Public Notice Analysis

Project Report

7th December, 2018

Contents

List of Figures

1 Objective

Process and analyze public notices to

- Explore interesting insights about the social, economical, cultural, political and other important factors for a set of population or different regions.
- Cross reference and relate patterns in notices with historical events.
- Statistical analysis of notices.

2 Challenges

Public notices are unstructured or at best semi-structured documents published by different agencies and organizations to conform with some legal compliance. The sheer amount of notices published in various circulars makes it difficult to prepare a search-able database from them, while there may be some very interesting event and trend forecasting or introspection that can be done with the help of information that they carry with them.

The task of analyzing public data comes with many challenges. The vocabulary that is used in these notices is not the same as we use in our daily language. This again varies with the domain that the notice belongs to. So, separate models will have to be specialized to deal with different types of notices. Another challenge would be to accommodate the model to understand the (non)-structure of the notices. The notices from same distribution might be in the similar format but to make a generic model that can segment the documents to segregate different relevant components of the document will need a more intelligent approach.

Another challenge is that the public notices are spread across so many different newspapers that it is difficult to accumulate data from all of them at the same time.

There are also not so many very evident patterns found across states, because we have different extent of the quantity of data extracted across states. This becomes even more challenging because almost all states has different sets of laws, macro-economic and macro-social parameters because of which there is no one common inference observed. For the same reason, this is not an easy thing to do even across counties within a state.

3 DataSet

3.1 Year-wise Notices:

We have year-wise data of notices from 2012 to 2018. We extracted the following information from the dataset:

- State : to get idea about the distribution of specific kinds of notices in particular regions.
- Date : (a) time of year vs type of notice or (b) year vs influx of certain types of notices might help predict/observe socioeconomic changes or trends.(based on foreclosure notices).
- Content : Actual text published in the newspaper. This is our main source of information extraction and work.
- Newspaper : get most popular newspaper according to region. Might help establish relationship between newspaper ad rates and choice of customers.
- Notice category : The dataset 2 is labelled with the notice categories and will be helpful in training the classifier for notice classification.

3.2 Socioeconomic Data:

We have various datasets for the following variables:

- **PopulationEstimates:** The data contains statistics about current populations, population change, immigration (domestic and international), etc.

- **Unemployment:** It contains the statistics about employment for various counties.
- **PovertyEstimates:** The data contains statistics related to the economy in different counties.
- **Education:** The data contains statistics about education attainment of various counties.
- **CountyVariables:** It contains demographic details of various counties. For e.g. Household Income, Median Age, etc.

4 Project work

4.1 Preliminary Data Analysis

4.1.1 Dataset

1. Dataset provided with the project.
2. Notices collected from external resource[1] which is used as training data for the classifier and additional data set for analysis.

4.1.2 Data cleaning

- For Dataset 1
  - Removed rows of data with null content.
  - Remove redundant columns : e.g. link, an unnamed first column, a redundant ID column.
  - Date formatting : added month/year/weekday columns
  - At the end of the prepossessing, all the entries are valid with proper notice contents to work with.
- For Dataset 2 (mypublicnotices)
  - Converted raw text files into analyzable csv format.
  - Assigned corresponding notice category to each entry.

4.2 Preparing Data for Analysis

We have endeavoured to extract some meaningful information from the notice texts through the means of natural language processing. We've chosen spaCy framework to base our pipeline upon and decided to train it according to our dataset. Major parts of pipeline used in analysis are :

4.2.1 Classifier

Notices can come from various categories as discussed in the domain survey section. It can be helpful to identify which category the notice belongs to in order to observe region wise socioeconomic trends when paired with corresponding geopolitical region. We have used a rule based(regex) classifier script to divide data into a number of broad categories. This classifier basically uses the most relevant keywords and token for each categories to determine where they should be put. Some of these categories are :

- Bids
- Auctions
- Default
- Foreclosure
- Adoption

- Parenting

  .....

The script was received from captain and we made some changes to it to incorporate into our analysis.
In order to deal with the huge data size, we classified it in chunks and took only the relevant columns of the dataset for analysis.

