615 Final Project

Yuetian Sun U55385536

December 15, 2017

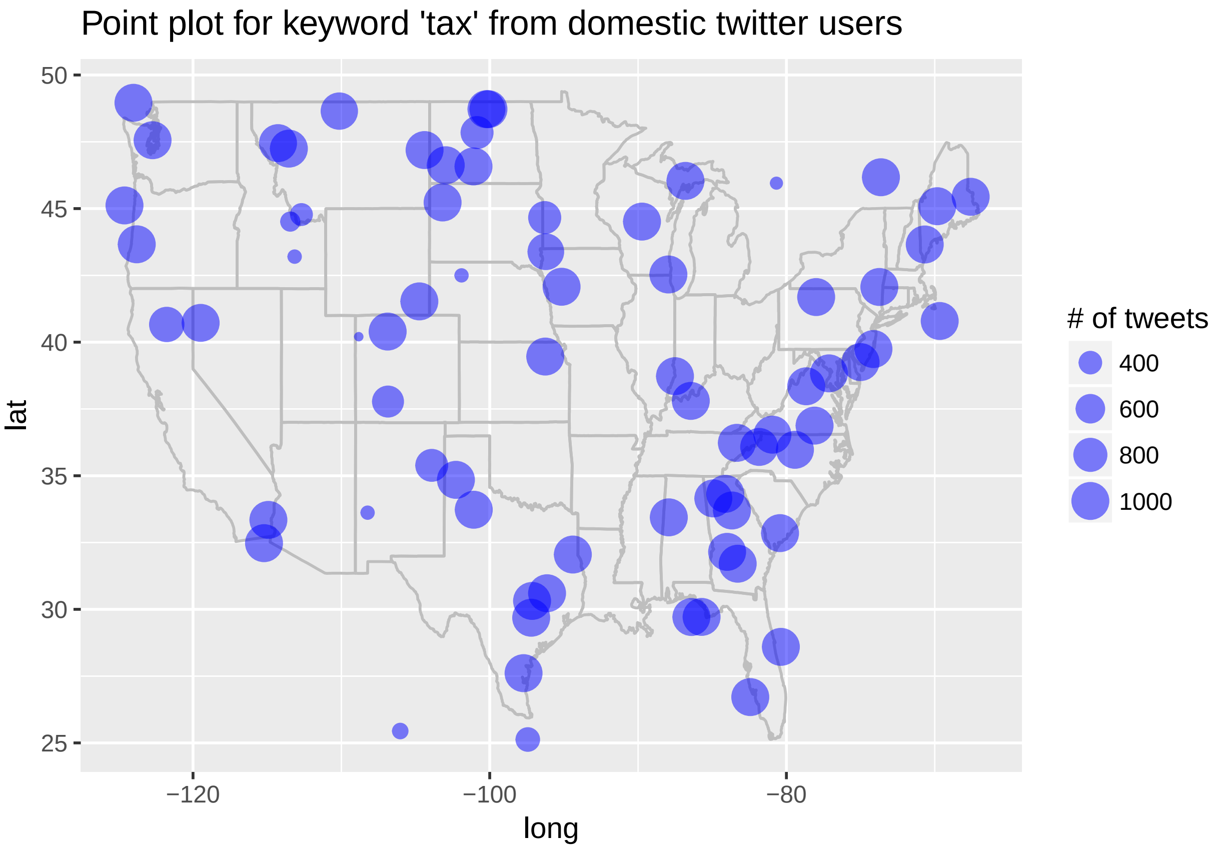
Introduction

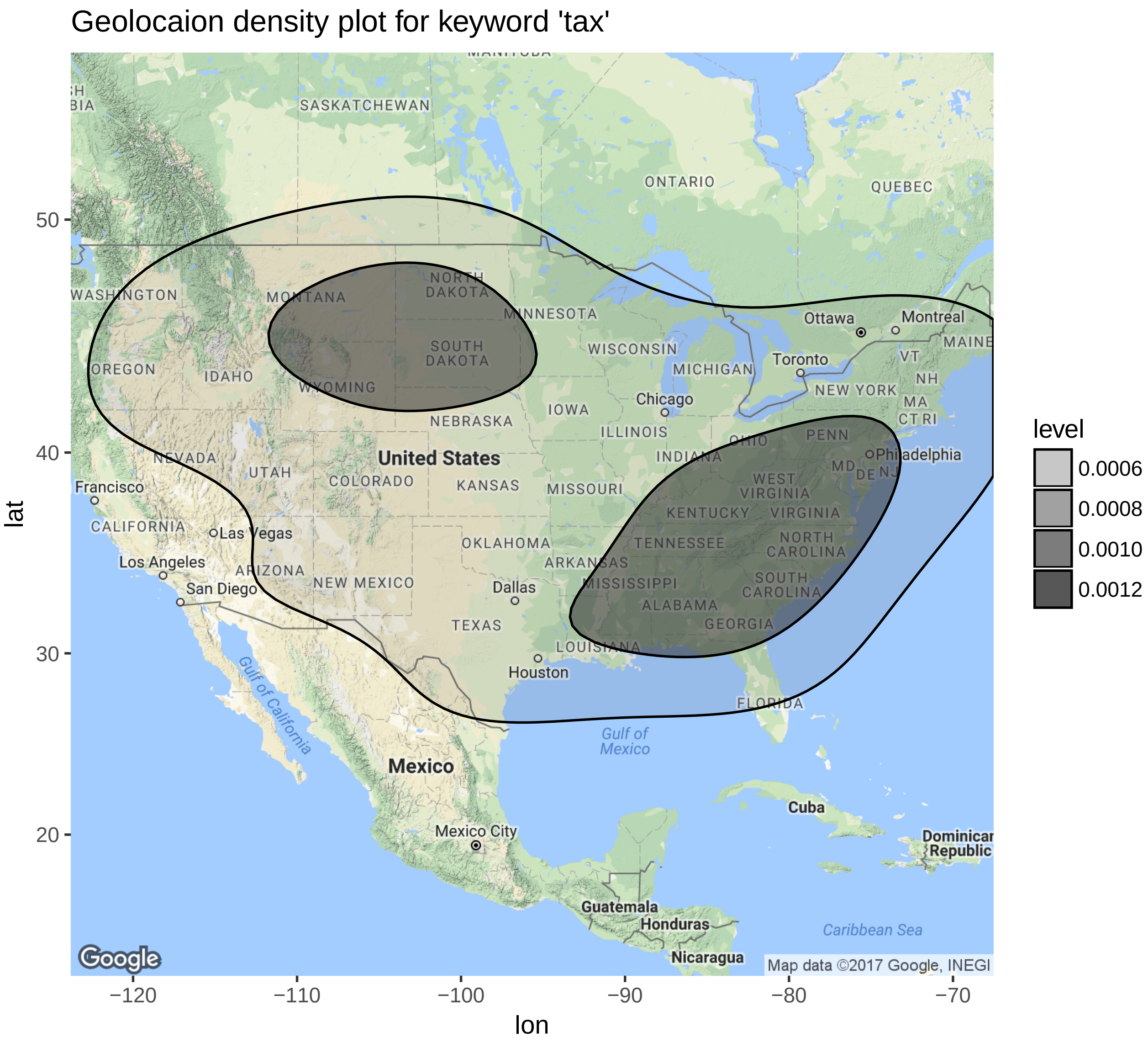
As we all know that President Trump has published a policy about reformation of tax these days. This policy has significant effect on thousands of people's life for example graduate student may have much heavier burdern about their tuition fee. This project aims to extract useful information when people twittering about tax finding out what people atitudes are and what they are most concerned about. This project is devided into several parts. Firstly, to have first impression of data, we plot the density graph and points on map about which area's people are more likely to twitter about tax. Then we conduct Emoji Analysis on these twitters trying to find people's opinion by analyzing emojis. Next we take a further step in the text part of twitters. We conduct Topic Modeling Analysis, Sentiment Analysis and make a perdict of American people thought when talking about tax. As a result of these, we can make an assumption of people satisfaction this policy.

Key words: Twitter, Emoji Analysis, Topic Modeling, Sentiment Analysis

1 Exploratory Analysis of Data

We randomly select 47474 of twitters with “tax”" from 11/24/2018 to 12/04/2018. Now let's take a first look of the data. We would like to figure out the general situation about twittering with “tax” across the US. Thus, we make a point plot and density plot on map to have a first impression.


Point plot for keyword 'tax' from domestic twitter users
50
45
40
35
30
25
lat
−120
−100
−80
long
# of tweets
400
600
800
1000



