Relevance of Text Analytics

At AlgoAnalytics

May 2, 2017

Outline

- About AlgoAnalytics
- Contracts Management
- Structured Document Decomposition
- Document Similarity in Text Analytics
- Predicting number of days for a case
- Other Text Analytics relevance
- Technologies

CEO and company Profile

About AlgoAnalytics

Analytics Consultancy

- Work at the intersection of mathematics and other domains
- Harness data to provide insight and solutions to our clients

Led by Aniruddha Pant

- +30 data scientists with experience in mathematics and engineering
- Team strengths include ability to deal with structured/ unstructured data, classical ML as well as deep learning using cutting edge methodologies

Expertise in Mathematics and Computer Science

- Develop advanced mathematical models or solutions for a wide range of industries:
- Financial services, Retail, economics, healthcare, BFSI, telecom, …

Working with Domain Specialists

- Work closely with domain experts – either from the clients side or our own – to effectively model the problem to be solved

Aniruddha Pant

CEO and Founder of AlgoAnalytics

PhD, Control systems, University of California at Berkeley, USA 2001

Highlights

- 20+ years in application of advanced mathematical techniques to academic and enterprise problems.
- Experience in application of machine learning to various business problems.
- Experience in financial markets trading; Indian as well as global markets.

Expertise

- Experience in cross-domain application of **basic scientific process**.
- Research in areas ranging **from biology to financial markets to military applications**.
- Close collaboration with premier educational institutes in India, USA & Europe.
- Active involvement in startup ecosystem in India.

Prior Experience

- Vice President, Capital Metrics and Risk Solutions
- Head of Analytics Competency Center, Persistent Systems
- Scientist and Group Leader, Tata Consultancy Services

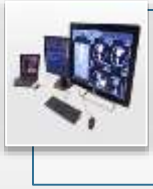
AlgoAnalytics - One Stop AI Shop

BFSI

- Dormancy Analysis
- Recommender System
- Credit/Collection Score

Retail

- Churn Analysis
- Recommender System
- Image Analytics



Healthcare

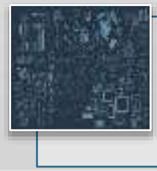
- Medical Image Diagnostics
- Work flow optimization
- Cash flow forecasting

Legal

- Contracts Management
- Structured Document decomposition
- Document similarity in text analytics

Internet of Things

- Predictive maintenance in ovens
- Air leakage detection
- Engine/compressor fault detection



Others

- Algorithmic trading strategies
- Risk sensing – network theory
- Network failure model

- We use structured data to design our predictive analytics solutions like churn, recommender sys
- We use techniques like clustering, Recurrent Neural Networks,

Structured Data

- We use text data analytics for designing solutions like sentiment analysis, news summarization and many more
- We use techniques like natural language processing, word2vec, deep learning, TF-IDF

Text Data

- Image data is used for predicting existence of particular pathology, image recognition and many others
- We use techniques like deep learning – convolutional neural network, artificial neural networks and technologies like TensorFlow

Image Data

- We use sound data to design factory solutions like air leakage detection, identification of empty and loaded strokes from press data, engine-compressor fault detection
- We use techniques like deep learning

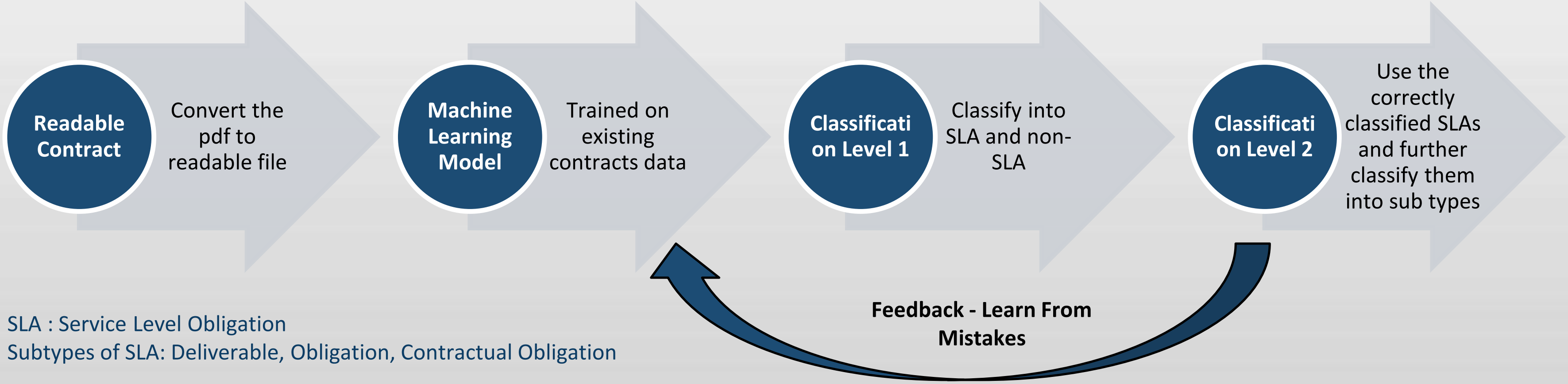
Sound Data

Contracts Management – Overview

Motivation:

