Case Study 5 (CS5): Text Mining

"On my honor, I pledge that I have neither given nor received help on this assignment."
Name: Sharath Chand PV
ID: vp4pa
Date: 11-7-2015

Introduction:

In this study we have analyzed the tweets originated in geographical boundaries of Chicago. In addition to understanding tweets' distribution we have applied topic-modeling techniques to extract inherent information from them. We then developed a methodology to use information from the topic modeling in crime prediction models.

Data

We have used two sets of data

1. Tweets collected in the geographical boundaries of Chicago
    a. Date range of 27th Oct 2012 to 15th April 2014
    b. 70,823 tweets are collected
    c. Each tweet is geo tagged (longlat locations are available)
    d. Data fields are Text, longitude, latitude, timestamp
2. *2014_THEFT.csv*
    a. All the theft incidents occurred in Chicago from Jan 1st 2014 to Dec 4th 2014
    b. Important fields are Date, time, location coordinates, location description, incident description, address, arrest made etc. are the important features

Preprocessing:

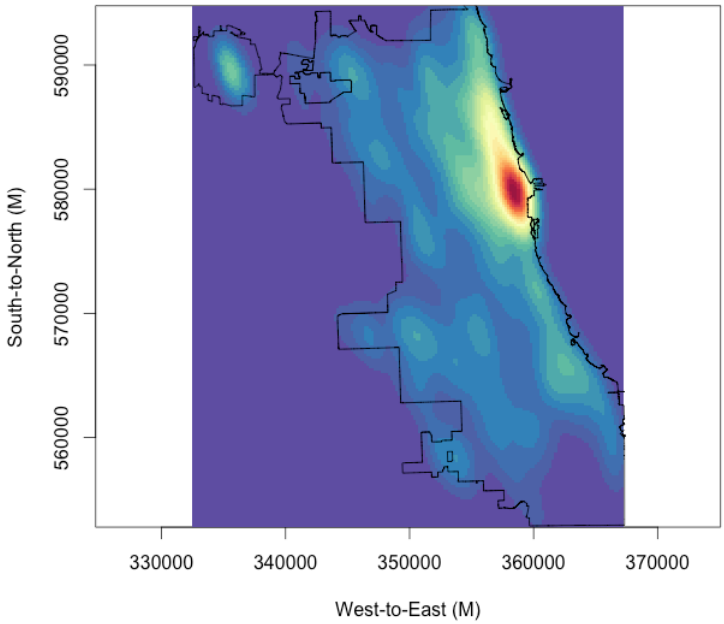
Following preprocessing steps are carried out

- Step1: Coordinate Ref system of city boundary is in meters. So the location coordinates of our data to be converted to the same CRS (epsg: 26971)
- Step2: parse the date variable (ex: 12/04/2014 11:30:00-05) to extract Hour of day, day of week and month etc.

Distribution of Tweets

Our hypothesis is tweets shall not be distributed uniformly across Chicago. To test the hypothesis we have calculated KDE estimate of tweets origination and plotted the same on city boundary map.

Following plot shows density estimate of tweets across Chicago. We can observe that one location has high density of tweets compared to rest of the city. That part of the city is downtown area and that explains the high tweet density



KDE plot of tweets originated in Chicago

As hypothesized, tweets are not uniformly distributed across the City.

Weekday Tweets Vs Weekend Tweets

To understand topical differences between weekday and weekend content we applied LDA (Latent Dirichlet Allocation) for topic modeling on tweets. We modeled topics on weekday tweets and weekend tweets separately. We compared words in the topics to understand topical differences.

Methodology:

We built topic models using LDA in which topic proportions are assigned following a Dirichlet distribution to a document and word proportions are assigned again following a Dirichlet distribution to topics.

Ti = Topic distribution for Doc(i)
Wk = Word distribution for Topic(k)
Both Ti and Wk are drawn from Dirichlet distributions with different prior parameters (as per the notation alpha and gamma respectively)

Step-by-step process:

a. Split the tweets in to two sets (weekday and weekend)
b. Run topic model on each set using LDA algorithm
    a. Create corpus from each set
    b. Clean the corpus and create term-document matrix
    c. Train topic model using LDA

d. Generate 100 topics for each dataset

c. Verify terms in each topic to understand any differences

d. We can compute the cosine similarity between words in the topic to measure the similarity between topics generated in weekday and weekend tweets. But we are restricting ourselves to verify the topics manually to understand topical differences

Comparison between topics:

Below table has first 12 topics obtained from topic modeling from weekend and weekdays tweets. We can observer the differences in content from the words in each topic.