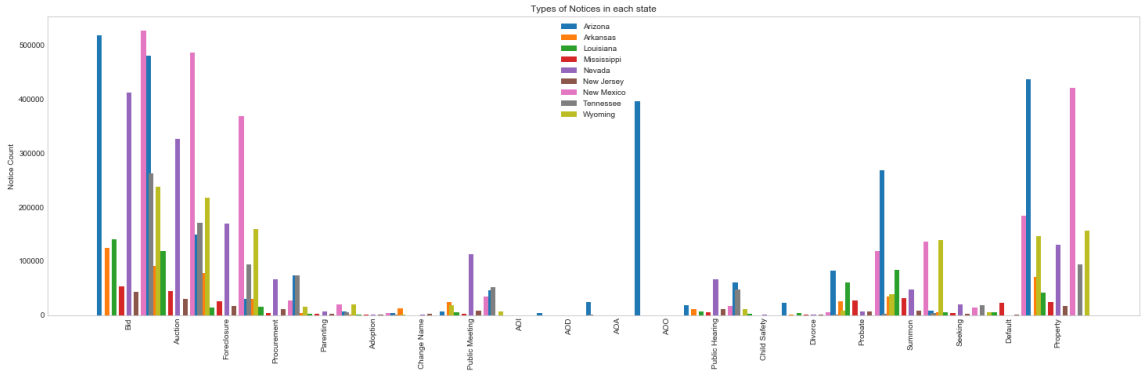

Figure 1: Distribution of types of notices in various states

4.2.2 Topic Modelling

We've given a shot to unsupervised classification by using topic modelling on the notice content. Topic modelling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.We've used gensim libraries for this purpose.
The idea was to discover any topics that we might have skipped in our classifier.We divided our data into 20 topics.The results of one such clustering are as follows:
$Topic : 0$
$Words : 0.014 * "storage" + 0.012 * "sale" + 0.011 * "notice" + 0.011 * "city" + 0.010 * "property" + 0.009 * "mexico" + 0.007 * "unit" + 0.007 * "county" + 0.006 * "district" + 0.006 * "sold"$
$Topic : 1$
$Words : 0.024 * "bids" + 0.014 * "city" + 0.010 * "time" + 0.010 * "project" + 0.008 * "county" + 0.008 * "received" + 0.008 * "contract" + 0.008 * "shall" + 0.008 * "parish" + 0.007 * "sealed"$
$Topic : 2$
$Words : 0.008 * "district" + 0.007 * "notice" + 0.006 * "parish" + 0.006 * "development" + 0.006 * "county" + 0.006 * "louisiana" + 0.005 * "state" + 0.005 * "operator" + 0.004 * "action" + 0.004 * "charles"$
$Topic : 3$
$Words : 0.017 * "shall" + 0.011 * "article" + 0.008 * "section" + 0.008 * "corporation" + 0.007 * "election" + 0.006 * "place" + 0.006 * "code" + 0.006 * "county" + 0.005 * "center" + 0.005 * "address"$
$Topic : 4$
$Words : 0.043 * "sale" + 0.032 * "trustee" + 0.028 * "arizona" + 0.017 * "trust" + 0.017 * "county" + 0.015 * "address" + 0.012 * "deed" + 0.010 * "said" + 0.010 * "notice" + 0.010 * "property"$
$Topic : 5$
$Words : 0.031 * "sale" + 0.026 * "trustee" + 0.019 * "county" + 0.012 * "trust" + 0.012 * "public" + 0.012 * "instrument" + 0.010 * "property" + 0.009 * "loan" + 0.008 * "notice" + 0.008 * "address"$

We observed a new cluster for Election notices:
$Topic : 3$
$Words : 0.017 * "shall" + 0.011 * "article" + 0.008 * "section" + 0.008 * "corporation" + 0.007 * "election" + 0.006 * "place" + 0.006 * "code" + 0.006 * "county" + 0.005 * "center" + 0.005 * "address"$

We included this category into the classifier by using appropriate phrases and tokens.

4.2.3 NER(Named Entity Recognition)

Initially we tried to build a Named Entity tagger in order to extract entities from within the notice text.We started with the pre-trained spaCy model for English language (en_core_web_lg) . Since public notice language is hardly English, the model was retrained according to the domain specific data. The idea was to be able to tag these entities for unseen notices post training. We trained it to detect some 2-3 entities like 'Address', 'county name' etc.
However, we didn't move ahead with it's expansion as it required much effort in manual labelling of training data.

