2021

Tags Generator for Legal Text Documents

Data Science Project

Rahul Sharma
Prasfur Tiwari
2/19/2021

Index

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:

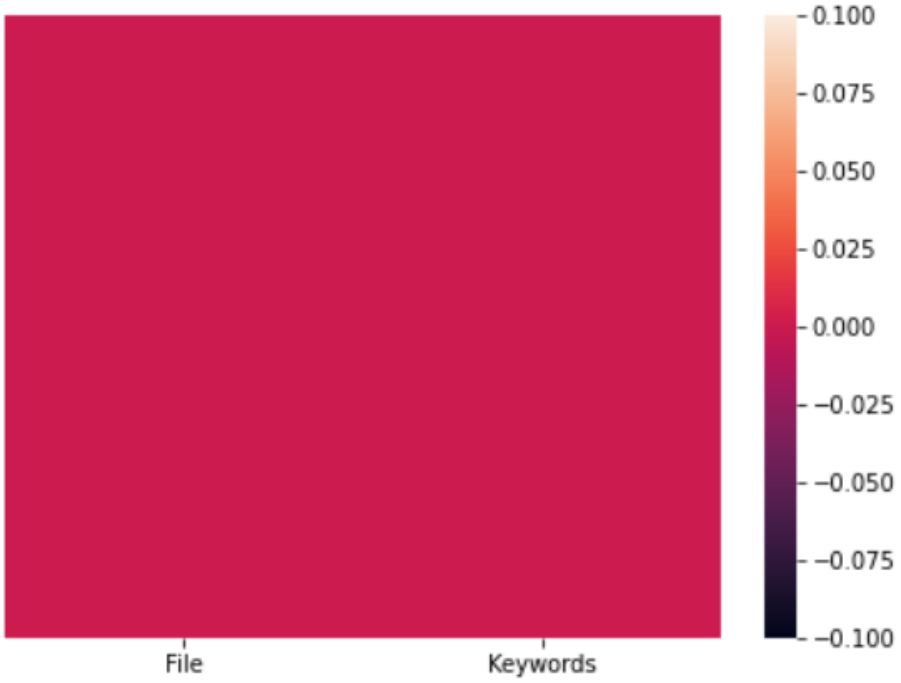
	File	Keywords
0	\n\nKurian Joseph, J.\n\n1. Leave granted in S...	Cause of Action
1	\n\nAbhay Manohar Sapre, J.\n\n1. Delay in fil...	Abetment, Abetment of Suicide, Absconding, Acc...
2	\n\nPinaki Chandra Ghose, J.\n\n1. This crimin...	Decision, Exemption, Exemption Notification, I...
3	\n\n1. This matter 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 adjoi
ning 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
```

Data Vectorization:

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

To convert the above processed text into vectors, TF-IDF vectorizer is preferred over Bag Of words (since TF-IDF is more accurate, especially when dealing with a project in which keyword generation is the key objective to be achieved.)

Data Visualization:

Word cloud for training data:

From the below visualization, we can see that the most common words used in the file column of the dataset are:

- **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:

Model	Accuracy (%)	Jaccard Index	F1-Score
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.