INDIUM
Digital | QA | Gaming

Success Story

Implementing Text Analytics Algorithms to benefit academia & researchers with relevant & progressive topic searches

Business:
Text Analytics, Topic Modeling, Document clustering

Domain:
Information Services,

Tools:
LDA, NER Algorithms, D3.js, HTML, Python, C++, Django, Docker, SpaCy

Key Highlights

Key Success:
» The Text Analytics solution attracts platform users to access most relevant and focused content for niche topics in the SEO-manipulated web world.
» Maximising information entropy on topic searches minimized user's read/search effort.

Data Scrapping:
Scrapy, Selenium, BeautifulSoup

Database:
PostgreSQL, Elasticsearch (Data Indexing)

Algorithms:
Latent Dirichlet Allocation (LDA),
Named Entity Recognition (NER)
Forced Acyclic Graph
TF-IDF, Word2vec, Cosine similarity

Client
The client is a leader in the $200B local services market in India and for NRIs in USA, Canada, UK, and UAE.

Overview
R&D involves an enormous amount of research which often entails scouring the web, journals, historic papers etc. for specific topics and then reading each document for a specific piece of information which can often be obscurely worded. Academic research is currently difficult and arcane due to the complexity and domain specificity with the material. The client commissioned Indium Software to develop a platform that would automatically analyze thousands of documents to return highly specific results in structured searches, identify all the topics a document touches upon, make connections between documents and other resources etc.

1 Status Quo

The client is an information aggregator platform powered by an insight network. The platform acts as an interface to the torrent of text data available on the web by adding an intelligence layer to it. The platform content takes a more relevant and logical form to the search engine data.

A typical user persona of the platform is research-oriented, knowledge gathering and exploring. The client envisioned an intelligent layer to the available web data that refines web search for academicians and researchers. The users of the platform will be able to find:

» Clear structured search results
» Trending topics within documents
» Related content for topics and similar documents
» A cluster of topics/documents for a progressive learning

2 Solution

Indium Software implemented an NLP and Text Analytics based solution for formulating the intelligent layer of the platform.

The Solution consisted of building,

1. A **documents cluster** which a user can check for any topic.
2. A **topics cluster** which a user can check for any document.
3. A **name entity recognition map** detailing the 7class recognizable entities.

For building the solution,

1. The public data's web links stored in PostgreSQL have to be **scraped** and stored.
2. The topics have to be discovered in every document using **topic modeling** algorithms.
3. The output clusters have to be visualized via appropriate **interactive graphs.**

3 Solution Modules

Data Scrapping

» Data gathering is achieved using Python scrapping packages such as BeautifulSoup, Scrapy and Selenium tool.
» Platform content repository is maintained up-to-date with automated scheduling and queuing of web content crawling.
» The platform holds a repository of about 120+ million documents from the web in various formats of PDF, doc, HTML pages.
» Content is updated real-time into a "Listener" and stored in the Database.

Topic Modeling

» Scraped data is indexed using Elasticsearch for efficient querying. Elasticsearch's inherent ability to digest text data and provide faster query results helped in the choice of database.
» LDA is implemented on C++ and called using python on the data to perform Topic Modelling. The outputs being a **cluster map of topics** within a document and a map of **documents related to a topic.**
» Network results – The group of documents and topics are organized in related clusters and visualized in directed graphs. D3.JS provided a rich visualisation map to represent these clusters and also helped in providing **interactivity within the cluster maps.**

Entity Recognition within documents

» Implemented Named Entity Recognition (NER) Algorithm to identify entities under the 7 **class classifiers** such as {Location, Person, Organizations, Money, Percent, Date, Time} for the document content.

» Implemented a Tree Graph representation to v**isualise in an orderly way giving insights about the entities** in the document.