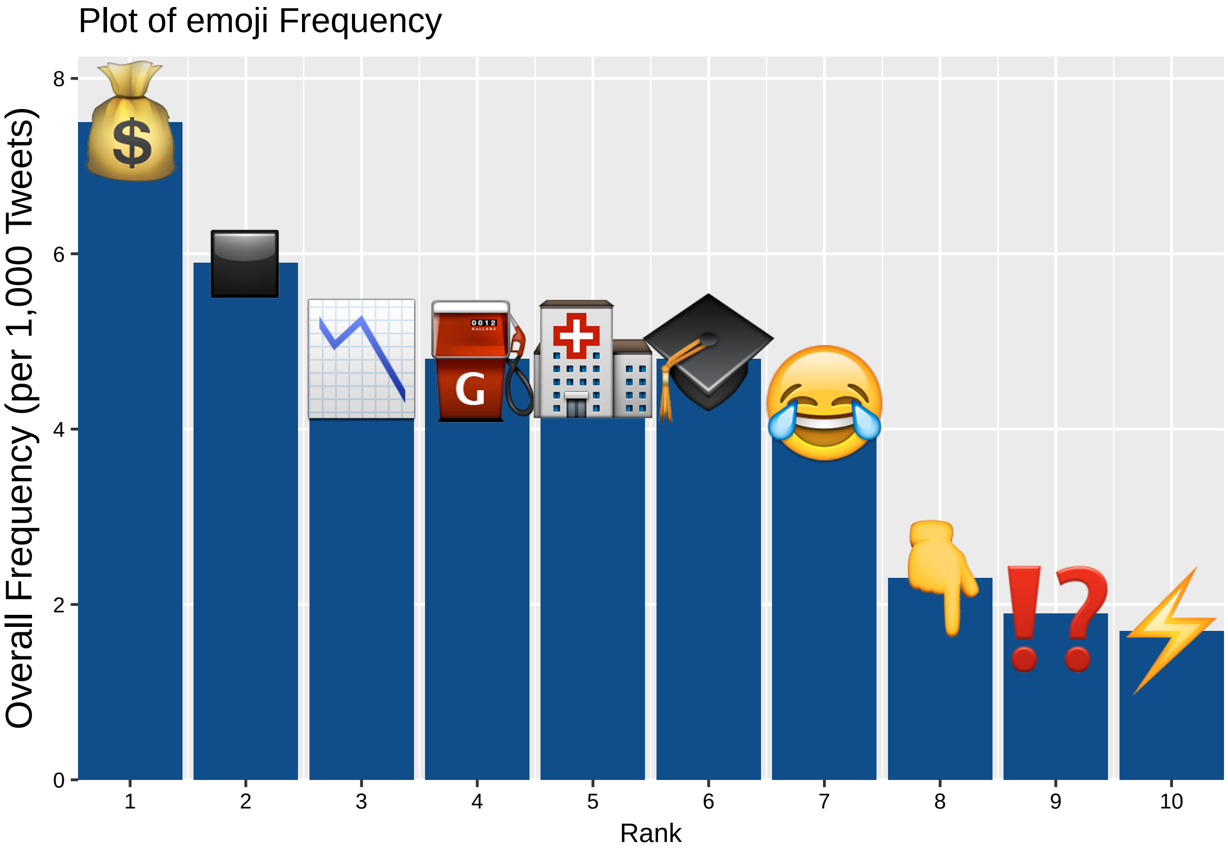
As we can see the plot above, the point plot and the density show that people in the east coast and north middle part of America have more twitters about "tax", which means in some sense they are more concerned about tax issue. We can guess that may be the east coast has lots of schools and has large population. The reformation of tax will have significant effect on this area. And also people live in the north middle parts of America don't have very high salary on average and the reduction of individual income tax can help them have better life.

2 Emoji Analysis

Now we want to take a deep step on people's attitude when twittering "tax". Let us take a looking of what emojis people would like to twitter about "tax". First of all, we need to clean the dataset and extract the emojis (emojis code) from the tweets we selected. Then by pairing the emojis code with emojis dictionary and emojis picture, we can get the original emojis. Then we calculate the frequency of these emojis and figue out which one people are more likely to use when twittering about "tax".

Table 1: Most frequently used emojis

name	dens	count	rank
money bag	7.5	355	1
black small square	5.9	278	2
chart with downwards trend	4.8	228	3
fuel pump	4.8	227	4
hospital	4.8	227	5
graduation cap	4.8	226	6
face with tears of joy	4.3	205	7
white down pointing backhand index	2.3	107	8
exclamation question mark	1.9	92	9
high voltage sign	1.7	79	10