- Automate / semi automate manual labor to read and extract information from legal contracts
- Classify the legal contract paragraphs in to SLA vs Non SLA
- Use Natural Language Processing to extract meaningful information like name, place, location, entity, dates, amounts etc.
- Similar approach can be used for any text classification problem

Overview:



Readable Contract – Convert the pdf to readable file

Machine Learning Model – Trained on existing contracts data

Classification on Level 1 – Classify into SLA and non-SLA

Classification on Level 2 – Use the correctly classified SLAs and further classify them into sub types

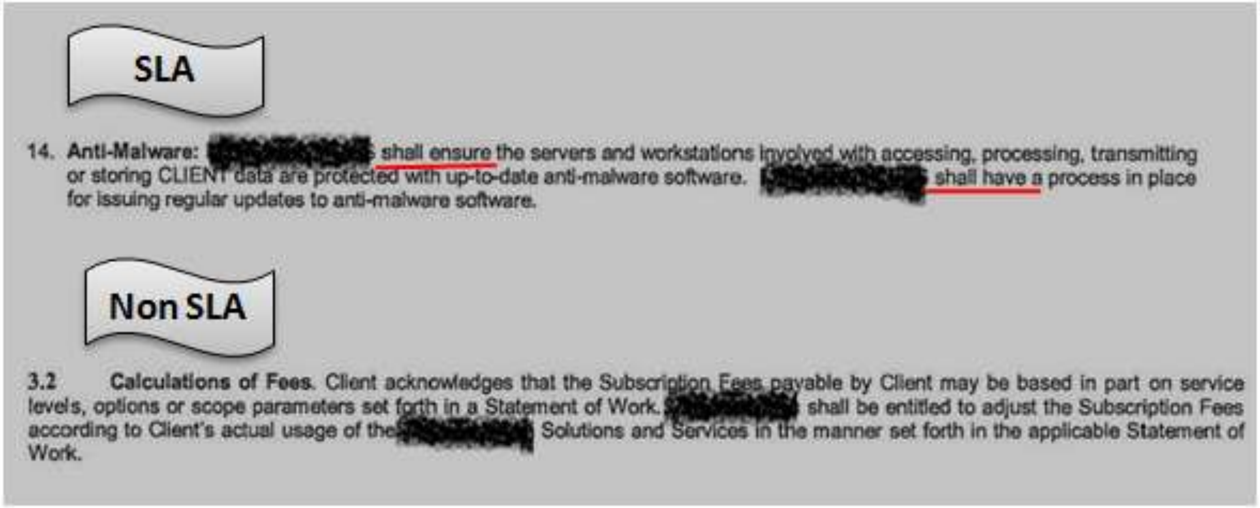
Feedback - Learn From Mistakes

SLA : Service Level Obligation
Subtypes of SLA: Deliverable, Obligation, Contractual Obligation

SLA vs Non SLA

SLA Vs. Non SLA

- Learn from structure and content of text
- Similar supervised learning problem can be designed for any structured text. E.g. Relevant text Vs. Irrelevant text
- Relevant text can then further be classified into more subclasses



SLA

14. **Anti-Malware:** [redacted] shall ensure the servers and workstations involved with accessing, processing, transmitting or storing CLIENT data are protected with up-to-date anti-malware software. [redacted] shall have a process in place for issuing regular updates to anti-malware software.

Non SLA

3.2 **Calculations of Fees**. Client acknowledges that the Subscription Fees payable by Client may be based in part on service levels, options or scope parameters set forth in a Statement of Work. [redacted] shall be entitled to adjust the Subscription Fees according to Client's actual usage of the [redacted] Solutions and Services in the manner set forth in the applicable Statement of Work.

