SYM-AM-24-078

# Excerpt from the Proceedings of the Twenty-First Annual Acquisition Research Symposium

## Acquisition Research:
## Creating Synergy for Informed Change

May 8–9, 2024

Published: April 30, 2024

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

Disclaimer: The views represented in this report are those of the author and do not reflect the official policy position of the Navy, the Department of Defense, or the federal government.

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School

# A Semiautomated Framework Leveraging NLP for Skill Identification and Talent Management of the Acquisition Workforce in the Department of Defense

**Dr. Jose E. Ramirez-Marquez—**is an Associate Professor at the School of Systems and Enterprises, Stevens Institute of Technology. He holds degrees from Rutgers University in industrial engineering (PhD and MSc) and statistics (MSc) and from Universidad Nacional Autonoma de Mexico in actuarial science. His research focuses on developing mathematical models to analyze and compute system operational effectiveness - reliability and vulnerability as the basis for designing system resilience. He has published more than 200 refereed manuscripts related to these areas in technical journals, book chapters, conference proceedings, and industry reports.

**Garry Shafovaloff—**is the Senior Advisor, Policy and Legislation, Defense Acquisition University. Previously, he served as the Director and Deputy Director of the Office of Human Capital Initiatives (HCI), responsible for defense acquisition workforce strategic planning and initiatives from 2010 through 2022. In 2022, he served as a senior lead for the $50 million Artificial Intelligence DoD Upskilling initiative in partnership with the Chief, Digital and Artificial Intelligence Office. He is currently detailed to the Office of the Assistant Secretary of Defense (Acquisition), serving as the Senior Program Manager for the Defense Civilian Training Corps (DCTC) pilot initiative.

**Mark Krzysko—**is the Principal Deputy Director for Acquisition Policy and Innovation directing acquisition data governance, data access, and data science enabling the Department of Defense sound business decision-making. Additionally, he served in the National Academies of Sciences, Engineering, and Medicine's Data Science Post-Secondary Education Roundtable discussion on data science education and practice, the needs of the community and employers, and ways to move forward. Krzysko holds a Bachelor of Science degree in finance and a Master of General Administration, Financial Management, from the University of Maryland University College and numerous certificates from Harvard University.

**Dr. Dinesh Verma—**received a PhD (1994) and an MS (1991) in industrial and systems engineering from Virginia Tech. Verma currently serves as the Executive Director of the Systems Engineering Research Center, a U.S. Department of Defense–sponsored University Affiliated Research Center focused on systems engineering research, along with the Acquisition Innovation Research Center. At Stevens, he has proposed research and academic programs exceeding $175 million. He has authored over 100 technical papers, technical monographs, and three textbooks. Verma has received three patents in the areas of life-cycle costing and fuzzy logic techniques for evaluating design concepts.

## Abstract

The Department of Defense (DoD) must address critical questions about talent management and workforce adaptability. This research introduces the potential for leveraging Natural Language Processing (NLP) techniques to address these challenges. The paper describes an NLP-based framework to analyze vast text data, including government, industry, and academic reports. The primary objective is to identify critical skills necessary within the DoD acquisition workforce efficiently and accurately. By automating this process, the DoD can swiftly pinpoint areas of expertise and allocate resources accordingly, ensuring the hiring and deploying of personnel with the right skills where needed most. With the insights derived from NLP analysis, decision-makers within the DoD can make informed choices regarding talent acquisition, training and development programs, and skill gap remediation. The ability to swiftly and accurately identify essential skills optimizes resource allocation, reduces skill gaps, and elevates operational efficiency. This newfound efficiency extends to talent management, enabling the DoD to nurture and develop critical skills proactively. Identifying and managing critical skills is pivotal for ensuring preparedness and resilience in a rapidly changing world order.