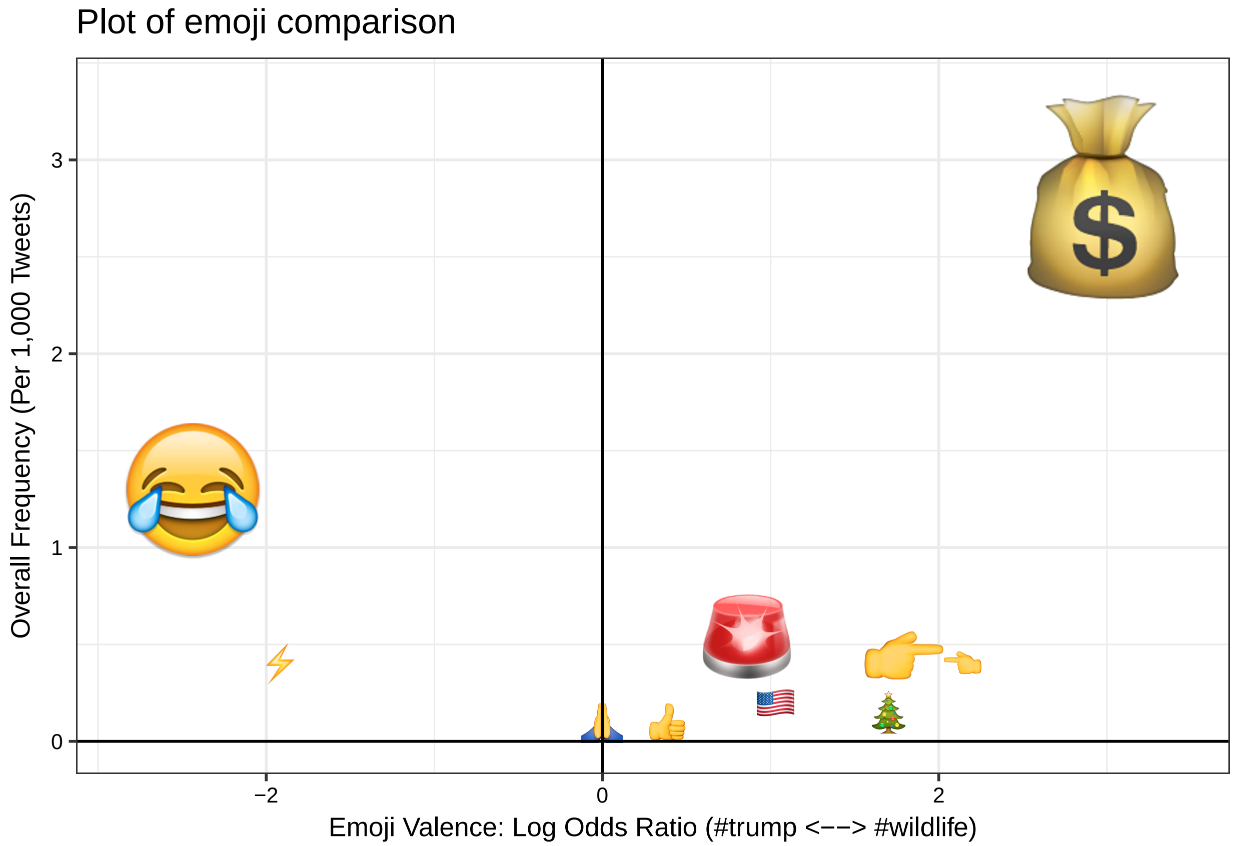


As we can see from the plot and the table above, the most frequently uesd emoji when talking about tax is money bag. This make sense because tax directly related to money and this should be the thing people most concerned about. The third emoji is chart with downwards trend. We can assume that this is a negative one and people using this to express their life going down or it may be a positive one used to express the tax reduction. Also the fourth to sixth shows people concern about hospital fee, tuition fee and many other fee related to tax policy. And the face with tears of joy can some how show the people's atitude about tax reduction.

The next step after visualizing the top emojis in the overall dataset is to compare emoji frequency between two different subsets of the data. To find out more information about it, we can give several pairs of important words people may twitter with "tax" and comparing which emoji they would like to use. We take keywords "trump" and "cut" as an example.

Table 2: Table of emoji comparison

name	dens.1	dens.2	freq.1	freq.2	dens.mean	logor
money bag	5.3	0.3	253	12	2.80	2.9723849
white left pointing backhand index	0.7	0.1	34	3	0.40	2.1690537
white right pointing backhand index	0.8	0.1	36	5	0.45	1.8191584
christmas tree	0.3	0.0	15	2	0.15	1.6739764
regional indicator symbol letter u/s	0.3	0.1	13	4	0.20	1.0296194
police cars revolving light	0.8	0.3	36	15	0.55	0.8383292
thumbs up sign	0.1	0.1	5	3	0.10	0.4054651
regional indicator symbol letter u/s	0.1	0.1	6	6	0.10	0.0000000
high voltage sign	0.1	0.7	4	34	0.40	-1.9459101
face with tears of joy	0.2	2.4	9	115	1.30	-2.4510051



There are multiple pairs of words you can choose for these graph. If you want to see the results of all these words, please use the shiny.app. Here let us interpret “trump” and “cut” example. As we can see above, when people mention “trump” in twitter with “tax”, they tends to use face with tears of joy while people mention “cut” are more likely to use money bag. This result makes sense because cutting tax is directly related to money. Also people who mention “cut” have higher probability of using police cars revolving light emoji. This may use to call people’s attention about the reformation of tax. In additon, people who mention “cut” or “trump” are equal likely to use person with folded hands. May be they use it to hope for a good life.

