Project: Network Analysis Tool for Defense Industry News

Overview

In my current role as a Data Scientist/Analyst contractor with the U.S. Air Force Acquisition (SAF/AQ) team, one of my tasks was developing a tool to assist leadership in staying apprised of industry trends and insights reported in daily news articles. These articles, often collected by a contractor, contain valuable data but can be overwhelming to review due to their volume. I built a network analysis tool to automate the extraction of key themes from these articles, making it easy for SAF/AQ leadership to make data-driven decisions and to be well prepared for industry discussions and engagements.

Problem

The U.S. Air Force Acquisition team receives daily email feeds of scanned news articles on defense industry trends. These articles are often lengthy and detailed, making it difficult for leadership to quickly identify relevant information and trends. The process of manually reviewing and organizing this information is time-consuming, leading to inefficiencies in identifying critical insights.

Solution

I developed an automated solution that processes these articles, extracts key themes, and presents them in an interactive network graph. The tool provides the following features:

1. **Article Parsing**:
    - Scanned news articles are processed into readable PDF files.
    - A Python script extracts sentences, identifies key themes, and categorizes them.
2. **Data Organization**:
    - The extracted data is stored in an Excel file, with each article's themes and entities organized.
    - The data is continuously updated to reflect new articles as they are received.
3. **Interactive Network Graph**:
    - A web-based interface presents a network graph that visualizes relationships between entities, articles, and themes.
    - Leadership can interact with the graph to explore specific articles, view detailed sentences, and download relevant PDFs.

4. **Search & Filter**:
    - A search bar allows users to filter articles based on multiple terms.
    - Date filters ensure the data presented is relevant to the specified time frame.
5. **Data Download**:
    - Users can download the processed Excel file containing all relevant article data and sentences for offline analysis.

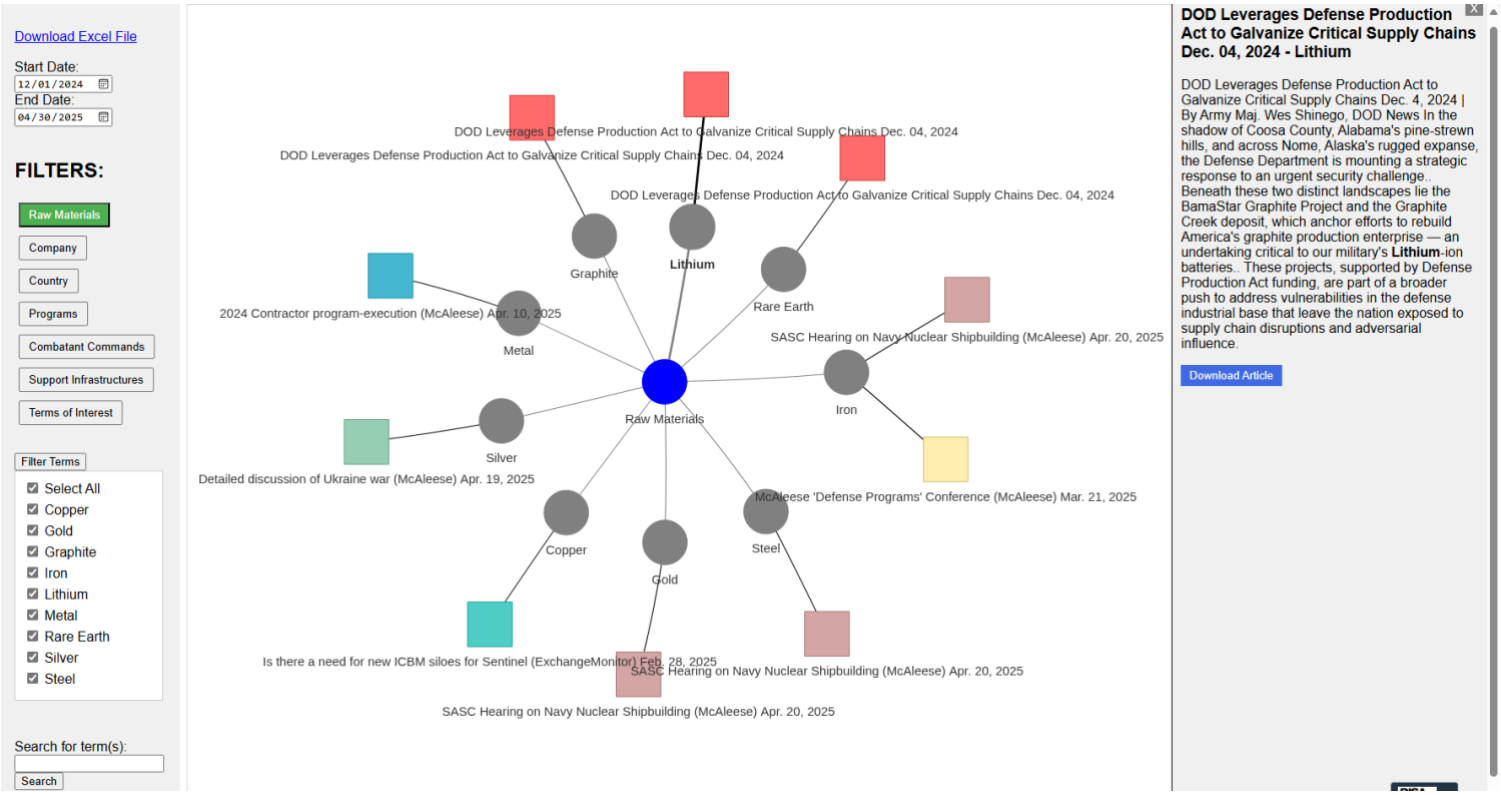
Technology Stack

- **Python**:
    - Used for data processing and theme extraction with libraries such as re, nltk, pandas, and pdfplumber.
- **Flask**:
    - Web framework used to build the interactive interface, allowing users to interact with the network graph.
- **Pyvis**:
    - A Python library for creating interactive network graphs. It's used to visualize relationships between articles, themes, and entities.
- **Excel & Pandas**:
    - Excel files are used to store and manage the processed data. Pandas is used for efficient data manipulation.
- **HTML/CSS/JavaScript**:
    - The web interface is built using standard web technologies. Custom JavaScript is used for dynamic behavior like search term filtering and node interactions in the network graph.

