Extracting Structured Numerical Data from Large Quantities of Memoranda

Laura Odell, Miranda Seitz-McLeese, and James O'Conner

The Challenge: The Defense Logistics Agency needed to understand the cumulative variances of actual and expected weights of natural resources stockpile materials. Physically reweighing the stockpiles to determine the differences between the expected and recorded amounts was not feasible.

As part of DLA's 2015 audit-readiness effort, it needed to understand the cumulative variances of actual and expected weights of natural resources stockpile materials.

Background

The Defense Logistics Agency (DLA) is tasked by executive order with managing the nation's stockpile of strategic materials. As part of DLA's 2015 audit-readiness effort, it needed to understand the cumulative variances of actual and expected weights of natural resources stockpile materials. Since the data were not readily available, auditors suggested that the DLA Strategic Materials Division (SMD), which oversees the strategic materials stockpile, reweigh the piles to determine the differences between the expected and recorded amounts. DLA SMD determined that reweighing the piles was not practical because reweighing would mean physically moving material, which is expensive, time-consuming, and labor-intensive, and could affect the environment. Also, for some materials, the reweighing process itself can cause material loss or degradation.

Instead, DLA turned to IDA for an alternative solution. Although DLA had not kept the data required to determine the difference between the expected and actual weights, it had maintained paper copies of 549 memoranda documenting the transaction or event details needed to assess whether material stockpiles were within a generally acceptable range of loss or gain when compared to industry benchmarks. IDA determined that, if we could extract numerical data from these memoranda, we could use those data to calculate whether the cumulative variances fell within industry standards.

Why Text Analytics?

Extracting the numerical data necessary to calculate variances from a stack of paper without a text analytics capability would have been labor-intensive. Someone would have had to read each memorandum, find the relevant numbers, and enter them into a spreadsheet. In addition, these data needed to be at an audit-ready level, which requires a small margin of error; the margin of error for manually entered data would be too high. Using IDATA allowed IDA researchers to produce high-quality data quickly.

The Process

DLA scanned the documents into JPEG files, but the scans had no text data associated with them, and IDATA cannot perform text analysis on documents that do not have text data. Our first step was to extract the text from the scanned documents using Adobe Acrobat's Optical Character Recognition (OCR) tool. Once the OCR tool extracted the text, we used IDATA's extractor tool to create structured data in the form of a spreadsheet. Extractors find information buried in text data according to certain patterns. For DLA's memoranda, IDA researchers wrote an extractor that pulled the dates of each memorandum based on their predictable structure (month, day, year). We wrote another extractor to find the reason given for any discrepancy based on the assumption that these reasons usually occurred in phrases such as, "…caused by [reason]."

Because DLA provided the entire data set, the search and discovery phase of IDA's information triage approach was not needed, and the data set moved to the exploratory analysis phase. IDA ran the text data through the extractors and entered the output into a spreadsheet. We reviewed the completed spreadsheet and adjusted the extractors to improve performance. We also had to input some data by hand because some of the documents were of such poor quality that the OCR tool could not read them. We also performed random spot checks to ensure the accuracy of the extractors and checked the completed spreadsheet for missing, incorrect, or duplicative entries, which we corrected manually. This process took two IDA researchers less than a week of work to complete.

Results

IDA used its information triage process to identify 469 instances of stockpile material measurements with associated reasons for weight differences from the 549 memoranda. We divided the causes into categories:

- Scale variance between measurements (i.e., equipment discrepancies)
- The environmental effect on the stockpile (i.e., snow, rain, type of ground)
- Administrative error (i.e., human error and equipment failure)
- Moisture evaporation over time
- Theft (this occurred only once, but we considered it important to include)
- Other (i.e., damage to equipment, no available explanation, comingling of piles).

Of the documents entered in the spreadsheet, none were missing the subject field, four were missing a cause statement, five were missing the shipment date, 10 were missing a dollar amount, two were missing a weight difference, 17 were missing the acquisition rate, and 78 were missing the percent over or under the original expected weight. We restricted further analysis to the 391 documents that had information about the variation in terms of the percentage of the original expected weight.

IDA used pivot tables and charts to determine the degree of variance between the expected weight and the actual weight and found that the variance was almost an order of magnitude lower than the industry standard. These results convinced DLA's auditors that reweighing the piles was unnecessary, which saved DLA significant time and money.