Results

SLA Vs. Non SLA Classification

Average Accuracy	99.26%
Average Kappa	97.85%
Average ROC	99.53%
Sensitivity	100%
Specificity	96.43%

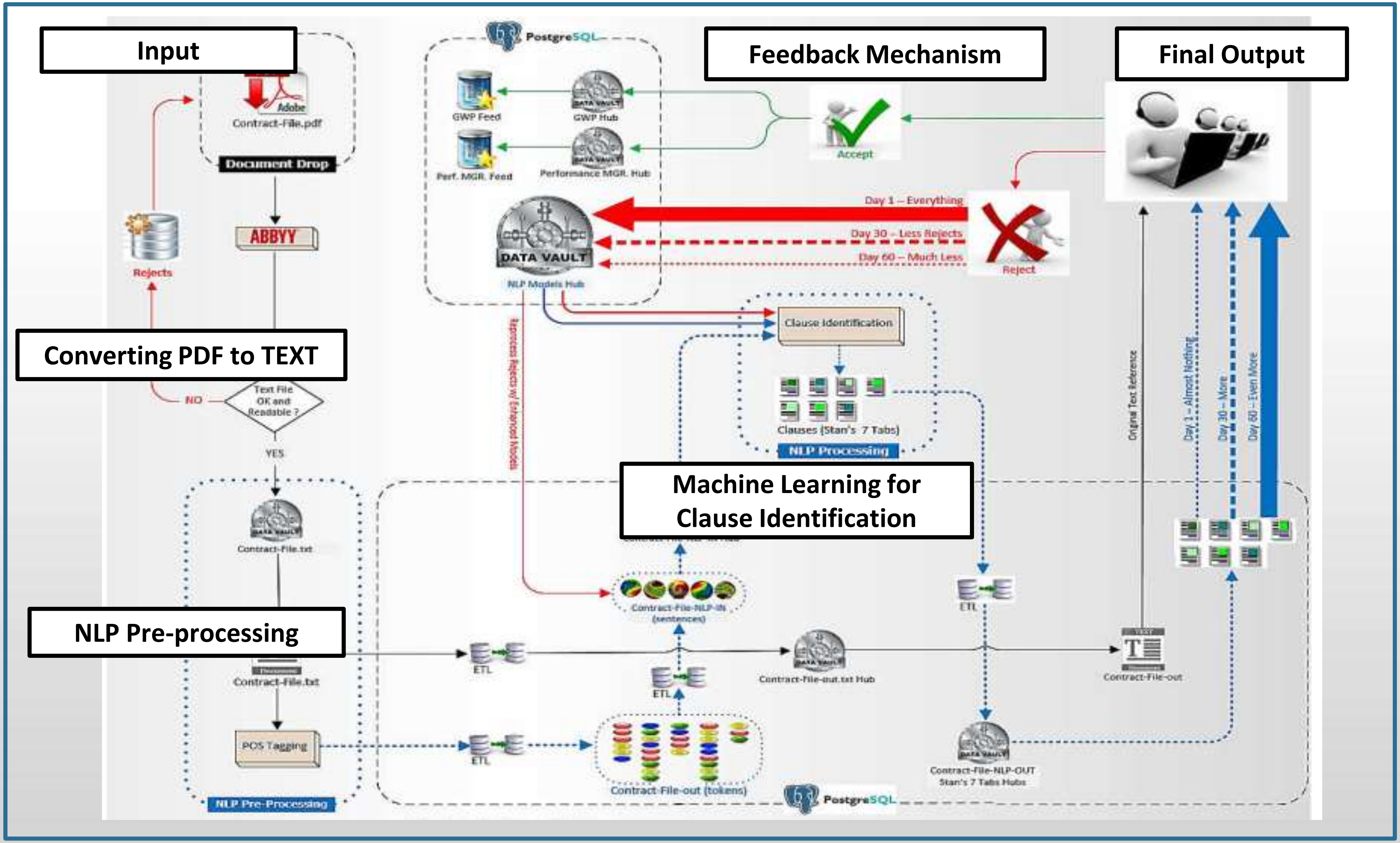
Confusion Matrix		
	SLA Predicted	NonSLA Predicted
SLA Actual	54	0
NonSLA Actual	2	213

Within a single contract

Average Accuracy	74.44%
Average Kappa	44.02%
Average ROC	78.77%
Sensitivity	95%
Specificity	46.43%

Confusion Matrix		
	SLA Predicted	NonSLA Predicted
SLA Actual	26	4
NonSLA Actual	30	73

