Business Intelligence With Raytheon

PURDUE UNIVERSITY | The Data Mine

Amritraj Binuraj, Ishaan Bhatt, Matthew Choi, Kyu Kim, Akshay Patel, Rachel Pfeifer, Vibha Polisetty

Raytheon Technologies

INTRODUCTION

COMPANY BACKGROUND

Raytheon Technologies is a multinational aerospace and defense company that develops, manufactures, and researches sophisticated technology products for the aerospace and defense industries.

PROJECT SUMMARY

- Our goal is to develop a application that takes contract opportunity and job data from multiple sources then analyzes and interprets them to give Raytheon indicators of what contracts could be advantageous to the company.

PROJECT PROCESS

- Our members originally started off with little coding, database, web scraping, Apache Drill, and Natural Language processing experience
- Last semester we were able to web scrape data from two different sites (sam.gov and usajobs) and began working to clean that data and get it into a database.
- Over time we learned how to extract information off the web, store data in JSON documents, create a dashboard that makes data visualization easier and use NLP for text simplification and summarization

RESEARCH METHODOLOGY

WEB SCRAPING

- Method of extracting information and data off web sources
- These two websites is most practical and cost-efficient method of data extraction
  - Process can be automated using code
  - Most websites can be scraped

DATABASE

- A database is an organized collection of structured information
- Hosted on a server so all team members could access data readily
- Explored Apache Drill as an alternative to MongoDB
- Decided to use MongoDB because:
  - Stores data in flexible, JSON-like documents
  - Fields are allowed to vary from document to document
  - Data structure can be changed over time.

NLP

- NLP uses research in linguistics to develop tools and techniques that approximate how humans deal with language computationally
- Important for topic modeling
  - Topic modeling is a form of text mining
  - Form of identifying patterns (topics) in a text
- Extraction of information or knowledge from narrative text

DASHBOARD

- A dashboard allows for easy visualization of analyzed data from by a user
- Using Tableau to create a dashboard due to its efficiency and user-friendly experience

WEB SCRAPING

METHODS

- Web scraping – utilized Beautiful Soup (Python)
- Automation utilized Helium (Python)

SOURCES

Multiple sources were used to diversify the data we collected. Both websites were free to use and scrape data from, practical and cost-efficient, while providing useful insights

- Sam.gov
- Usajobs.gov

RESULTS

- Retrieved over 6 million contract opportunities
  - Contract opportunities are procurement notices from federal contracting offices. Anyone interested in doing business with the government can use this system to search opportunities.
- Successfully implemented the ETL (Extract, Transform, Load) process
- Wanted both: Contract opportunity display the amount of budget invested in certain field, Job opportunity displays areas that are in high demand
- Gathered over 13 thousand government job opportunities

SAM.GOV

USAJOBS

SHAPE AMERICA'S FUTURE

Flask

FLASK API

MONGO DB

DOCKERS

KUBERNETES

DATABASE

GOALS

- Get data that was extracted via Web Scraping and upload and store it in a MongoDB database for later usage and analysis.

LIBRARIES AND INFRASTRUCTURE

- **MongoDB** a non-SQL database that can accept and save data
- **MongoDB Compass:** an application that allows for direct access to a MongoDB database
- **Python Flask API:** a website framework that can handle the functionality for saving and reading scraped data from the web
- **Dockers:** a software that converts our code into a container (a package of our code and all its dependencies so the application can run quickly and reliably from one computer to another).
- **Kubernetes:** a library that acts like a manager to run and operate these containers.

PROCEDURE

- Load the data into a Mongo database that runs on Kubernetes container in Purdue's Geddes server.
- Automatically upload data to the database using the MongoDB address

MongoDB ALTERNATIVE: APACHE DRILL

- Explored as an alternative and backup to MongoDB in case infrastructure issues prevented the use of MongoDB
- MongoDB data source can be connected to Apache Drill
- Apache Drill has applications with MongoDB
- Query data/data sets, configure storage, execute SQL queries

Business Intelligence With Raytheon

Amritraj Binuraj, Ishaan Bhatt, Matthew Choi, Kyu Kim, Akshay Patel, Rachel Pfeifer, Vibha Polisetty

NLP

GOALS

- Use natural language processing to extract data from the database to explore topic modeling
- Topic modeling – way of identifying patterns in text; topics are these patterns