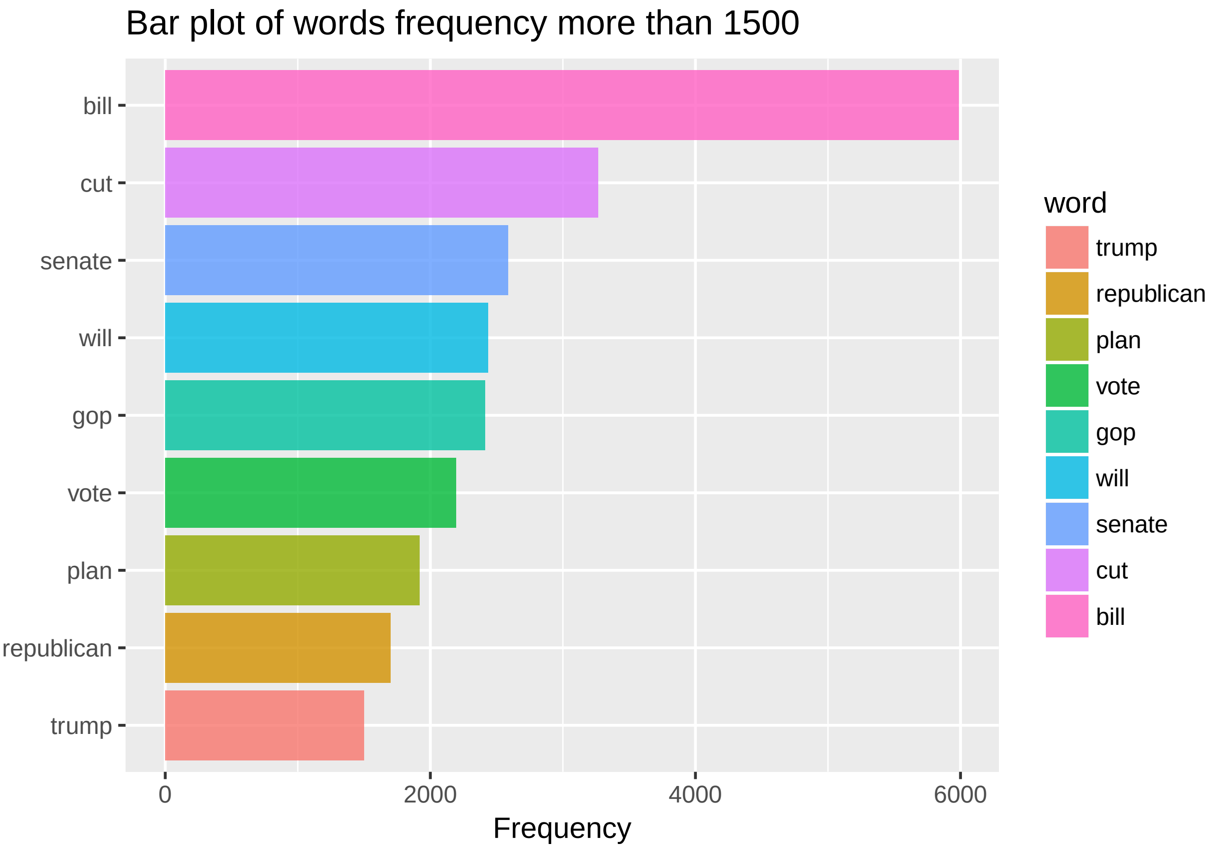
After exploring the emojis, we now focus on the text part of twitters.

3 Text Analysis

Now let us take a look of what words people would likely to use when twittering with “tax”. To make the result intuitively, we make a table and a frequency plot of words showing up more than 1500 times.

Table 3: Table of words frequency more than 1500

word	n
bill	5989
cut	3268
senate	2587
will	2437
gop	2413
vote	2193
plan	1918
republican	1702
trump	1501



From the bar plot and the table above, We can find that people are concerned about cuting their bill, republican and voting things about tax. These make sense because the reducation tax policy is passed in senate and is comed up by Trump, Repulican Party. And the aim of these policy is to cut the tax and promote American’s life.

Now let us continue finding out what topics people would likely to talk about tax.

3.1 Topic Modeling

To take a further step, we conduct Topic Modeling and analyze what topic people will talk when twitter about tax.

Firstly, we would like to introduce what is Topic Modeling and how it works on twitter. In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear equally in both.

Latent Dirichlet allocation (LDA) is a particularly popular method for fitting a topic model. It treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language.

In this report, we focus on using LDA to conduct Topic Model. And we would like to choose Gibbs Sampling. Here is how Topic Modeling works mathematically.

The topic distribution for each twitter is distributed as

$$\theta \sim Dirichlet(\alpha)$$

where $Dirichlet(\alpha)$ denotes the Dirichlet distribution for paramter α.

The word distribution on the other hand is also modeled by a Dirichlet distribution, just under a different parameter η.

$$\phi \sim Dirichlet(\eta)$$

The utmost goal of LDA is to estimate the θ and ϕ which is equivalent to estimate which words are important for which topic and which topics are important for a particular document(twitter), respectively.

For each document(twitter) d, go through each word w (a double for loop). Reassign a new topic to w, where we choose topic t with the probability of word w given topic t * probability of topic t given document(twitter) d, denoted by the following mathematical notations:

$$P(z_i = j | z_{-i}, w_i, d_i) = \frac{C^{WT}_{w_i j} + \eta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\eta} * \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

$P(z_i = j)$: The probability that token i is assigned to topic j.

z_{-i} : Represents topic assignments of all other tokens.

w_i : Word (index) of the ith token.

d_i : document(twitter) containing the ith token.

For the right side of the equal sign:

C^{WT} : Word-topic matrix.

$\sum_{w=1}^{W} C^{WT}_{wj}$: Total number of tokens (words) in each topic.

C^{DT}: Document-topic matrix.

$\sum_{w=1}^{W} C^{DT}_{d_i t}$: Total number of tokens (words) in document(twitter) i.

η : Parameter that sets the topic distribution for the words, the higher the more spread out the words will be across the specified number of topics (K).

α : Parameter that sets the topic distribution for the documents, the higher the more spread out the tweets will be across the specified number of topics (K).

W : Total number of words in the set of twwets.

T : Number of topics, equivalent of the K we defined earlier.

After we're done with learning the topics for 1000 iterations, we can use the count matrices to obtain the word-topic distribution and document-topic distribution.

To compute the probability of word given topic:

$$\phi_{ij} = \frac{C^{WT}_{ij} + \eta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\eta}$$

Where ϕ_{ij} is the probability of word i for topic j.

$$\theta_{dj} = \frac{C^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dk} + T\alpha}$$

Where θ_{dj} is the proportion of topic j in document d.