5 Results and Inferences:

1. **Arizona - Decrease in Procurement Notices with election of the new governor, Doug Ducey**:
   We took procurement notices for all counties in Arizona with an intention to see how much development is being done by the government there. Also, we wanted to correlate the development work with the population change in the county. We made our plots as time chart for all counties separately to avoid the disparity due to size of the counties. This experiment gave us very interesting results. We observed that for almost every county in Arizona, these notices decremented significantly in numbers after 2015. When we looked for a reason for this change, we found out that in Jan,2015, Doug Doucey replaced Jan Brewer as Governor of the state and made significant changes for his budget.Brewer had previously reduced business property and equipment taxes and corporate income tax.Ducey, on the other hand made significant budget cuts in administration and employment freeze with significant firings, which can easily reflect on the pace of infrastructure development and other service procurement done by the government.[3] Additionally, While the procurement notices are not very strongly correlated with the change in population, there count is certainly a function of the total population of the county. Below are some selected plots which reflect the above 2 analysis.

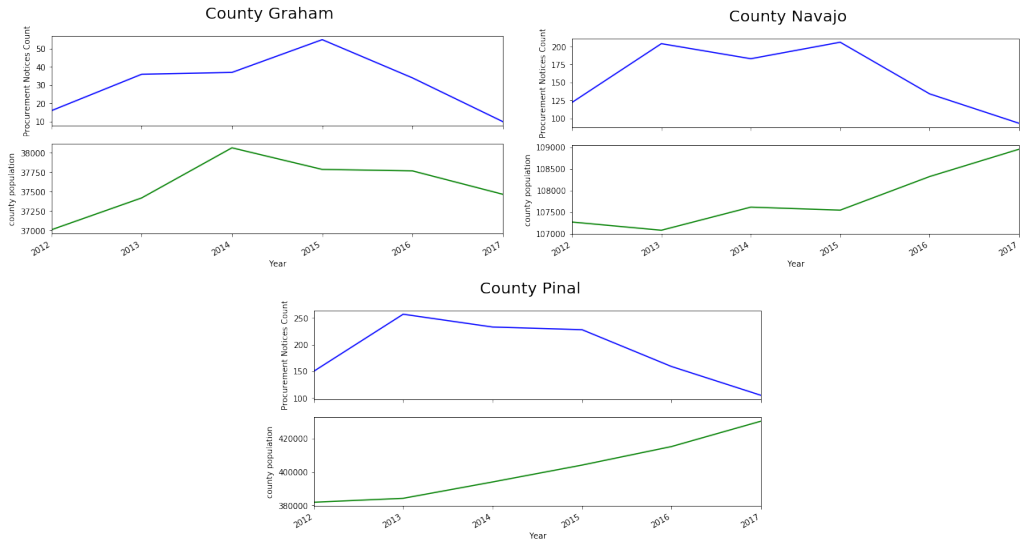

Figure 2: Procurement Notices and County Population per year

2. **Election Notices Analysis**
   We took out the distribution of Election notices for each state, each year. Also, we tried to predict the election times based on these notices. Here are the observation made from the election notices' analysis:

(a) Total election notice count correctly reflects the major election year across all states.e.g. year of the presidential elections.(Figure 3)

Election Notice Distribution (2015-16)

	state	count
0	Arizona	1222
1	Arkansas	147
2	Louisiana	2688
3	Mississippi	147
4	Nevada	134
5	New Jersey	12
6	New Mexico	590
7	Tennessee	586
8	Wyoming	206

]

(a) Election Notice Count (2015-16)

Election Notice Distribution (2017-18)

	state	count
0	Arizona	25
1	Arkansas	4
2	Louisiana	20
3	Mississippi	1
4	Nevada	9
5	New Mexico	19
6	Tennessee	4
7	Wyoming	3

(b) Election Notice Distribution (2017-18)

Figure 3: Election Notice Distribution

(b) We took line plots for 2 states : Arizona and Louisiana, for year 2015-16.
For both the states, the plot shows a peak in notices towards 2015 end owing to the gubernatorial elections and during Feb-April,2016, in line with the "Presidential Preferential Primary" and "Municipal General"[4][5][6] elections.(Figure 4)

