Examining the Death Row Inmates of Texas

Gonzalo Bustos
Duke University
gonzalo.bustos@duke.edu

Drew Jordan
Duke University
drew.jordan@duke.edu

Sarah Normoyle
Duke University
sarah.normoyle@duke.edu

Megan Robertson
Duke University
megan.robertson@duke.edu

Abstract

We investigate the similarities and differences among the inmates who have been executed in Texas using text mining and clustering techniques. The crimes and last statements of each inmate are assigned a topic through Latent Dirichlet Allocation (LDA). The k-modes clustering algorithm is used to group similar inmates and determine if similar inmates touch on the same subjects in their last statements. This project provides insight into the characteristics of executed prisoners in Texas, and extracts information from the descriptions of their crimes and last statements.

1 Introduction

The United States has executed 1,421 inmates since 1976 and 26 have been killed so far in 2015.[1]. The state of Texas is responsible for the most executions in the United States since 1982.[2] There have been 531 inmates executed in Texas since 1976.[3].

We use topic models, specifically Latent Dirichlet Allocation (LDA), in order to learn more about the descriptions of the crimes and the last statements of inmates who have been executed in Texas. LDA allows us to classify the crime and each last statement by a topic. We also use clustering methods to determine which inmates are similar based on demographic data available on the Texas Department of Criminal Justice website. Thus, we are able to examine the similarities between inmates in each cluster and the topics of their last statements.

This project provides us with the opportunity to better understand the psychological nature of inmates sentenced to Death Row. Analyzing the text of last statements and crime descriptions can potentially provide insight into the recurring themes of last statements as well as the crimes committed that result in the death sentence. Our project opens new research venues into the extremes of human behavior and as well as provides information about human behavior before death.

[1] Death Penalty Information Center, ”Execution List 2015”, http://www.deathpenaltyinfo.org/execution-list-2015

[2] Death Penalty Information Center, ”Texas”, http://www.deathpenaltyinfo.org/texas-1

[3] Texas Department of Criminal Justice, ”Executed Offenders”, https://www.tdcj.state.tx.us/death_row/dr_executed_offenders.html

2 Methods

2.1 Latent Dirichlet Allocation

The first step of analysis was to assign each inmate a topic for the crime that they committed and their last statement through topic modeling. This method is used to analyze documents that do not contain labels, and the topic is a probability distribution of a collection of words.[4] LDA results in a model that "describes how the documents in a data set were created".[5]. The model used in LDA is a Bayesian mixture model that assumes that the topics are uncorrelated.[6] For more information about topic models and LDA see the References section.

LDA requires that the number of topics in the documents be specified beforehand. This is an open question for which multiple strategies have been proposed. In order to determine the number of topics for both the crimes committed by the inmates and their last statements, the R package "ldatuning" was used. This package implements four proposed measurements for determining the number of topics (found in Arun, 2010; Cao Juan, 2009; Deveaud, 2014; Griffiths, 2004). The measurements proposed by Arun et al. and Cao Juan et al. should be maximized at the proper number of clusters whereas the other two measurements should be minimized. For more information on the package and references of the measurements, see Murzintcev Nikita's page in the Reference section.

2.2 K-Modes Clustering

In order to create clusters of the inmates, the k-modes algorithm was used. This method was chosen as opposed to hierarchical clustering because we did not believe that our data has a hierarchical structure. The k-modes algorithm extends the k-means algorithm to be used with categorical data.[7] The k-modes algorithm has three significant differences from the k-means algorithm. These modifications are the use of k-modes instead of k-means, different dissimilarity measures, and updating modes using a frequency approach.[8] For more information on the k-modes algorithm, see Zhexue Hang's "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining."

Most of the categorical variables in the data set were used in the k-modes clustering. K-modes requires that the number of clusters be determined beforehand. This is an open question, so we arbitrarily chose K = 4. We also examined K=3, but decided to define that there were four clusters.

2.3 Logistic Regression

Logistic regression was used in order to predict whether the last statement of an inmate would be a certain topic. Five new binary variables were created for each of the five topics so that the presence of a topic could be a response variable. Logistic regression models were fit for various combinations of the covariates, including all of the available covariates as well as covariates that we considered to have potential significance in determining topic assignments.

3 Results

An LDA model with five topics was fit for the last statements and an LDA model with three topics was fit for the descriptions of offenders' crimes. The methods used to select the number of topics are described in the Appendix, and the topics are described in the

[4] Colorado Reed, "Latent Dirichlet Allocation: Towards a Deeper Understanding, 2

[5] Reed, "Towards A Deeper Understanding", 2

[6] Bettina Grun and Kurt Hornik, "topicmodels: An R Package for Fitting Topic Models", 1-2.