PROCESS

- Retrieve data from the database in the form of text files
- Import packages (Python libraries with various functions)
  - NLTK, gensim, NumPy, pandas, matplotlib, PIL
- Import specific NLTK packages (for NLP)
- Clean data (stopwords, word_tokenize)
- Model topics using LDA (LdaModel function)
  - Latent Dirichlet Allocation (LDA) – Python function that organizes the text in a document by topic; it is an example of a topic model

RESULTS

- Our primary objective for this NLP code was to organize data for topic modeling
- This allowed us to identify multiple pattens in our data, group it properly and find trends within our data

FIGURE 1

- Figure 1 (bottom left) is a data visualization was produced by querying the data in the database utilizing MongoDB Compass
- Shows the data sorted and analyzed by Department Agency

FIGURE 2

- Figure 2 (to the right) is a visual of topic modeling
- We aggregated data from the 2018 and 2021 National Defense Strategy and performed topic modeling using natural language processing (NLP)

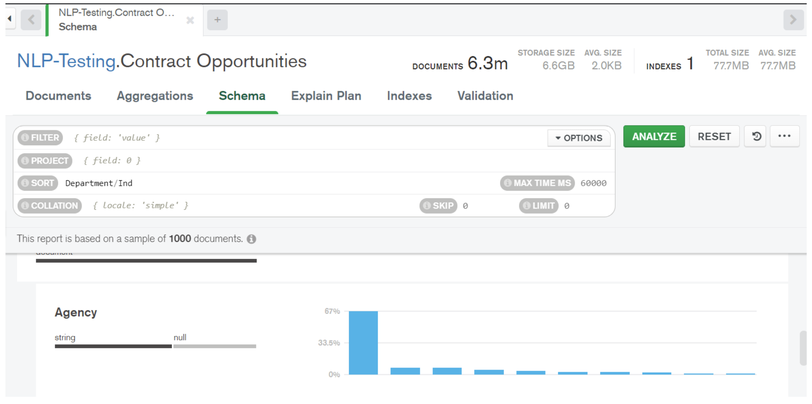


FIGURE 1

DASHBOARD

- We used Tableau to create an interactive dashboard to easily view the analyzed data in a user-friendly manner
- The dashboard is simply a way to visualize interesting correlations with the data.
- Figure 3 (below) shows various visualizations of the data ranging from job counts, Departments/Agencies vs. Sum of Awards, and dollar maps.

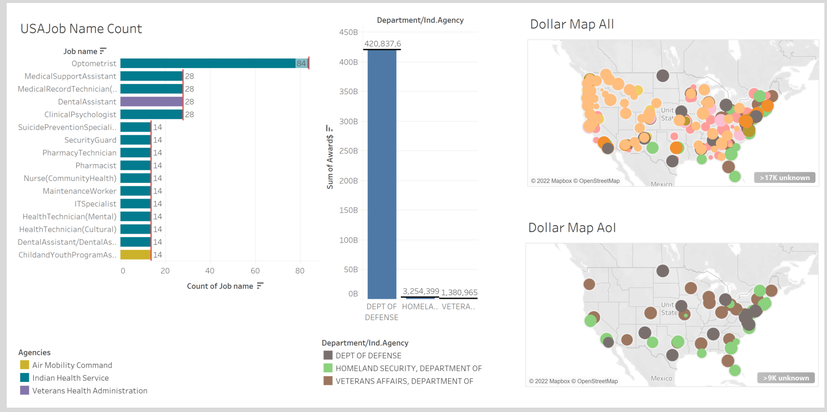
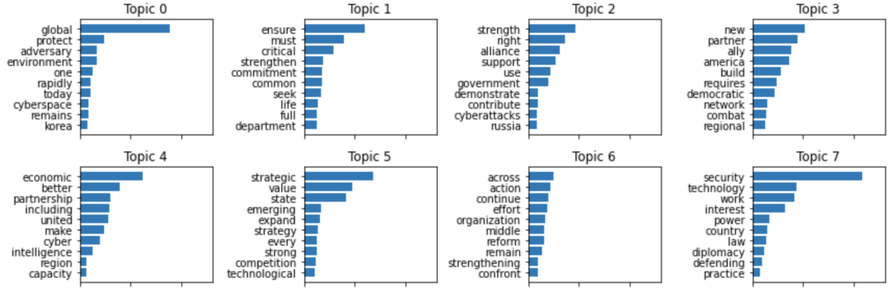


FIGURE 3 (ABOVE)

FIGURE 2 (BELOW)



ACKNOWLEDGEMENTS

We would like to thank **Raytheon Technologies** for the opportunity to take part in this project, and our Corporate Partners Mentor **Mike Douglass** for all his guidance and support throughout the development of our project. We would also like to thank the **Purdue Data Mine Staff, Dr. Ward, Maggie Betz, and Kevin Amstutz**.

