

2021

Tags Generator for Legal Text Documents

Data Science Project

Aayush Singhal, N277
Kashish Tiwari, N292



Index

Introduction	3
Problem Statement	3
Data Analysis	3
Pre-processing	4
Data Cleaning	4
Lemmatization	5
Vectorization	6
Data Visualization	6
Data Modeling	7
Results	8
Prediction	8

Tags Generator for Legal Text Documents



Introduction:

Tags have been a relevant part of information retrieval system. They help the system to get relevant information in most effective way. These tags have been generated manually by human annotators and makes the system more expensive for the clients. The only way to make these technologies affordable is to find a way to automatically generate tags that resembles to human annotations

Problem Statement:

We are provided with a legal text document in Training set documents. In Training Tag documents, the relevant tags have been manually created by human annotators respectively to the files of Training set documents. The aim of this project is to create a solution that automatically generate tags for the Testing set legal text document.

Data Analysis:

The data used was provided by a law firm. The data on which the model is trained consists of the tags used by a functioning law firms for the particular text.

Data Pre-processing:

Since the data was present in individual text files, the first task was to store the data from an individual text file into a single data frame. On performing the required code, the following was obtained:

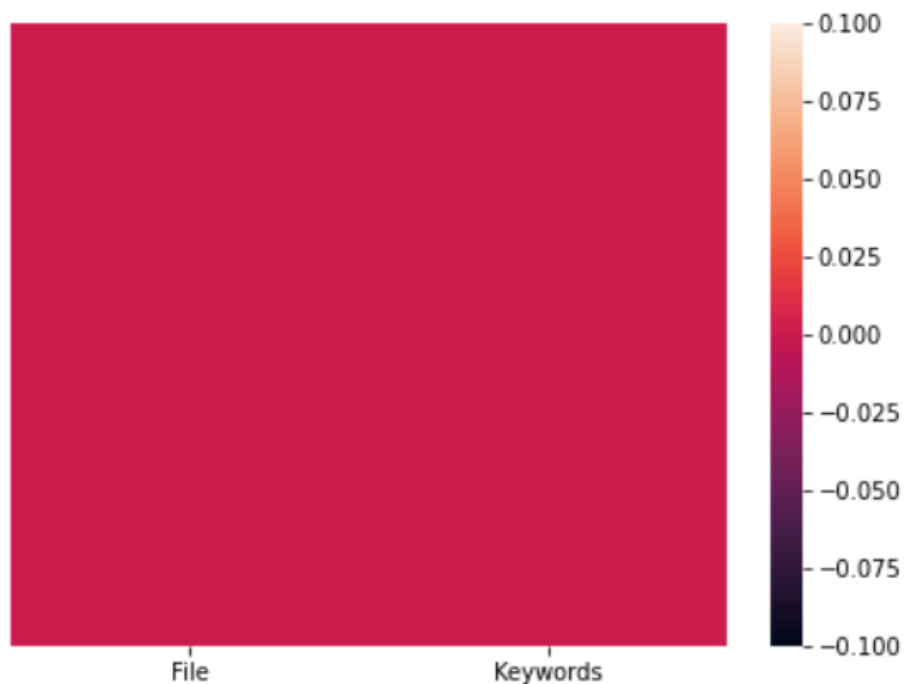
	File	Keywords
0	\n\nKurian Joseph, J.\n\n1. Leave granted in S...	Cause of Action
1	\n\nAbhay Manohar Sapre, J.\n\n1. Delay in fil...	Abetment, Abetment of Suicide, Absconding, Acc...
2	\n\nPinaki Chandra Ghose, J.\n\n1. This crimin...	Decision, Exemption, Exemption Notification, I...
3	\n\n1. This matter is placed before us as a Be...	Child Labour, Compensation, Fundamental Right,...
4	\n\n1. We have heard learned Counsel for the p...	Account, Auditor, Authentication, Commercial, ...

