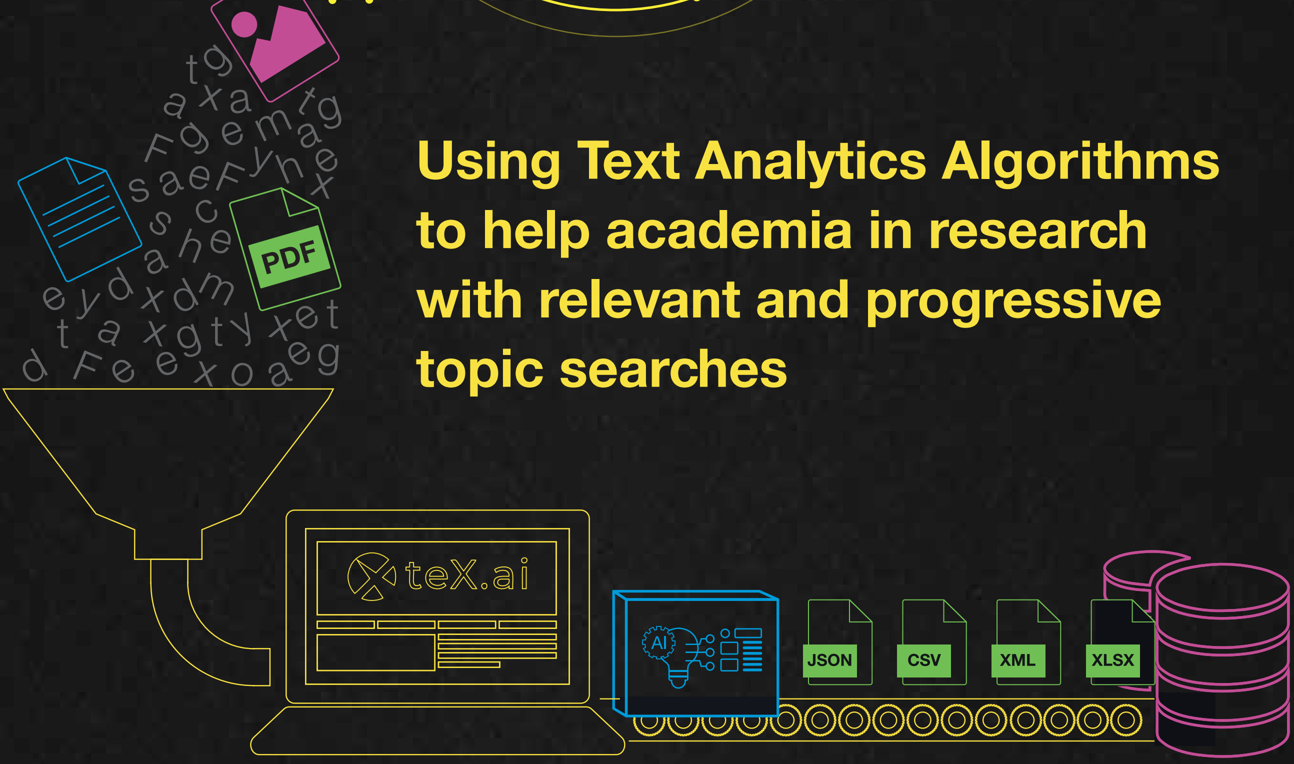

Using Text Analytics Algorithms to help academia in research with relevant and progressive topic searches

- Reduced user bounce rate by 20%
- Reduced document search time for users by 30%

teX.ai™

Customer Background

The client is an information aggregator platform powered by an insight-driven network. R&D involves an enormous amount of study which often entails scouring the web, journals, historic papers and other sources for specific topics and then reading each document for a specific piece of information which can often be obscurely worded.

Business

Text Analytics

Domain

Information Services

Tools

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

Data Scraping - Scrapy, Selenium, BeautifulSoup

Database - PostgreSQL, Elasticsearch (Data Indexing)

Key Highlights

- The Text Analytics solution attracts platform users to access most relevant and focused content for niche topics in the SEO-oriented web world the bounce rate by 20%
- Maximizing information entropy on topic searches minimized user's read/search effort

Business Requirement

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 for 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 progressive learning.

Solution Overview

teX.ai was used to implement an NLP and Text Analytics based solution to formulate the intelligence layer of the platform. The solution consisted of building:

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

Solution Modules

Data Scraping

- Data gathering is achieved using Python scrapping packages such as BeautifulSoup, Scrapy and Selenium. The platform content repository is maintained with up-to-date information 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 such as PDFs, doc and 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 the database.
- The output is a cluster map of topics within a document and a map of documents related to a topic which is visualized in direction graphs.

Entity Recognition and Documents

- Implemented Named Entity Recognition (NER) Algorithm to identify entities under the 7 class classifiers such as - Location, Person, Organizations, Money, Percent, Date and Time - for the document content.
- A Tree Graph representation was implemented to visualize in an orderly manner providing insights about the entities in the document

Business Impact

The text analytics solution attracts platform users to access most relevant and focused content for niche topics in the SEO-oriented web world reducing bounce rate by 20%.

Document search and knowledge gathering is significantly faster and efficient as users can access a voluminous range of data and develop and high-level understanding of focused topics quickly.

A wide array of topics within a document and related content for any document is generated in tandem for a topic search reducing search time by 30%.

Maximizing information entropy on topic searches minimizing user effort in reading and search.

Get in touch to see how teX.ai can help you!

✉ info@tex-ai.com

☎ +1 (888) 207 5969 (Toll-free)

🌐 www.tex-ai.com

Suite 100, 19925 Stevens Creek Blvd,
Cupertino, CA — 95014

teX.ai ™