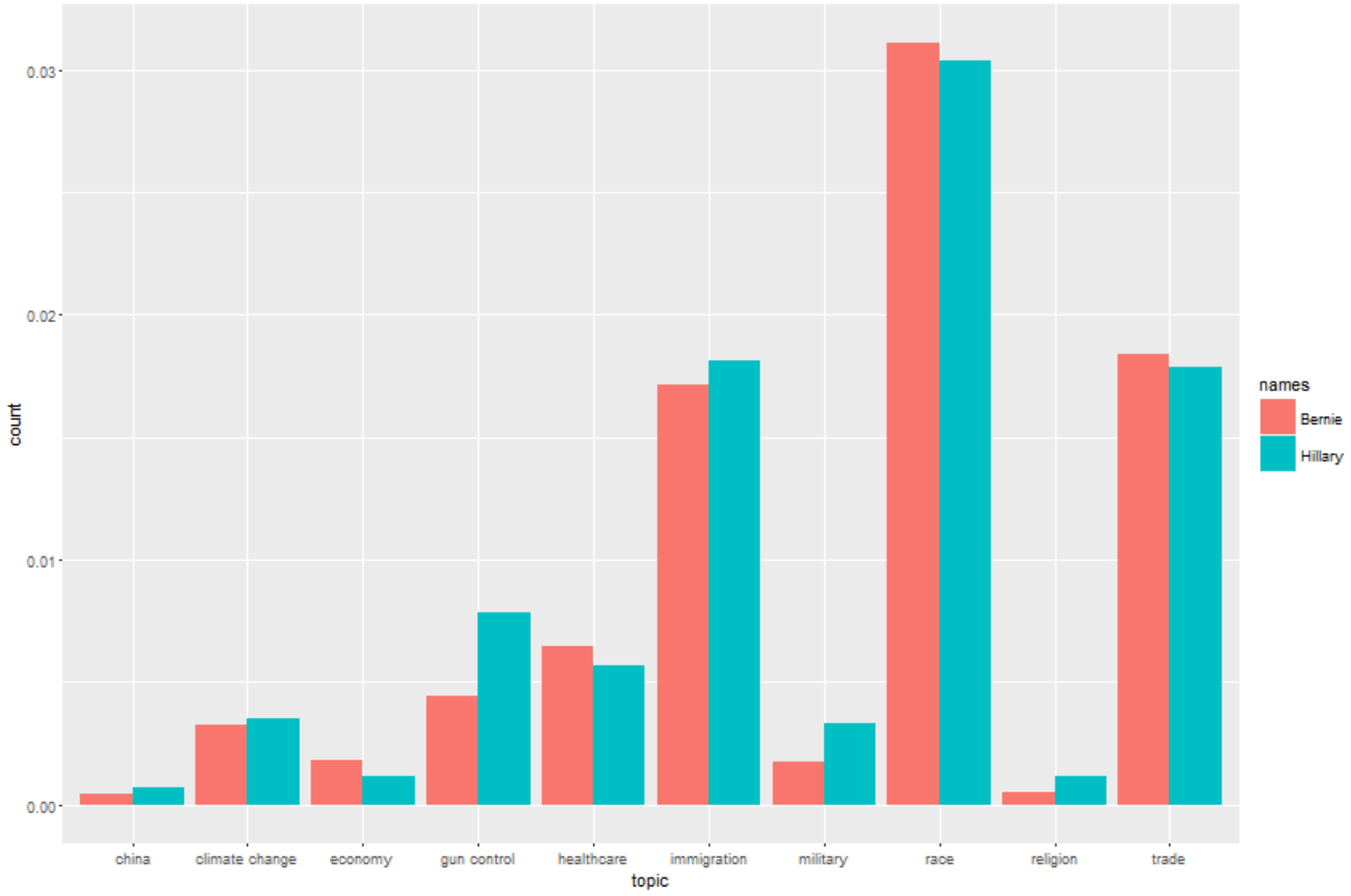
4/9/2016

Visualizations:

Two static visualizations comparing the levels of the usage of each of the *topics* for

