Sfile Technology

1990 Post Oak Blvd,
Space 1990 D
Houston, TX 77056
1-(866) 997-3453

Sfile Complex Redaction

Providing excellent redaction at unprecedented speeds

Challenge

One of the top ten biggest oil and gas companies was faced with a complex redaction and only ten days to get it done. Thousands of private files containing private drilling site locations needed to be redacted. We needed to identify if a number string was a responsive site identifier, and if not to redact all the related information. The normal method produced lots of inconsistent results and site identifiers were easily missed. It took 4 hours of a reviewer's time to redact a single spreadsheet, and there were over 2,000 Excel files to be redacted. They needed a trusted team with the ability to get exceptional results on a very tight deadline.

Our Solution

Our E-Discovery platform is second to none in its ability to solve these complex redaction challenges in tight windows. WIthin 24 hours, the data was tokenized and full-text indexed where distinguishable patterns were uncovered. The spreadsheets were able to be clustered into sets of similarly related files. If a spreadsheet was derived from a master template, all spreadsheets created from that template would be found to be similar and clustered together for analysis. Once significant clusters are identified, data scientists analyzed a bag of terms representing the responsive site identifiers. using the full text index of the Excel files, an analyst uncovered the complexities of the data. Using the Sfile discovery platform, analysts were able extract column and row positions where site identifiers were present. These represented pivot points in a machine model for predicting the existence of site identifiers where machine learning techniques were leveraged to calculate the probability of a valid site identifier. Several models had to be created to address the various patterns that existed in the population of Excel files.

Results

After several training exercises, we were able to create a multiple convolution neural network for processing the various machine models created to identify and predict where site identifiers existed. Once identified, another model was created to discover the extents of the site related data and redact and site information not part of the responsive identifiers. The neural network is very tunable and flexible so that if new patterns in the Excel files were identified, a new machine model could be created and added to the neural network. Despite not having time to develop a parallel processing engine for this job, processing in a synchronous batch took 10 hours. now with the ability to process in parallel, it would take closer to one hour for 2,000 Excel files.

Sfile was able to quickly create the patterns and identify clusters ahead of schedule. With a ten day deadline, we finished it in eight. While it would have taken close to four hours for a person to do just one Excel sheet, it took our machine model ten hours to do over 2,000. The job was performed in a fraction of the time and at a substantial savings to the company.

As an added benefit, redactions are no longer subjectively determined by a human operator, who is prone to inconsistencies and potentially costly errors. With the automated engine, redactions are consistent and logically performed across the entire production set making the redactions very defensible. Only Sfile provides energy companies the ability to find insights into their data, while cutting labor costs and operating at the fastest speed in the industry.