Detailed Schema

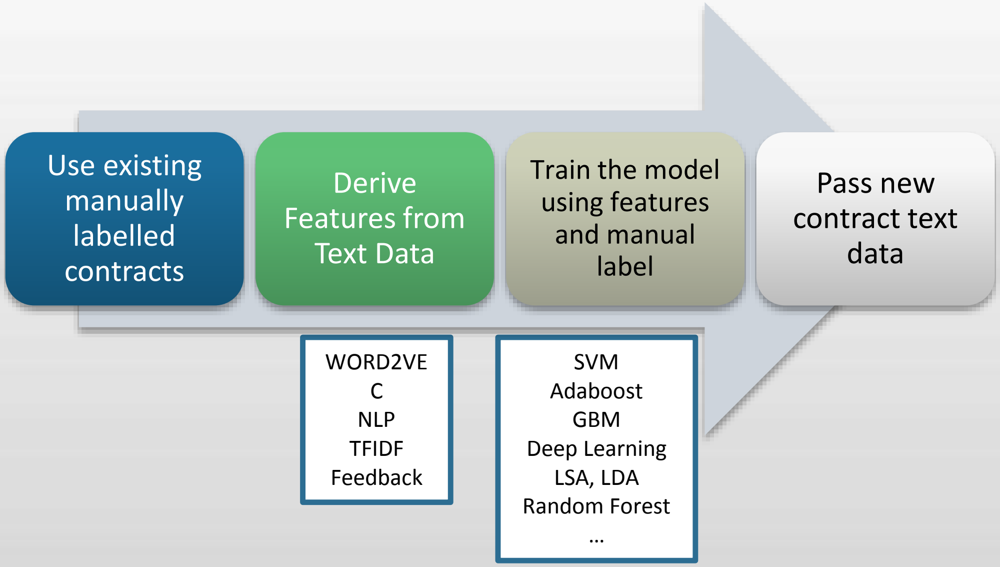


AlgoAnalytics

Applying Machine Learning and NLP

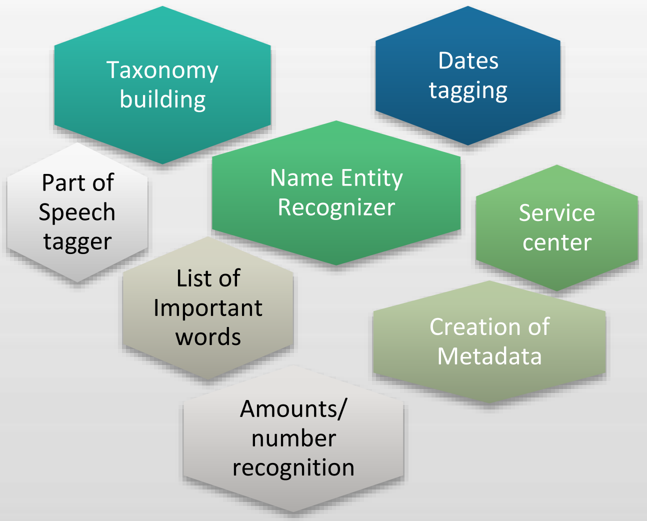
Machine Learning and NLP

- Machine learned features – TFIDF and latest Word2Vec
- Human feedback for misclassified text will also be used as features
- NLP has been used in Cleaning of Text, Topic Detection , Keyword Extraction, Summarizing the text ,etc.
- The name and entity recognition can be effectively used in any text application.
- Sentiment Analysis has been extensively used in risk event detection.



Use existing manually labelled contracts
Derive Features from Text Data
Train the model using features and manual label
Pass new contract text data
WORD2VEC
NLP
TFIDF
Feedback
SVM
Adaboost
GBM
Deep Learning
LSA, LDA
Random Forest
…

Machine learning process overview



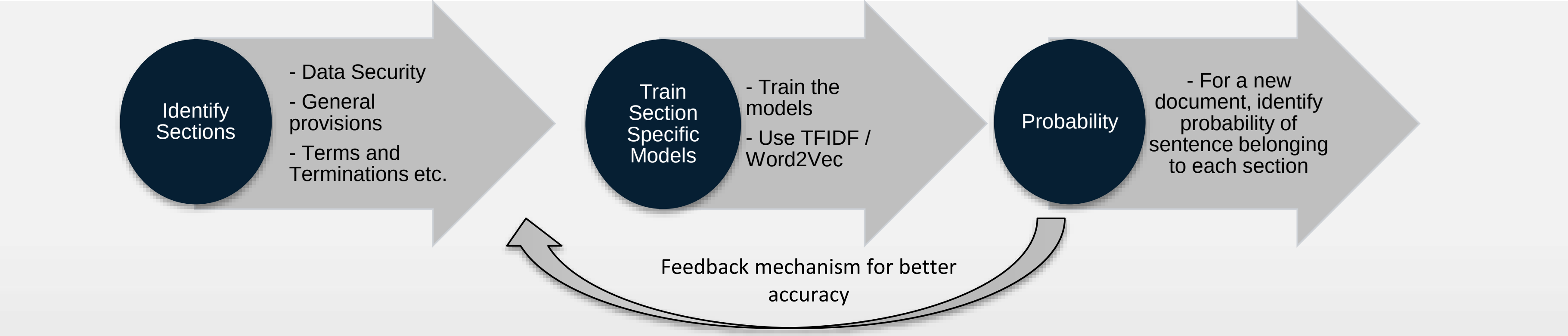
NLP

Structured Contract Decomposition : Motivation and Basic Schema

Motivation:

- Build section specific models and intelligence for an individual section
- The combination of models gives the probability of sentence belonging to that section
- The decomposition methodology framework can be extended to any structured document / text.

