NC State University
Learning, Design, & Technology
STUDENT INTERNSHIP REPORT

Summary of Tasks Performed: The Summer Conference on Applied Data Science (SCADS) is an annual 8-week workshop hosted by the NC State University Laboratory for Analytic Sciences (LAS). The overarching vision is to bring together industry, academic, and government professionals to collaboratively attack a Grand Challenge in the machine learning (ML) and artificial intelligence (AI) development space. The Grand Challenge (described below) is a somewhat lofty, 5 to 10-year research and development goal that can greatly influence knowledge workers' analysis practices in all sectors. Stepping-stone problems on the way to achieving the Grand Challenge offer near-term value potential.

For the inaugural conference, the Grand Challenge framework was proposed as a method to develop ML capabilities that proactively provide individuals (or organizations) with information relevant to their needs. The objective of a tailored daily report (TLDR) describes a relatively short report, auto-generated, perhaps on-demand or on a schedule. This report contains new information of high interest to the user, drawing simultaneously from any number of sources and weighing the value of source materials relative to the user's objectives and interests. The US Department of Defense (DoD) Intelligence Community, our target audience, manually produces a daily report of this nature for senior government decision-makers, summarizing new updates of intelligence value. Conceptually, natural language processing (NLP) technology, in the form of specific ML models, could produce something similar for a host of alternative use cases.

During the first week of the conference, LAS leadership facilitated onboarding sessions that reinforced the primary goals driving the SCADS program: grand challenge progress, partnerships, and learning. The program is comprised of approximately 40 participants with diverse experiences, skill sets, and research interests. While we were all encouraged to pursue research that interests us personally, the overarching grand challenge was to be kept in the foreground of our work. The conference was purposefully organized to maximize knowledge transfer across academia, industry, and government boundaries. The DoD recognizes the efficiencies to solve these complex challenges at scale lie in partnering with like-minded organizations working on similar problems. Lastly, as SCADS is hosted by a university research laboratory, hands-on learning was critical to expanding a base of knowledge that can be passed from this conference to the next.

During the conference's second week, participants self-organized into four working groups, each focused on a contributing aspect to the text summarization grand challenge. The working groups were comprised of the following:

- Human-Computer Interaction (HCI),
- Knowledge Graphs (KG),
- Recommender Systems, and
- Text Summarization (TS)

By the end of week two, I had selected the text summarization group as my focus area. Based on a group brainstorming session, I produced a process view (Figure 1, below) describing how each working group contributes to the larger process of text summarization. The **HCI** team focused on how knowledge workers engaged with information through data tools designed to capture analysts' key interests and focus areas. A **knowledge graph** was chosen as a method to organize the large amounts of data and information that IC analysts could access to understand and update their daily work tasks. **Recommender systems** query the knowledge graph to determine analyst preferences and return documents whose summaries may satisfy their queries. The **text summarization** team developed methods to process the recommended documents to deliver a concise summary that would answer specific analyst questions and/or add to the knowledge graph through a feedback loop. At multiple stages throughout the process, results can be refined through additional recommender or topic modeling layers.

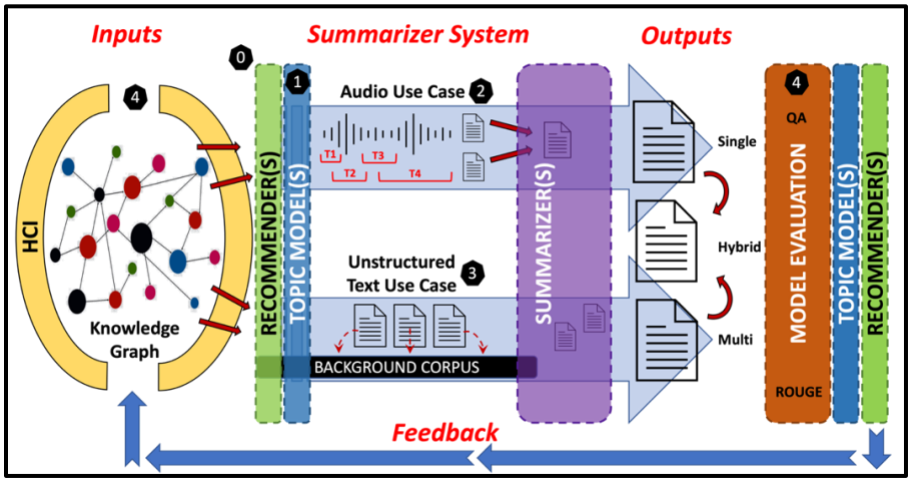

Figure 1 - Summarization Process Overview

For this first SCADS iteration, the TS group prioritized three use cases for experimentation: one for audio files and two for unstructured text documents. Those use cases are highlighted by the numbered polygons (#2 and #3) in Figure 1. The additional polygons represent work group efforts that either enabled the primary use cases (#0 and #1) or built on their results to deliver feedback during the output phase of the process (#4). Those specific efforts are detailed in Figure 2:

