Case Study 5 (CS5): Text Mining

"On my honor, I pledge that I have neither given nor received help on this assignment."

Name: Prateek Agrawal
Computing ID: pa7sb
Course: Data Mining

Situation

The Chicago Police Department (CPD) is interested in gathering information from social media content. They believe that the discussions transpiring online might reveal leading indicators for crimes of different types. However, they do not have the manpower to manually analyze social media content on a large scale. They seek answers to the following questions via automated methods:

1. Are tweets distributed uniformly across Chicago? Show evidence?

S*olution*:-

Data:

A large sample of tweets corresponding to the year 2014, located within the boundary of Chicago.

The following features of the tweets are used for analysis

a. Time-Stamp
b. Location (longitude and Latitude)
c. The text in a tweet

Approach:

1.) First of all, the longitude and latitude are converted in x and y (in meters) for the purpose of plotting the tweets on a map, using '*proj=+init=epsg:26971*', the projection value for the city of Chicago.

2.) Converting the time stamp to a more readable format of 'YY-MM-DD hh:mm:ss'.

	timestamp	text	longitude	latitude	x	y	hour	day.of.week	month
1	2013-05-19 22:47:15	@alltimejeffy ughhhh I can't stand her	-87.78322	41.78633	345728.6	568517.4	22	1	5
2	2013-03-16 16:54:18	Call me @_lovepinky_	-87.66325	41.75029	355732.2	564585.2	16	7	3
4	2014-03-21 17:56:42	â€œ@Lodizzle09: I hate all of you ðŸ˜Šâ€	-87.72923	41.74553	350248.1	564015.0	17	6	3
7	2013-05-03 08:16:56	There's a woman in a suit and tie #Radicals	-87.90173	41.97662	335771.2	589596.3	8	6	5
8	2014-02-07 16:49:54	@asanchez_12 I had to eat doing all that work without...	-87.63954	41.83818	357625.6	574362.2	16	6	2
10	2013-06-19 09:48:01	The Bloomingdale Trail Gets A Re-Branding: "The 606,...	-87.62627	41.88805	358682.2	579909.7	9	4	6
11	2013-01-10 08:01:36	Day 5 no caffeine! I'm doing great! Just a bit more slee...	-87.74841	41.99030	348467.8	591190.4	8	5	1
12	2013-07-23 13:00:24	What's the use of having that ass if you ain't even goi...	-87.81620	41.93573	342887.3	585093.1	13	3	7
14	2014-01-04 02:01:46	Well I whine over everything I WHINE ALOT	-87.58862	41.78304	361907.6	568273.2	2	7	1
15	2013-06-12 12:57:59	I'm at @CardinalFitness (Chicago, IL) http://t.co/Q8G...	-87.66026	41.88148	355866.4	579157.8	12	4	6
16	2013-09-12 17:41:36	I'm at Pauly's Pizzeria - @paulyspizza (Chicago, IL) [pi...	-87.62741	41.87259	358601.6	578192.8	17	5	9
17	2013-08-13 14:55:53	Paul, Omar & Frenchy enjoying DJ Scrabble. http:...	-87.66866	41.93623	355122.5	585233.8	14	3	8
20	2014-01-03 22:51:05	I used to love vodka.	-87.61836	41.81014	359410.6	571261.8	22	6	1

Fig1: the Dataframe with tweets

3.) Filtering all tweets based on the boundary of Chicago. i.e removing all tweets that don't lie within the boundary of Chicago city.

4.) Plotting the tweets on the map of Chicago

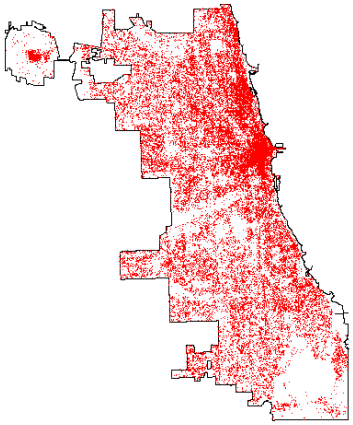

Fig2: Tweets spread across the city of Chicago

→ *We observe that the tweets are not spread uniformly within the city, as the southern part of the town shows no tweets shared and even the spread of tweets is quite low near the airport, while the downtown area has a very dense concentration of tweets.*

5.) For better evaluation we plot a Kernel Density Function for the tweets for better evaluation of its spread within Chicago.