[7] Zhexue Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining", 1.

[8] Huang, 3

tables below. The potential occurring themes were determined by examining the fifteen most popular words in each topic (see Appendix for top fifteen words for each topic). Table 1 shows the resulting topics for the last statements.

Table 1: Last Statement Topics

Topic	Potential Occurring Themes
1	religion, praying, referencing Jesus
2	mentioning family members, ready for execution
3	not as clear as other topics, life, family, time
4	apologizing for pain caused, family
5	maintaining innocence
6[9]	No last statement

The topics that resulted from the crime descriptions were more difficult to define based on the most frequent words. These topics are described in Table 2.

Table 2: Last Statement Topics

Topic	Potential Occurring Themes
1	vague topic, crimes involving cars
2	murders, possibly in victims' homes
3	robberies, shootings,

After completing K-modes clustering with K=4, there were some notable features of the different clusters. These features were examined by looking at the resulting proportion of characteristics in each cluster. It is interesting to note that the clusters appear to be separated by race with clusters 1 and 4 being predominately black. The clusters vary on other features noted in the table below.

Table 3: Characteristics of K-modes Clusters

Cluster 1	Black, prior criminal record, White victims, smaller counties
Cluster 2	Hispanic, Hispanic victims, low proportion of apologetic last words, high proportion of family references in last words
Cluster 3	Black, no prior criminal record, low proportion of White victims, from Harris County
Cluster 4	White, skilled workers, White victims, smaller counties, low proportion of apologetic last words

In most of the logistic regression models that were fit, there were not any predictors that were found to be significant in predicting if an inmate's last statement belonged to a certain topic. However, it is interesting to note that the topic assignment from the summaries of the crimes was the most significant predictor (had the lowest p-value for one of the models) among the covariates. Using a Chi-Square Goodness of Fit Test, we tested whether or not a model with crime topics and prior prison record (a binary variable indicating whether or not an inmate had been to prison before) fit better than a model with only an intercept. The Goodness of Fit Test resulted in a p-value for the model with two co-variates of .439. This indicates that our model with the two predictors does not fit significantly better than the model containing only an intercept. However, this is not surprising because it is difficult to predict the topics found in the last statement based on solely the demographic information.

4 Discussion

The LDA analysis resulted in assigning one of five topics to the last statement of each inmate. Four of these topics seemed to be well defined based on their most often occurring words. On the other hand, LDA did not seem to perform as well for defining topics for the descriptions of offenders' crimes. This may have occurred because of the nature of the crimes that result in an inmate being sentenced to Death Row. Most of these prisoners will have committed multiple crimes at the same time or multiple capital crimes over time. Thus, the descriptions of their crimes will contain language that is associated with many different crimes. LDA may also not have worked very well because there are only so many crimes that a person can commit and receive the death penalty for. Therefore it is more difficult to define distinct topics in the crime descriptions using LDA.

The cluster analysis resulted in some separation of different groups of offenders that were executed on death row. Even though there were noteworthy differences, these results did not provide significant insights into our data. However, the clusters did provide evidence of potential relationships that could be further explored. For example, we could look into the relationship between the different types of the crimes and offenders within each racial group.

For some of models fit in logistic regression, the covariates for only the topics of crime gave some weak evidence to predict the last statement topic assignment. However, given the results of the models fit, we are not comfortable determining whether or not these covariates influence the topic of an inmate's last statement. The logistic regression models may have performed poorly because the variables in the data set only provide so much information about an inmate. Information such as family history, health, socioeconomic status and more could inform last statements.

There are multiple areas where this project can be expanded and improved. One area to investigate further is the LDA method. As mentioned above, LDA assumes that the topics in documents are uncorrelated. This may not be a reasonable assumption as prisoners talk about many different subjects (family, innocence, etc.) in their last statements. In addition, many inmates commit multiple offenses (armed robbery, murder, rape, etc.) that occur in the descriptions of their crimes. Thus, it would be beneficial to explore other topic models such as a correlated topics model (CTM). Another area that can be further explored is the number of topics that are specified when the LDA model was fit. In addition, the project can be improved by doing more research into determining the best value of K for use in the k-modes algorithm. There are also potential relationships found in the clusters to examine more. Further research in these areas could result in more accurate definitions of clusters and topics.