After computing the probability that each document(twitter) belongs to each topic (same goes for word & topic) we can use this information to see which topic does each document(twitter) belongs to and the more possible words that are associated with each topic.

Table 4: Ten topics for twitter tax

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
people	cut	gop	will	like	reform	new	get	vote	bill
pay	corpor	plan	american	now	back	year	can	bill	senate
money	class	trump	say	take	america	time	make	scam	pass
dont	rich	republican	care	state	one	democrat	know	support	reform
think	break	lie	increas	deduct	talk	call	read	pleas	republican
work	middle	presid	million	incom	economi	just	see	yes	hous
want	give	show	health	right	nation	congress	good	follow	gop
help	busi	never	ite	great	stock	stop	need	parti	via
just	wealthi	win	deficit	credit	also	today	even	just	major
estatee	benefit	claim	famili	student	point	everi	let	ask	news

As we can see from the above table, this is the ten topics people would likely to talk about when twittering "tax". Now let's indicate the first three topics. we can find that topic 1 contain words like "rich", "poor", "middle" and "class". Thus we can make an assumption that people care about how tax affects on different classes. The second topic has "trump", "reform", "thank" and "lie" in it. Thus we can indicate that this may relate to topic about President Trump and reformation of tax. In the third part, we have words "vote", "people", "support" and "follow". Therefore, we can view the third topic as people supporting degree on tax reduction.

3.2 Sentiment Analysis

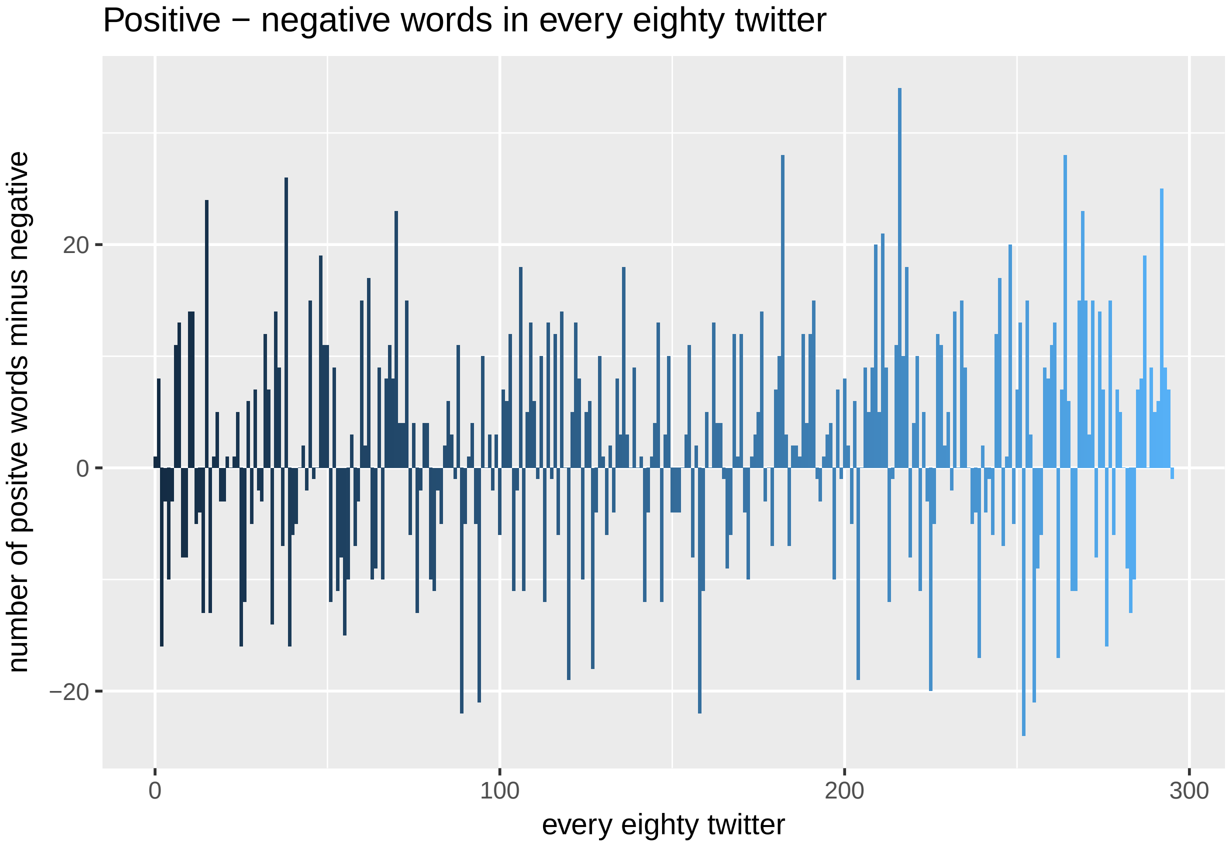
Now let us take a look on what people attitude on tax by conducting sentiment analysis. To start with, let us see what words people use to express their emotion. We can choose emotion like joy, angry, disgust. Here we only show joy as an example. If you want to see other emotional words result, please use the shiny.app. There are multiple choices in shiny.app.