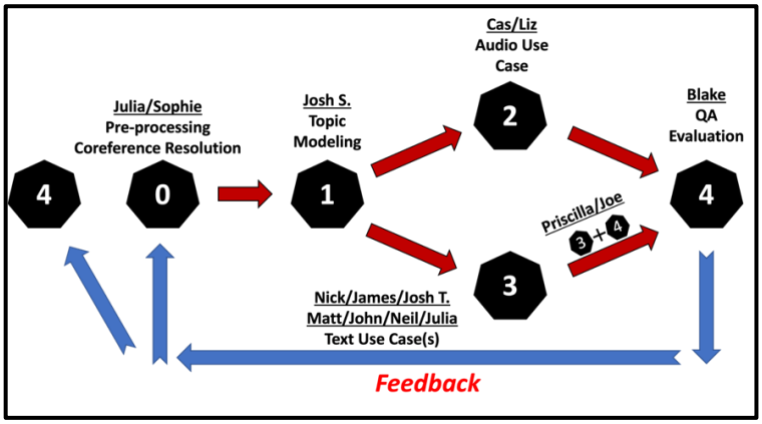

Figure 2 - Text Summarization Workflow

Figure 2 depicts how the text summarization group allocated team member resources to the multiple facets of the challenge. Beginning at #0, participants worked on pre-processing data we would normally obtain via the knowledge graph and recommender groups as input for the summarization use cases. A topic model was applied for the unstructured text documents to determine potential document groupings prior to the summarization model. Post-summarization, three group members applied question answering (QA) modeling to measure the ability of the summaries to answer specific user requirements. This resulted in a feedback loop that informed updates to the HCI and knowledge graph teams.

In week three, my work group developed our research path following the traditional data science process. The process phases determined the key tasks we would follow for the remainder of the project: prepare, wrangle, analyze, model, and communicate. To prepare, we conducted a thorough literature review to understand the current state of automated text summarization. From there we explored various data sets to identify the top contenders for ingestion into our ML models (week 4). Once the data was obtained, we conducted numerous statistical tests to better understand the quality of our data (week 5). The results of this analysis ensured we chose the best data for the modeling phase where we fine-tuned various parameters on an extractive summarization algorithm (week 6). Finally, we consolidated our findings and prepared to present them to the rest of the conference and some key project sponsors (wees 7 & 8).

As this was the inaugural SCADS, our goals were more about exploration and learning that solving the Grand Challenge. To that end, we made great strides in identifying good and bad data sets as well as some key characteristics we would like to see in future data. Within the specific ML model (OCCAMS), we identified specific term weight parameters that can be adjusted to deliver more accurate training models. Lastly, we presented alternatives on how to better evaluate model results. Each of these outcomes and the processes that produced them was delivered in a 60-page research report and companion presentation that was delivered on the final day of the conference.

Lessons Learned: The first "challenge" I experienced in the project also revealed my first lesson, which was identifying the research area in which I had the most interest. I was forced to define what I wanted to learn as well as what I thought I could produce. All the workgroups intrigued me, but I didn't have the capacity to work with them all. In the end, text analysis touched on multiple interest areas so that's where I focused my energy. From the first brainstorming session through the completion of our paper, the entire project was a case study in identifying the strengths of each team member. Rarely will you ever get to personally select/hire all your team members. As a result, it's a good practice to quickly figure out who does what the best and then reinforce those strengths. In our group, we had coders, researchers, writers, and mathematicians. Each played a key role in the experiment and presenting our insights. This practice also had the added effect of building buy-in across the group as they saw their unique expertise being valued as the project progressed. Lastly, I experienced firsthand the advantage of integrating professional development into the execution of the project. We had regular "lunch & learn" sessions to shore up specific skills across the cohort. From basic data science methods to Python coding examples, we got to request (and even present) topics that were needed across the larger research group. Not only did this shore up some knowledge gaps, but it also provided time to network and better understand what other workgroups were discovering.

The conference reinforced some of my nascent data science skills while also giving me some practice in new areas. Previously, all my data analytics experience relied on coding in "R." This project used Python and SAS, so I got some experience learning alternatives for processing ML and topic models. For Python, I was also able to explore Jupyter Notebooks and the Visual Studio Code development environments. I was fortunate to be on a team with an expert in library science. She introduced me to better knowledge management techniques for keeping track of our literature review sources as well as gave us insights into multiple research databases. My biggest skill enhancement from the conference was learning how important dataset selection and curation were to the rest of the project. There is no shortcut to a thorough review of previous work and the lessons learned on which datasets are suitable for your research goals.

Overall Evaluation: I would highly recommend this conference, or any opportunity like it, to someone looking to apply learning analytics processes to real-world problems. The breadth of experience across the 40-person cohort was immense, from educators to statisticians to mathematicians to coders. This gave me tangible insights into multi-disciplined problem-solving that can't really be replicated in the classroom. I established relationships with other students, professional academics, government analysts, and industry experts that will serve me well as I look to further my goals beyond graduate school.

While this project focused on finding a better text summarizer, I gained valuable insights into applications to the field of learning design:

1. Literature reviews: Researchers can use summarizers to increase the pace of reviewing project-specific articles, papers, or books.
2. Evaluating student work: Summarizers can be used to quickly process student papers while QA models can compare it to a rubric for quality.
3. Resource discovery: Topic modeling can assist in growing a personal learning network by linking students or professors that are studying similar areas.
4. Knowledge management: Knowledge graphs are great tools for efficiently organizing data across a class, school, or district, enabling much broader opportunities for sharing.

I pursued this internship to explore the fields of educational technology and digital literacy, specifically to better understand how text mining and ML tools could decrease the burden on learners.

This project accomplished that while also allowing me to network with the DoD and intelligence communities as I prepare for opportunities after grad school.