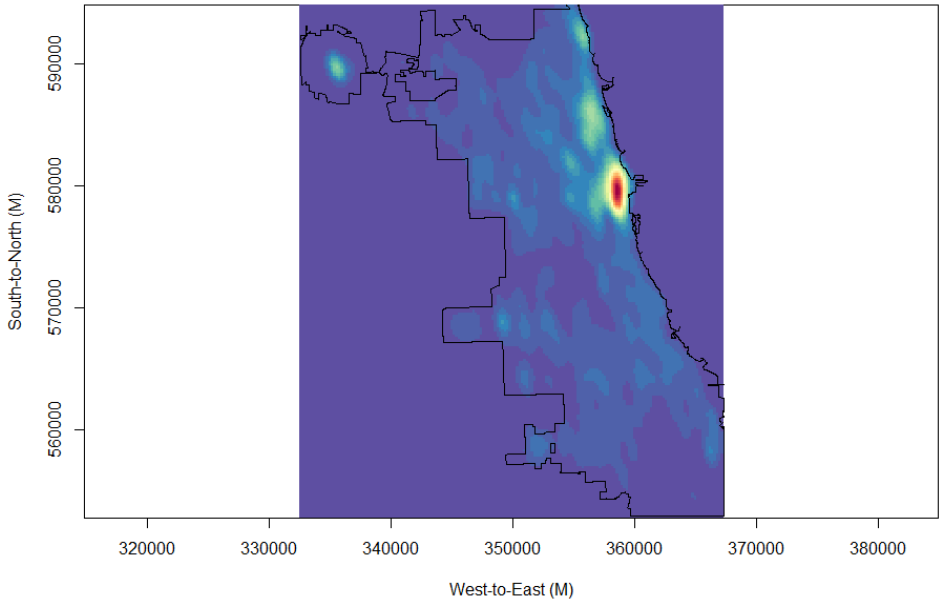

Fig3: KDE of tweets in the city of Chicago

→*This confirms our earlier hypothesis that downtown area of the city has the highest concentration of tweets.*

Observation:

This trend is mostly observed as teenagers and tourists are the maximum users of social media and as there are higher numbers of schools and landmarks situation in the downtown area therefore, it has the highest concentration of tweets.

---

2. *Are there topical differences between weekday content and weekend content? Evidence should come from topic analyses*?
Solution:

Initial Guess:

There will be a tropical difference between the weekend contents of topics found and the weekday content. As during weekdays most of the people with their offices and school while weekends people tend to relax and go out for seeing and bunch of other stuff.

Data:

The tweets from the year 2014 in the city of Chicago are divided in two sets based on the day of the week, while weekdays contains all tweets from Monday to Friday, weekend tweets contain all tweets on Saturday and Sunday.

Assumption:

I have considered each tweet as one document as even if the number of words in each document is less (max 140 words), but mixing different tweets together can lead to unintentional content mixing, i.e. tweets belonging to different topic could be clubbed as one and after pre-processing of the document, the output can lead to different topics, then it would actually have meant to be in. Therefore, each topic is considered as individual document to maintain its identity.

Approach:

1.) *Loading Data*; Different corpuses are created, one with tweets on a weekday and other on weekends.

2.) *Pre-processing of Text*; this step allows text for analysis, the steps involved are:

a) *Removing punctuations*: punctuation and other special characters look like more words to the computer therefore, they need to be removed
b) *Removing number*: The computer also considers numbers as words; therefore, they need to be removed as they do not influence topics.
c) *Stripping whitespaces*: white spaced between words just add to the dimensionality of the matrix and which is of no use therefore they are removed
d) *Stop Words/ Common Words*; many words e.g. is, are, the etc are frequently use in English language which are used for sentence formation but does not necessarily have an individual identity, therefore they are removed. Also along with stop words, some other words with are irrelevant to the analysis can also be removed in this step.
e) *Removing Capitalization*; all the words are kept to lower cases otherwise the computer will treat the same word differently. and most importantly,
f) *Removing Special and Alphanumeric Characters*; as there are many special characters used by people e.g. [â€œ@ðŸ˜Šš€?ðŸ˜™ÿ.~#^:/]. So removing them before analysis is important as I have observed that these alphanumeric and special character can affect the analysis drastically.
g) *Stemming*; removing common word endings like ing, es, s etc, as these words mean the same but are just ended in different manner.

