rises.io
software solutions

Legal Tech: Advanced Analytics Applied Verticalized Solution in Legal Domain

13 Dec 2016
Version: Ver 0.5

LegalTech

Similarity Search and Auto Summarization Solution

Needs and Challenges

- Time consuming and inefficient processes to find similar documents
- Case summarization take lot for effors in law firms as well as contracts in BFSI
- There is no predictability of the case outcome, it impacts the planning
- Low utilization of analytics and cloud

Approach

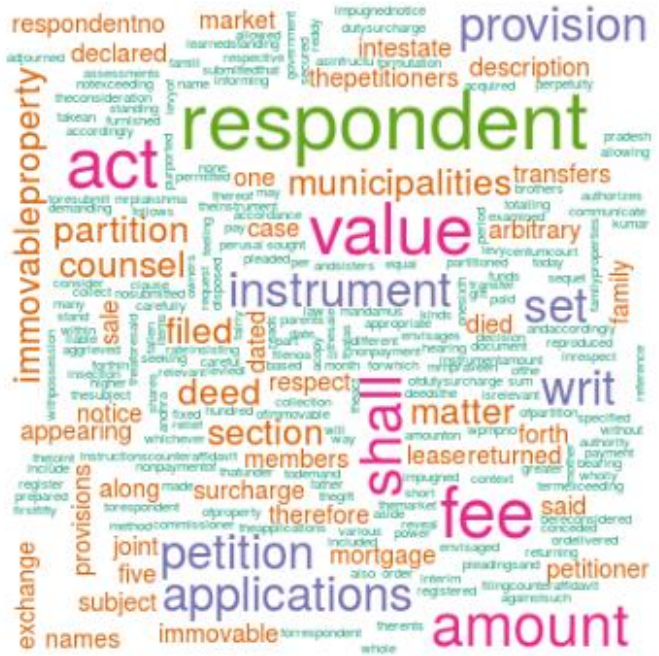
- Advanced analytics techniques to search similar docs
- Use of NLP / Text Analytics to automatically generate the summary of legal doc / contract
- Use of predictive analytics
- Risk solution - Improve the understanding of risk, analyse entire portfolio of documents

Techniques

- Machine learning / AI based Algos
- Natural Languages Processing / Text Analytics
- Predictive Analytics

Solution

- e.g. Finding nearest matching docs using various techniques



Clientele

- LPO/KPO, Legal Departments in BSFI

The Difference

Same workflows but different and more relevant solutions

Finding Relevant Documents

Current Search

- Text based
- A “word” is searched through all documents and ANY document containing the “word” is presented

Improved Search

- Theme based
- Documents are pre-populated with themes, which are matched with the query “word”

Current Outcome

- Slow, mostly manual processing
- Brings many un-important results
- Weeding through them takes time

Improved Outcome

- Faster
- Results based on ranked importance
- The top results are most relevant

Finding Similar Documents

Current Method

- Search based, mostly manual
- A “word” is searched through all documents and ANY document containing the “word” is presented

Current Outcome

- Slow
- Bring many un-important results
- Weeding through them takes time

Improved Similarity

- All documents are pre-populated with pair-wise similarity scores
- For a given document, similar documents list is readily available

Improved Outcome

- Faster
- Results based on ranked importance
- The top results are most relevant

Finding Summary of a Document

Current Method

- Summary prepared by Experts
- “Head Notes” of a judgement

Current Outcome

- Slow
- Not scalable
- Expensive and subjective

Improved Summary

- Sentences are extracted, ranked, ordered and presented
- Domain knowledge is incorporated