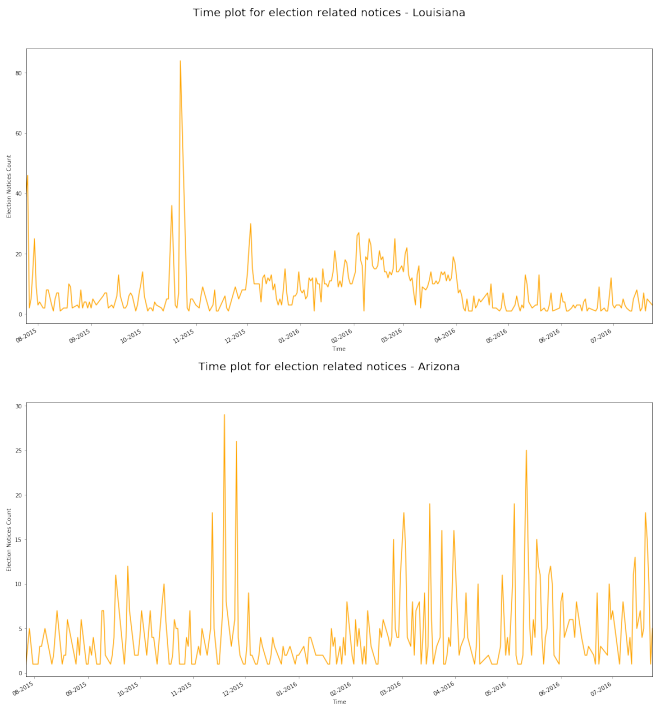

Figure 4: Election Notices time chart for year 2015-16

3. **Analysis of name-change notices**
We analyzed the ratio of the number of name change notices to the total number of notices, and found

out that out of the top 20 counties of the results, 85% of the counties were from New Mexico, rest 12% were from Colorado which is located just above New Mexico. This tells us that a very large ratio of notices from New Mexico area is attributed towards name change notices, which could be because of relaxation in name changes, the high crime rate or even the flow of immigrants from the neighboring country since it is a border state. Here are the top few results for counties with state names.

Rio Arriba - New Mexico
Santa Fe - New Mexico
Sandoval - New Mexico
Taos - New Mexico
Los Alamos - New Mexico
San Miguel - Colarado
San Juan - Washington
Chaves - New Mexico
Mineral - Colarado
Otero - New Mexico
Socorro - New Mexico
Dona Ana - New Mexico
Luna - New Mexico
Cibola - New Mexico
Lea - New Mexico

4. **Bankruptcy Notice count Analysis**
We tried to get an idea about the bankruptcy notices published for each state. We found significantly high count for "Tennessee" state. Within Tennessee too, it's particularly high and increasing for Shelby county, although the unemployment rate is ever decreasing. This can be contributed to the relaxed bankruptcy filing laws in the state of Tennessee. This observation is supported by the facts available about trends in bankruptcy filings among American states. Shelby county is indeed one of the top counties which in terms of bankruptcy filings.Other counties of the Tennessee make the list too but we have included the visualization for just Shelby county.