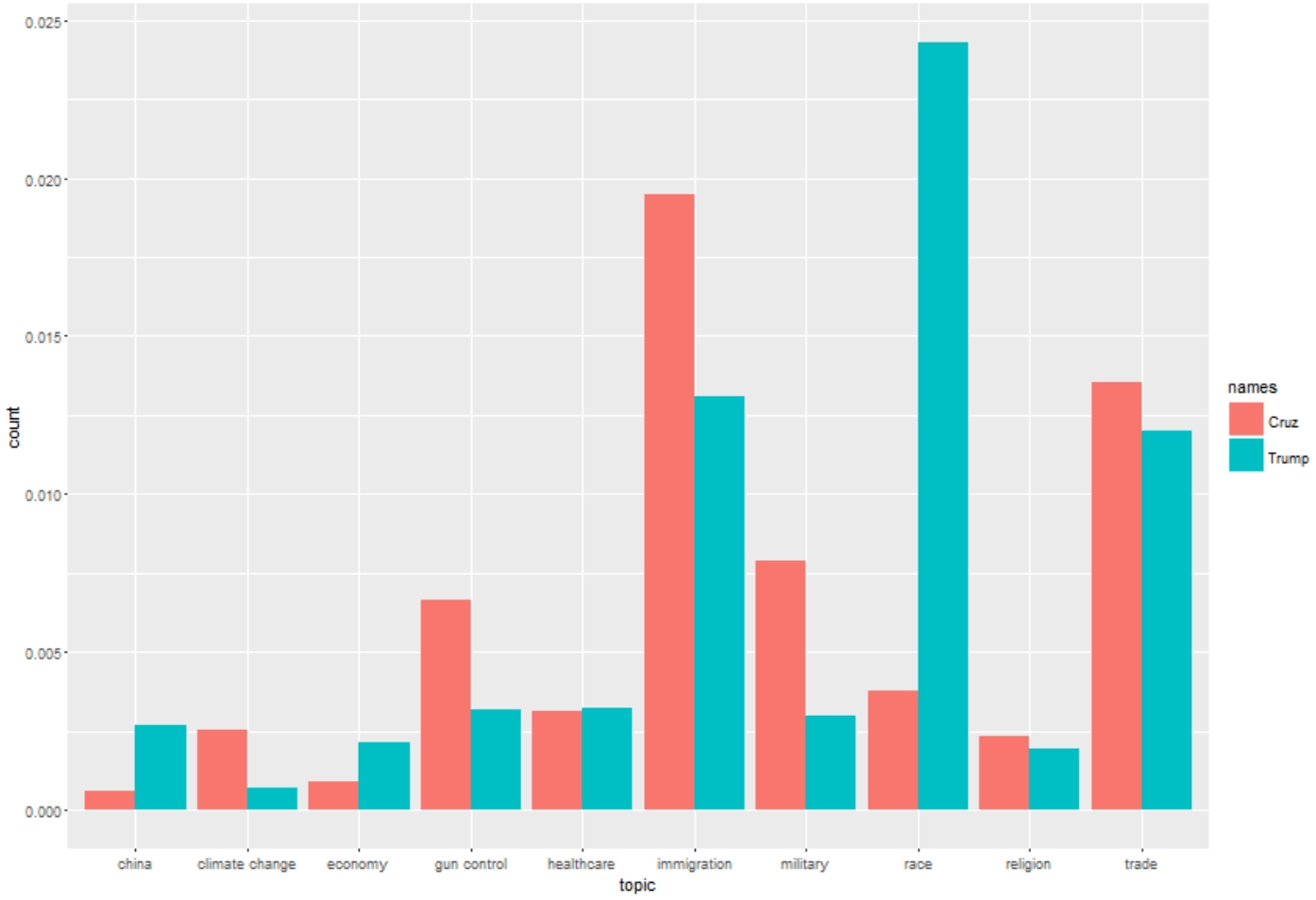
a) Between the two Democratic candidates

It is important to note that I adjusted the number of counts through dividing it by the total number of tweets by a given candidates. This will show us a fair representation of what topics are used regardless of the total amount of tweets (which could have skewed the results towards the candidates with a bigger online presence)



