**TACC HPC Allocation Proposal**

**Elements: Data: Sustaining Modern Infrastructure For Political And Social Event Data**

**(NSF OAC-** **1931541)**

**November 1, 2022**

**PI: Patrick T. Brandt (Political Science, University of Texas, Dallas,** [**pbrandt@utdallas.edu**](mailto:pbrandt@utdallas.edu)**)**

**Co-PIs: Latifur Khan (Computer Science, University of Texas, Dallas,** [**lkhan@utdallas.edu**](mailto:lkhan@utdallas.edu)**) and Vito D’Orazio (Political Science, University of Texas, Dallas,** [**vxd152030@utdallas.edu**](mailto:vxd152030@utdallas.edu)**)**

**Research Objectives**

This request is to support the NSF funded project (OAC-1931541) to collect news reports and generate event data from them. Event data are constructed from very large corpora of news reports on who did or said what to whom coded using the Conflict Analysis and Mediation Event Ontology (CAMEO). The overall project creates a general research platform to study civil protests, international conflict, and civil unrest using texts from English, Spanish, and Arabic news reports.

This expands the development of our programs, data and services available for coding regional verbal, material conflict and cooperation methods beyond the current English-only approaches. This will enable research that will advance new approaches to core questions in the social, behavioral, and economic sciences. The project's data and methods help make data-driven decisions about foreign policy, civil war prevention, human rights policies, and the effects of other factors such as environmental or economic policies on these phenomena.

Our project machine encodes the news reports using a series of open-source natural language processing (NLP) algorithms, more specifically the BERT language representation model, and a custom event encoder. The following are the basic steps in out process:

1. Scrape news articles from a list of nearly 400 English news agencies, 50+ Spanish news agencies and 200+ Arabic news agencies and store the results in a text database
2. Using NLP tools, parse, lemmatize, and deploy a series of semantic and part of speech tags to the sentences in the news reports. This is done using the BERT model depending on the language of the news reports. This is the most computationally intensive step in our event coding process for the news reports.
3. Store the results for subsequent analysis, visualization, and make them available to other researchers

Given our need for additional resources in current allocation, the Jetstream2 cloud environment allows us to scale up to what is needed. It includes a mix of various nodes, great GPU capability, storage, and VM resources that meet the needs of our project.

**Storage, Code Performance, Scaling, and resource choice**

All of our storage requirements are calculated based on the currently stored data in our UTD resources. Based on current allocation experiences, Table 1 lists the different Tasks and the nature of their computational and storage requirements.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tasks | Computation (CPU) | Computation (I/O) | Storage | Memory |
| Web Scraper | Moderate | High | High | Moderate |
| ConfliBERT Text Processing | Very High | High | Very High | Very High |
| Political Event Coding | Moderate | Moderate | Moderate | Moderate |

**Table 1: Computational, Bandwidth, Storage and Memory requirements of each of the modules**

Based on our current infrastructure, after data collection, we are storing around 100 GB of original news reports text data, and processed metadata. We collected around 3 million news reports (~6500 documents per day) in English. Annually it becomes 2.4 million documents and occupies 81 GB of storage.

To parse English language articles from our nearly 400 sources we will require about 168 GB of data storage. To add Spanish and Arabic language sources, we will incur a similar number of articles as English therefore a simliar data storage.

Table 2 highlights all of the datasets and their storage requirements with projected usages.

|  |  |  |
| --- | --- | --- |
| **Dataset Name** | **Current Size** | **Projected Size in a year** |
| ICEWS ? | 10.5 GB | 10.5 GB |
| Cline Center Datasets ? | 2.5 GB | 2.5 GB |
| TERRIER Dataset ? | 30 GB | 30 GB |
| English collect dataset | 30 GB | 30 GB |
| Spanish collected dataset | 8 GB | 30 GB |
| Arabic collected dataset | 5 GB | 30 GB |
| **Total** | 86 GB | 133 GB |

**Table 2: Showing storage requirement for different event datasets.**

|  |  |
| --- | --- |
| **Items** | **Required Storage** |
| Original Raw Text + Metadata (i.e parses) for English | 168 GB |
| Original Raw Text + Metadata (i.e parses) for Arabic | 168 GB |
| Original Raw Text + Metadata (i.e parses) for Spanish | 168 GB |
| Dataset Storage (From Table 2) | 133 GB |

**Table 3: Showing storage requirement for entire project.**

So, in total we will require ~ **637 GB** of data.

**Justification of Service Unit Request**

Our justification for resource request is driven by an empirical study over one year of collection and processing tasks held on our UTD resources.

Here is how we calculated SU as per Jetstream2 documentation available:

*Number of SU required for a VM annually = number of cores \* 24 \* 365.*

Table 4 lists component wise required SUs. Table 5 points to prospective new module and related SUs required.

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Number of Cores | Service Units/Day | Service Units/Year |
| Web Scraper | 16 | 384 | 140160 |
| BERT Training English | 30 | 720 | 262800 |
| BERT Training Arabic |
| BERT Training Spanish |
| Political Event Coding | 10 | 240 | 87600 |
| **Current Total** | 56 | 1344 | 490560 |

**Table 4: Task and related SU usage summary**

From the data presented in Tables 4, our annual estimate for required SUs is 490560. Table 6 shows the requested e SUs and Storage estimated for 1 year.

|  |  |
| --- | --- |
| **Number of SUs per year** | 490560 |
| **Storage requirement** | 637 GB |

**Table 5: Project-wide required SU and storage requirements**

We believe that these estimates of the SU request and the machine sizes will allow us to continue to produce real-time political event data for a low cost, with cloud redundancy. This will allow us to serve our 90+ API users who include students, industry such as Facebook and Google, and the national security community and its contractors (U.S. Department of State, Accenture, etc.).

Finally, given the infrastructure nature and broad scope of this project, we should be able to pursue additional resources to fund further development of this project. This includes work to code new kinds of political events, allowing other users to input and apply their own actor and verb dictionaries to our corpora across languages, and other extensions. This would then move the SPEC proposed allocation here from just being the implementation of the *services* proposed here to a science *gateway*.