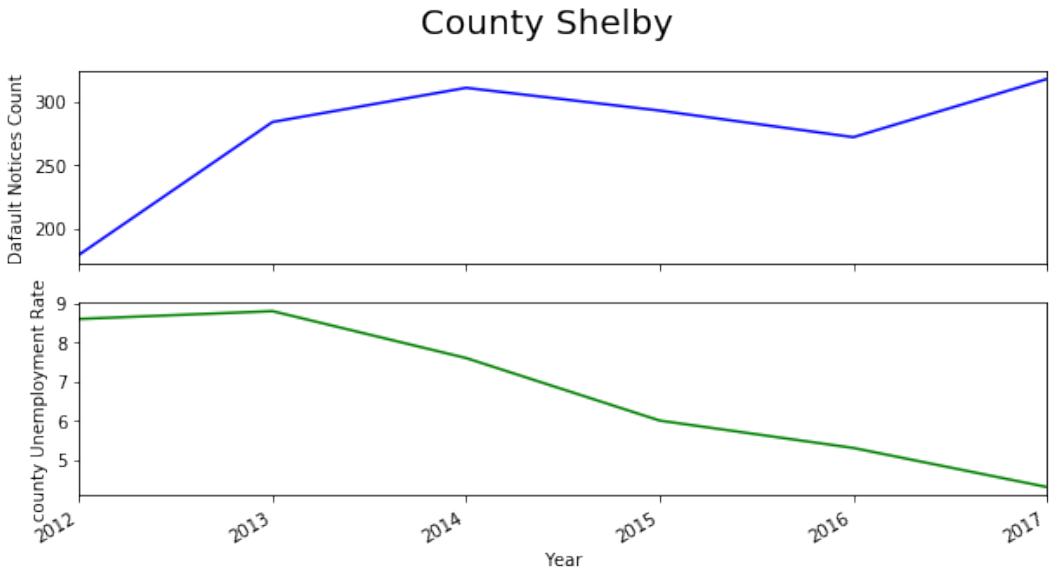

Figure 5: Default notice count and unemployment rate for Shelby county, Tennessee

5. **Proximity to a normal distribution (mean < std. dev)**

We tried to map the number of public notices across counties in a state(and the country as well) to a normal distribution, and found out that plotting the Cumulative Density Functions of the notice count distribution against a normal distribution of mean 1700, and standard deviation 5000, the map almost overlaps. Here, we observe that since, the mean is much lesser than the standard deviation, we analyze that the distribution of notices is not uniform across the counties in a state.(or US overall), i.e. there are few counties which have humongous amount of public notices compared to the many others, which could be due to their huge population or because of their urbanization. Below is an example for the state of Tennessee:

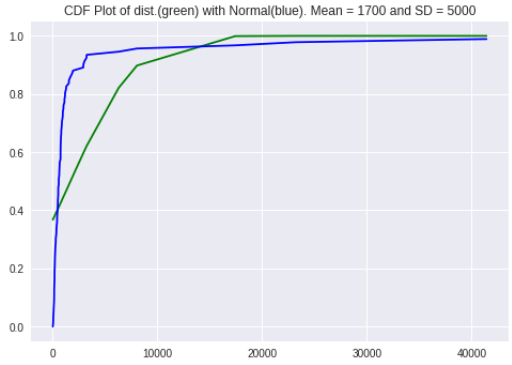

Figure 6: Proximity to a normal distribution (mean < std dev)

