

STA312H5S Tutorial 4

Matthew C. Scicluna

University of Toronto Mississauga

February 10, 2015

What We Are Going to Cover

A bit more information about this

Today we are going to discuss the answers to assignment 2 and talk about project 2

Assignment 1 Part 1

```
#This is the code I used to clean up the data
mycorpus<-Corpus(dataset, readerControl=
list(language="eng", reader=readPlain))
mycorpus <- tm_map(mycorpus, removePunctuation)
mycorpus <- tm_map(mycorpus, removeNumbers)
mycorpus <- tm_map(mycorpus, stemDocument)
stopwordseng=stopwords(kind = "en")
mycorpus <- tm_map(mycorpus, removeWords, stopwordseng)
mycorpus <- tm_map(mycorpus, content_transformer(tolower))
```

Assignment 1 Part 1

#just print the SK matrix! Easy money.

```
> round(SK)
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]
[1,]	87	0	0	0	0	0	0	0	0	0	0	0	0
[2,]	0	41	0	0	0	0	0	0	0	0	0	0	0
[3,]	0	0	36	0	0	0	0	0	0	0	0	0	0
[4,]	0	0	0	32	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	26	0	0	0	0	0	0	0	0
[6,]	0	0	0	0	0	25	0	0	0	0	0	0	0
[7,]	0	0	0	0	0	0	24	0	0	0	0	0	0
[8,]	0	0	0	0	0	0	0	21	0	0	0	0	0
[9,]	0	0	0	0	0	0	0	0	21	0	0	0	0
[10,]	0	0	0	0	0	0	0	0	0	19	0	0	0
[11,]	0	0	0	0	0	0	0	0	0	0	17	0	0
[12,]	0	0	0	0	0	0	0	0	0	0	0	16	0
[13,]	0	0	0	0	0	0	0	0	0	0	0	0	15

Assignment 2 Part 2

This part is a bit more tricky, but to show relationships between the clusters I partitioned the data by using the logical vectors Cluster1vec and Cluster2vec to index the rows based on what cluster they came from.

```
>Cluster1Vec=unlist(KMEANSTEST[1])==1
```

```
>Cluster2Vec=unlist(KMEANSTEST[1])==2
```

```
> head(data[Cluster1Vec,])
```

	Last.Name	First.Name	TDCJ.Num	Age	Date	Race	Co
1	Ladd	Robert	999237	57	01/29/2015	Black	S
3	Paredes	Miguel	999400	32	10/28/2014	Hispanic	P
4	Coleman	Lisa	999511	38	09/17/2014	Black	Tar
5	Trottie	Willie	999085	45	09/10/2014	Black	Ha
6	Villegas	Jose	999417	39	04/16/2014	Hispanic	Nu
7	Hernandez	Ramiro	999342	44	04/09/2014	Hispanic	

Assignment 2 Part 2

This is how I tested to see if race was significantly different in each cluster.

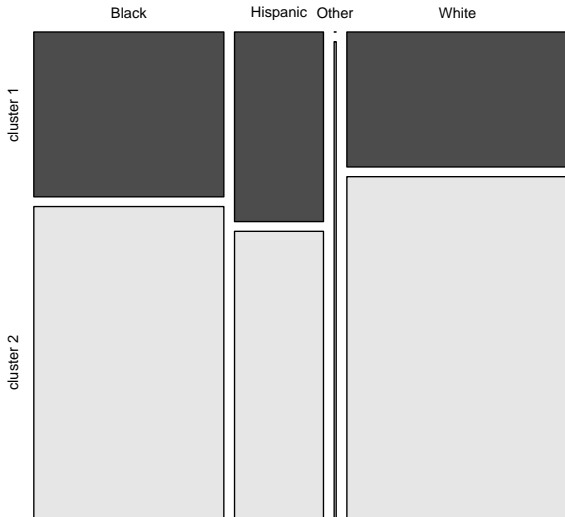
```
> cbind(table(racecluster1), table(racecluster2))  
      [,1] [,2]  
Black    65  123  
Hispanic  35   53  
Other     0    2  
White    62  157  
> fisher.test(cbind(table(racecluster1),  
table(racecluster2)))
```

Fisher's Exact Test for Count Data

```
data:  cbind(table(racecluster1), table(racecluster2))  
p-value = 0.1642  
alternative hypothesis: two.sided
```

You Could Also Use a Picture, Too

Mosaic Plot for Race Vs Cluster



Assignment2 Part 2

This is how I tested to see if county was significantly different in each cluster

```
#Get rid of the counties with less than 10 inmates
>x=table(countycluster1)
>x=x[x>10]
>y=table(countycluster2)
>y=y[y>10]
>z=cbind(x,y)
> z
```

	x	y
Bexar	12	27
Dallas	16	33
Harris	33	82
Tarrant	14	21

Assignment 2 Part 2

```
> fisher.test(z)
```

Fisher's Exact Test for Count Data

```
data:  z
```

```
p-value = 0.6421
```

```
alternative hypothesis: two.sided
```

```
#No evidence to suggest county is affecting the clustering
```

Assignment 2 Part 2

Now to check if the date of execution has any effect on the clustering

```
#The year of the execution was extracted from the exact date  
>year=substr(data[,6],7,10) #Gets the year  
>yearcluster1=year[unlist(KMEANSTEST[1])==1]  
>yearcluster2=year[unlist(KMEANSTEST[1])==2]  
>t.test(as.numeric(yearcluster1),as.numeric(yearcluster2))
```

Assignment 2 Part 2