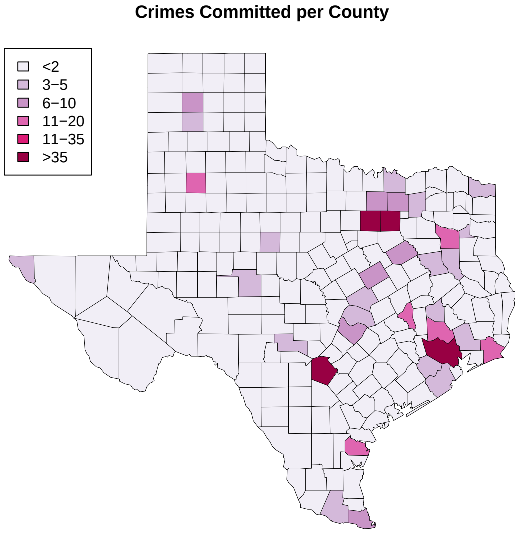
5 Appendix

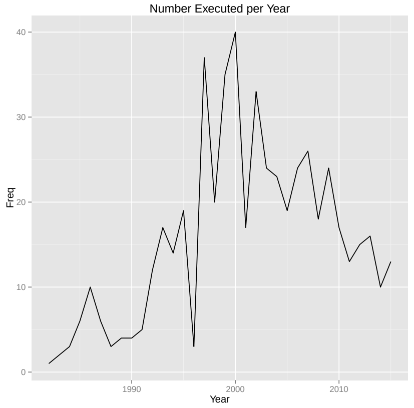
5.1 Collecting the Data

The data was scraped from the Texas Department of Criminal Justice website that lists the executed offenders [10] using the rvest and httr packages. The offender information that was contained in images was entered into Excel manually once it was determined that the Optical Character Reader would not be effective strategy to collect this data. The data was collected for 529 inmates who have been executed in Texas. One inmate was executed between the due date of this project and when the data collection process began, so he is not included in our data. One observation was dropped during the data collection possibly because there was missing data or their identifiers were different in the last statement data and demographic data. Some inmates were missing observations either because the website did not contain the data or there were errors in data collection. There was a lot cleaning and manipulation required to achieve the final data set. For example, many inmates had their education level as GED, and these were replaced with 12 in the data set. Some of the categorical variables were reduced because they had too many levels. Some of these reductions were subjective, for example categorizing the offenders' prior occupations, whereas others depended on defining regions of the country and globe (Native State). For more information about the data collection, see the code.

5.2 Exploratory Data Analysis

After the data was collected and cleaned, exploratory data analysis was conducted on the available demographic information of the offenders. Various summary statistics were also calculated. Below are some examples of interesting plots and figures that supplement the discussion and analysis of the various covariates.