6. **Correlation between Default and Auction Notices**

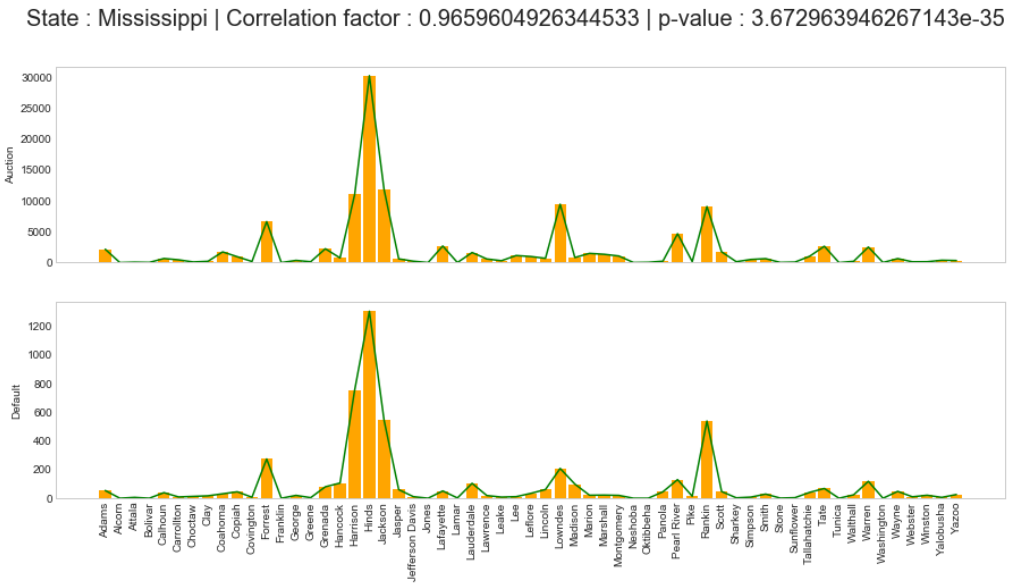

Figure 7: Count of Default and Auction notices per county in Mississippi

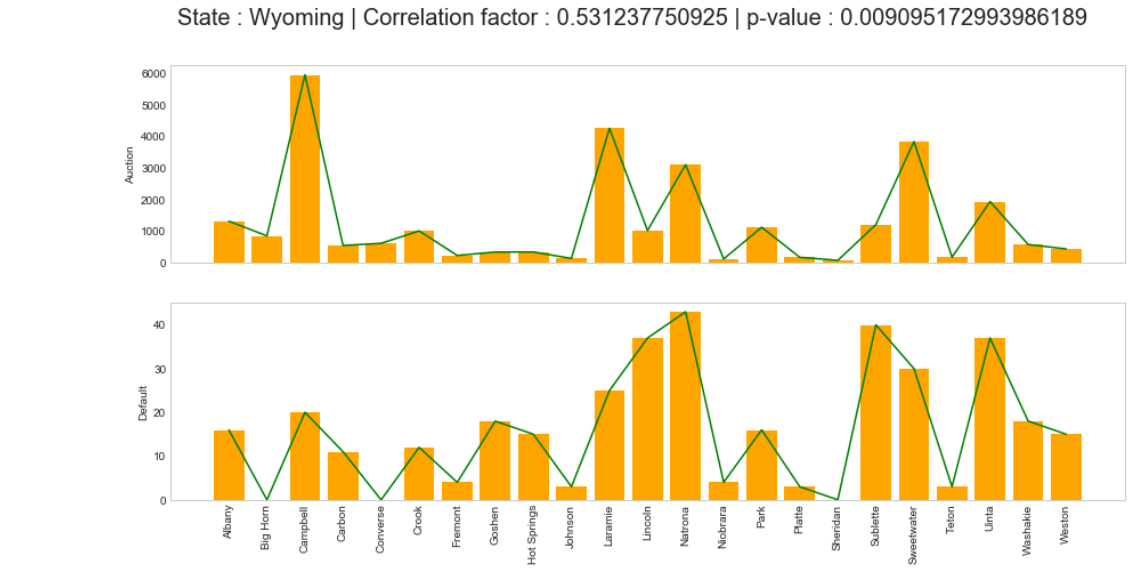

Figure 8: Count of Default and Auction notices per county in Wyoming

We've found that share of auction notices from total count of notices is correlated positively to the share of default notices in the states : Mississippi, Arkansas, Tennessee, Nevada and New Mexico.
Figure 7 shows the share of Auction and Default notices in Mississippi with high correlation and significant p-value.
Figure 8 shows the share of Auction and Default notices in Wyoming that has low correlation.