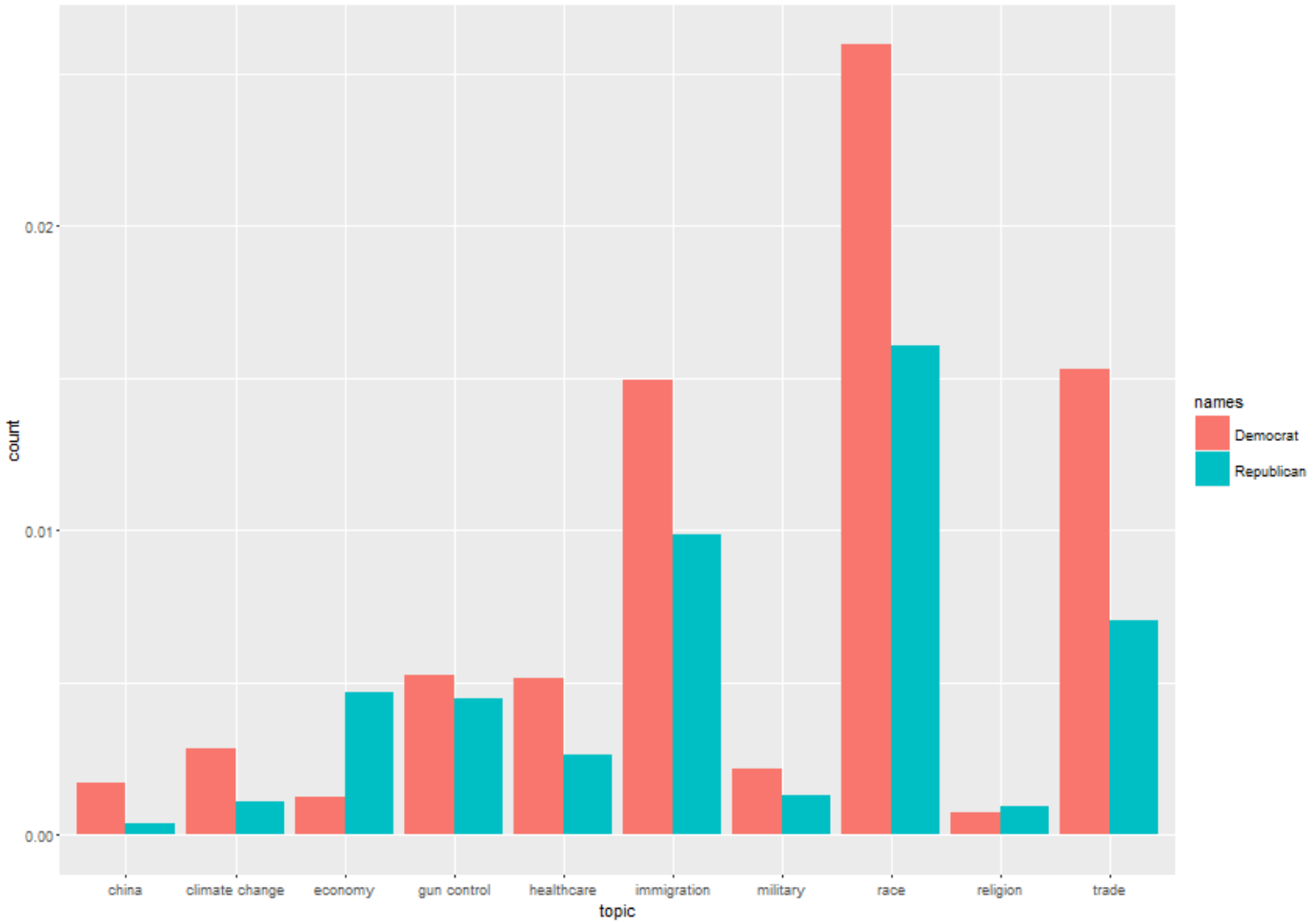
It is clear that the discussion about both democratic candidates involves almost proportional amount of usage of immigration, trade and race topic. Which are usual popular topics with liberals.