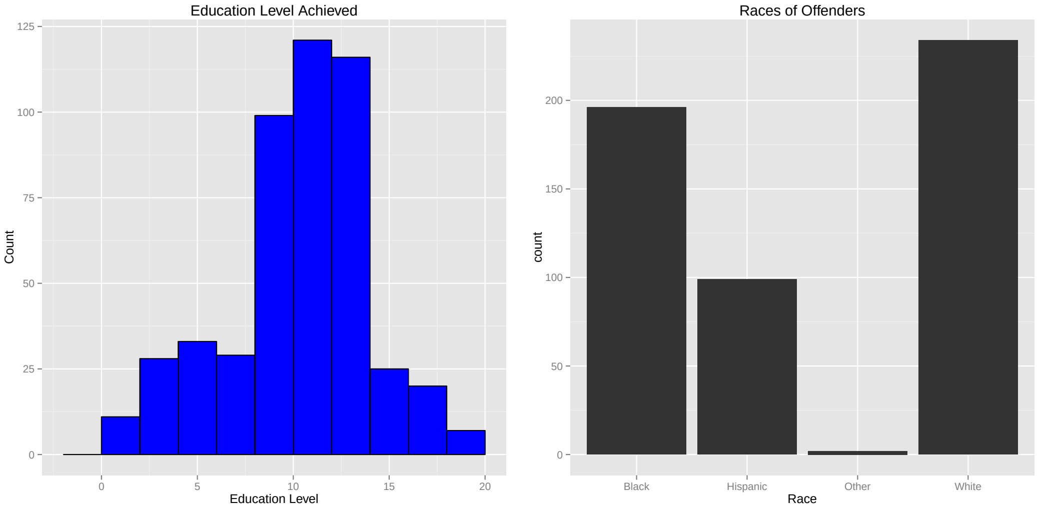




(b) Offenders Executed per Year

(a) Map of Offenses per County

[10] `https://www.tdcj.state.tx.us/death_row/dr_executed_offenders.html`



(a) Education Level

(b) Race Breakdown of Offenders

5.3 Latent Dirichlet Allocation

Figure 3: Metrics for Determining Number of Topics - Last Statements

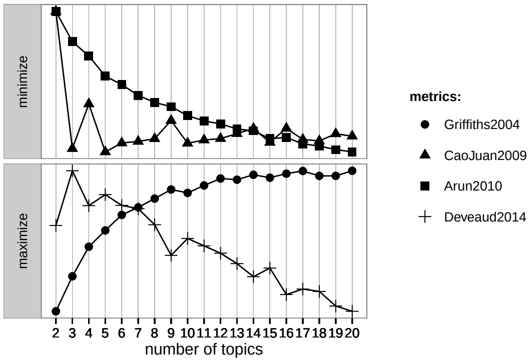
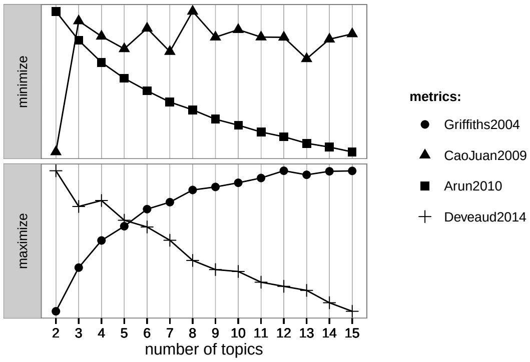


Figure 4: Metrics for Determining Number of Topics - Last Statements



As shown in Figures 3 and 4, there was not a single number of clusters that both maximized and minimized the appropriate values, therefore reasonable values for topic numbers were selected through a process combining the four measurements as well as observing the resulting topics from different numbers. Five topics were selected for the last statements because this number had a small Cao Juan metric and a larger value for the Griffiths metric. Larger number of topics performed better in terms of these metrics, but LDA models fit with larger numbers resulted in topics that did not make sense to us. We did not want to use less than five topics for LDA as four topics resulted in worse metrics for every metric.

The metrics were not very helpful in determining the number of topics for the crimes. This is probably a result of the fact that the offenders committed multiple crimes at the same time. Therefore the description of their crimes contained words that related to more than one crime. The topic that appears to be related to robbery (Topic 3), seemed to appear in LDA models with more than three topics. This could have happened because the words associated with robberies may not be as similar to the words that are associated with sexual assaults and murders that do not involve robberies. The number of topics selected for the crimes did not really consider the metrics but instead resulted from the process of fitting models with different numbers of topics and selecting the number of topics that seemed to fit the documents well based on resulting common words in each topic.

Table 4: Last Statement Topics

Topic	Top Fifteen Words
1	god, forgiv, lord, will, ask, father, jesus, life, pray, bless, come, day, pleas, christ, one
2	love, thank, tell, take, yes, warden, friend, readi, see, care, support, home, good, mom, sister
3	know, want, yall, say, dont, just, everybodi, let, time, now, help, didnt, life, done, back
4	famili, sorri, like, hope, peac, can, one, give, pain, thing, cause, heart, find, year, becaus
5	will, peopl, keep, get, kill, happen, innoc, strong, got, make, man, way, everyth, ani, right

Table 5: Crime Topics

Topic	Top Fifteen Words
1	victim, year, old, head, codefend, time, car, texa, apart, male, resid, white, femal, assault, sexual
2	death, murder, home, convict, kill, yearold, two, found, stab, later, bodi, day, insid, rape, also
3	shot, robberi, shoot, arrest, store, convict, pistol, polic, three, dure, offic, houston, calib, money, rob

6 References

1. Bettina Grun and Kurt Hornik, "An R Package for Fitting Topic Models".
2. Colorado Reed, "Latent Dirichlet Allocation: Towards a Deeper Understanding", January 2012.
3. Death Penalty Information Center, "Texas", `http://www.deathpenaltyinfo.org/texas-1`
4. Death Penalty Information Center, "Execution List 2015", `http://www.deathpenaltyinfo.org/execution-list-2015`
5. Murzintcev Nikita, "Select number of topics for LDA model", 2015-09-09, `https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html`.
6. "Remove empty documents from DocumentTermMatrix in R topicmodels?", `http://stackoverflow.com/questions/13944252/remove-empty-documents-from-documenttermmatrix-in-r-topicmodels`
7. "Spatial maps and geocoding in R," http://bcb.dfci.harvard.edu/ aedin/courses/R/CDC/maps.html.
8. Texas Department of Criminal Justice, "Executed Offenders", `https://www.tdcj.state.tx.us/death_row/dr_executed_offenders.html`.
9. Zhexue Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining", `http://www.ict.griffith.edu.au/~vlad/teaching/kdd.d/readings.d/huang97fast.pdf`.

We also consulted some websites for coding questions, etc and would like to thank Professor Steorts for her assistance on the project.