CONCLUSION

STRENGTHS

- We created a Web scraping process that was powerful and efficient allowing for capturing millions of jobs and contract opportunities.
- In-lab communication was great in terms of dividing up work based on each member's strengths.

CHALLENGES

- Hosting the database on a server was a huge challenge and setback in terms of time and lack of knowledge and experience.
- Sub-team communications could be improved in the future, as one sub-team's code can often be used in another sub-team's work as well.
- Outside-of-lab communication can be improved in terms of checking in with each other on our progress

Overall, this semester, were able to successfully create the main Web scraping and Database infrastructure and created a solid base for further analysis and visualization as well.

FUTURE PLANS

PLANS

- More data sources (ex. FRED St. Louis Fed)
- Continuous stream of data/automated pipeline
- Complete Flask database
- More descriptive analytics
- Even more detailed visualization tool with Tableau

REASONING

- Incorporating new data sources will provide more data
  - Leads to more descriptive analytics and better insights from the tool
- Automating the transfer of data will allow for constant pulling of data online and sending to database
  - Removes manual work
- Completing the Flask database gives a more desired alternative to the current MongoDB
- Since we have all the pieces, we simply need to put it together and create a final dashboard to display findings

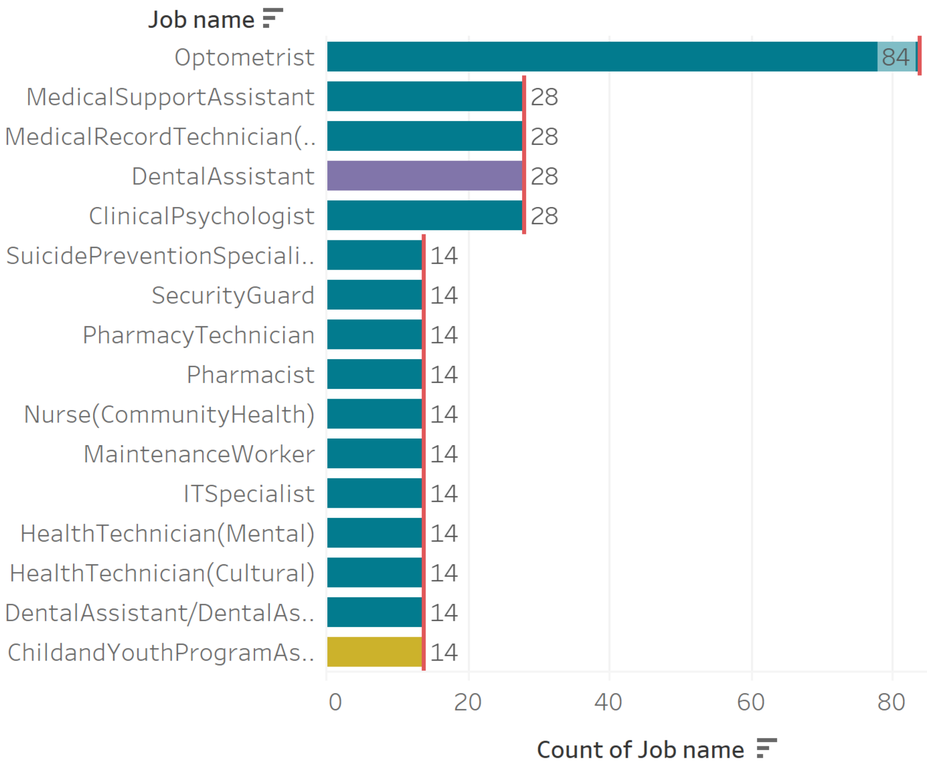
Business Intelligence With Raytheon

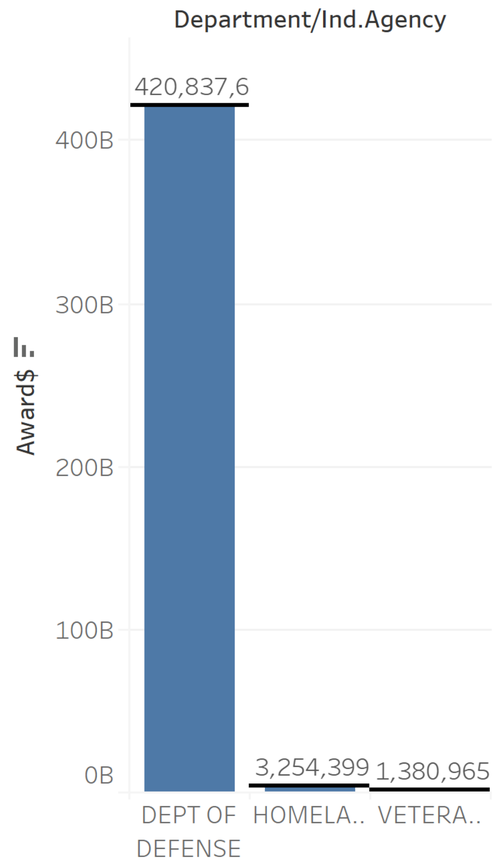
Amritraj Binuraj, Ishaan Bhatt, Matthew Choi, Kyu Kim, Akshay Patel, Rachel Pfeifer, Vibha Polisetty

PURDUE UNIVERSITY | The Data Mine

Raytheon Technologies

USAJob Name Count

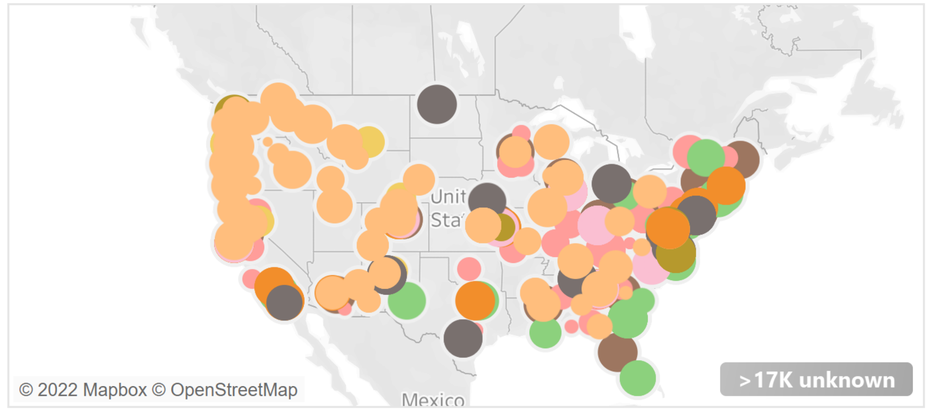




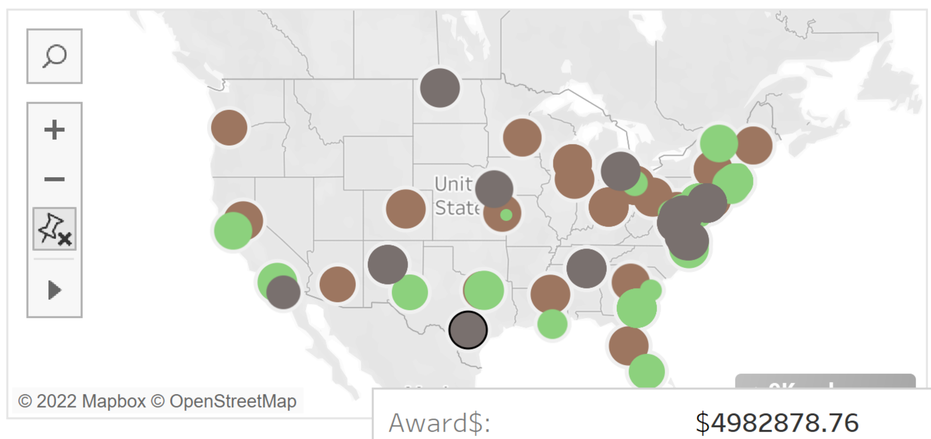
Agencies
- Air Mobility Command
- Indian Health Service
- Veterans Health Administration

Department/Ind.Agency
- DEPT OF DEFENSE
- HOMELAND SECURITY, DEPARTMENT OF
- VETERANS AFFAIRS, DEPARTMENT OF

Dollar Map All



Dollar Map Aol



Award$:	$4982878.76
Department/Ind.Agency:	DEPT OF DEFENSE
Zip Code:	78234