Table 5: Table of top 10 joyful word

word	n
vote	2193
pay	976
money	719
good	424
save	243
hope	190
child	170
love	134
deal	122
true	114

As we can see from the above table, the most frequently used joyful words are 'vote', 'pay', 'money' and 'good'. Thus, we can guess that people with positve opinion about 'tax' are likely to talk about vote for the tax policy and the money things.

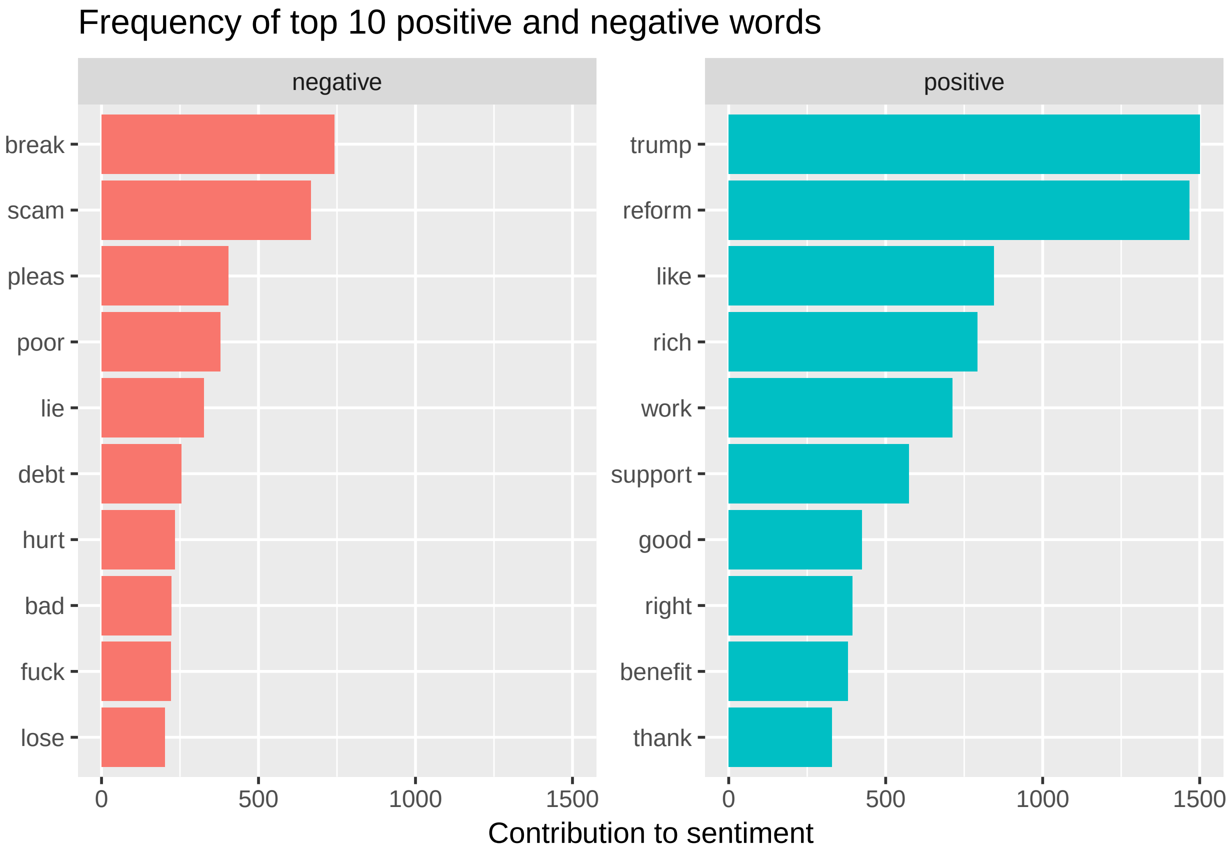
Now let us analyze people atitude by comparing the positive and negative words.



As we can see from the positive - negative plot, the lines above horizontal line means the positve words are more than negative words in every eighty twitter. This graph show that the case positive words are more than negative happens more than the other case. That is to say, the number of positive words is larger than the number of negative words. Thus we can indicate that the on the average people show positive atitude when twittering about tax. We are now interested in what positive and negative words people most frequently used.

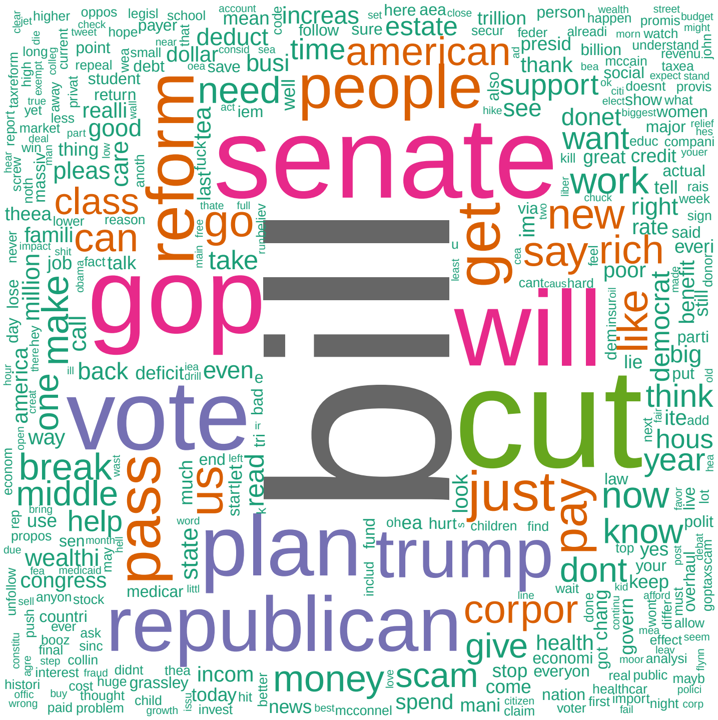
Table 6: Table of top 10 positive and negative words