Data Cleaning:

As it is visible from the above image, the data in the “File” column contains \n, which can affect the accuracy of the model. Hence, “\n” was replaced and the final outcome of the dataset is as follows:

	File	Keywords
0	Kurian Joseph, J.	Cause of Action
1	1. Leave granted in Special Leave Petition (Ci...	Cause of Action
2	2. Around 46.93 acres of Land was acquired by ...	Cause of Action
3	3. Learned Counsel for the Appellants submitte...	Cause of Action
4	4. Shri Sanjay Kumar Tyagi, learned Additional...	Cause of Action
5	5. Learned Counsel appearing for the Appellant...	Cause of Action
6	6. Prior to amendment Act 68 of 1984, the amou...	Cause of Action
7	Section 25. Rules as to amount of compensation-	Cause of Action
8	(1) When the applicant has made a claim to com...	Cause of Action
9	(2) When the applicant has refused to make suc...	Cause of Action

Heatmap: To identify null values:



Since the heatmap is uniform, this represents the presence of **no null values** in the dataset.

Data Lemmatization:

The “File” column of the data frame contains a lot of stop words and words which are irrelevant for model creation.

The pre-processing of the data is performed by using the libraries re and nltk. The unnecessary elements have been removed to produce an accurate model, using Lemmatization.

```
['kurian joseph ',  
 'leave granted in special leave petition civil no 12495 of 2015  
,  
,  
'around 46 93 acre of land wa acquired by the respondent state of  
haryana initiating the proceeding by notification dated 19 09 1983  
issued under section of the land acquisition act 1894 the purpose  
of acquisition is residential and commercial for panchkula sector  
21 the acquired property is in village fatehpur in respect of the  
same development we have seen that this court in many case ha base  
d the fixation of the land value based on acquisition proceeding i  
nitiated in 1981 in village judian those property in village judia  
n had access to state highway and the value fixed by this court is  
r 250 per square yard in respect of property situated in the adjo  
ining village of the appellant namely devi nagar we have fixed land  
value at the rate of r 250 per square yard that wa the acquisition  
initiated in the year 1987 and that property had extensive nationa  
l highway frontage ',  
'learned counsel for the appellant submitted that in all the adjo  
ining village for the property acquired for the same purpose this  
court having fixed the land value at r 250 per square yard and abo
```

Since our model needs numeric data, we need to transform the processed sentences into vectors.

Data Visualization:

[illegible]

- **Court**
- **Section**
- **Act**
- **Appellant**
- **Case**

Model Fitting:

These are the algorithms used in this project:

1. **Bernoulli's Naive Bayes**: It implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable.
2. **Support Vector Machine**: In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.
3. **Passive Aggressive Classifier**: The Passive-Aggressive algorithms are a family of Machine learning algorithms that are not very well known by beginners and even intermediate Machine Learning enthusiasts. However, they can be very useful and efficient for certain applications.
4. **k-Nearest Neighbour**: It is a type of classification, where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.
5. **Multinomial Naive Bayes classifier**: It is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts.
6. **Decision Trees**: They are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

Results:

	Accuracy (%)	Jaccard Index	F1-Score
Model			
Bernoulli Naive Bayes	86.87	0.767909	0.878655
Support Vector Machine	96.72	0.936446	0.968658
Passive Aggressive Classifier	96.77	0.937393	0.968360
K-Nearest Neighbors	96.67	0.935500	0.968565
Multinomial Naive Bayes	96.06	0.924217	0.960877
Decision Tree	94.45	0.894762	0.948486

Since the **SVM** model shows the best results on the training set i.e. **96.72%**, along with comparatively better F-1 Score, we choose it for predicting the keywords for the test set.

Prediction:

The same SVM model is used for the prediction of the test set's keywords.

The data from the text set cleaned, lemmatized, vectorized and then passed in the model. The generated keywords are then exported to a separate excel file.