Words like music, beach, lake Michigan, nap, bed, dope, church, movie, play, bike etc. are prominent in **weekend topics**

Words like bore, learn, tire, service, support, due, tie, create, period, report, matter etc. are prominent in **weekday topics**

	WEEKEND TOPICS											
	TOPICS 1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
[1,]	cant	best	street	like	call	god	can	bitch	year	will	face	guy
[2,]	wait	**music**	havent	**bed**	servic	hour	gone	hand	**movi**	**play**	minut	caus
[3,]	summer	check	seem	cta	femal	bring	**church**	breakfast	next	song	**river**	die
[4,]	cri	money	**beach**	flight	later	buy	kick	realiz	shot	front	east	ima
[5,]	**hotel**	soon	complet	smile	answer	town	crib	fat	cold	five	catch	dat
[6,]	forev	video	**lake**	**dope**	block	definit	shoe	possibl	drop	stick	cloud	coffe
[7,]	tower	wat	**michigan**	driver	convo	heat	con	motiv	took	alot	thot	read
[8,]	drake	ago	**nap**	cab	custom	daddi	fri	argu	min	solo	messag	yet
[9,]	mid	control	king	chines	squar	door	addict	extrem	special	**bike**	nbs	anyon
[10,]	tear	glass	bridg	honey	kiss	weak	oop	tongu	order	goodby	normal	met
	WEEKDAY TOPICS											
	TOPICS 1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
[1,]	week	hot	think	tweet	bitch	find	amp	movi	next	phone	hoe	fuck
[2,]	everi	ladi	keep	art	nigga	okay	lie	**tire**	hey	guess	face	done
[3,]	stay	shop	mani	restaur	wanna	soon	**servic**	sex	favorit	iphon	anoth	excit
[4,]	away	**bore**	wont	**paper**	dat	may	**support**	**due**	meet	bag	awesom	whole
[5,]	finna	**learn**	train	possibl	alon	hahaha	lock	pub	seen	**report**	wow	smile
[6,]	half	save	end	**institut**	drunk	part	step	cab	**finish**	unless	came	lay
[7,]	fresh	**mood**	minut	justin	dnt	**number**	sandwich	limit	apart	molli	rememb	cus
[8,]	updat	arent	women	wild	whos	via	wors	loft	bruh	sooooo	ugh	foot
[9,]	wall	creat	matter	appar	anyway	chees	cooki	tie	toward	war	far	toma
[10,]	mari	period	thot	tha	luv	irrit	**rule**	breweri	awak	florida	piss	boot

Below table has 13 to 24 topics obtained from topic modeling from weekend and weekdays tweets. We can observer the differences in content from the words in each topic.
Words like holiday, season, yay, chillin, sleepin, smoke, dance, Sunday, cheer, entertainment etc. are prominent in **weekend topics**
Serious words like follow, comment, student, news, power, trade, goal, whitesox, comment, mac etc. are prominent in **weekday topics**

	WEEKEND TOPICS											
	13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	
[1,]	come	love	see	still	mad	win	first	done	new	**night**	lol	mo
[2,]	start	beauti	though	show	asl	brother	turn	**sunday**	listen	find	bull	eve
[3,]	**season**	downtown	wasnt	walk	close	pick	christma	**danc**	ampamp	anyth	serious	wo
[4,]	cup	**yay**	swear	lot	drive	seen	stuff	water	suppos	sometim	mouth	cha
[5,]	view	gtgtgtgt	tap	earli	speak	oscar	fact	top	parad	woke	share	exc
[6,]	**holiday**	cooki	**sleepi**	**smoke**	chocol	glad	homework	futur	step	luck	rose	pla
[7,]	figur	beyond	wed	act	lord	main	annoy	shout	earth	straight	seat	for
[8,]	**chillin**	grow	count	dress	voic	public	woman	troubl	silli	lolla	scream	soo
[9,]	born	grown	behind	cut	dri	pub	chitown	fish	gorgeous	thru	quick	join
[10,]	stress	**vacat**	longer	knock	huge	road	past	across	molli	**entertain**	**cheer**	pull
	WEEKDAY TOPICS											
	13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	
[1,]	**follow**	still	beauti	that	best	lol	free	white	time	come	everyth	ive
[2,]	show	outsid	sorri	guy	hour	morn	wit	proud	can	thought	bring	son
[3,]	enjoy	woman	**reason**	eat	**team**	dream	theyr	sox	anyon	happen	bye	wo
[4,]	jus	whatev	super	food	must	idk	nobodi	basebal	shes	kill	plane	wea
[5,]	lord	spot	cloth	**power**	seem	bae	hawk	mlb	market	car	downtown	fou
[6,]	**student**	jordan	booti	cat	beach	**news**	speak	squar	clear	tryna	mile	kne
[7,]	miami	pack	color	grab	loop	instagram	spend	child	page	sleepi	quot	sou
[8,]	**comment**	goal	omfg	corner	drug	block	kiss	respond	respect	footbal	selfi	wo
[9,]	april	neighborhood	carri	mac	folk	ampamp	air	except	paint	posit	bathroom	me
[10,]	youtub	sun	general	whenev	**trade**	notic	sis	**whitesox**	**system**	faith	million	coa

