Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 10, Issue, 08, pp.72661-72665, August, 2018

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

ANOMALY DETECTION IN LEGAL DOCUMENTS USING MACHINE LEARNING

***Deep Modi, Harsh Parekh, Darshan Rathod and Kushal Doshi**

India

ARTICLE INFO

Article History:
Received 29th May, 2018
Received in revised form
10th June, 2018
Accepted 07th July, 2018
Published online 30th August, 2018

Key Words:
Anomaly detection, machine learning,
Porter Stemming,
K means clustering ,
Agglomerative clustering

ABSTRACT

Legal documents have always been lengthy and it is difficult to read and understand them completely. In this project, we devise a system which is Machine Learning (ML) based tool that takes in document and highlights anomalies in the text. The document can be given as soft copies. To our knowledge, some categories of legal documents contain duplicated information that do not require our attention. However, manually extracting non-duplicate information from documents requires considerable amount of effort. Thus, we want to use machine learning algorithms to pick up unordinary sentences for us. For this purpose, we propose a set of algorithms that filters out duplicate information and returns useful information to the user. We are able to train a learner that can mark unordinary parts of a legal document for manual scrutiny. Scikit and NLTK are open source module of python which have been used to develop this tool that we've created. Flask modules have been used for the simple User Interface. This project contains a simplified architecture which has various algorithms and methods that have been implemented successfully.

Citation: ***Deep Modi, Harsh Parekh, Darshan Rathod and Kushal Doshi*****, 2018.** "Anomaly Detection in Legal Documents using Machine Learning", *International Journal of Current Research*, 10, (08), 72661-72665.

INTRODUCTION

The cases where unfair agreement are made to be signed by the company is not rare in the past. To achieve their goals, they make the document long and tedious which makes people to stop reading it midway and miss the important points. There are many cases where we have Europcar Rental AGREEMENT where in Australian Competition and Consumer Commission v CLA Trading Pty Ltd [2016] FCA 377 the Federal Court decided that certain clauses of the Europcar Rental Agreement were unfair terms within the meaning of section 12BG of the ASIC Act, and were therefore void pursuant to section 12 BF(1) of the ASIC Act. The terms purported to allow Europcar to impose, in a way which is unfair, no fault liability on the customer and to make the customer's liability unlimited in cases of breach of contract, no matter how trivial or removed from the loss suffered by Europcar. And then we have another case of CHRISCO HAMPERS wherer In ACCC v Chrisco Hampers Australia Ltd [2015]- FCA 1204 the Federal Court found that Chrisco included an unfair contract term in its 2014 lay-by agreements relating to its "HeadStart Plan", which allowed Chrisco to continue to take payments by direct debit after the consumer had fully paid for their lay-by order and declared the provision void. Consumers were required to "opt out" in order to avoid having further payments automatically deducted by Chrisco After their lay-by had been paid for. The case where in George Mitchell (Chesterhall) Ltd v Finney Lock Seeds Ltd -Finney Lock Seeds Ltd agreed to supply George Mitchell (Chesterhall) Ltd with 30 lb of Dutch winter cabbage seed for £201.60. An invoice sent with the delivery was considered part of the contract and limited liability to replacing 'any seeds or plants sold' if defective (clause 1) and excluding all liability for loss or damage or consequential loss or damage from use of the seed (clause 2). 63 acres (250,000m2) of crops failed, and £61,513 was claimed for loss of production. Examples of potentially unfair terms include those that:

- Charge the consumer a large sum of money if they don't fulfil their obligations under the contract or cancel the contract (e.g. a consumer does not pay an insurance premium or mortgage repayment on time)
- Tie a consumer into the contract, while letting the firm decide whether or not to provide the service
- Require the consumer to fulfil all their contractual obligations, while letting the firm avoid its own.
- Other unfair terms:
- Limit a firm's obligation to honour its agents' commitments to the consumer (e.g. whole agreement clauses)
- Allow the firm to transfer its consumer obligations to a third party without the consumer's consent
- Mislead the consumer about the contract or their legal rights

Corresponding author:* **Deep Modi
India.
DOI: https://doi.org/10.24941/ijcr.32030.08.2018

Problem Definition

Most of the documents especially Legal documents more likely to being lengthy and verbose. There is possibility that legal documents contain redundant data that is not important. But, manually reading the whole document and finding the relevant non duplicate information from documents is tedious job for anyone. So we proposed to implement some well known machine learning algorithms to find relevant sentences.We are planning to implement it for not only for online documents but also offline paper with the help of OCR

Literature survey

Semi-supervised machine learning for textual anomaly detection

Authors- Carl Steyn and Alta De Waal

Firstly the need of anomaly detection has been mentioned. As it helps to solve the problem in text analysis where large set of irrelevant document are been used for particular scenario. This paper discuss way of finding a solution to the above problems, for the specific domain at hand. The general problem faced in anomaly detection is that the performance of each algorithm or method largely depends on the domain it is applied to. So the wide variety of following semi supervised algorithm has been used in this paper-

A) Naïve Bayes Text Classification: Here each documents have been divided into an ordered list of words. Now each unique word is used. We can assume that there is independence between words once we know the class that it belongs to. The portion of training data that belongs to each class is termed as Class Prior distribution. Theta will be the set of all unique words in the vocabulary and the Class Prior probabilities. Then the estimation of the parameters in our model by using Maximum a Posteriori (MAP) estimation. Zero probability problem is also solved beforehand.

B) Expectation Maximization: It is an iterative algorithm used to estimate parameters when dealing with incomplete data. By iterating over (E) and (M) steps, the algorithm maximizes the log-likelihood of the model by convergence, which is used to find the MAP parameter estimates. Two types of data are present in the training data - some labelled documents and large amounts of unlabelled data. Missing labels in our training set can make our data incomplete but it is main source for collecting result in this algorithm. Instead of the original log-likelihood of the model, We will use a different one which is exactly same but without the sum of logs. There is (E) and (M) steps, in which one is used to determine the expected values of Z given the current parameter value i.e. Theta and the other is used to determine new parameter with help of updated value of Z. Initially theta is found by using naïve bayes on label data.

Drawbacks of Existing System

Carl Steyn and Alta De Waal proposal to use Naive Bayes and Expected Maximum seems interesting. But it does not give the intended result in case of limited training data set.

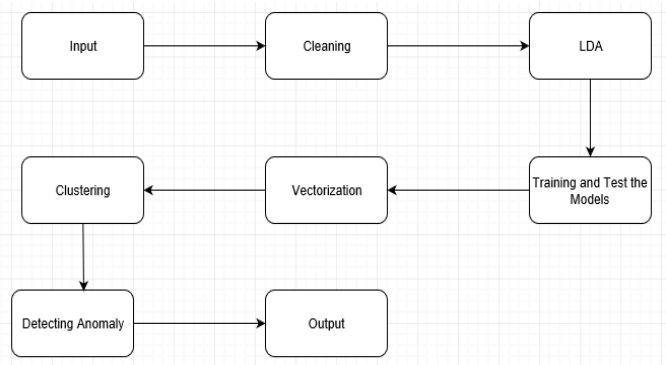
Draft of proposal

Legal documents usually very verbose. It is a very tedious task to read all the content in the document. Moreover the language in the documents is so convoluted that a layman is not able to understand the specifics of the terms. So it is essential that the user knows that part of the document might cause him any mishap in the future before signing.The solution is an automatic highlighter which highlights the anomalous content in the documents. The solution proposes a model for offline documents. The User can upload the document. Then the algorithm analysis the document through various techniques and uses the trained model to find out text to be highlighted. The sentences are the highlighted as the solution.

Expected modules:

- Perform preprocessing on the data to remove redundant content, noise, etc.
- These techniques include lemmatizing, Porter stemming, etc.
- Perform Contextual analysis on the data.
- Perform Word to Vector operations.
- Perform Hierarchical Agglomerative clustering.

Proposed Architecture



Architecture Description

The various modules depicted in Figure 1 are explained as below:

- **Input:** Collect all documents and build a corpus. In this process, we scrap the web to find a maximum number of software documents related to a common domain.
- **Cleansing:** Clean the documents by dealing with suffixes, Remove prepositions, punctuations, and numbers.
- **LDA Modelling:** Perform LDA to get common topics. We need to fine tune the model to get finite number of topics. After running LDA on all the documents, we get a list of topic words. Perplexity score can be used to limit the number of topics.
- **Training set and Testing:** Train and test model get a bunch of common topics.
- **Vectors:** convert the remaining words to create sentence vectors. After this, we establish a mapping from words to vectors.
- **Clustering:** cluster sentence vector using agglomerative clustering. Clustering gets us clusters which are common clusters
- **Outlier Detection:** Use local outlier factor to get outlier

which is our anomaly. Outliers sentence vector is the anomaly we are searching for.

- **Output:** Highlight the anomalous sentence which is our result

Features of Proposed System

1. **Platform Independence:** The transformation performed on the code are independent of the nature of application and are applied on high-levelcode.
2. **Time Saving:** No Software / Hardware installing needed.
3. **Wide Range:** Diversity is provided because of wide range of documents available
4. **Low cost:** With the automation of various transformations and compatibility with existing systems, the code involves low maintenance cost and efficient use of resources. The only cost that will be incurred is that of aGPU.