Improved Outcome

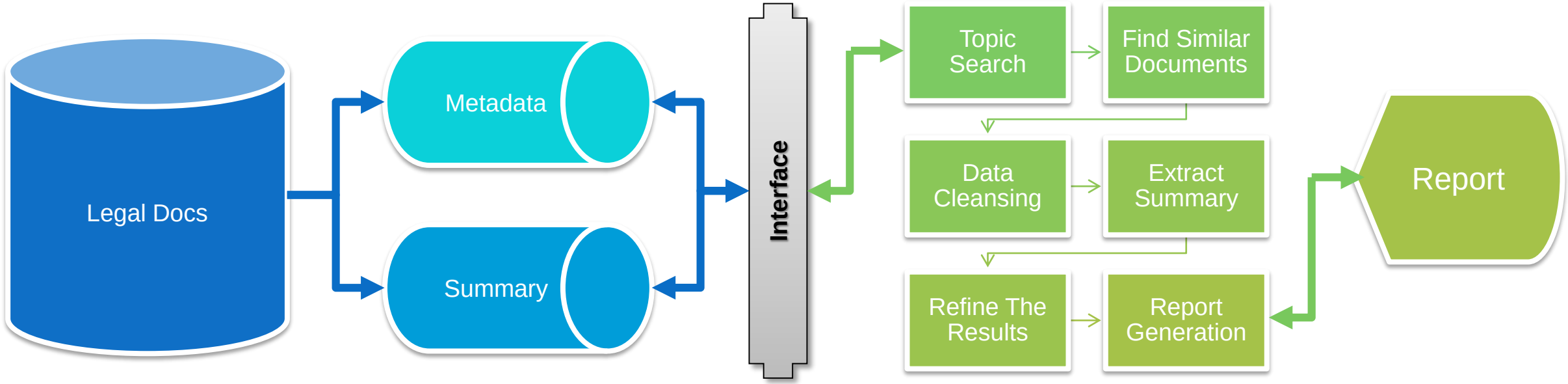
- Faster
- Highly scalable
- Consistent and customizable

The Approach

Incorporating new technologies, algorithms and Deep Learning

Solution Design

Natural Language Processing (NLP) Pipeline applied to LegalTech



Collection of Supreme Court and High Court Judgments

Title, Names of Appellant-Respondents-Judges, important keywords-theme, Core of the judgement