b) Between the two Republican candidates



We can see a clear distinction between the topics used in tweets about Donald Trump and Ted Cruz. When it comes to race topic. Trump is ahead of everyone. This might be because a lot of online commentators regard him as a racist person, who has been saying racist comments about Mexicans and Muslims. Thus this topic is prevalent. Interestingly enough, when it comes to immigration Ted Cruz is involved much more. My guess would be it is because he is a son of a Latino immigrant himself and this is an important issue for him and for people who discuss him online. Within the key terms in the immigration topic I have included the word “wall” in the reference to the Trump’s idea to build a wall, but as we can see Ted Cruz proportionally is discussed much more with the regards to immigration topic. It is also necessary to mention that the choice of words is very important.

One static visualization comparing the levels of the usage of each of the *topics* between the democrats (as a whole) and the republicans



From the above visualization I can conclude that overall, the discussions on the 10 topics and 50 keywords identified by me mainly involves democratic candidates. It seems that the conversation about Republicans could be about other topics. According to the graph, democratic candidates are mentioned in a discussion about immigration, race and trade much more than their Republican counterparts. The only topic that is prevalent with the republicans in Economy.

- Six dynamic visualizations for each of the candidates, and for democrats and republicans as a whole, showing the *daily* change in the level of the usage of each of the topics during the time period you are considering (March 16 to March 25)

After examining the data (same one that we used in past) between March 16 and March 25 I realized that the mined tweets only cover the time of 24 hours. Thus there is no way to show the daily change, for the data is encompassing only one day to begin with. Thus I have decided to build an interactive Shiny App that will show the change between different candidates and parties within the 24 hour frame. Since the content is dynamic I am going to include print screens of the shiny app implementation and I will also upload the Shiny App as a part of my assignment. Please note that for dynamic visualization I decided to include raw amount of counts (not adjusted to the tweets) because it is also interesting to compare the actual amount of counts. If a candidate has less amount of tweets to begin with, this does not mean this fact should be disregarded. In fact this fact should find its realization on the visualization. Less tweets, equals less count. Also since we are comparing different sets of data dynamically it is important for the scale to remain constant for ease of comparison. Thus reaching a balance between adjusted counts in static visualizations and raw counts for dynamic visualization.

~/Semester 2/Data Visualization/ps4/Shiny - Shiny

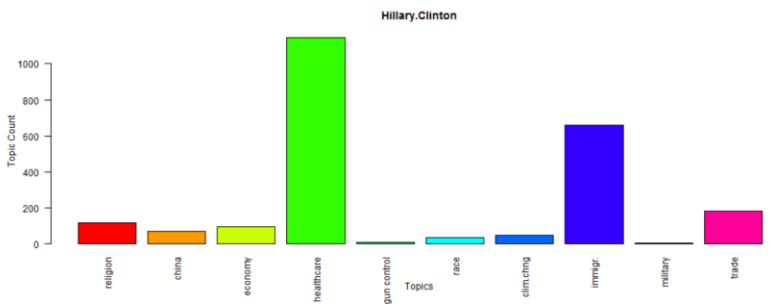
http://127.0.0.1:6131 | Open in Browser | Publish

Topic Count by Day

Name:

Hillary.Clinton

Data from topics and dates



Search the web and Windows

ENG 12:39 AM 4/10/2016

~/Semester 2/Data Visualization/ps4/Shiny - Shiny

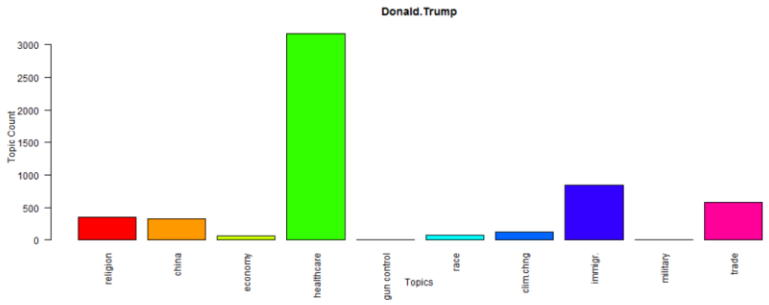
http://127.0.0.1:6131 | Open in Browser | Publish