Resource Requirement

A) Hardware Requirements

- PC Scanner – It used as input provider.
- RAM – It should be based on a computer having minimum 8GB RAM
- Graphics Card- high end card like NvidiaGEForce 840x

B) Software Requirements

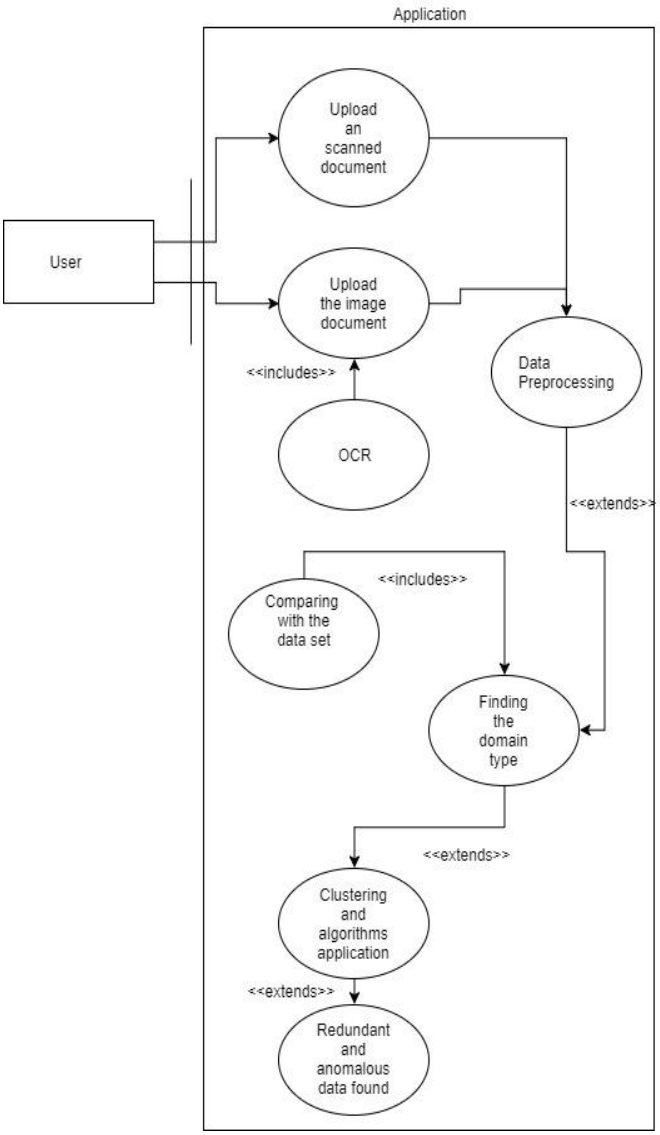
- Scikit – It is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- Python 3.3-The main program language will be python.
- NLTK- NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries

C) OperatingRequirements

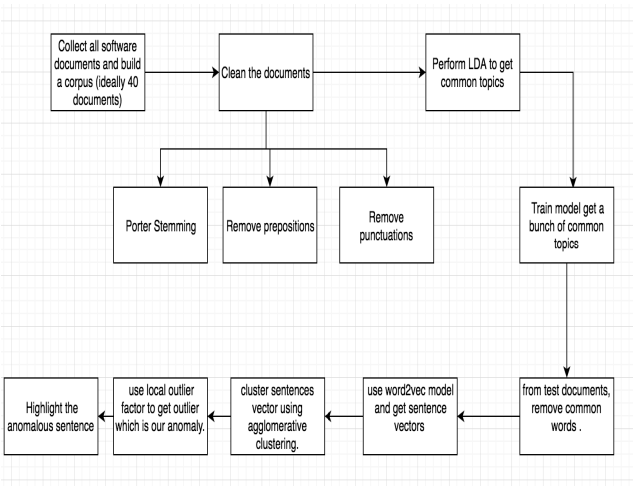
- Windows 8 andabove

Since it is an application, it can be accessed from anywhere in any operating environment

Use Case Diagram



System Architecture



Component Description

The various modules depicted in Figure are explained as below:

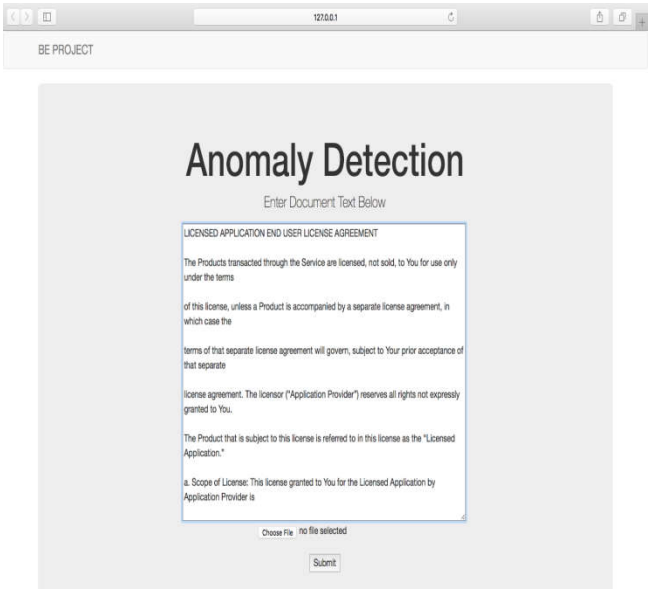
- **Data Collection:** Collect all software documents and build a corpus (ideally 40 documents). In this process, we scrap the web to find a maximum number of software documents related to a common domain.
- **Cleansing:** Clean the documents by performing porter stemming by dealing with suffixes. Remove prepositions, punctuations, and numbers. For example, in our model, "I have an iPhone", and ," I have iPhone", mean the same.
- **LDA Modelling**: Perform LDA to get common topics. We need to fine tune the model to get finite number of topics. After running LDA on all the documents, we get a list of topic words. Perplexity score can be used to limit the number of topics.
- **Training set**: Train model get a bunch of common topics.

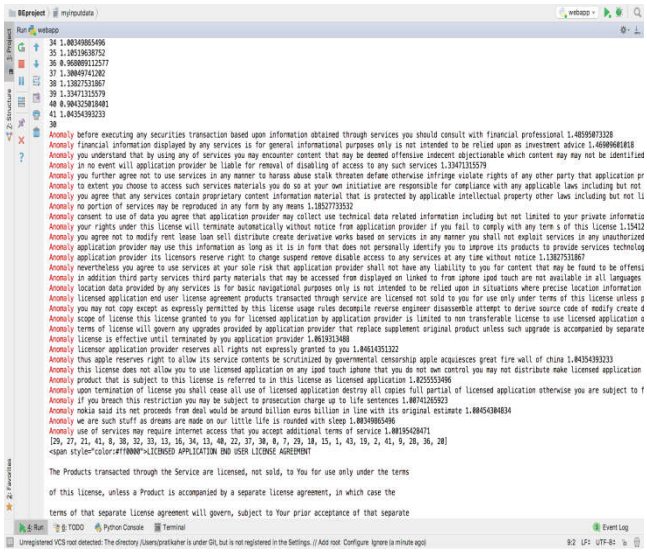
- **Common words removal**: From test documents, remove common words. We observe that these topic words do not contribute to the "specialty" of a sentence. Therefore, we remove the topic words from the test document.
- **Vectors**: use word2vec model to convert the remaining words to create sentence vectors. After running Word 2Vec, we establish a mapping from words to vectors.
- **Clustering**: cluster sentence vector using agglomerative clustering. Clustering gets us clusters which are common clusters
- **Outlier Detection**: Use local outlier factor to get outlier which is our anomaly. Outliers sentence vector is the anomaly we are searching for.
- **Detection**: Highlight the anomalous sentence which is our result

Algorithms

- **Porter Stemming**: For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing., it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set. The goal of both stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.
- **Latent Dirichlet allocation (LDA)**: it is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics
- **Word2vec**- It is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space
- **K-means clustering:** It is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
- **Agglomerative Clustering:** Itis a method of cluster analysis which seeks to build a hierarchy of clusters. This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Output





Conclusion

Tedious methods of reading a long legal statement no longer appeal to anyone. Many fraud have been conducted in this way by adding unfair condition. This is going on for years. Lately, a need to counter this has been in demand because of rise of technology. For a document like Legal paper, which involves a lot of facts, a novice may find it very difficult to find various important details and hence making him/her a victim in some cases. To overcome these hurdles, to prevent the customers from this and became a ray of hope in this dark clouds of tedious long documents, Our Project was developed. Via this project, Reductant data is removed. With this, more unique statement are highlighted which have a higher probability to be a harmful to the customer. Moreover, this project will goes through the entire documentation - from introduction to conclusion and shortens the length for the customers. With increasing advances being made in the world of Machine Learning, new algorithm for anomaly detection will come and become even more effective and efficient. This is already being seen as new process like lemmization has appeared as an alternative to stemming. Thus, Machine Learning is the future of anomaly detection. We have created this project as symbol of how effective these algorithm are for the above purpose.

REFERENCES

Amogh, Mahapatra, Nisheeth, Srivastava and Jaideep, Srivastava, Contextual anomaly detection for text data; Publisher: MDPI

Carl, Steyn; and Alta De Waal, 2013. Semi-supervised machine learning for textual anomaly detection; Publisher: IEEE

Jude, Shavlik, and Sriraam, Natarajan. Anomaly Detection in Text: The Value of Domain Knowledge; Publisher: Twenty-Eighth International Florida Artificial Intelligence Research Society Conference.

Mikolov, Tomas, *et al.* 2013. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems.

Radim Rehurek, Optimizing word2vec in gensim. Nov 07, 2015

****