3.) *Document-Term Matrix*: we create a document term matrix, which stages the data for further analysis.

4.) *Word Frequency*: Through this method we see the words which have a freq of more than 50, and removing the irrelevant words, by repeating the Pre-Processing phase in which all with stop common words, we added the corresponding words along with the stop words for removal from the analysis.

5.) *LDA*: latent Dirichlet Allocation is a topic modelling approach which is used to find the topic to which each document belongs to. It uses the Dirichlet distribution of Document-Topic matrix and Topic-Word matrix.

a) Within the LDA function I have set "*seed=300*" to maintain the reproducibility of the topics. (i.e but setting seed' s We can reproduce the same topics again and again)

6.) We observe the 10 most likely words in each topic and compare the terms spread in each topic from the models.

7.) Using the *posterior distribution* we can find the hidden distribution (Document-Topic Matrix, and Topic-Term Matrix), using which I have named each topic as, the top 5 most likely words appearing in that Topic.

	lol,love,like,amp,girl	lol,now,like,day,fuck	shit,like,got,now,want	feel,good,one,can,back	love,time,like,come,need
1	0.2045584	0.1976358	0.1999214	0.2001284	0.1977560
2	0.2016771	0.1982123	0.1971921	0.1964285	0.2064901
3	0.2036323	0.1971317	0.2017307	0.1990843	0.1984211
4	0.2020571	0.1955557	0.2031334	0.2049771	0.1942767
5	0.2011313	0.1995160	0.1888398	0.2071462	0.2033668
6	0.2001080	0.2003166	0.1987601	0.2008188	0.1999964

Fig4: Topic distribution for first 6 documents while showing the topic names for Weekdays contents

	love,need,day,got,back	like,lol,fuck,girl,love	lol,make,amp,like,time	now,fuck,guy,work,can	good,amp,like,love,one
1	0.1995593	0.2001334	0.2005151	0.2004395	0.1993527
2	0.2034816	0.2005484	0.1984625	0.1986689	0.1988385
3	0.1960929	0.1970466	0.2041673	0.2044928	0.1982004
4	0.1987183	0.2016575	0.1989337	0.2002357	0.2004548
5	0.2026666	0.1995522	0.1973415	0.1975984	0.2028413
6	0.1912014	0.1882673	0.1923713	0.1928352	0.2353249

Fig5: Topic distribution of first 6 documents while showing the topic names for Weekend contents

Inference:

As per our initial guess, both the contents are different this can be shown as the term distribution and term type of each topic for both corpus was different, which proves that the topics in both the corpus will be different (as the words inside topics are different), which implies that the content of both the corpus will be different.

```
          ## Weekday Content ##
     Topic 1 Topic 2 Topic 3 Topic 4 Topic 5
[1,] "lol"   "lol"   "shit"  "feel"  "love"
[2,] "love"  "now"   "like"  "good"  "time"
[3,] "like"  "like"  "got"   "one"   "like"
[4,] "amp"   "day"   "now"   "can"   "come"
[5,] "girl"  "fuck"  "want"  "back"  "need"
[6,] "know"  "amp"   "peopl" "see"   "see"
[7,] "ass"   "night" "back"  "fuck"  "day"
[8,] "good"  "can"   "one"   "got"   "fuck"
[9,] "bitch" "one"   "other" "think" "know"
[10,] "best"  "make"  "right" "amp"   "got"
```

Fig6: The word in each topic

```
          ## Weekend Content ##
     Topic 1 Topic 2 Topic 3 Topic 4 Topic 5
[1,] "love"  "like"  "lol"   "now"   "good"
[2,] "need"  "lol"   "make"  "fuck"  "amp"
[3,] "day"   "fuck"  "amp"   "guy"   "like"
[4,] "got"   "girl"  "like"  "work"  "love"
[5,] "back"  "love"  "time"  "can"   "one"
[6,] "shit"  "need"  "one"   "wait"  "know"
[7,] "nigga" "peopl" "think" "other" "now"
[8,] "fuck"  "know"  "see"   "come"  "other"
[9,] "great" "other" "back"  "let"   "shit"
[10,] "know"  "think" "bitch" "feel"  "ass"
```