In general we can observe the tone of weekend topics from the words distribution is lighter, cheerful and representative of weekend activities. On the contrary, weekday topics are more serious, less of cheerful words and representative of work related activities

The above analysis shows that tweets can give important information about people, situations and surrounding.

Analysis of Tweets around Geographical Locations

To understand topical relationship of tweets with a given geographical location type, we have obtained topics discussed in the tweets originated in the vicinity of the location.

For this purpose, we have analyzed tweets around two spatial factors **hospitals** and **farmers markets**. The following shapefiles downloaded from Chicago data portal

1. farmers_markets_2012 (points shape file)
2. Hospitals (points shape file)

Following is the methodology that we followed for this analysis

a. Get all the tweets originated with in 50 meters radius from a given geographical location type (Hospital or Farmers market)
b. Build topic models on the extracted tweets
    a. Create corpus
    b. Clean the corpus and create term-document matrix
    c. Train topic model using LDA
    d. Generate optimum number of topics based on relevance of their word composition
c. Check whether the words relevance is improved by changing the radius and number of topics
d. Verify terms in each topic to understand relationship with the given location type

Tweets around Hospitals

We have tried the radii of 100 meters, 50 meters and 20 meters. With 50 meter radius our topics obtained from LDA started making better sense. We have modeled 8 topics on the tweets originated with in 50 meters radius from all the hospitals in Chicago. Following is the words distribution in those topics

We can observe that **topic 8** relevance to hospitals with words like rehabilitation, institution, hospital, check, clinic, build, center, drug, doctor etc.

Most likely words in Topics around hospitals							
1	2	3	4	5	6	7	8
episod	gtgtgtgtgt	hospit	alway	feel	adida	like	hospit
girl	ean	later	christmas now	doesnt	chicago	nickfriedel	chicago
work	just	tell	countdown	happi	now	waddleandsilvi	institut
told	dat	ask	cuepan	like	anybodi	account	rehabilit
amp	good	bayarea	done	nay	anyth	babybossx	alreadi
ass	head	bitch	eat	adapt	avenu	becar	april
bethani	amazing and	call	everi	aliv	blow	better	best
colortheroynerd	avail	can	exhaust	beemahri	brother	busi	bitch
come	best	chicago	favorit	birthday	buy	canelo	build
detox	caus	children	forreal	chrome	can	delet	center
different but	chadsand	compliment sessssssi	free	come	center	demetriousdailyy	check
disney	chelsculv	cook	friday	day	charact	exact	clinic
easi	chicago	counti	give	dem	code	fail	day
elmurf	china	die	got	didnt	determin	father	doctor
even	danc	dont	hahaha	enjoikri	dope	follow	drug
follow	decemb	farley	here	explod	everi	gone	fbi
funni	doesnt	girlpost	hey	girlbandrock	far	gonna	forgot
give	dog	give	httptcojbuxiqabzi	happybirthday	ghetto	hella	friend
good	dontanswerthat	gotten	hungri	httptcosqnlmlmdj	good	httptcoisuskilxiz	httptcombbygasa
got	float	home	ideservemor	imedina	hear	igetbucket	httptcomoqeyynul

Tweets around Farmers Markets

We have modeled 8 topics on the tweets originated with in 30 meters radius from all the farmers markets locations in Chicago. Following is the words distribution in those topics

We can observe that **topic 4** relevance to food and drinks with words like coffee, starbucks, barista, cocktail, cigarettes etc. In general across all the topics we can see food and market related content.