4 Business Impact

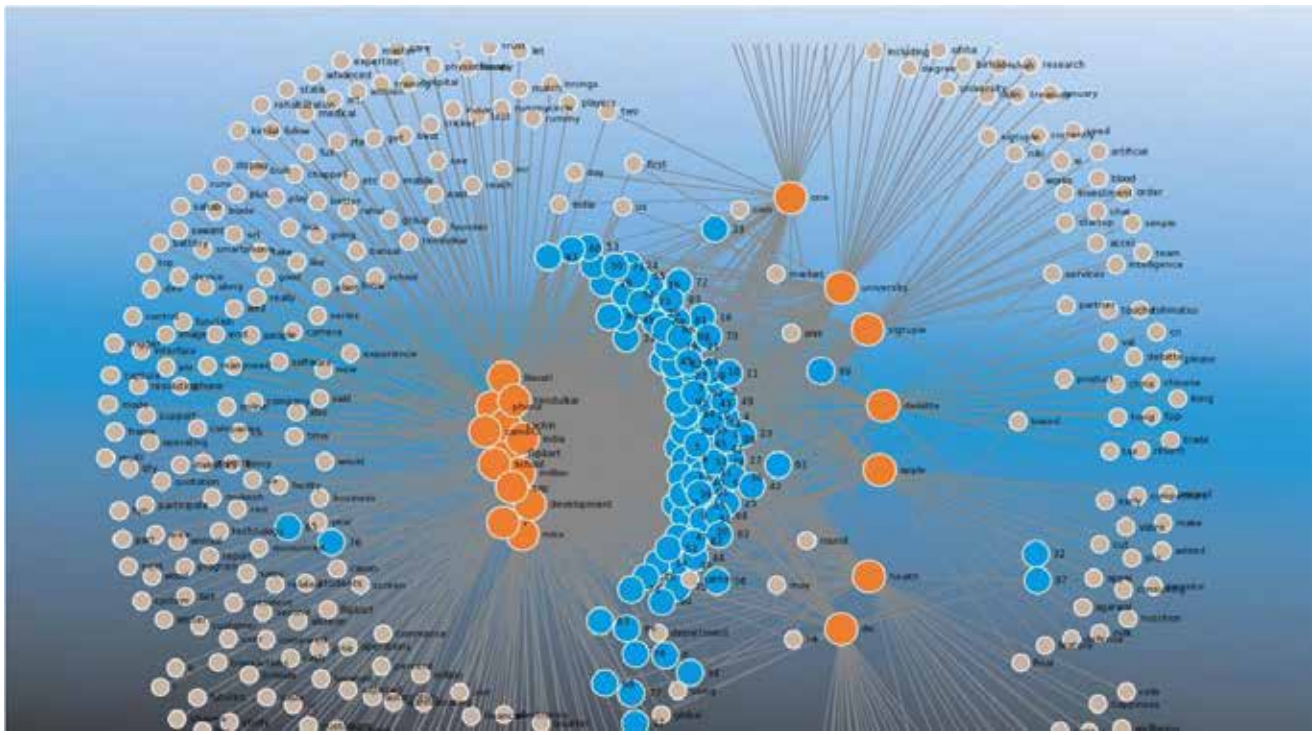
» The Text Analytics solution attracts platform users to access most relevant and focused content for **niche topics** in the SEO-manipulated web world.

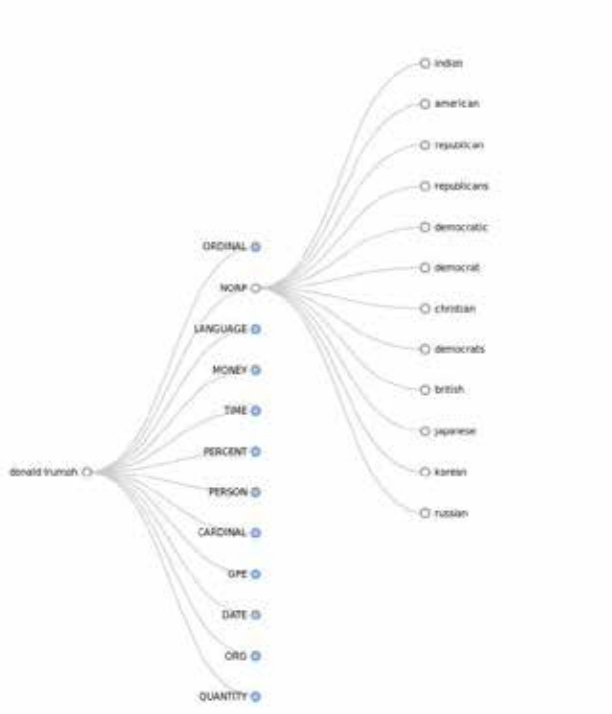
» Document search and knowledge gathering is significantly faster and efficient, users can access a voluminous range of data and develop a high-level understanding of focused topics at a quick rate.

» A wide **array of topics** within a document and **related content** for any document is generated in tandem for a topic search.

» **Maximising information entropy** on topic searches minimizing user's reading/ searching effort.

5 Sample Data Charts





INDIUM
Digital | QA | Gaming

INDIA
Chennai
+91 44 6606 9100
Bengaluru
+91 80 4645 7777
Mumbai
+91 022 6215 4028

USA
Cupertino | Princeton
Toll-free: 1 888 207 5969

SINGAPORE
+65 9630 7959

UK
London
+44 773 653 9098

General Inquiries : info@indiumsoftware.com
Sales Inquiries : sales@indiumsoftware.com