Fig7: The word in each topic

---

3. Select a type of spatial entity to analyze (e.g., bars, schools, or something else) -- any shapefile provided by the Chicago Data Portal. For your selection, analyze the topical content of tweets posted within a 50-meter radius. For example, if you choose bars, present an analysis of tweets posted with 50 meters of a bar. Your analysis must use topic modelling, with the key output being the topic-word distributions. Are any of the topics interpretable? If so, what are they? If not, try reducing the radius to focus in on tweets that are closer to your target venues. Or try varying the number of topics. If you still cannot obtain interpretable topics, simply present your analysis and list the topic-word distributions.

Solution:

Shapefile: I have used the School's Shapefile, as teenagers are the most prominent users of twitter therefore, it makes more sense to capture their uses and as seen in last case study, the areas near schools show a linear relationship to crime.

Assumptions:

1.) I have captured tweets posted within a 100-meter radius rather than 50. Tweets within 100 meters makes more sense, as it better covers the area near the school and provides us a bigger sample to work on.

2.) As earlier, each tweet is considered an individual document (despite less text) to maintain the identity of each tweet, as different content tweets can get mixed up if different tweets are mixed.

Approach:

The approach remains the same as elaborated above. The only step added is:

Using the posterior of LDA (which provides us with the hidden variable i.e. Document-Topic matrix and Topic-Term Matrix). We find the Topic-word probability distribution i.e. probability of each word occurring in every Topic.

```
> topic.schl$terms[1:5, 60:65]
                          academ      academi       accept       access    accessori        accid
need,just,amp,see,good   2.063167e-04 0.0002562907 5.887177e-04 5.942820e-05 9.273117e-06 9.377322e-06
fuck,got,look,today,can  5.693691e-05 0.0006438486 1.824156e-06 1.117672e-04 1.123662e-04 2.300112e-07
like,shit,love,lol,now   1.435594e-04 0.0008955265 5.082820e-06 9.711182e-06 1.025168e-04 1.047651e-05
lol,like,want,bitch,good 9.383239e-05 0.0002604696 1.525863e-06 1.208721e-04 3.542674e-05 1.771550e-04
just,lol,make,got,even   1.225752e-04 0.0001249355 2.637620e-05 9.906953e-06 5.198274e-05 4.257779e-04
```

Fig8: The topic word probability distribution, with topic represented as top 5 most likely words in each topic

```
> terms(tweet.lda.schl, 10)
      Topic 1 Topic 2 Topic 3 Topic 4 Topic 5
 [1,] "need"  "fuck"  "like"  "lol"   "just"
 [2,] "just"  "got"   "shit"  "like"  "lol"
 [3,] "amp"   "look"  "love"  "want"  "make"
 [4,] "see"   "today" "lol"   "bitch" "got"
 [5,] "good"  "can"   "now"   "good"  "even"
 [6,] "never" "like"  "today" "fuck"  "like"
 [7,] "lol"   "bitch" "watch" "one"   "day"
 [8,] "time"  "ass"   "back"  "just"  "know"
 [9,] "readi" "day"   "amp"   "come"  "school"
[10,] "can"   "work"  "look"  "girl"  "fuck"
```

Fig9: Top 5 most likely words in each topic.

Inference:

I have chosen 5 topics for this problem, as we can see from the above analysis, we can interpret Topic 2, Topic 4 and Topic 5 as bad/susceptible topics, if we could follow the individuals with those documents/tweets, and we might be able to prevent some criminal activities. While individuals tweeting text belonging to Topic 1 and Topic 3, can be categorised in non-threat individuals.

Recommendations:

We can further continue our research to predict crime locations by clustering to find the tweets belonging to topic 2, 4 and 5 and then these clusters to find the location of hotspot area of the crime.

```
$`1`
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 619

$`2`
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 556

$`3`
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 607

$`4`
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 617

$`5`
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 628
```

Fig10: Total number of Documents in each Topic.

