Text Analytics of Selected Reports from the NASA Technical Reports (NTRS)

NIFS Internship: Summer 2016 *Jefferson Ridgeway, 2016 NIFS Summer Intern*

Manjula Ambur, Jeremy Yagle, Ted Sidehamer, NASA Langley Research Center

Background:

Currently at NASA, NTRS (NASA Technical Reports) houses scientific research documents from the NASA Scientific and Technical Information (STI) Program. NTRS allows subject matter experts to search documents via a graphical user interface (GUI) online and see immediate search query results and subsequent metadata in XML format. However, with the advancement of technologies and software tools, this method of searching documents has now become inefficient in comparison of other tools that are currently within the industry. One of the recent tools being used in the past few years is IBM's Watson Content Analytics (WCA) software.

Goals:

Through using the WCA software tool, the goals of this project included:

- Investigating methods of locating, extracting, and ingesting a subset of documents from NTRS into WCA
- Writing python scripts to prepare XML files from NTRS, Endnote, Mendeley, for ingestion into WCA proper format
- Document processes related to adding another collection to the Watson system at NASA LaRC

Methodology:

The subsequent methodology conducted for this research project is as follows:

- Acquired a subset of NTRS subject matter search queries from NTRS Personnel in XML or CSV format
- Created a .java class that can be used in the runnable .jar file that will process NTRS and ingest into WCA
- Visually Reviewed the XML files to make sure the files are in appropriate WCA ingestion format
- If XML files were not in appropriate format or need to be edited, a python script was written that will insert or delete tags and write a new XML files
- If received a CSV file format, a python script was written to convert CSV to XML format for WCA ingestion
- Created a modularized process for taking directories of XML files and applying previous python script
- Built WCA Components (crawler, facet structure, index fields) and then pointed crawler to config, .jar, and source files
- Run ingest of XML Files into WCA software tool and visually inspect and validate WCA Collection

Results and Next Steps:

The results from conducting this research show the power that WCA provides to the user that include, unlike more general search query tools, key insights and patterns, ability to identify trends and connections, and visualize networks of experts in a given domain. The next steps within this research would be to ingest all of NTRS documents into a WCA collection.