Basic schema:



Identify Sections

- Data Security
- General provisions
- Terms and Terminations etc.

Train Section Specific Models

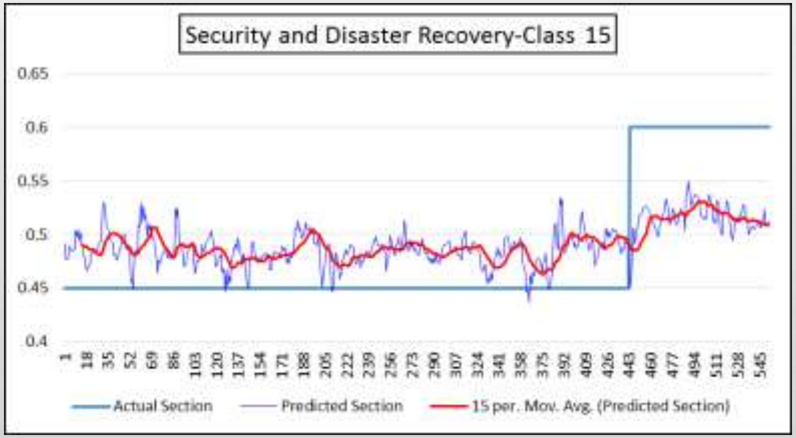
- Train the models
- Use TFIDF / Word2Vec

Probability

- For a new document, identify probability of sentence belonging to each section

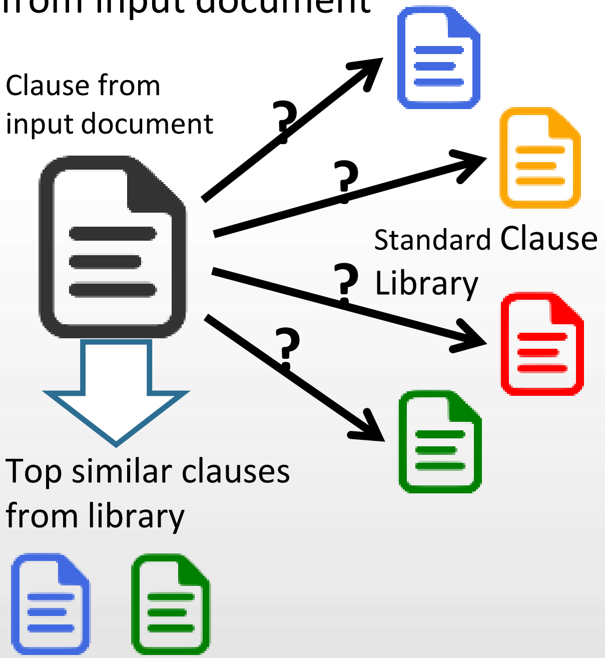
Feedback mechanism for better accuracy

- The structured document can be decomposed using machine learning and text analytics methods .
- The method can also tell us ALL section of a document where assigned section topic is being mentioned.
- Example shows how we can separate out the Security and Disaster Recovery sentences and effectively the entire section
  - X-Axis : Sentences
  - Y-Axis : Probability
  - Probability < 0.5 Not belonging to section
  - Probability >0.5 Belongs to a section



Document Similarity in Text Analytics

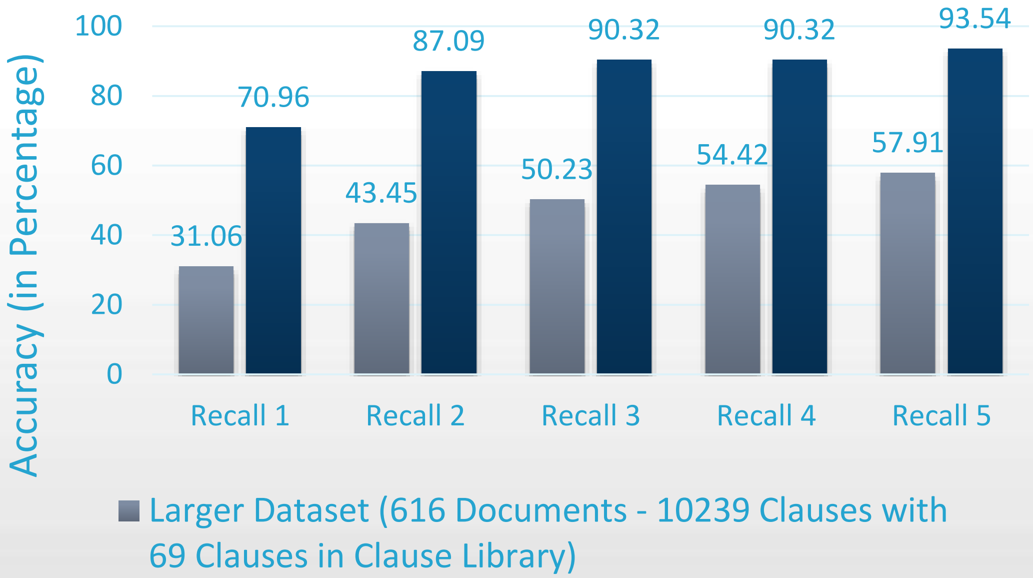
- **Problem Statement -** Finding semantically similar clause from standard clause library for each clause from input document