- ➔ We can see the document distribution as per the most likely topic (We can also use K-means clustering).
- ➔ We also see the topic distribution, with topic 2,4 and 5 coloured are red and Topic1 and Topic3 coloured as green while the schools are coloured in Yellow. We can see the relationship of schools with crime location hotspots here.

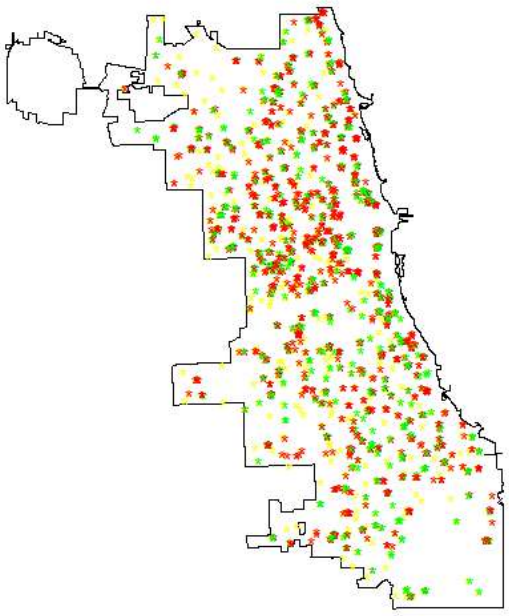

Fig11: Topic distribution along with schools in the city of Chicago

---

Bonus Opportunity

Integrate your topic analysis into a linear model of crime. The hypothesis to be tested is that topics generated at the current time are predictive of the occurrence of crime at a following time. You may select a single crime type and any time interval you wish. You do not need to consider any additional factors beyond topics. Successful completion of this task requires a complete description of the analysis performed as well as a listing of correlation coefficients for topics as they relate to the occurrence or non-occurrence of crime. If completed successfully, your lowest case study score will be dropped.

***S**olution:*

Data:

1. The data is collected from the City of Chicago data portal https://data.cityofchicago.org/Public-Safety/Crimes-2014/qnmj-8ku6

The data contains the reported cases of thefts cases from 01/01/2014 12:00:00am to 12/31/2014 11:58:00pm. It has 22 different fields ranging from the crime ID, Case Number, ICUR code, Time, Location, Longitude, Latitude, Description of Crime etc. The data contains 273348 reported crime cases.

2. A large sample of tweets corresponding to the year 2014, located within the boundary of Chicago.

The following features of the tweets are used for analysis

a. Time-Stamp
b. Location (longitude and Latitude)
c. The text in a tweet

Assumptions:

I have considered each tweet as one document as even if the number of words in each document is less (max 140 words), but mixing different tweets together can lead to unintentional content mixing, i.e. tweets belonging to different topic could be clubbed as one and after pre-processing of the document, the output can lead to different topics, then it would actually have meant to be in. Therefore, each topic is considered as individual document to maintain its identity.

Analysis:

1.) For the purpose of analysis I have taken the crime type as Theft.
2.) For training the data the theft reports in the month of January to March is considered and predictions are made on the month of April.
3.) Similarly tweets in the month of January to March are used for training the Topic model to create a Topic-Term Probability matrix.

Approach:

1.) A training data frame is created containing theft reports for the month of January to March.
2.) The tweets data frame is created containing tweets for the month of January to March.

3.) *Loading Data*; A Corpus is created containing the tweets from the month of January to March.

4.) *Pre-processing of Text*; this step allows text for analysis, the steps involved are:

a) *Removing punctuations*: punctuation and other special characters look like more words to the computer therefore, they need to be removed
b) *Removing number*: The computer also considers numbers as words; therefore, they need to be removed as they do not influence topics.
c) *Stripping whitespaces*: white spaced between words just add to the dimensionality of the matrix and which is of no use therefore they are removed
d) *Stop Words/ Common Words*; many words e.g. is, are, the etc are frequently use in English language which are used for sentence formation but does not necessarily have an individual identity, therefore they are removed. Also along with stop words, some other words with are irrelevant to the analysis can also be removed in this step.
e) *Removing Capitalization*; all the words are kept to lower cases otherwise the computer will treat the same word differently. and most importantly,
f) *Removing Special and Alphanumeric Characters*; as there are many special characters used by people e.g. [â€œ@ðŸ˜Šš€?ðŸ˜™ÿ.~#^:/]. So removing them before analysis is important as I have observed that these alphanumeric and special characters can affect the analysis drastically.
g) *Stemming*; removing common word endings like ing, es, s etc, as these words mean the same but are just ended in different manner.

