**Name**: Mayank Baheti

**Topic**: Text mining and insights from US regulatory space

**Aim of Exercise**:

It is widely believed that some administrations have different focuses for their regulatory endeavours. The Federal Register contains a list of all of the proposed rules and final rules that government agencies issue to manage the US economy. Final rules are rules that have taken effect and present the true impact an administration has had on the regulatory structure.

Hence our aim is to find the focus of each of the government of George Bush (January 20, 2001 to January 20,2009) and Barak Obama (January 20, 2009 to January 20, 2017) . So, here our task is to find out the departments and issues for which they made the rule and got the implemented. Also, we need to build a classifier model so that we can abstract the data and hidden information from the rules that were being made by them.

**Data Description**:

We were supposed to collect the data using the Federal Register API, hence in R, I have used the “federalregister” package to collect the data. But while doing so, faced a lot of problem as this package will extract one page at a time (in all max 10 pages can be extracted) and 1000 records per page, doing so this package tried to hit the server many times, and it‘s request later on was considered as spam and gave up error like HTTP :429 error. Hence as a workaround I have used a URL(as shown line 56) to extract the data in which I have scrapped 50 pages by specifying the time period of each President as the range of publication date and type as “RULE” for the final rules. Below is the data description of the data being collected.

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Description |
| title | chr | Documentation Subject Line |
| type | factor | Type of document (only final rule considered) |
| agency\_names | factor | Publishing agency name |
| abstract | chr | A brief write-up of the rule |
| document\_number | chr | Document Serial Number |
| html\_url | chr | Link to access the complete document |
| pdf\_url | chr | Link to access the complete document in pdf form |
| publication\_date | date | Date of Publication |

Using the different Dataset for both Bush and Obama we started with Topic modelling using Text2vec package below are the distribution of the documents

* Bush’s Document distribution and top 3 topic inferences

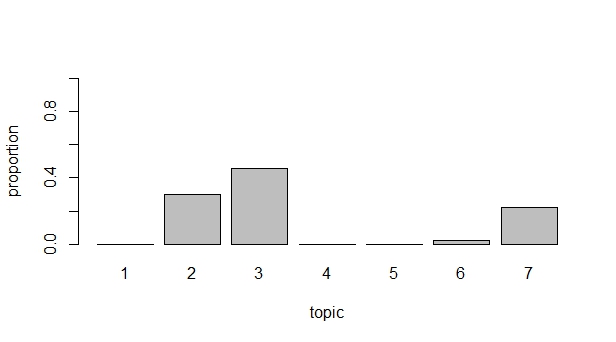


Fig.1 – Bush Document Distribution among the seven topics we have specified.

Now let’s see what are these 3 major topics on which Bush government has emphasised.

Below is the screenshot of the top 20 terms occurring in top 3 topics which are Topic 2,3 and 7 respectively.

**Code**: lda\_model\_bush$get\_top\_words(n = 7, topic\_number = c(2L, 3L, 7L), lambda = 1)



Fig.2 - Top 20 terms occurring in top 3 topics

We can make the following inferences from these topics

Topic 2: Mainly for Medical and healthcare

Topic 3: Regarding Banking and Real-estate

Topic 7: Regarding aviation and security. Also due to 9/11 attack there can be a lot of regulation which would have been implemented.

* Obama’s Document distribution and top 3 topic inferences

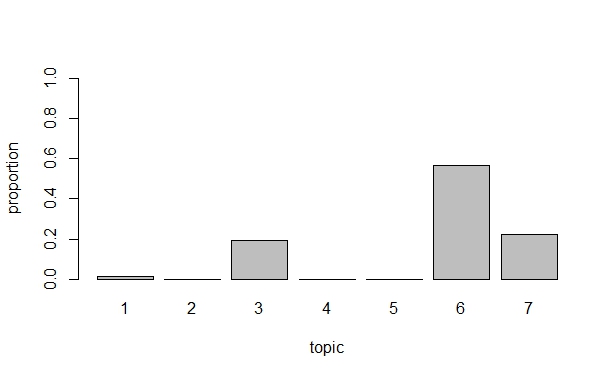


Fig.3 – Obama Document Distribution among the seven topics we have specified.

Now let’s see what are these 3 major topics on which Obama government has emphasised.

Below is the screenshot of the top 20 terms occurring in top 3 topics which are Topic 2,3 and 7 respectively.

**Code**: lda\_model\_Obama$get\_top\_words(n = 7, topic\_number = c(3L, 6L, 7L), lambda = 1)



Fig.4 - Top 20 terms occurring in top 3 topics

We can make the following inferences from these topics

Topic 3: Mainly for Environment and safety

Topic 6: Regarding Shipping restriction and prohibition to reduce the crime on water bodies like attack on US boats from Somalian pirates.

Topic 7: Regarding Rules and Regulation for civil people.

**Concepts Used for Topic Sampling:**

**LDA (Latent Dirichlet Allocation)** model also decomposes document-term matrix into two low-rank matrices - document-topic distribution and topic-word distribution. Bit it is more complex non-linear generative model. We won’t go into gory details behind LDA probabilistic model, reader can find a lot of material on the internet.

There several important hyper-parameters:

1. n\_topics - Number of latent topics.
2. doc\_topic\_prior - document-topic prior. Normally a number less than 1, e.g. 0.1, to prefer sparse topic distributions, i.e. few topics per document.
3. topic\_word\_prior - topic-word prior. Normally a number much less than 1, e.g. 0.001, to strongly prefer sparse word distributions, i.e. few words per topic.

**Classification Model:**

Now after doing the topic modelling we should be able to differentiate between the two President’s document. So for this we have following approach:

**Feature Engineering and Modelling:**

1. Extract the data from the URL.

2. Bind the two datasets of Bush and Obama with their name as target.

3. Now merge the title and the abstract columns together and do random sampling and divide the data into train and test.

4. Using the text2vec package do the pre-processing on this data like making the text into corpus, bringing into lower case, removing stop words, punctuations, and numbers, stripping the white spaces and doing the lemmatization.

5. Now on this pre-processed corpus make a Document Term Matrix separately for Train and for train use the vocabulary of train to make DTM using TF-IDF.

6. Best Model which worked gave the best accuracy was GLMNET.

**Other Models I have tried:**

1. Using the Gibbs sampling, perform Latent Dirichlet Allocation (LDA), to get K (K=7 in this case) topics out of this Document Term Matrix of the combined text of Obama and Bush.

2. Collect the Topic probabilities as the feature for each document and bind it to the DTM matrix so that we can use it as rich features for each document.

3. Divide the data randomly into train and test.

4. Start with several classification model like GLM, decision tree, SVM and Random forest.

5. Also, as a part of Feature Engineering I did SVD (singular value decomposition) over DTM matrix to get the Document and it’s hidden feature space which will be U matrix, so that we can get more rich feature about the documents and append it the existing data frame and used it in Random forest to get 30 important attributes and built the model again to reduce the complexity and more generalized model.

**Evaluation Metric**: For this problem, I have used accuracy, as our aim is to classify the documents correctly.

Best Accuracy achieved: 85% GLMNET -Lasso Regression

**Future Scope**: extracting more data as due to spamming issue on federal register site was not able to extract much data. So need to explore more alternative to avoid this issue.