Clause from input document

Standard Clause Library

Top similar clauses from library



Recall Performance

Methods

A. Frequency based similarity
B. Unigrams and bigrams modeling
C. TF-IDF
D. Latent Semantic Analysis
E. Word2Vec model

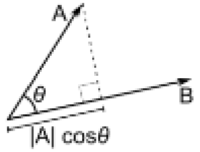
Ensemble Technique: Combination of above models to improve performance

Limitations of Unsupervised Approach

(1) Large no. of clauses (2) Unusual size of clauses
(3) Noisy data (4) Idiosyncrasies of data

Overview of Modeling Process

I. Removing punctuations and special characters
II. Stop-words removal
III. Tokenization: paragraphs as list of words



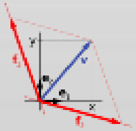
Cosine similarity metric for finding similar vectors

Clause Segmentation → Data Cleaning → Vector Space Model → Similarity Function → Model Evaluation

Dividing text documents into set of paragraphs (clauses)

Vector representation of text data

1. Frequency based
2. Unigrams and Bigrams
3. TF-IDF representation
4. word2vec model



Recall@N metric to evaluate performance of various models

Predict number of days for a case to get approval

- Predict the number of days a case will take to get approval, based on given question-answers for all cases
- Find most similar cases from the dataset (question-answers for all cases) using distance scores

Pre-Processing :

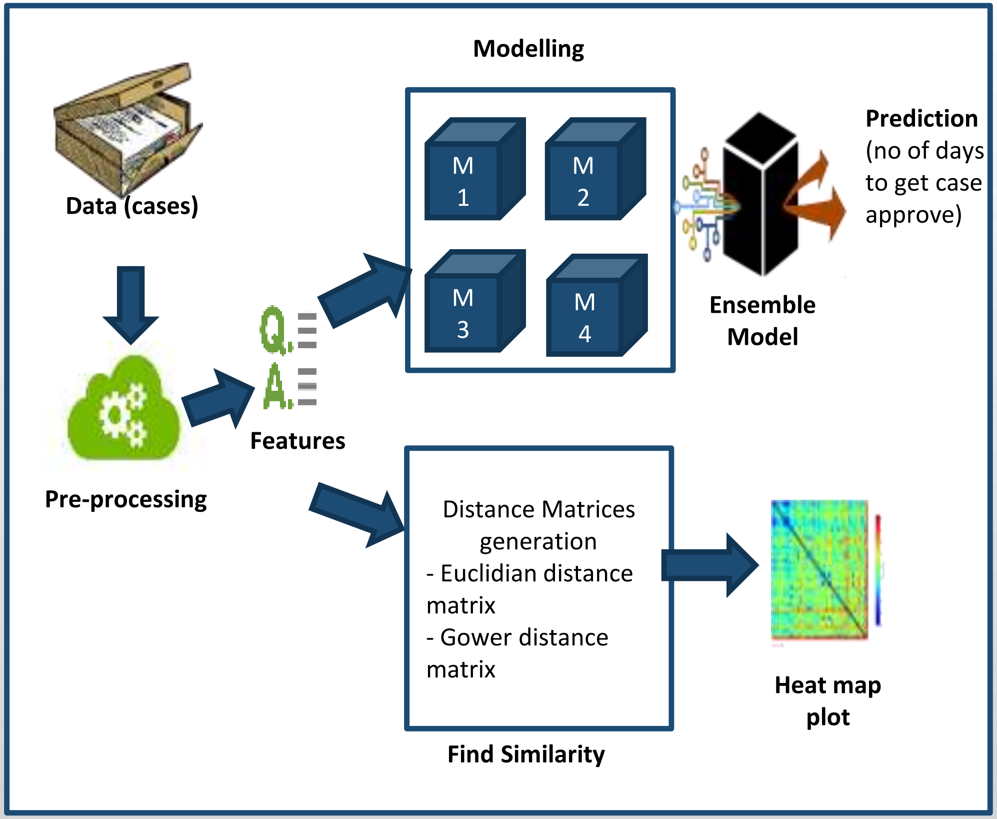
Only those questions for which **at least 65%** of the documents had answers were considered for analysis

Questions with more than 5 possible answers were ignored

Empty or NAs were replaced with “Not Apllicable”