5.) *Document-Term Matrix*: we create a document term matrix, which stages the data for further analysis.

6.) *Word Frequency*: Through this method we see the words which have a freq of more than 50, and removing the irrelevant words, by repeating the Pre-Processing phase in which all with stop common words, we added the corresponding words along with the stop words for removal from the analysis.

7.) *LDA*: latent Dirichlet Allocation is a topic modelling approach which is used to find the topic to which each document belongs to. It uses the Dirichlet distribution of Document-Topic matrix and Topic-Word matrix.

a) Within the LDA function I have set "*seed=300*" to maintain the reproducibility of the topics. (i.e but setting seed's We can reproduce the same topics again and again).
b) The value of alpha is chosen to be '1', as it promotes the significance of more dominant topic over each document/Tweet.
c) Three different models of LDA is created containing 3, 5, 8 topics.

8.) We observe the 5 most likely words in each topic for each model and topic distributions over documents to check the sparsity.

9.) To create a logistic regression model, first I have created a data frame with documents in rows/observations and topic probability distribution in the columns.

Response Variable:

The idea is that, every tweet that has been made close to crime location during the period under survey is considered as influential and assigned a response value of '1', while all the tweets that are far away from the crime location are allotted a response value of '0'. Therefore,

1. Using the theft locations I have captured all the tweets with located that are less than or equal to 150 meters for the month of January to March.
2. This is done for all the three models with number of topics 5, 8 & 3.
3. A logistic regression model is created for all the three models

For K= 5

The model had multi-co linearity which is confirmed by the use of 'VIF' and 'ALIAS' functions and therefore Topic 5 is removed.

The new model created is:

Response = 1.526 - 0.53*(Topic 1) -0.504*(Topic 2) + 0.061*(Topic 3) - 0.502*(Topic 4)

The coefficients are:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.52616    0.08460  18.040  < 2e-16 ***
Topic1      -0.53626    0.12887  -4.161 3.16e-05 ***
Topic2      -0.50476    0.13101  -3.853 0.000117 ***
Topic3      -0.06182    0.13091  -0.472 0.636767
Topic4      -0.50262    0.12947  -3.882 0.000104 ***
```

Fig12: Coefficients for the model with K=5

Checking the multi-collinearity

```
> vif(glm.fit2)
  Topic1   Topic2   Topic3   Topic4
1.619817 1.604639 1.642235 1.607172
```

Fig13: VIF values for the model to check multi collinearity

The F and P- values for the model.

```
> summary(aov(Response~Topic1+Topic2+Topic3+Topic4, data = tweet.data))
              Df Sum Sq Mean Sq F value   Pr(>F)
Topic1         1      1  1.3951   7.846 0.005097 **
Topic2         1      2  1.6159   9.087 0.002576 **
Topic3         1      0  0.4889   2.749 0.097307 .
Topic4         1      3  2.6455  14.878 0.000115 ***
Residuals  26483   4709  0.1778
---
```

Fig14: The F and P- values for the model with K=5.

For K= 8

The model had multi-co linearity which is confirmed by the use of 'VIF' and 'ALIAS' functions and therefore Topic 5 is removed.

The new model created is:

Response = -1.2313 - 0.1319*(Topic 1) -0.2260*(Topic 2) + 0.9781*(Topic 3) + 0.2752*(Topic 4) + 0.5723*(Topic 5) + 0.1965*(Topic 6) – 0.1858*(Topic 7)

The coefficients are:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.7707     0.1256   6.138 8.34e-10 ***
Topic1        0.3056     0.1900   1.609   0.1077
Topic2        0.3458     0.1947   1.776   0.0758 .
Topic3        0.9045     0.1894   4.775 1.80e-06 ***
Topic4        0.4128     0.1943   2.124   0.0337 *
Topic5        0.9722     0.1880   5.171 2.33e-07 ***
Topic6        0.2182     0.1905   1.145   0.2522
Topic7        0.3133     0.1942   1.613   0.1067
---
```