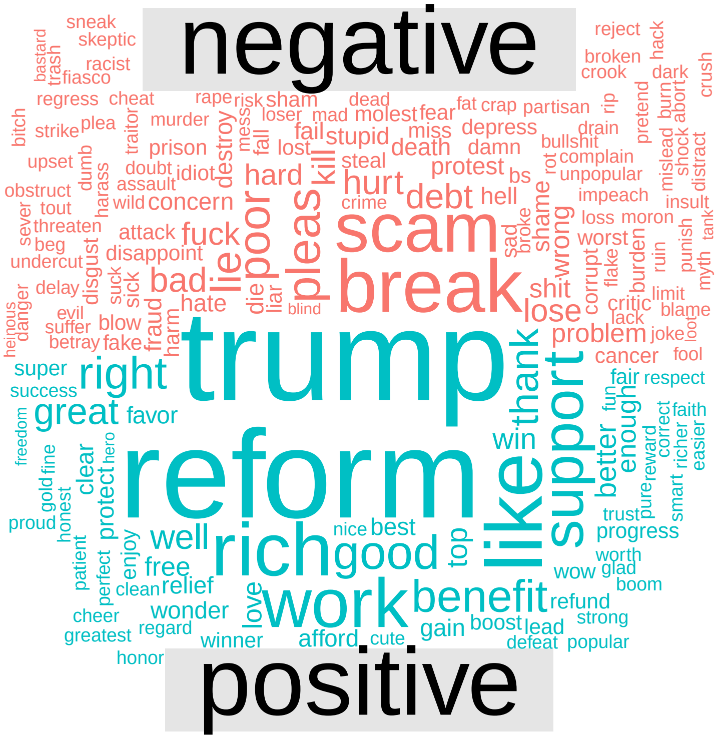
word	sentiment	n
trump	positive	1501
reform	positive	1467
like	positive	845
rich	positive	792
break	negative	742
work	positive	713
scam	negative	667
support	positive	574
good	positive	424
pleas	negative	404
right	positive	394
benefit	positive	379
poor	negative	378
thank	positive	329
lie	negative	325
debt	negative	254
hurt	negative	234
bad	negative	222
fuck	negative	221
lose	negative	201



As we can see from the plot and table above, the three most frequently used words are all positive. People would likely to use 'reform', 'trump' and 'rich'. We can indicate that people tend to believe the reformation of tax published by Trump can help people become rich and wealthy. But as we can also find that some people use negative words like ‘scam’, ‘poor’ and ‘break’. Thus, although lots of people support the policy, there are some people think it is a scam and will make people poor.

To have a more intuitive way of the words and its frequency, I make word clouds of top 500 frequently used words and top 200 positive and negative words.





As we see from the two word clouds above, it is very clear tha people concern about the about bill and money staff most and then they also care about the policy itself like the voting things, senate, repiblican, Trump and so on. Besides, people with positive attitude would likely to think about the reform of tax will make them rich while people with negative attitude think the reduction of tax is a scam.

Conclusion:

In this project, we care about how people think about the tax through twitter. We find that people living in east coast and north middle part show more concern about tax. Besides, by analyzing the emojis, we figue out that people are most concerned about money stuff. They care about hospital fee, tuition fee and many other fee related to tax policy. To narrow down, we use "trump" and "cut" as our second keywords. When people mention "trump" in twitter with "tax", they tends to use face with tears of joy while people mention "cut" are more likely to use money bag. In additon, people who mention "cut" or "trump" are equal likely to use person with folded hands. For the test part, from the bar plots, word clouds and tables, we can find that people are concerned about cuting their bill, republican and voting things about tax. And by conducting Topic Modeling, we find that people would likely to talk about how tax affects on different classes, President Trump and reformation of tax, and supporting degree on tax reduction. Additionaly, by conducting the Sentiment Analysis, we find that people show more postive attitude when twittering about tax and they tend to believe the reformation of tax can bring people good life.

To sum up, when twittering about tax, people are most concerned about the reduction policy and how it affect their money and their life. In general, although some people think this policy is a scam, lots of people do believe this reformation will make them wealthy and they support Trump policy of cuting tax.

Reference:

[1] URL: https://deanattali.com/blog/building-shiny-apps-tutorial/#before-we-begin

[2] URL: http://ethen8181.github.io/machine-learning/clustering_old/topic_model/LDA.html

[3] URL: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation