John Snow LABS | VAKIL SEARCH

Vakilsearch Understands Scanned Legal & Tax Forms Using John Snow Labs

Vakilsearch, a leading technology-driven legal and tax compliance company, found in John Snow Labs' Spark NLP a viable solution to address its pain points of text processing, identity card information extraction, and document classification in the Indian milieu. John Snow Labs helped Vakilsearch find lasting solutions to its issues by applying natural language processing and cutting-edge language models to integrate a plethora of data into a scalable system.

Vakilsearch: simplifying legal norms

Among the various services that Vakilsearch offers to both startups and enterprises are: incorporation, government registrations, regulatory filings, bookkeeping, documentation, and annual compliance. Vakilsearch also provides a wide range of services, including property agreements and tax filings, to individuals and groups. Vakilsearch aspires to provide a one-click solution to legal and professional needs.

The Problem

Indian industries are accustomed to using physical documentation; the amount of paperwork required varies by industry. Many industries rely on scanned document images (which typically contain non-selectable text) to obtain information for important index fields to carry out their daily duties.

Vakilsearch, too, faced similar hurdles.

1. The first significant step is to index various types of documents to sort out data quality issues, which aids in the extraction of information and meta-data from a wide range of complicated documents. This task is known as **Document Classification**. The first challenge was to identify how advanced machine learning and natural language processing (NLP) techniques could solve this significant issue.

2. Second, Vakilsearch faced difficulty in **extracting parts of mailing addresses**. This is complicated because Indian addresses lack a standardized structure for breaking down an input address text into extracted pieces using rule-based parsers.

3. Third, Indian addresses may **contain text in multiple languages** – such as English, French, Hindi, and vernaculars. Languages can even be mixed in a single form.

4. Fourth, the company had difficulty **extracting and classifying information from photos of identity cards**. Different data fields are dispersed across multiple spots on an identity card –each with their own format, font, size, and style. These are frequently superimposed over complex background images, some containing language, interfering with an image recognition pipeline's ability to interpret the information effectively.

Solution

Vakilsearch handled more than **10 distinct paper types – such as electricity bills, bank statements, passports, and Aadhaar numbers**. Vakilsearch built an automatic procedure that can process many lengthy documents quickly and accurately. Working with paper-based records adds another layer of complexity to document classification. The paper must first be scanned, and then written, or typed text must be extracted for further examination. A technology required for this must detect text and layout from photos and scans, allowing for converting paper documents into digital format for subsequent classification.

Working with complex document classification and information extraction problems necessitates strong Natural Language Processing (NLP) technology. Vakilsearch built a document classifier pipeline using the Spark NLP deep learning model, which supports easily training & tuning models for custom document types to optimize accuracy.

For identity card information extraction, Vakilsearch employs an image processing pipeline using Spark OCR, and then extract data fields by applying a John Snow Labs' Named Entity Recognition (NER) model trained specifically to extract identity card information. The image processing pipeline is the extraction of meaningful information primarily from digitally stored images using a variety of techniques, each one applied to a specific task because current generic image analysis methods do not match human accuracy. Both the identity card and mail address extraction pipelines were built with Spark NLP using BERT embeddings and a Bidirectional-LSTM deep learning architecture.

Results:

- **The document classification pipeline outperformed the comparable pipelines by 9% and attained 96% accuracy.**
- **The identity card information extraction pipeline achieved an accuracy of 87%.**
- **The mailing address resolution pipeline achieved an accuracy of 89%.**

> What differentiates Spark NLP from the other available frameworks is that it provides a simple and elegant pipeline. This simplicity allows us to achieve best performing accuracy, scalability, and higher data literacy. We are building more products using Spark NLP to enhance the user experience at Vakilsearch.
>
> **Rajkumar Ganapathy, CTO, Vakilsearch**

Conclusion

John Snow Labs' Spark NLP library provides cutting-edge accuracy and speed, while supporting easy training and tuning for specialized document types. The library includes production-ready versions of state-of-the-art embeddings and deep learning models for a variety of common NLP tasks – such as form understanding, named entity recognition, and document classification. Spark NLP is not only more accurate than other libraries, but also optimized for scalability so that standard NLP pipelines run orders of magnitude quicker than what legacy libraries allow.