Investing in Legacy: Cleaning and Categorizing a United States Government Information Technology Budget Dataset to Understand Public Sector Legacy Systems Investments

Undergraduate Researcher: Natalie Simkins (Data Analytics)
Graduate Mentor: Julia Stachofsky (Information Systems)

Undergraduate Research in Progress

BACKGROUND

Government agencies often use legacy information technology (IT) systems that limit their ability to respond to the needs of constituents. Recently the United States (US) House of Representatives have begun to address this with the Legacy IT Reduction Act (S.3987, 2022) in which agencies will have to inventory their legacy IT assets along with strategic modernization plans.

For this research we aim to better understand the state of legacy systems in the US government with the following research objectives:

RO1. Create a legacy IT dataset from secondary US government budget data

RO2. Understand the overall scope and focus of legacy IT projects in the US government

METHODS

1. Dataset Construction
    1. We exported 11 public data feeds from the US government IT dashboard (itdashbord.gov) which includes data from 2020 to November 2022 about all government IT project budgets.
    2. Next, we documented every field in each data feed, recorded the type of data, and a short description of what the field is tracking
2. Dataset Coding
    a. Since this dataset includes all IT projects not just legacy projects, we analyzed each project's description to determine if it was related to a legacy project or not.
    b. Projects were coded as: Legacy, Non-Legacy, or Maintenance projects
3. Data Analysis
    a. Initial descriptive statistics were created in Excel to understand the *scope* of legacy IT projects
        i. We analyzed project counts, software vs non-software projects, project counts by agency, and adherence to project schedule and cost variance
    b. Next, we conducted a Latent Dirichlet Allocation (LDA) topic analysis (Deborotoli et al. 2016) of the legacy IT project descriptions to determine the *focus* of legacy IT projects
        i. LDA is an unsupervised text-mining approach that takes freeform text as an input, applies probabilistic weights of related words that are assigned to different topics (Deborotoli et al. 2016)
        ii. We used scikit-learn, a machine learning library in Python to conduct the LDA topic analysis (Kapadia, 2022)

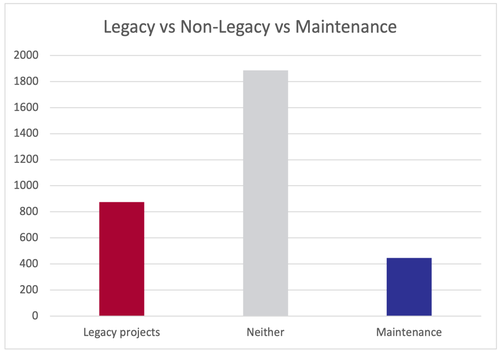

Figure 1: Legacy vs Non-Legacy vs Maintenance projects for the years 2020-2022

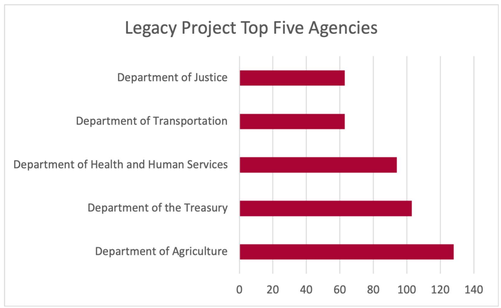

Figure 2: Number of legacy projects per agency for the top five agencies in years 2020-2022

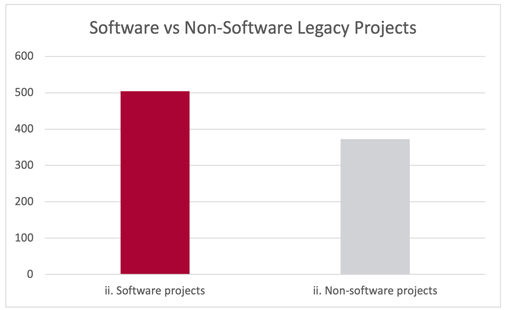

Figure 3: Proportion of Software to Non-Software Legacy Projects for 2020-2022

RESULTS

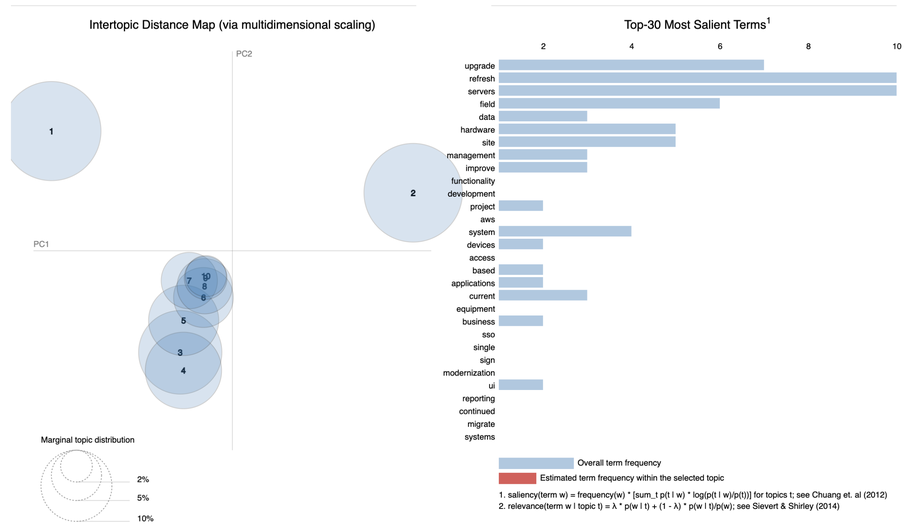

Figure 4: On the left, the intertopic distance map found via multidimensional scaling. On the right, the top-30 most salient terms. Both pertain to the description sections of the legacy projects.

Figure 5: Word cloud for legacy projects

FUTURE WORK

Our next step will be to analyze and theoretically explain the topics produced from the topic modeling analysis. Additionally, now that this dataset is created, we can do further statistical analyses such as panel regressions on legacy project CIO evaluations, government investment, and project types as well as potentially integrating our data with other public government datasets.

ACKNOWLEGEMENTS AND FUNDING

The authors are grateful for funding from The Griffiss Institute with support from award no. SAA 10012021MM0336, a VICEROY Project entitled Northwest Virtual Institute for CyberSecurity Education & Research (CySER)

References
Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. Communications of the Association for Information Systems, 39, 110–135. https://doi.org/10.17705/1CAIS.03907
Kapadia, S. (2022, December 23). Topic modeling in Python: Latent dirichlet allocation (LDA). https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0
S.3987 – 117th Congress (2021-2022) Legacy IT Reduction Act of 2022, S.3987, (2022, December 22). https://www.congress.gov/bill/117th-congress/senate-bill/3897