Machine Learning, Natural Language Processing Algos

Keywords/topics, Stats, Summary Doc and Nearest Docs

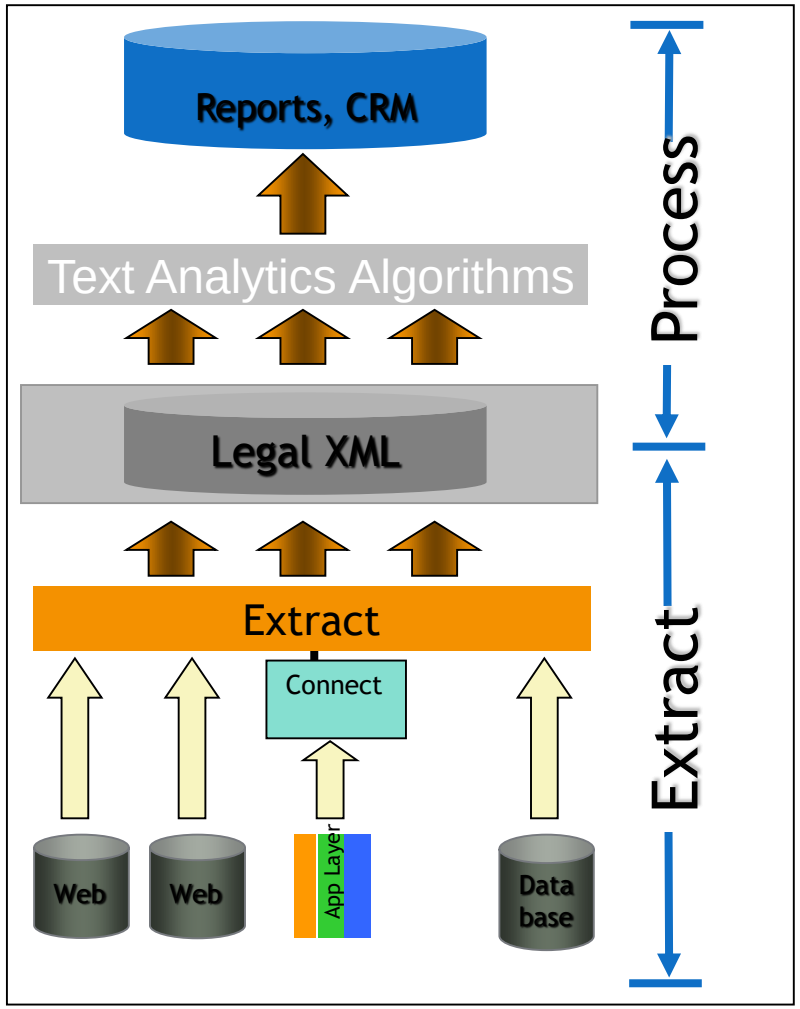
Big Data Platform

Analytics Platform

Visual

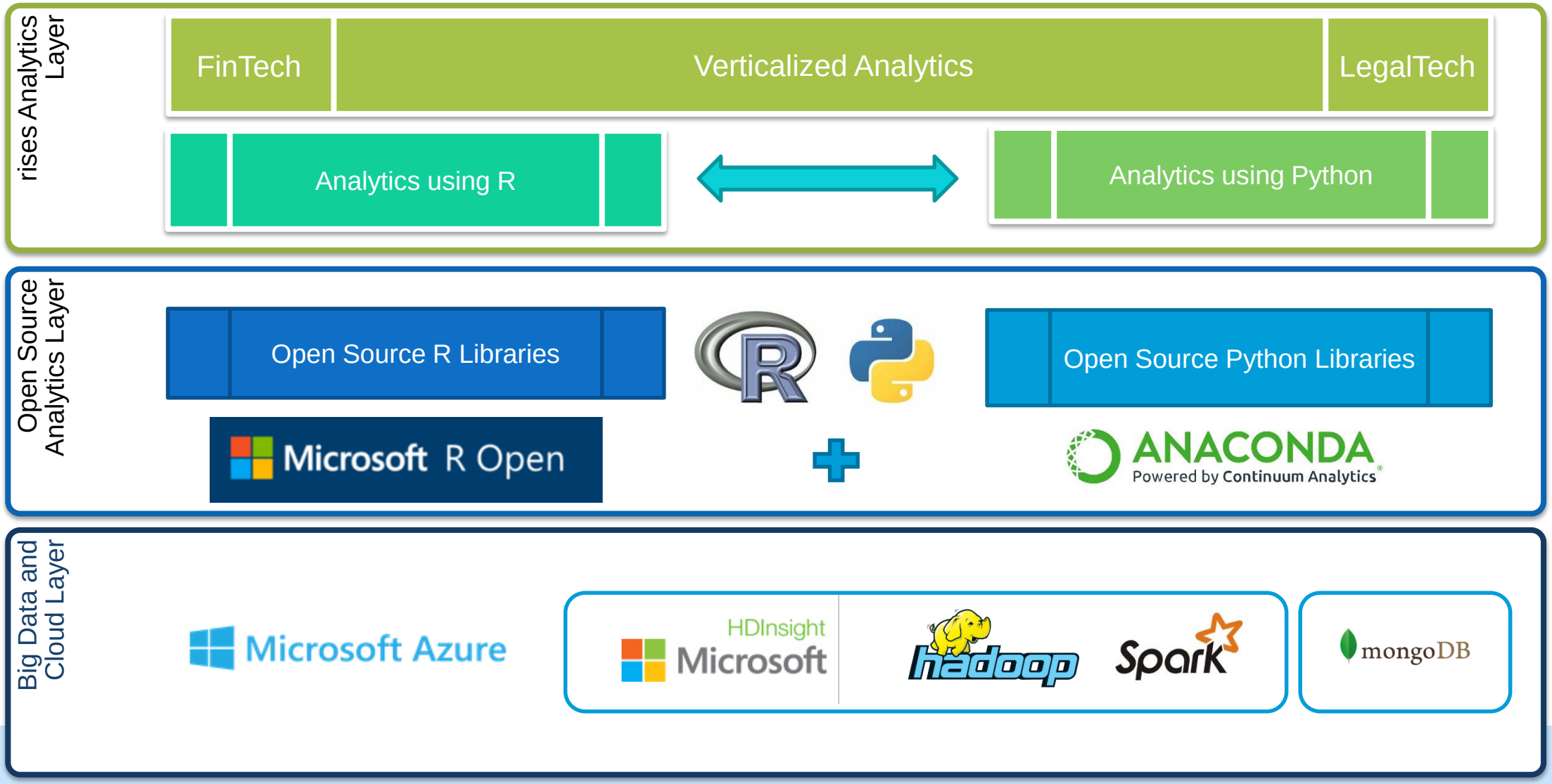
Legal Analytics Workflows

- Legal Text Sources : Web, Databases, Local
- Pre-processing
  - Html -> Text
  - Db -> Text
- Linguistic Pre-Processing
  - Tokenizing
  - Stop word removal
- Text Algorithms: Search, Similarity, Summary
- Results reporting