Finally, questions with less than 5% variance in answers were discarded e.g. if a question contains all the NOs then it will not be considered

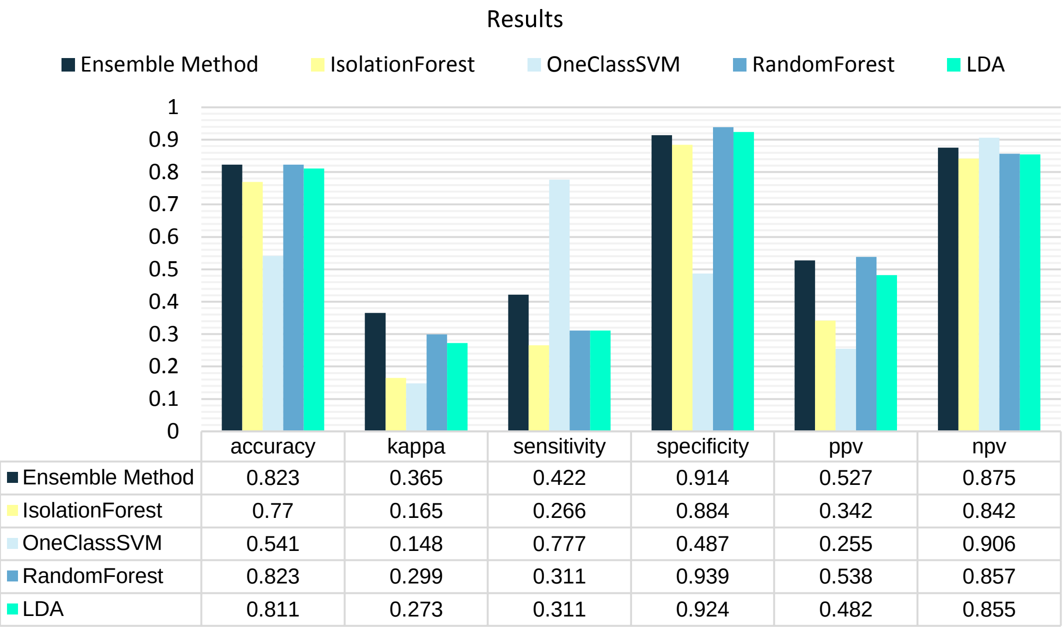
Methodology :



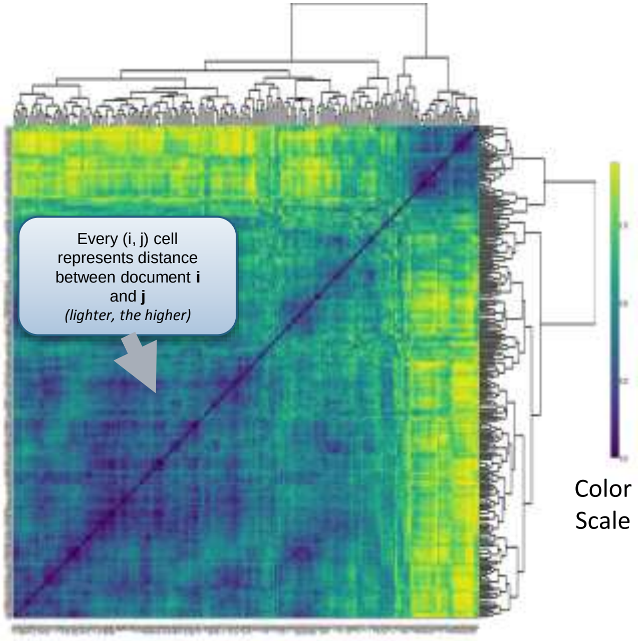
Distance Matrices generation
- Euclidian distance matrix
- Gower distance matrix

Results

- Models used for prediction :
  - Outlier Detection : 1. IsolationForest 2. OneClassSVM
  - Classification Method : 1. RandomForest 2. LinearDiscriminantAnalysis
- Methods used for distance scores:
  - 1. Euclidian distance 2. Gower distance
- Input Data:
  - 244 cases with 18 features(question-answers)



	accuracy	kappa	sensitivity	specificity	ppv	npv
Ensemble Method	0.823	0.365	0.422	0.914	0.527	0.875
IsolationForest	0.77	0.165	0.266	0.884	0.342	0.842
OneClassSVM	0.541	0.148	0.777	0.487	0.255	0.906
RandomForest	0.823	0.299	0.311	0.939	0.538	0.857
LDA	0.811	0.273	0.311	0.924	0.482	0.855



Every (i, j) cell represents distance between document **i** and **j** *(lighter, the higher)*

Color Scale

AlgoAnalytics

Other Supporting Text Analytics Work

Twitter Analytics

- Identify, process and group together relevant tweets using machine learning methods

News Analytics

- Access, identify and analyze relevant news article given a topic

- News summarization

App Development

Download, analyze twitter feeds of stocks to get sentiment and topic detection

Multi-language Sentiment Analysis

- Model can be used to get similar words.

