When numbers get complicated: mining thoughts

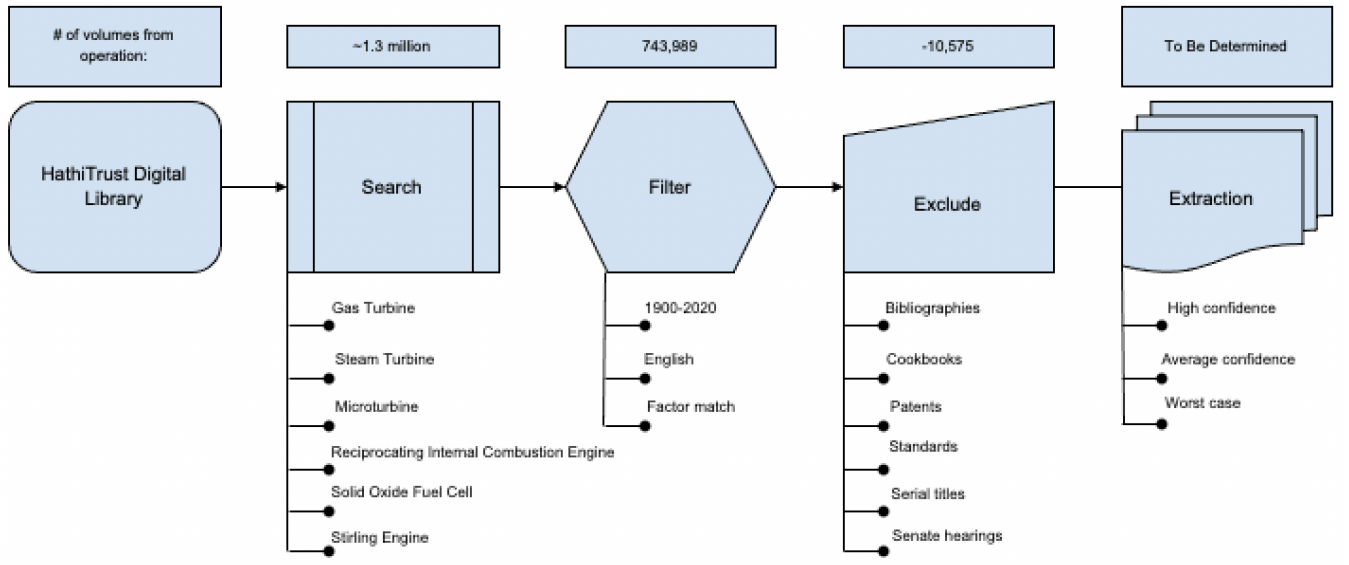
Project update for *Surveying Applicability of Energy Recovery Technology for Waste Treatment*

Project investigator: Aduramo Lasode

Energy, specifically renewable energy has gained a perpetual spotlight as weather patterns continue to change and social pressure for public policy action increases. Renewable energy conversations are expanding beyond traditional solar and wind sources, to evaluate other options for decreasing greenhouse gas (GHG) emissions. One of the industries that offers an opportunity to mitigate GHG emissions is waste treatment, through energy recovery processes. Energy recovery leverages prime mover technologies, for example turbines, fuel cells and internal combustion engines, in power or heat generation from biogas obtained during waste treatment. A preliminary study I conducted leveraged a data-driven approach to recommending prime mover technologies based on efficiency, an important factor for consideration. During this first take, some challenges presented an opportunity to explore resources at the HathiTrust Research Center (HTRC). HTRC offered a wide range of data sources, with an automated mining appeal.

A crucial first step in kicking off the Advanced Collaboration Support (ACS) project was activating an interdisciplinary team, with expertise for this type of problem and study approach. My background is in mechanical engineering and combustion applications through the Thomas E Murphy Engine Research Lab at the University of Minnesota. During the course of my application to the HTRC ACS program, I was able to enlist the help of a Digital Media Librarian at the University of Minnesota Libraries who is familiar with text mining. It was important to clearly define the scope of text mining within the multiple volumes of HathiTrust, and devise ways to identify areas of interest. In addition, the team from HTRC helped in the beginning to understand digital library structure and terminology, quite fascinating I must add. Understanding how search terms work, how scanned text and figures are classified using metadata, how volume contents are represented for internal catalog systems, all complimented my technical knowledge in creating search input and extraction logic.

The proposed project scope explores six prime mover technologies: gas turbine, steam turbine, microturbine, reciprocating internal combustion engine, solid oxide fuel cell and Stirling engine. Some pre-extraction steps include building a list of synonyms for each of the technology names and factors of interest. On average, there were two variations of each technology name and units were considered for search terms regarding efficiency, cost and fuel terms. The project requires a nested search approach that first matches technology occurrence before doing a streamlined search for efficiency cost and fuel terms. The relevant volumes can then be filtered to reduce the chances of obtaining false positives and potentially reduce computational cost. Spot checks are done on a random collection of volumes within the filtered list to identify catalog groups that could be further excluded. For example, cookbooks for an acronym match for RICE or fiction and musicals for name matches. An initial single term search produced technology matches in more than 1 million volumes. A filtering based on relevant dates, language and efficiency, cost or fuel factor matches reduces the volumes of interest to under 750 thousand. An exclusion of volumes based on performed spot checks eliminates another 10 thousand volumes. The following illustration maps out the approach adopted by the team working on this project.



Text mining is an exciting adventure for data-driven studies like mine. One word that makes this adventure quite challenging is “association”. The notion that the search features an intrinsic relationship between factors of interest. An example of the ideal data entry is an instance in a volume page or sentence that names a single technology of interest, states its power rating and states the efficiency (cost or fuel) information for the same unit. Linking three factors in this mining process presents unique challenges when compared with text mining that might identify and extract single-term search matches. This introduces a matter of confidence in accuracy of data extracted, which is of high priority when considering information to be extracted from volumes in the library that are not available for full-view due to copyright reasons. One approach this project team is exploring leverages logic-based algorithms that leverage knowledge of how the information to be extracted is typically presented in published volumes. For example, presentation styles could imply proximity of search terms in documents. With a high confidence in the extracted data, further analysis can be conducted to achieve the broader impact of this project- recommending prime mover technology using a data driven approach to navigate intricate decision factors of technology efficiency, cost and fuel utilization.