rises Analytics Platform

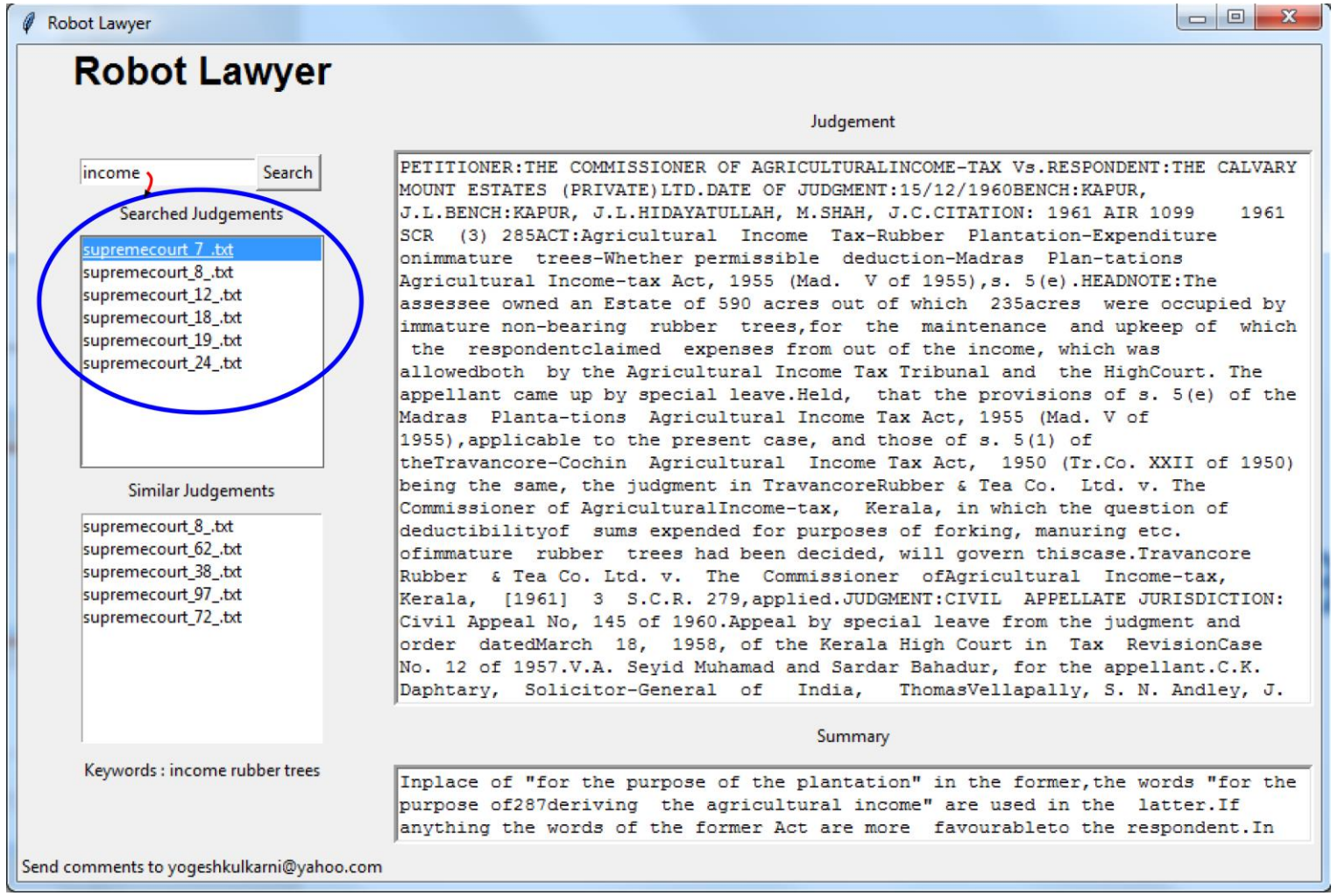
Platform Architecture



Searching

Matching not just a “word” but to the topics or themes of the document

Searches by Topics



Topics by LDA (Latent Dirichlet Allocation)