- Trained model can learn proximity of words

Topic Summary & Concept Detection

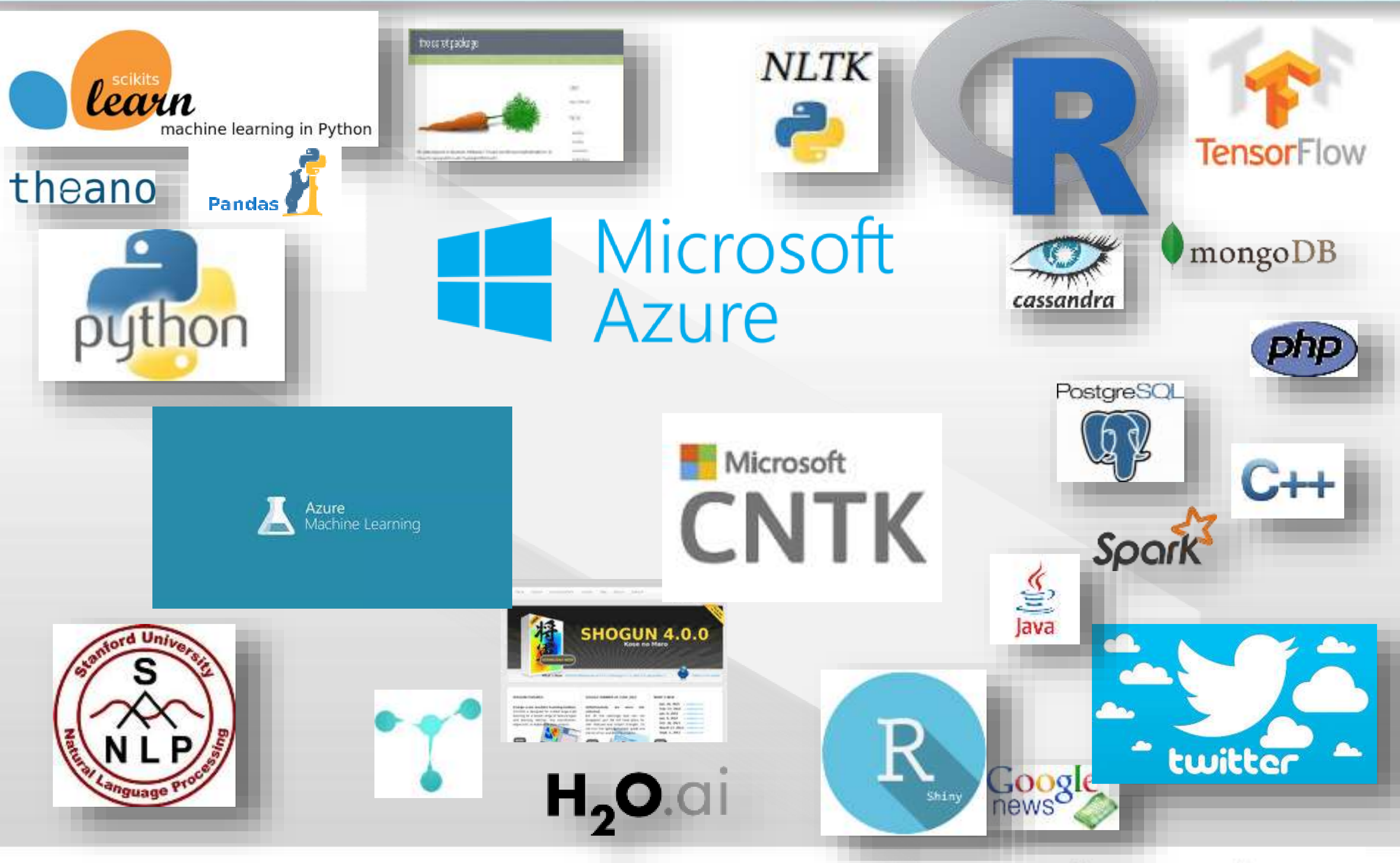
- Keyword extraction

- Summary extraction

- Topic detection

-Words Importance

Technology:


scikits learn
machine learning in Python
the caret package
NLTK
R
TensorFlow
theano
Pandas
python
Microsoft Azure
cassandra
mongoDB
php
PostgreSQL
Azure Machine Learning
Microsoft CNTK
C++
Spark
Java
SHOGUN 4.0.0
Kose no Maro
Stanford University
S
NLP
Natural Language Processing
H2O.ai
R
Shiny
Google news
twitter