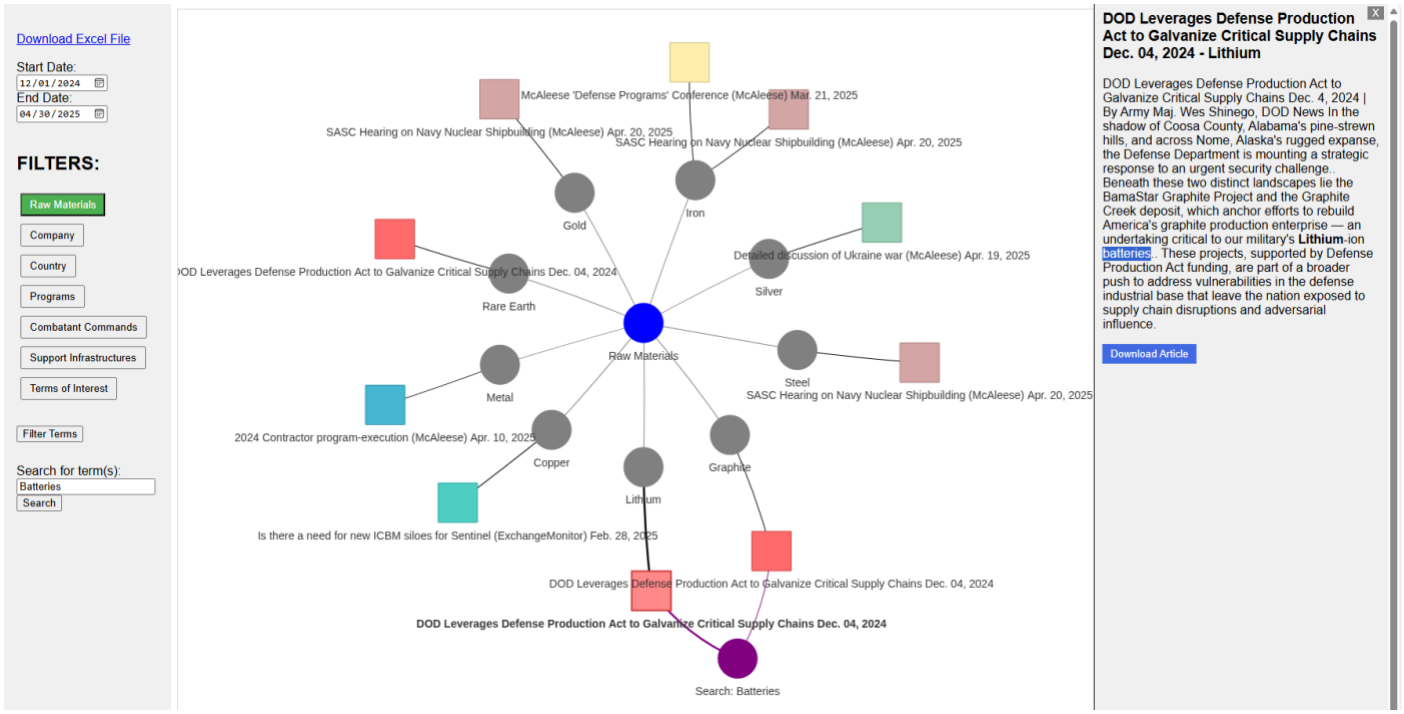
How It Works

1. **Data Ingestion**:
    - Articles are sent via email and processed into readable PDFs.
    - These PDFs are parsed using Python to extract relevant sentences, entities, and themes.

2. **Data Processing**:
    - Using Python libraries, the tool processes the extracted text to identify key entities (e.g., companies, projects, locations).
    - This information is organized in a structured format (Excel) and stored in a repository.
3. **Visualization**:
    - The processed data is visualized using Pyvis, where articles, themes, and entities are represented as nodes.
    - Users can interact with the graph to explore the connections between entities and articles.
4. **Interactivity**:
    - Clicking on nodes reveals detailed information about the associated article and its sentences.
    - Search functionality allows users to filter data by specific terms or date ranges.
5. **Results Display**:
    - The resulting network graph is embedded in a web page where users can visually explore the data.
    - Additional functionality is included to allow downloading both the network graph and raw data for offline analysis.



This screenshot shows the network graph with the "Raw Materials" filter applied, illustrating how different articles are connected by shared themes like "Lithium," "Rare Earth," and "Copper." The sidebar allows filtering by date and topics, and clicking on a node reveals detailed article sentences.



This screenshot demonstrates how the network graph dynamically updates when a search term (e.g., "Batteries") is applied. The term is highlighted in the graph, and the article details are filtered to show only those related to the search term.

Results

The tool has greatly streamlined the process of reviewing industry reports for SAF/AQ leadership. By automating the extraction of key themes and making them easy to visualize, the tool enables leadership to focus on strategic decision-making rather than spending time sifting through pages of articles. It also provides a transparent, repeatable process for trend analysis and ensures that all team members have access to the same up-to-date information.