SSA GROUP

Advanced PDF Parsing Software for Leading Polymer Science Firm

Client Background

Our client is a leading American polymer science and engineering innovations company that provides innovative proprietary packaging solutions to the customers worldwide, using patented adhesive and hardware technologies. The team of highly experienced polymer chemists, chemical engineers, researchers, and technicians leverage a high throughput research platform and sustainable development approach to create solutions that bring long-term value for their customers' business.

Business Challenge

The business mission of our client is to provide their customers with supreme quality, cost-effective packaging solutions. Using a materials database of over 20 million data points and 15.000+ formulations and receipts, our client needs to constantly keep an eye on the price of every component included into chemical formulas in order to provide their customers with products of best cost price on the market.

The client reached out to Db Devs with the request to develop a solution capable of scraping a web-site www.ulprospector.com and collecting the data on the raw materials used in ink formulations which are an indispensable part of many chemical compositions. Particularly the client was looking for the following data attributes:

- Product name;
- Product description;
- Check-summed number;
- Percent solids;
- Viscosity;
- Molecular weight;
- Appearance;
- Color, APHA;
- Density;
- Elongation;
- Functionality;
- Modulus;
- Oligomer weight;
- Refractive index;
- Tensile strength;
- Glass transition temp.

The raw materials data provided by different manufacturers, that contained text and numeric descriptions of products was expected to be downloaded in .pdf format data sheets, which later had to be parsed, saved into text format, and then processed with the aim to find values for the attributes required. Furthermore, the client was willing to have the retrieved data organized in a MS SQL database with ability to add new materials every few months.

Project Description

Before engaging into the project development, Db Devs engineers looked through the CSV document provided by the client that contained over 7.000 unique URLs to the pages with raw materials the client was looking for. Each product page on the website contained a link to the pdf document where the manufacturers specified all the relevant information about their products.

During the first project stage engineers developed a crawling solution for scraping PDF files from the pages on www.ulprospector.com, by using links provided by the client.

In order to access the pdf documents with products' data, it was necessary to log in to the website. However, due to a complex registration and verification process, it was a very challenging matter. Furthermore, engineers faced with the restriction of the number of pdfs they could scrape using one account, therefore they needed to create additional accounts for scraping all the products from the client's list.

Most manufacturers used a single style of specification sheets across all their product ranges, but each manufacturer differed in certain moments. Db Devs engineers parsed the scraped PDF files, performed a full text search, and created specific advanced algorithms to identify key words from the page for every single manufacturer and assigned encoded words to them. The encoded words were later used to find the required values while processing the text.

The last step was devoted to processing the text for matching values with the key words which were identified earlier and delivering the retrieved data in Excel format to the client.

Value Delivered

The Db Devs team delivered a custom web scraping solution capable of collecting, analyzing, and converting massive amounts of data for providing the client with the exhaustive and relevant information. Thanks to the implemented software, our client could receive the valuable information regarding any cost fluctuations of the raw materials used in ink formulations to have the opportunity to react immediately for keeping their manufacturing prices on the competitive level.

About SSA Group

SSA Group is a software development company that designs, implements, and supports cutting-edge digital solutions, enabling customers to unlock their business potential.

With over 10 years of experience, SSA Group creates innovative products and modernize complex legacy systems that shape today’s digital and business landscape.

SSA Group is a trusted partner for the world’s industry leaders, consistently turning their ideas into reality.

Contact us
Email: contact@ssa.group
USA Toll Free: +1 866 263 99 03

Visit our website
https://ssa.group/

Reliable IT Partner
for Your Business