7. **Correlation between Adoption Notices and Poverty**

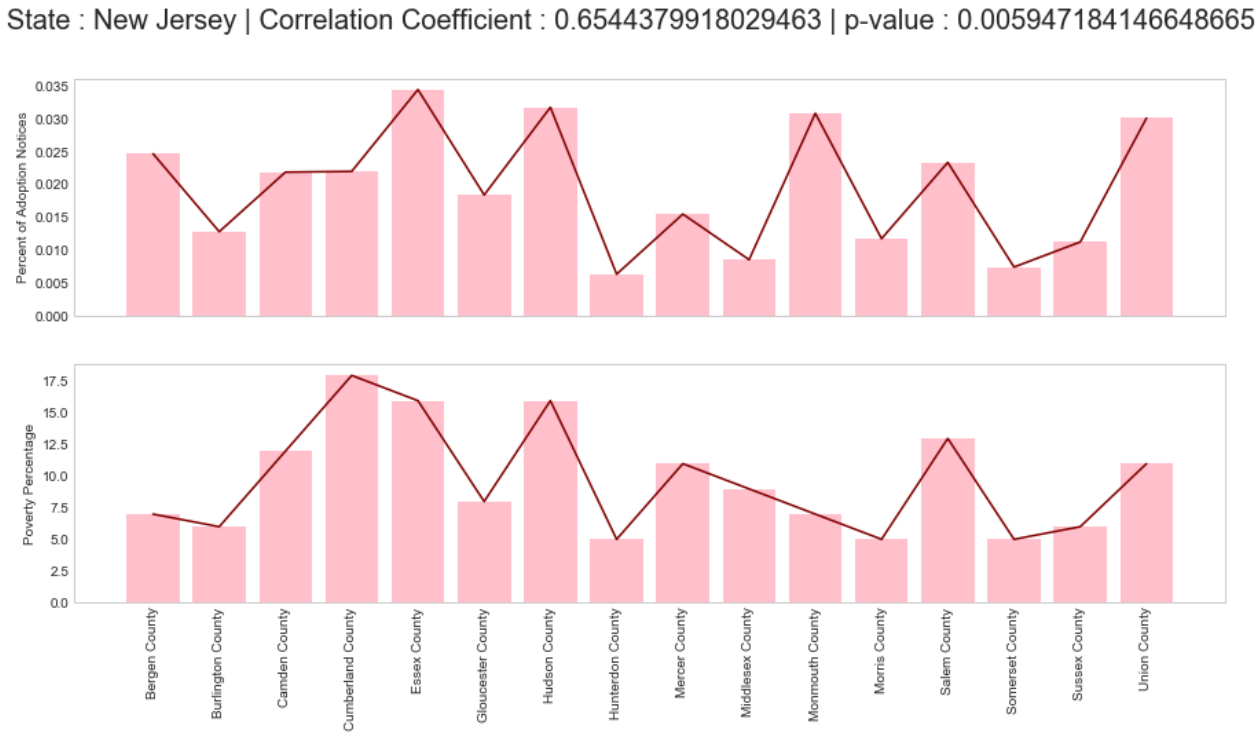

Figure 9: Percentage of Adoption Notices and Poverty in various counties of New Jersey

We found that notices of type adoption are correlated with Poverty percentage in various counties of state New Jersey. The correlation and significance (p-value) is as shown in Figure 9. This correlation seems logical considering child care laws implemented by State.

8. **Correlation between Parenting and Divorce Notices**

Year : 2017 | Correlation Coefficient : 0.766168398065627 | p-value : 0.026619805593122308

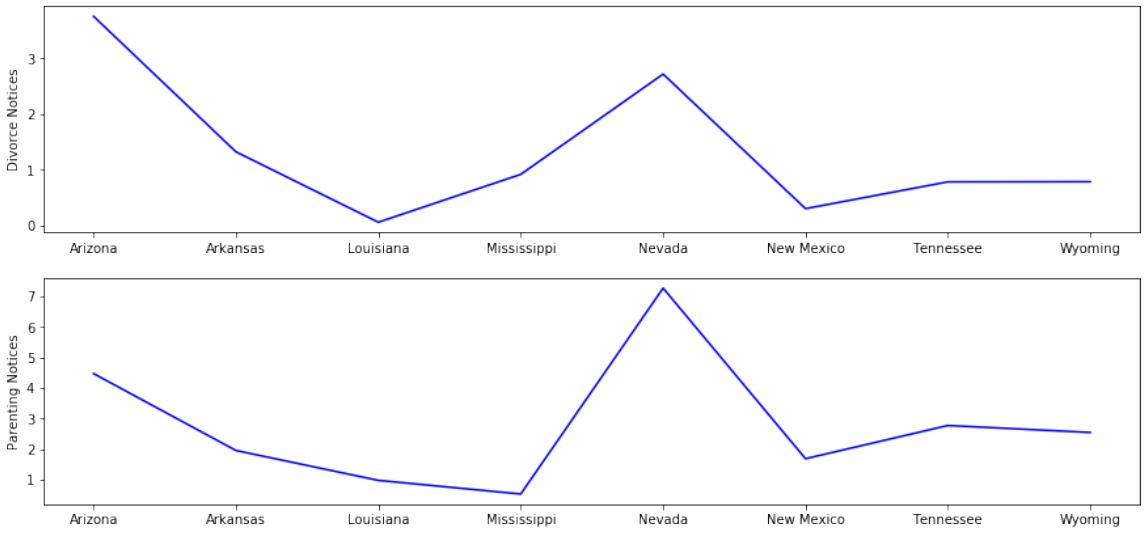

Figure 10: Percentage of divorce and parenting notices in each state

We tried to correlate the share of divorce notices and parenting notices from total count of notices, and found positive results with a p-value significantly less. The states with high percentage of divorce notices have relatively higher percentage of parenting notices.