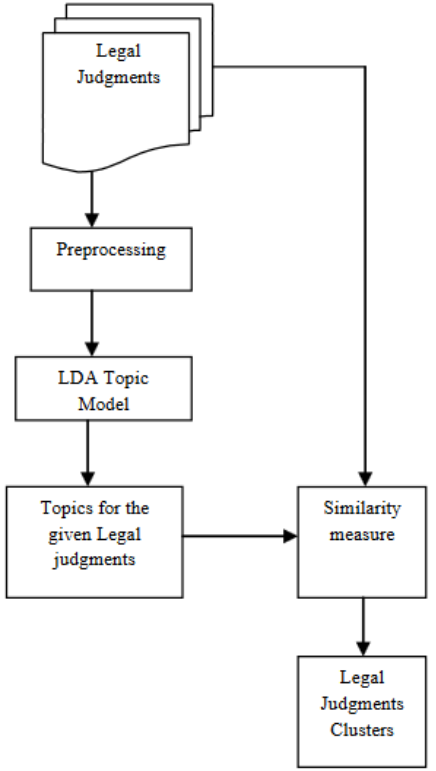
•Discovering hidden topical patterns that are present across the collection

Doc 1: Income generated from Rent was not disclosed.
Doc 2: Present litigation does not come under Rent control act.
Doc 3: Appellant violated the Sales agreement.
Doc 4: Proceeds of the Sale of the Motor were shown as Income
Doc 5: Sales and disclosed Income did not match.

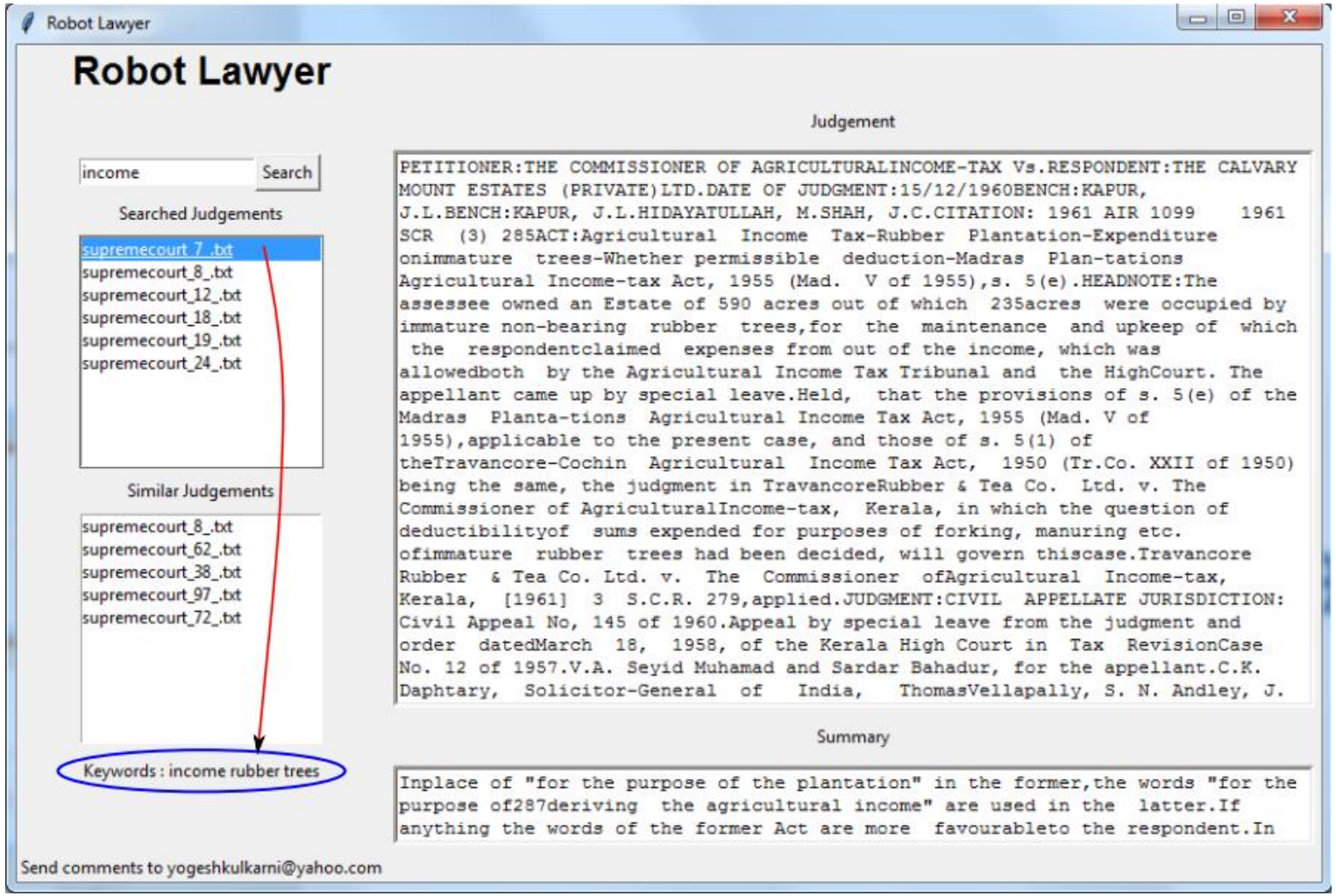
Topic 1: 30% Income, 15% Rent,
Topic 2: 10% Sales, 10% Motor, 5% Income