Most likely words in Topics around hospitals							
1	2	3	4	5	6	7	8
bank	chicago	free	haha	day	christkindl market	plaza	christkindl market
know	stop	got	cant	anyway	year	chicago	other
ohm	bar	headph on	dude	bar	happi	printer	aint
whoiswill i	love	ill	empti	bcschampi onship	□\u26c4□	row	andersonvi ll
good	properti	special	everyon	beauti	actual	hackney	anoth
chicago	solut	time	man	birthdayho od	asymmetr	peopl	beltran
atm	wicker	amp	outsid	bradi	boobc	yes	carlo
big	park	aweso m	total	callmekier aaa	can	other	carlosyesca
chicagoni ghtlif	amp	bad	coffe	caus	christkindl markt	alchi	center
close	arianakissedm	bonus	amass	champions hip	classi	barloui	citi
daaaanc	basic	bottl	amaz	convert	daleyplaza	beer	come
danc	begin	brad	barista	cost	day	biketowor kweek	comedi
dinner	better	bradley	bummb	dcantu	design	bikram	copernicus
disappoi nt	call	bring	busi	download	deutsch	blain	crazi
djcharlie boy	christlkindlmar ket	bus	can	dude	dont	booth	debat
droppin	citi	butchm cguir	caus	eek	dress	cargobik	els
effin	colleg	calebza hm	christm aschi	even	easter	chuggin	fair
fan	come	central	cigarett	fight	excit	church	german
festiv	cornersdamen milnorth	char	cocktail serv	friend	eye	cold	guitar

Bonus Problem

The idea is the information generated from the tweets can help us predicting occurrence of crime. We hypothesized that topics discussed 1 month prior to a given day has correlation with the occurrence of crime on that day. We have considered theft cases in our crime prediction model.

We have tweets data available from Oct 2012 to 15th Apr 2014, but the Thefts data is only for the year 2014. So, we considered following durations for training and testing purposes

Training Data:

1. Response: Thefts occurred on 1st March 2014 are our responses along with randomly selected non-crime points
2. Predictors: The topics distribution of tweets originated in the 1 month prior to 1st Mar 2014 (i.e February 2014 tweets)

Validation Data:

1. Response: Thefts occurred on 1st Apr 2014 are the ground truths against which our theft predictions are validated.
2. Predictors: The topics distribution of tweets originated in the 1 month prior to 1st Apr 2014 (i.e Apr 2014 tweets)

To understand whether the tweets could help in prediction of crimes, we have built a logistic regression model with topics extracted from tweets as predictors. We the evaluated the model strength using validations and surveillance plots as mentioned below

Step-by-step Methodology:

Predictors (Base data) – Feb 2014 tweets data
Response (Train data) – 1st March 2014 theft data
Predictors (Validation base data) – March 2014 tweets data
Ground truths (Validation data) – 1st April 2014 theft data

Following is the step-by-step approach taken to build models

1. Train data preparation:
    a. In our train data (1st March theft) we have got only those locations (coordinates) where the theft has occurred, called as positive observations. To train our classification model we have added some locations where the theft has not occurred, called as negative observations.
    b. Add a response variable, 1 for positive observation and 0 for negative observation
2. Adding topics extracted from tweets to train data:
    a. Get all the tweets in the vicinity of each point in the crime data which will be called as document of the point

   b. The document is all the tweets originated with in 500 meters from the point
   c. Create a corpus with all the documents of train points and build topic model using LDA
   d. We have built 10 topics with a given control seed.
   e. Get probability distribution of topics for each document. It means probability of each topic out of 10 topics discussed around a train data point
   f. The extracted probabilities of topic distribution (matrix of n rows X 10 columns) will be our predictors for the training model, where 'n' is the number of points in train data
   g. Note that sum of topical probabilities of a document is 1. Intuitively, the topic with highest probability for a given point / document can be considered as representative of tweets' content in that neighborhood
3. Build logistic regression model for binary classification, Response as response variable and topical probabilities as predictors.
   a. Check coefficients' significance
   b. Identify and review the topics with lower p-values
4. Validate the model using 1st April 2014 thefts as ground truths and March 2014 tweets data as predictors
   a. Get prediction points – Uniformly selected points across the city
   b. Get March 2014 tweets data
   c. Get topics probability distribution for tweets around prediction points
   d. Prepare validation dataset for prediction.
   e. Predict using the Logistic regression model built in the above steps
5. Check the prediction effectiveness using Surveillance plots