12Keywords_test_set.xlsx - Excel

File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

Calibri 11 A A

B I U L Font Wrap Text Merge & Center

General Number Conditional Formatting Format as Table Styles

Normal Bad Neutral

Insert Delete Format Clear Sort & Find Filter Select

Clipboard Font Alignment Number Conditional Formatting Format as Table Styles

AutoSum Fill Clear

Editing

A6 the fact in brief are the respondent assessee is engaged in the business of tea spice etc during the assessment year 1985 86 previous year ending on 31 1985 the assessee wrote back in it account sum of r 14 65 997

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	File	Keywords												
2	venkatarama reddy	Attesting Witness, Consideration, Investigation, Notice, Partition, Pending, Possession, Property, Right, Sale, Sale deed, Transferred												
3	the opinion recorded by the kerala high court in nr 16 of 1997 has given rise to this appeal filed by the chief	Acknowledgement, Assurance, Auction, Bill, Buyer, Complainant, Complaint, Construction, Consumer, Consumer Dispute, Default, Delay, Delivery												
4	whether on the fact and in the circumstance of the case the tribunal is right in law and fact in holding that r 02	Adjudication, Administrative Power, Back Wage, Compensation, Competent Authority, Conditions of Service, Continuous Service, Discharge, Disci												
5	the high court accepted the view of the tribunal which partly allowed the appeal of the assessee and answered	Acknowledgement, Assurance, Auction, Bill, Buyer, Complainant, Complaint, Construction, Consumer, Consumer Dispute, Default, Delay, Delivery												
6	the fact in brief are the respondent assessee is engaged in the business of tea spice etc during the assessment	Adjudication, Administrative Power, Back Wage, Compensation, Competent Authority, Conditions of Service, Continuous Service, Discharge, Disci												
7	on further appeal by the assessee the tribunal set aside the addition of r 02 758 which was upheld by the appeal	Adjudication, Administrative Power, Back Wage, Compensation, Competent Authority, Conditions of Service, Continuous Service, Discharge, Disci												
8	it may be noted that the provision was made in the book of account towards purchase tax which was under dispute	Adjudication, Administrative Power, Back Wage, Compensation, Competent Authority, Conditions of Service, Continuous Service, Discharge, Disci												
9	the learned senior counsel appearing for the Income Tax department to contend that the assessee itself too	Adjudication, Administrative Power, Back Wage, Compensation, Competent Authority, Conditions of Service, Continuous Service, Discharge, Disci												
10	the decision of this court in commissioner of income tax sundarlam lyengar and son md narsu sc 1251 1999 assult, Circumstantial Evidence, Corroboration, Criminal Intimidation, Demeanour, Evidentiary Value, Eyewitness, First Information Report, Garg													
11	for the reason aforesaid we affirm the opinion expressed by the high court and dismiss the appeal filed by the i	Constitutional Validity, Defamation												
12	pattnaik	Absence, Accommodation, Amendment, Appeal, Applicability, Application, Arrear, Arrears, Arrears of Rent, Building, Case, Cause of Action, Claim												
13	this batch of special leave petition are by the state of tamil nadu directed against the judgment of the division i	Absence, Accommodation, Amendment, Appeal, Applicability, Application, Arrear, Arrears, Arrears of Rent, Building, Case, Cause of Action, Claim												
14	in accordance with the aforesaid excise policy and the provision of the act and the rule the exclusive privilege i	Absence, Accommodation, Amendment, Appeal, Applicability, Application, Arrear, Arrears, Arrears of Rent, Building, Case, Cause of Action, Claim												
15	the existing holder of the privilege in question who had obtained licence for carrying on the business for the ex	Absence, Accommodation, Amendment, Appeal, Applicability, Application, Arrear, Arrears, Arrears of Rent, Building, Case, Cause of Action, Claim												