Docs 1 and 2: 100% Topic 1
Docs 3 and 4: 100% Topic 2
Doc 5: 70% Topic 1, 30% Topic 2

International Journal of Applied Information Systems
Foundation of Computer Science FCS, New York, USA
Volume 2-No 6, 2012 – www.ijais.org



Topics extraction – presented as keywords

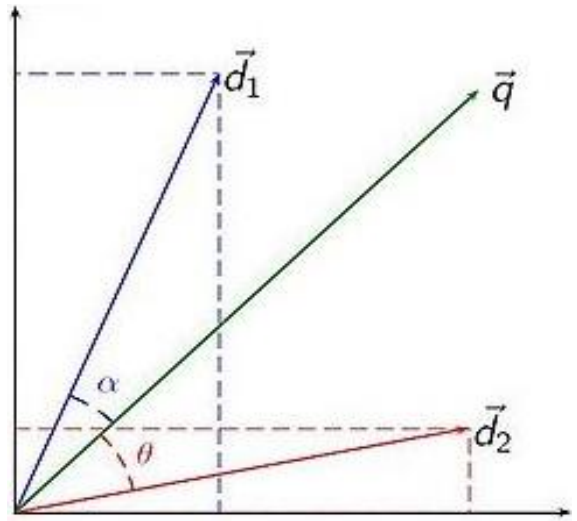


Document Similarity

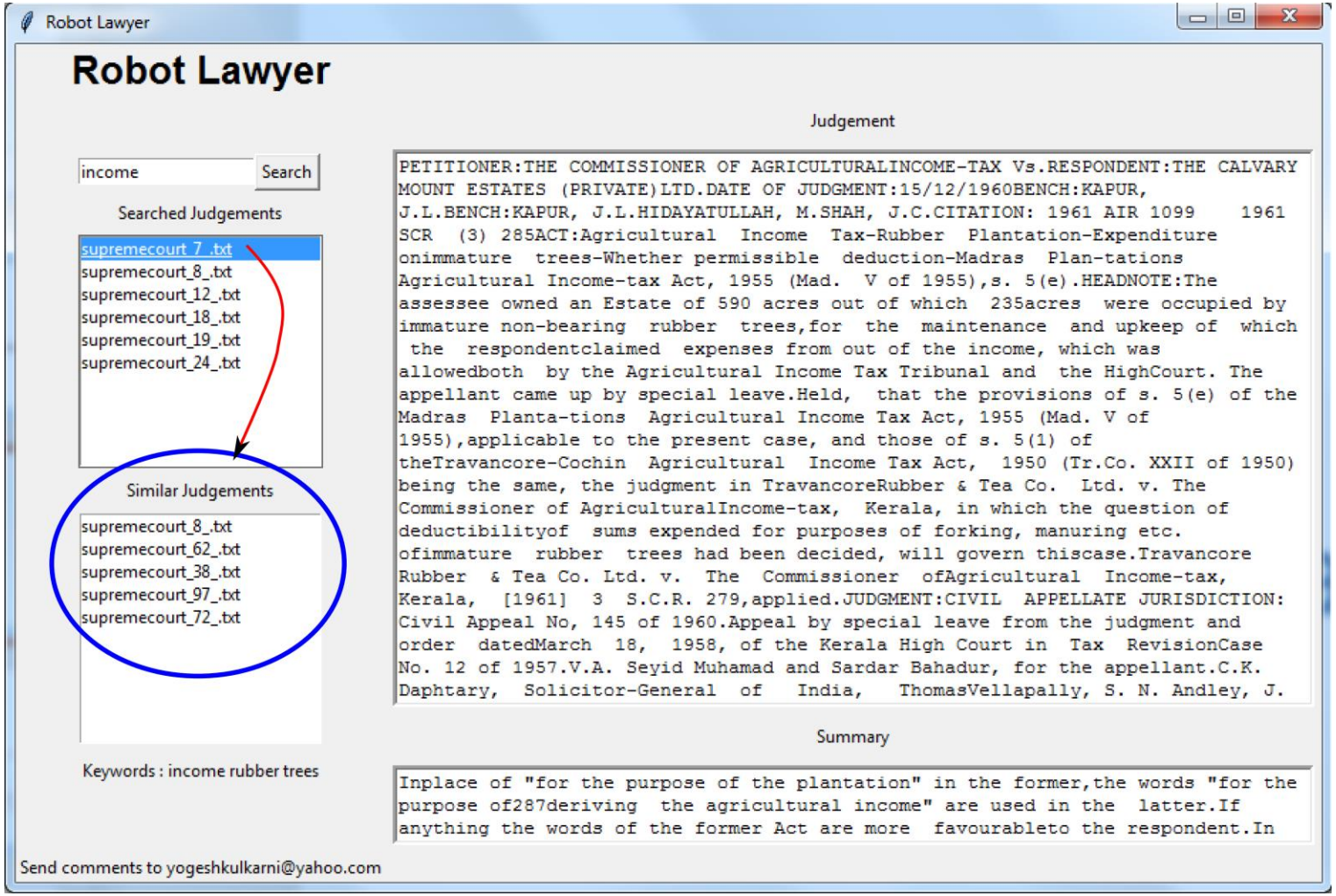
Preparing pair-wise similarity scores

Similarity by TF-IDF and Cosine