Logistic regression:
Following is the summary of logistic regression model built using topics as predictors and theft occurrence as response

We can notice that Intercept, topic 2, topic 6 and topic 12 are seem to be significant given their p-value

```
Call:
glm(formula = response ~ ., family = binomial, data = train.model[,
   -c(2:4, ncol(train.model))])

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.6548  -0.7299  -0.5729  -0.4997   2.0827

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.80423    0.56282  -3.206  0.00135 **
V1           2.23049    1.38754   1.608  0.10794
V2           4.09755    2.22139   1.845  0.06510 .
```

```
V3        0.01231    0.67199   0.018  0.98538
V4        0.70451    0.85408   0.825  0.40944
V5       -0.11811    0.70349  -0.168  0.86667
V6        2.11294    1.22486   1.725  0.08452 .
V7        0.12204    0.71298   0.171  0.86409
V8        0.65116    0.63511   1.025  0.30524
V9        0.24886    0.69300   0.359  0.71951
V10       0.75553    1.10837   0.682  0.49546
V11      -0.24721    0.76273  -0.324  0.74585
V12       1.77815    0.86531   2.055  0.03989 *
V13       1.14901    0.78912   1.456  0.14538
V14       3.30965    1.73855   1.904  0.05695 .
V15       0.65169    0.66833   0.975  0.32950
V16       1.48166    0.95120   1.558  0.11931
V17       0.54736    0.63744   0.859  0.39052
V18      -0.01985    0.73160  -0.027  0.97835
V19       1.05692    0.69998   1.510  0.13106
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 650.14  on 625  degrees of freedom
Residual deviance: 615.45  on 606  degrees of freedom
AIC: 655.45
```

10 most likely terms in topic 6 and 12:

Topic 6	**Topic 12**
just	just
just	like
businessmgmt	love
amp	night
day	dont
like	amp
still	can
tweetmyjob	time
now	got
manag	great

Log of (probability of a location has theft / probability of location doesn't have theft)

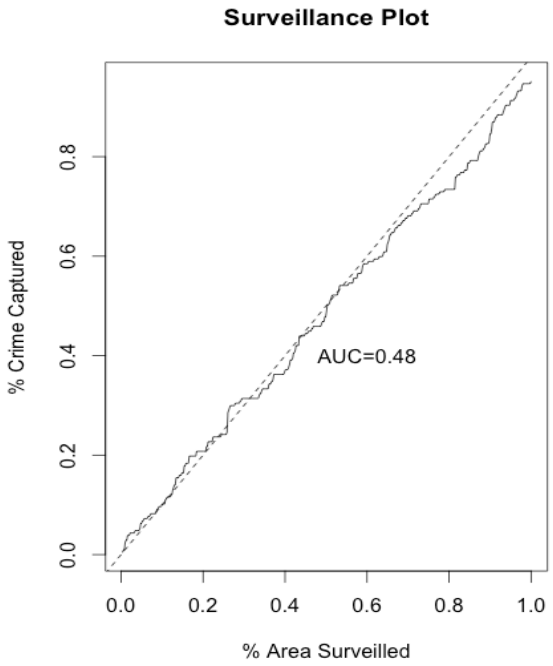
= -1.80423 + 4.09755 * V2 + 2.11294 * V6 + 1.77815 * V12 + 3.30965 * V14
Vi = Topic i

Intuitively it can be interpreted as, if the topics discussed on the tweets around a particular point has topic similarity with Topic 2, 6, 12 or 14 then the risk of theft near that point is higher

Testing Challenge:
The topic positions obtained in the test might vary w.r.t training topics. Since the coefficients in the logistic regression are as per the topic positions of the train dataset, it will mislead our predictions if test topic probabilities obtained from the LDA are directly used for the prediction.

We have used **cosine similarity technique** to arrange the test topics in the same topical sequence as train topics. It will ensure the consistency of coefficients of topics used to build the model and used in the prediction

We have predicted theft occurrence using topics from March 2014 tweets as predictors and validated the predictions against actual theft occurrence on 1st APR 2014. Following is the AUC curve



Though the predictive performance of the topic model is not significant as appeared in the above curve, it can be used along with other strong predictors such as historical crime occurrence (KDE), spatial factors, and other risk factors specific to demographics. Our analysis establishes that information generated on social media (twitter, facebook etc.) can be considered in improving the traditional crime prediction models.

References:
Predicting Crime Using Twitter and Kernel Density Estimation. Matthew S. Gerber