```
>t.test(as.numeric(yearcluster1),as.numeric(yearcluster2))
```

Welch Two Sample t-test

```
data:  as.numeric(yearcluster1) and  
as.numeric(yearcluster2)
```

```
t = 1.6236, df = 337.634, p-value = 0.1054
```

```
alternative hypothesis: true difference in means is not  
equal to 0
```

```
95 percent confidence interval:
```

```
-2.051887 21.455352
```

```
sample estimates:
```

```
mean of x mean of y
```

```
2004.735 1995.033
```

```
#The date seems to have no effect either on the clustering
```

Assignment 2 Part 2

Now to test age

```
>age=data[,5]
>agecluster1=age[unlist(KMEANSTEST[1])==1]
>agecluster2=age[unlist(KMEANSTEST[1])==2]
>t.test(agecluster1,agecluster2)
```

Welch Two Sample t-test

```
data: agecluster1 and agecluster2
t = -0.6505, df = 311.356, p-value = 0.5158
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -2.099171  1.056017
sample estimates:
mean of x mean of y
 39.02469  39.54627
#Hmmm doesn't seem like age affected the cluster either??
```

What could have caused the clustering then!?!?

Assignment 2 Part 3

Lets dig a little deeper into what kinds of words convicts in each cluster were using. We will build term document matrices for each cluster and look at word counts.

```
TDMC1=TDMMat[,unlist(KMEANSTEST[1])==1]
```

```
TDMC2=TDMMat[,unlist(KMEANSTEST[1])==2]
```

Assignment 2 Part 3

Lets look at the rows of TDMC1 and TDMC2 to see if we can spot a pattern...

>TDMC1

Docs																		
Terms	1	3	4	5	6	7	10	15	17	19	20	21	22	23	24	26	29	3
and	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ask	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
can	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
caus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
come	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
death	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
declin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
done	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	

Assignment 2 Part 3

Lets look at the rows of TDMC1 and TDMC2 to see if we can spot a pattern...

```
>TDMC2
```

Docs

Terms	2	8	9	11	12	13	14	16	18	25	27	28	31	32	33	34	38
and	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
ask	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
can	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0
caus	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
come	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0
death	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0
declin	0	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0
done	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Assignment 2 Part 3

#How often on average did the word love appear in each document...

```
> mean(TDMC1[27,])
```

```
[1] 3.123457
```

```
> mean(TDMC2[27,])
```

```
[1] 0.6776119
```

Assignment 2 Part 3

```
> t.test(TDMC2[27,],TDMC1[27,])
```

Welch Two Sample t-test

```
data:  TDMC2[27, ] and TDMC1[27, ]  
t = -10.4758, df = 190.772, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not  
equal to 0  
95 percent confidence interval:  
 -2.906371 -1.985319  
sample estimates:  
mean of x mean of y  
0.6776119 3.1234568  
#Finally, statistical significance...
```

Assignment 2 Part 3

#How often the word 'declined' appeared in each document,
on average

```
> mean(TDMC1[7,])
```

```
[1] 0
```

```
> mean(TDMC2[7,])
```

```
[1] 0.3014925
```

And the Moral of the Story is...

- ▶ All the offenders who declined to make a last statement were clustered together, apart from the inmates who used the word 'love' a lot.
- ▶ Of course, each cluster is different, so whatever you get will be different each time you run the algorithm!

A Final Bit of Fun

I transformed the cluster centres to see what they would 'look like' in word space!

```
clustercentres<-as.vector(KMEANSTEST[2])  
c1<-clustercentres$centers[1,1:13]  
c2<-clustercentres$centers[2,1:13]  
c1r<-t(as.matrix(c1))%*%SK%*%t(TK)  
round(c1r)  
c2r<-t(as.matrix(c2))%*%SK%*%t(TK)  
round(c2r)  
#Cool huh?
```

So What do They Look Like?

```
> round(c1r)
      and ask can caus come death declin done dont famili fo
[1,]    0    0    0    0    0    0    0    0    0    0    1
      give god heart hope jesus just keep know last let life
[1,]    0    0    0    0    0    0    0    0    1    0    0
      make offend one pain peac peopl readi say see sorri st
[1,]    0    0    0    0    0    0    0    0    0    0    0
      take tell thank that the thing this want warden will y
[1,]    0    1    1    0    0    0    0    0    1    0    1
```

So What do They Look Like?

```
> round(c2r)
      and ask can caus come death declin done dont famili fo
[1,]    0    0    0    0    0    0    0    0    0    0    1
      give god heart hope jesus just keep know last let life
[1,]    0    0    0    0    0    0    0    0    1    0    0
      make offend one pain peac peopl readi say see sorri st
[1,]    0    0    0    0    0    0    0    0    0    0    0
      take tell thank that the thing this want warden will y
[1,]    0    0    0    0    0    0    0    0    0    0    0
```

A Bit About Project 2

- ▶ You will be implementing some code that samples movies from IMDB.
- ▶ You will analyze the results and using techniques learned in class and in tutorial
- ▶ You will be writing a report and submitting it to me.
- ▶ You will have to do a unique analysis on the code, this is an opportunity to explore some of the techniques taught to you from tutorial (i.e. kmeans, LSA, etc...)

Some Pseudocode for Program Used in Project 2

```
Input starting movie;  
Initialize Excel file to write to;  
For i in 1 to 10000 do:  
    Go to the movies IMDB webpage;  
    Run beautifulsoup on the page;  
    Record movie information in Excel file;  
    Pick an actor/actress in the movie at random;  
    Run beautiful soup on their webpage;  
    Pick a movie at random they acted in;
```