- TF-IDF (Term Frequency Inverse Document Frequency)
- Input: Set of documents, each Vectorised by TFIDF
- Similarity between documents d_1 and d_2 is angle between them
- If angle is 0, they are similar (exact)
- If angle is 90 they are different.
- Query q can also be evaluated to see if its closer to d_1 or d_2 then return the closer one.



Pair-wise Similarity Result



Documents Summarization

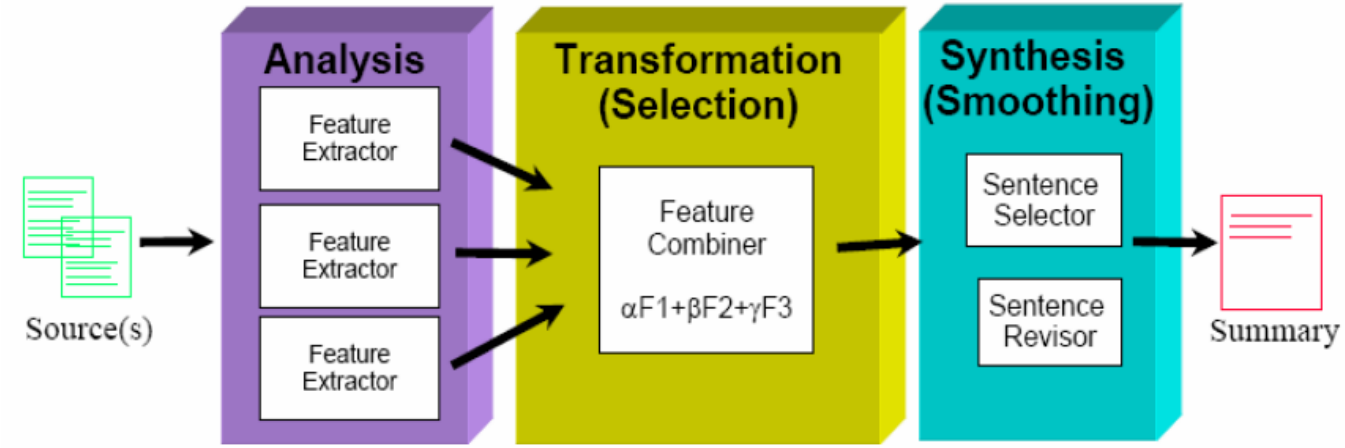
Judgements to Headnotes, Case reports to abstracts