Topic Count by Day

Name:

Donald.Trump

Data from topics and dates



Search the web and Windows

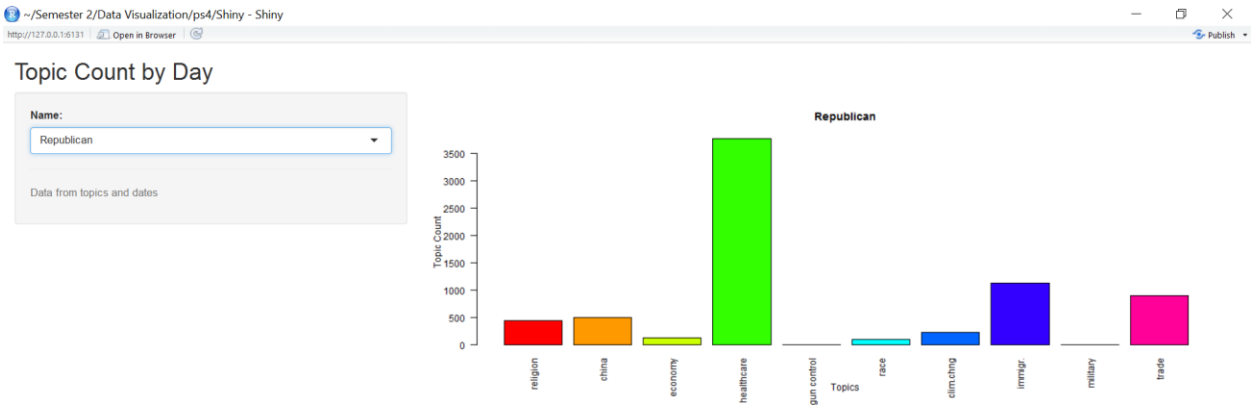
ENG 12:39 AM 4/10/2016


~/Semester 2/Data Visualization/ps4/Shiny - Shiny
http://127.0.0.1:6131
Open in Browser
Publish
Topic Count by Day
Name:
Ted.Cruz
Data from topics and dates
Ted.Cruz
Topic Count
0
100
200
300
400
500
600
religion
china
economy
healthcare
gun control
race
clim.chng
immigr.
military
trade
Topics

Search the web and Windows
ENG
12:39 AM
4/10/2016


~/Semester 2/Data Visualization/ps4/Shiny - Shiny
http://127.0.0.1:6131
Open in Browser
Publish
Topic Count by Day
Name:
Ted.Cruz
Data from topics and dates
Ted.Cruz
Topic Count
0
100
200
300
400
500
600
religion
china
economy
healthcare
gun control
race
clim.chng
immigr.
military
trade
Topics

Search the web and Windows
ENG
12:39 AM
4/10/2016



It seems that on the day when the tweets were calculated there was a big debate or discussion about the healthcare topic. It seems that every single candidate and party were being mentioned in connection with the topic of health care. That is the one case that stand out right away.

Extra Credit Attempt:

Because of the lack of time, I included the key words for the topics arbitrarily including my own view of the topic. One of the ways to go about this would be to split the corpus into sub sets by each topic. Do text mining of that subset and find 4-5 top words that are used together with the topic name, and use those words for the topic search.

These are the words that I initially chose for the lexicon. They could be improved:

words	topic
economy	
growth	
jobs	
gdp	
unemployment	economy
immigration	
refuges	immigration

syria	
border	
wall	
hospital	
healthcare	
cancer	
obamacare	
health	health_care
military	
defence	
army	
iraq	
syria	military
gun	
nra	
shooting	
amendment	
safety	gun_control
china	
beijing	
debt	
pollution	
innovation	china
tpp	
trade	
agreement	
partnership	
free	trade
black	
race	
blacklivesmatter	
police	
hispanic	race
climate	
environment	
warming	
melting	
science	climate_change
islam	
christianity	
atheism	
church	
religion	religion