

# STA312H5S Tutorial 3

Matthew C. Scicluna

University of Toronto Mississauga

February 3, 2015

# What We Are Going to Cover

A bit more information about this

Today we are going to go over assignment 2 and talk about Latent Semantic Analysis

# Firstly, what is Latent Semantic Analysis?

The Document Term Matrix can be thought of as a large vector space, where each axis represents the frequency of each particular word. Points in this space correspond to documents. If a word occurs frequently in a particular document, that document will have a large projection onto the axis representing the frequency of that word. Basically LSA is finding a lower dimensional representation of this space using a singular value decomposition. Specifically we do this to the transpose of  $X$ , the Term Document Matrix.

*Whats that?*

Suppose we have a  $m \times n$  Term Document matrix  $X$  (this means we have  $m$  terms from  $n$  documents). We can do a singular value decomposition to get  $X = U \Sigma V'$ , where  $U, V'$  orthogonal and  $\Sigma$  diagonal. The  $k$  largest singular values are selected from  $\Sigma$  and these can be used to approximate  $X$ .

We can check that this makes sense since  $U$  is  $m \times k$ ,  $\Sigma$  is  $k \times k$  and  $V'$  is  $k \times n$ , so  $U \Sigma V'$  is  $m \times n$

Finally, to make a  $k$  dimensional representation of a document (ie a row in  $X$ , which is normally  $m$  dimensional) we can do the simple transformation:

$$\Sigma^{-1}U'X$$

And notice that  $\Sigma^{-1}$  is  $k \times k$ ,  $U'$  is  $k \times m$  and  $X$  is  $m \times n$ . so we have that this new matrix has dimension  $k \times n$

**See how we went from having  $m$  dimensions to having  $k$ ??**

# The lsa Package

A short introduction to the lsa package

<http://cran.r-project.org/web/packages/lsa/lsa.pdf>

Always read the source documentation of a package before you use it!

## Assignment 2

You are going to analyze the prisoners data, you have been provided code that builds the corpus object from the excel file with the last statements in it. You are going to do an LSA on it, and then run a kmeans on it to try to find 2 clusters. You will then try to figure out what caused the 2 clusters. Most of the code to do this has already been provided for you.

## Look at the LSA object

The LSA object returns a list of the matrices from the singular value decomposition.

```
TK=as.matrix(as.data.frame(LSAOBJ[1]))  
#In notes, I called this U  
DK=as.matrix(as.data.frame(LSAOBJ[2]))  
#I called this V  
SK=as.matrix(as.data.frame(LSAOBJ[3]))  
#I called this sigma  
#This next bit of code makes SK a diagonal matrix  
SK=diag(SK[,1])  
# This code gets the lower dimensional reconstruction  
recon=solve(SK)%*%t(TK)%*%TDMred
```



# K Means Clustering

**K means Clustering** is an algorithm that groups data into k categories. In our case we want to cluster it into 2 categories.

```
#This code will perform the kmeans algorithm on k=2  
KMEANSTEST=kmeans(t(recon),2)
```

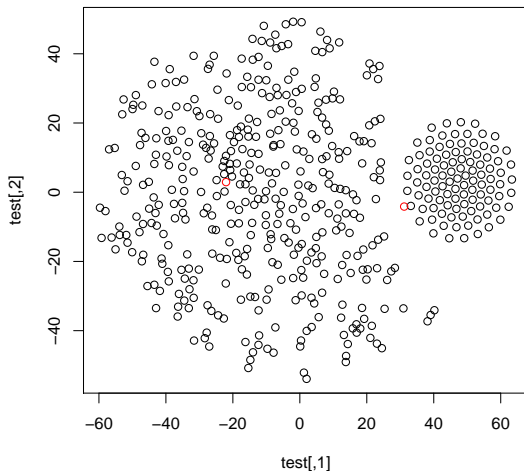
```
#This code will make 2 vectors that indicate which  
documents are in which cluster
```

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

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

## A Surprising Result

I used an alternative dimensionality reduction technique called tsne and reduced the dimension to 2. I then plotted the cluster centres and found this interesting clustering pattern.



# Grading Scheme

(3 marks) Run an LSA successfully and report to me the value of your  $\Sigma$  matrix.

(5 marks) Did County, Race, Date Of Execution or Age seem to affect which cluster each prisoner belonged to? Provide some proof.

(2 marks) Find out what the principle difference is between inmates in each cluster.

*good luck!*