Extractive Rank-based Summarization

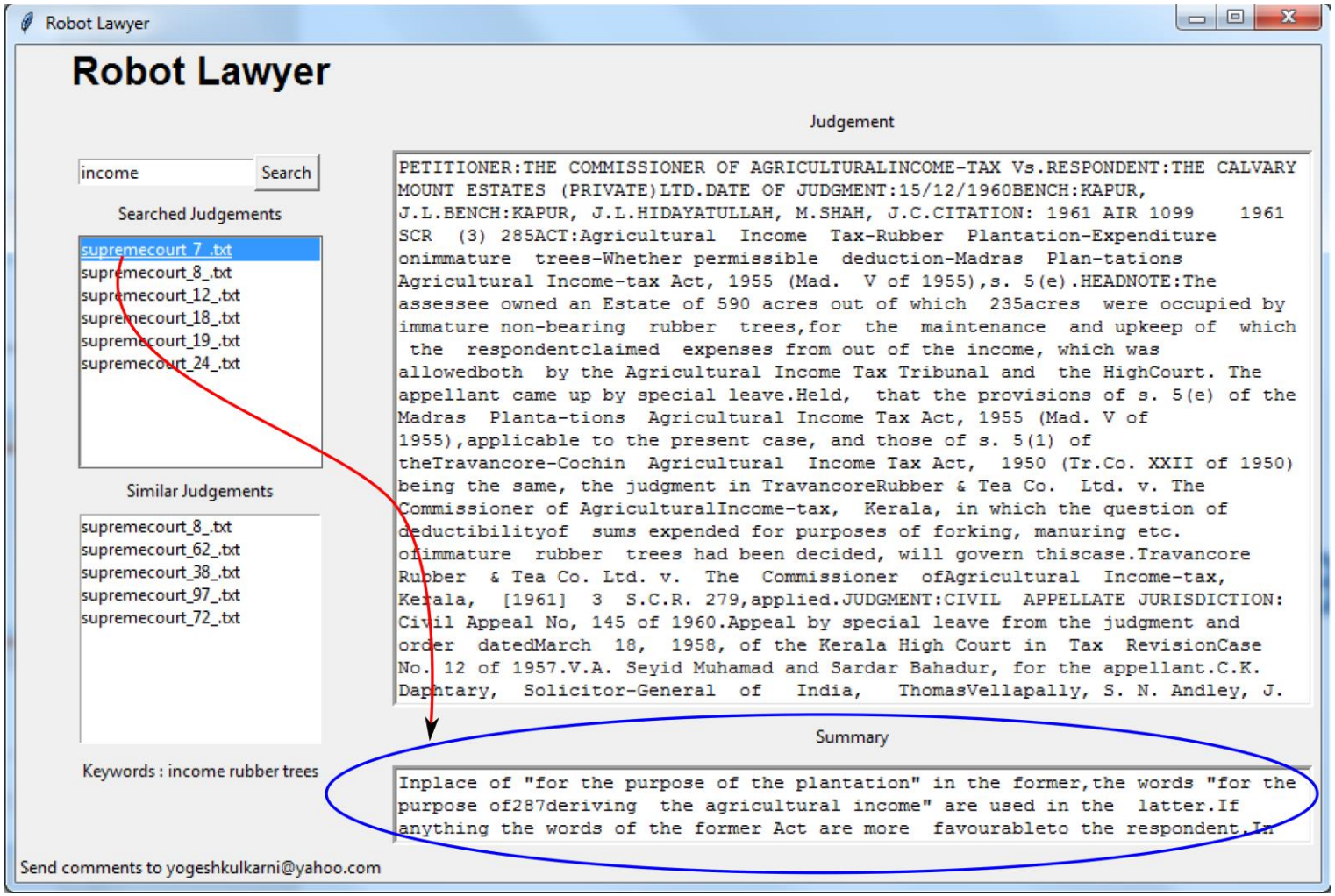
- The process of condensing a source text into a shorter version preserving its information content is called summarization.

- The motivation Here is to build such tool which is computationally efficient and creates summaries automatically.

- Current Applications
  - Multimedia news summaries
  - Abstracts of Research papers
  - Minutes of the meetings
  - Book reviews



Rank-based Extractive Summarization



The Road ahead

Challenges to get to the target

Needs

- Most NLP results are subjective
- Need “training data” (input-output pairs) to train the Deep Learning model.
- Need Cases, Judgements, Reports along with their correct summaries.

Challenges

- Parsing Text : Finding citations, rules, decisions, issues, meta data like dates, judges – appellant - respondent names, etc. Formats vary thus extraction becomes complex.
- Algorithmic complexities: Need to build more optimized searches, theme extraction, Deep Learning model development, etc.
- Infrastructure: Soft (Hadoop/Spark, Databases) and Hard (Cloud/Servers)

rises.io
software solutions

Company: rises software solutions

- Website: **rises.io**

Contact Info: Pune, India

- Ajit Deshpade, Founder
- e-mail: ajit.deshpande@rises.io
- Phone: +91 98224 48602

Thank you!