16	the government is at liberty to go ahead with the grant of privilege of retail vending of indian made foreign liqu	Absence, Accommodation, Amendment, Appeal, Applicability, Application, Arrear, Arrears, Arrears of Rent, Building, Case, Cause of Action, Claim												
17	it but the government shall adhere to the place of retail vending which have been licensed for the excise year 2	Account, Account Book, Accounts, Act, Amin, Any Order, Assumption, Award, Bench, Book, Books, Case, Challenge, Civil Appeal, Civil Court												
18	in the above facility of renewal to the petitioner shall be made available if the petitioner remit the requisite am	Constitution of India, Criminal Liability, Defamation, False Statement, Freedom of Speech, Fundamental Right, Innocent Person, Judicial Notice, Li												
19	in for any reason if there is delay in renewal the petitioner shall be entitled to vend the indian made foreign liqu	Account, Against Any Liability, Any Person, Authorised Insurer, Award, Awarded Compensation, Bodily Injury, Breach Of Condition, Carriage, Certi												
20	the government the commissioner and all the district collector shall be entitled to re locate the shop out of 00	Account, Account Book, Accounts, Act, Amin, Any Order, Appeal, Assumption, Award, Bench, Book, Books, Case, Challenge, Civil Appeal, Civil Court												
21	it is this order of the division bench of the madras high court which is the subject matter of challenge in all the	Absence, Accommodation, Amendment, Appeal, Applicability, Application, Arrear, Arrears, Arrears of Rent, Building, Case, Cause of Action, Claim												
22	after hearing mr venugopal the learned senior counsel appearing for the state of tamil nadu at great length an	Absence, Accommodation, Amendment, Appeal, Applicability, Application, Arrear, Arrears, Arrears of Rent, Building, Case, Cause of Action, Claim												
23	the learned senior counsel appearing for the state contended that there is no inherent right in it no inherent	Absence, Accommodation, Amendment, Appeal, Applicability, Application, Arrear, Arrears, Arrears of Rent, Building, Case, Cause of Action, Claim												
24	mr chidambaram the learned senior counsel appearing for the existing licensee on the other hand contended i	Absence, Accommodation, Amendment, Appeal, Applicability, Application, Arrear, Arrears, Arrears of Rent, Building, Case, Cause of Action, Claim												
25	we have carefully considered the rival submission at the bar as well as the decision cited in support of the	Absence, Accommodation, Amendment, Appeal, Applicability, Application, Arrear, Arrears, Arrears of Rent, Building, Case, Cause of Action, Claim												
26	mr chidambaram very fairly stated that none of the respondent have any grievance to be governed by the	Accusation, Advocate General, Alcoholic Liquor, Amending Act, Ancillary Matter, Animal Husbandry, Application of Article, Attorney General, Basic												
27	the competent authority the state government shall consider the application for renewal of the licence in acc	Attesting Witness, Consideration, Investigation, Notice, Partition, Pending, Possession, Property, Right, Sale, Sale deed, Transferred												
28	mr venugopal had referred to an affidavit which had been filed in this court by the secretary to the	Accusation, Advocate General, Alcoholic Liquor, Amending Act, Ancillary Matter, Animal Husbandry, Application of Article, Attorney General, Basic												
29	10 these special leave petition are accordingly dismissed with the modulated direction a state earlier	Absence, Accommodation, Amendment, Appeal, Applicability, Application, Arrear, Arrears, Arrears of Rent, Building, Case, Cause of Action, Claim												
30	after hearing the learned counsel for both the party at length we find ourselves in agreement with the	view tak Account, Accounts, Act, Adjustment, Agent, Agreement, Allotted, Anatomy, Appeal, Appropriate, Approval, Assessee, Assessing Officer, Assessment												
31	this appeal is directed against the judgment and order passed by the high court of karnataka in criminal appeal	Decision, Exemption, Exemption Notification, Import, Importer, India, Larger Bench, Notification, Order, Order of Reference, Reference, Stateme												

12Keywords_test_set