9. **Standard Deviation of notice count for each state**
We've tried to visualize the std deviation of count of notices across counties for different states and found that it's very high for some of the states which states few counties in those states have extraordinarily high amounts of public notices, indicating those counties' huge population or greater urbanization index. For the plot below, darker color indicates the evenness in spread of the number of public notices. The scale mentioned is in the increasing order of standard deviation of - the ratio of number of notices across counties to the total number of notices in the state for each state. We have taken such a ratio so as to normalize for the extent of quantity of data extracted from each state, which may be different for different states.

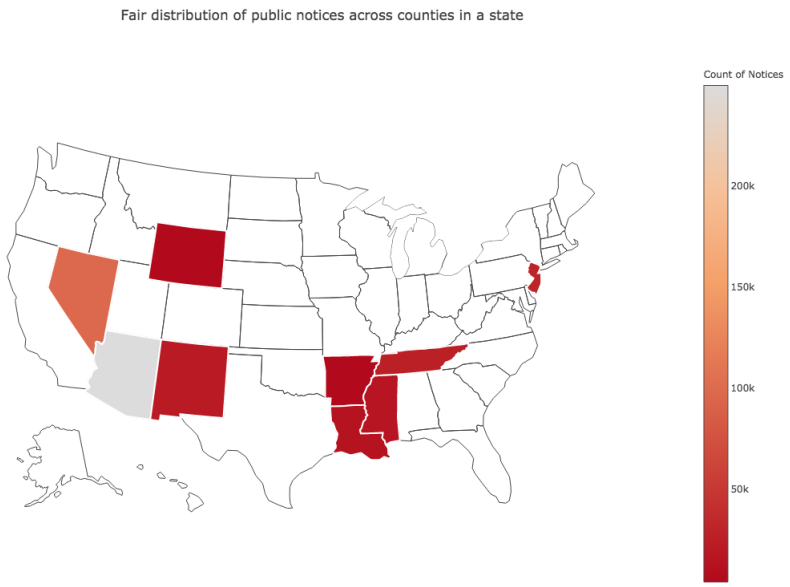

Figure 11: Distribution of public notices across counties among states

10. **Distribution of top 10 newspaper**

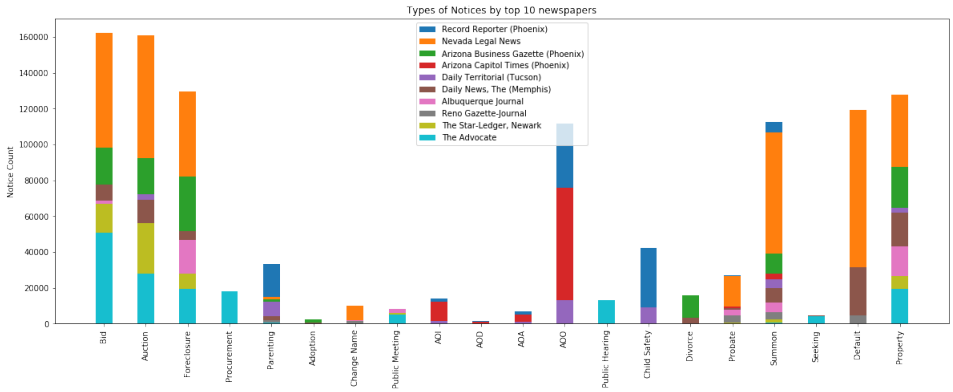

Figure 12: Distribution of types of public notices across major newspapers

Figure 12 shows the distribution of types of notices by the top 10 newspapers in terms of frequency of notices. It can be seen that a lot of notices in the type Article of Dissolution, Article of Amendment, Article of Organization and Article of Incorporation are by "Arizona Capitol Times".

References

[1] My public notices : Aggregated notice database http://www.mypublicnotices.com/PublicNotice.asp

[2] spaCy : Industrial strength Natural Language Processing
https://spacy.io/

[3] https://www.nerdwallet.com/blog/credit-cards/highest-bankruptcy-rates-states-counties/

[4] https://www.sos.la.gov/ElectionsAndVoting/PublishedDocuments/ElectionsCalendar2016.pdf

[5] https://www.sos.la.gov/ElectionsAndVoting/PublishedDocuments/ElectionsCalendar2015.pdf

[6] https://apps.azsos.gov/election/2016/Info/ElectionInformation.htm

[7] https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24