Fig15: Coefficients for the model with K=8

Checking the multi-collinearity

```
> vif(glm.fit4)
  Topic1   Topic2   Topic3   Topic4   Topic5   Topic6   Topic7
1.738016 1.695750 1.712681 1.678846 1.735765 1.705978 1.671757
>
```

Fig16: VIF values for the model to check multi collinearity

The F and P- values for the model.

```
> summary(aov(Response~. - x -y -text -Topic8, data = tweet.data))
              Df Sum Sq Mean Sq F value   Pr(>F)
Topic1         1      0   0.255   1.436 0.230733
Topic2         1      0   0.173   0.975 0.323368
Topic3         1      2   2.165  12.180 0.000484 ***
Topic4         1      0   0.002   0.011 0.915274
Topic5         1      5   4.811  27.064 1.98e-07 ***
Topic6         1      0   0.033   0.187 0.665633
Topic7         1      0   0.483   2.716 0.099356 .
Residuals  26480   4707   0.178
```

Fig17: The F and P- values for the model with K=8.

For K= 3.

The model had multi-co linearity which is confirmed by the use of 'VIF' and 'ALIAS' functions and therefore Topic 5 is removed.

The new model created is:

Response = 1.431 - 0.403*(Topic 1) -0.273*(Topic 2)

The coefficients are:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.43159    0.05231  27.370  < 2e-16 ***
Topic1      -0.40355    0.08259  -4.886 1.03e-06 ***
Topic2      -0.27322    0.08463  -3.228  0.00124 **
---
```

Fig18: Coefficients for the model with K=3

Checking the multi-co llinearity

```
> vif(glm.fit6)
  Topic1   Topic2
1.37191 1.37191
```

Fig19: VIF values for the model to check multi collinearity

The F and P- values for the model.

```
> summary(aov(Response~. - x -y -text -Topic3, data = tweet.data))
              Df Sum Sq Mean Sq F value  Pr(>F)
Topic1         1      3  2.5571   14.38 0.00015 ***
Topic2         1      2  1.8122   10.19 0.00141 **
Residuals  26485   4711  0.1779
```

Fig20: The F and P- values for the model with K=3.

Predictions:

As the correlation coefficient for all the three models is very close, we tend to prefer choosing the model with K=5, which also have the highest correlation coefficient.

Using the theory of ROC curves,

In which we define a grid and during our model training we sort the grid location from high crime probability grids to low crime probability grids.

We then fit the new crime locations into this grid to check what percentage of crime location actually satisfies our grid allocations. In doing so we use AUC as a metric through which we can compare different models.

Working on those lines, In our previous ROC curves in past case studies we have taken grid sizes to vary from 200 to 300meters i.e. covering an area around 90000m^2per grid but to here we will use the area around the influential tweet as 100 meters in radius.

Using the previous logistic regression we can predicted the tweets which belonged to near crime location, in doing so we had assigned a response value of ‘1’ making them influential tweets. Therefore, area around an influential tweet will act as a high crime probability area.

We compare our model for accuracy, we will find the percentage of new crime points that are located in the high crime occurring probability area. The higher the percentage the better the accuracy and this will act as out *METRIC*.

The area covered by each location obtained is taken to be 100 meters but as we increase the area the accuracy will also increase.

Benefits:

1.) This model provides us concise locations as compared to using grids, where the area is too large.
2.) Due to smaller area for predicting crime locations, several geographic factors like parks, lakes which get ignored when grid area is large in the previous analysis will now be taken into consideration.
3.) The influential tweets not only provide locations but also the most likely topic in these tweets can act as a fundamental topic of predicting crime. i.e. every tweets that belongs to this topic belong can be considered as a potential threat irrespective of the area and immediate action can be taken if further efforts are put into the analysis.

Observation:

For k=5 and considering 100 meters of radius, we get a 66% accuracy in predicting the threats.