**Keywords**: Skills, Data Science, Natural Language Processing, Data Visualization

## Introduction

The Department of Defense (DoD) stands as one of the most expansive and intricately structured organizations globally, charged with safeguarding the national security interests of the United States. In an era marked by evolving geopolitical tensions, particularly concerning the escalating influence of Russia and China, the DoD's capacity to effectively harness and optimize its human resources is paramount in maintaining a competitive advantage. Talent management transcends conventional recruitment and retention concepts; it encompasses identifying, cultivating, and deploying critical skills and expertise vital to addressing contemporary security challenges. As adversaries continually enhance their military capabilities and extend their influence, the agility and adaptability of the DoD hinge significantly on its ability to leverage advanced technologies, such as Artificial Intelligence (AI) and Natural Language Processing (NLP), to discern and cultivate the requisite skills necessary to outpace adversaries. Consequently, talent management strategies integrating cutting-edge NLP techniques can prove instrumental in ensuring the DoD's agility, responsiveness, and readiness in navigating complex and dynamic global threats.

The effective management of a workforce as expansive and multifaceted as the DoD is essential for upholding the nation's military readiness and global influence. By harnessing NLP and other advanced technologies, the DoD can streamline skill identification, align human resources with strategic objectives, and empower decision-makers to make well-informed decisions regarding recruitment, training, and skill enhancement initiatives. Within this framework, talent management emerges as a potent force multiplier, enabling the DoD to adeptly confront the nuanced and evolving challenges Russia and China pose. Ultimately, the efficacy of talent management strategies within the DoD significantly contributes to the United States' ability to assert itself globally and navigate the intricate dynamics of the contemporary international security landscape.

### Problem Description

Amidst escalating geopolitical tensions, particularly with nations like Russia and China, the DoD grapples with myriad challenges in effectively identifying crucial skills and managing its workforce. The evolving landscape of modern warfare and rapid technological advancements necessitate continually adapting the DoD's workforce to anticipate and counter emerging threats. However, identifying these critical skills is complex due to the DoD's diverse composition, encompassing various military services, civilian roles, and contracted personnel. Geopolitical tensions introduce unpredictable dynamics, demanding a nimble workforce capable of addressing traditional military challenges alongside emerging threats like cyber warfare, information warfare, and hybrid conflicts. Balancing long-term skill development with the imperative for immediate readiness in a dynamically changing global environment further compounds this challenge.

Effectively managing the DoD's workforce in such circumstances demands a nuanced approach, considering demographic shifts, technological progress, and geopolitical realities. Furthermore, in an era where recruitment and retention strategies extend beyond talent attraction to include talent retention amidst heightened competition, the DoD faces aligning its strategies with its mission. Addressing these challenges is a matter of organizational efficiency and a critical component of national security, enhancing deterrence capabilities and ensuring a credible defense posture. Thus, amidst geopolitical tensions, the DoD's ability to identify and manage essential skills is pivotal in bolstering the United States' preparedness and resilience in an ever-evolving global context.

This research developed a framework that implements AI and NLP techniques to identify critical skills within the DoD workforce. By harnessing the capabilities of NLP, the project endeavors to enhance talent management, workforce planning, and skill development strategies within the DoD, ultimately contributing to a more agile and effective defense organization. The project’s overarching objectives are as follows:

1. Skill Identification: Utilize NLP algorithms to analyze extensive textual data, including industry, government, and academic reports, to identify critical skills within the DoD workforce automatically.
2. Decision Support: Provide DoD decision-makers with actionable insights and recommendations from NLP analysis, empowering them to make well-informed decisions regarding talent acquisition, training programs, and skill gap remediation.

### Importance

The proposed framework uses NLP techniques to identify pivotal skills within the DoD workforce to enhance defense operations, efficiency, and readiness. Against the backdrop of escalating geopolitical tensions demanding swift response and adaptability, the expeditious and accurate identification of essential skills holds immense importance in bolstering defense operations. By automating skill identification via NLP, the DoD can promptly identify areas of expertise and allocate resources accordingly, ensuring optimal deployment of personnel with the requisite skills to areas of utmost need. This streamlined process enhances resource allocation efficiency and augments readiness by mitigating skill gaps and elevating overall force preparedness.

The efficiency gains derived from NLP-driven skill identification extend beyond defense operations to encompass talent management and career development within the DoD. The research fosters strategic talent management initiatives, empowering the DoD to proactively nurture and cultivate critical skills and enabling targeted training programs and skill enhancement strategies. Ultimately, this fosters individual personnel effectiveness and contributes to cultivating a more agile and responsive defense organization adept at confronting the evolving challenges posed by geopolitical tensions. The ripple effect of the project’s impact extends across various facets of defense, culminating in heightened operational efficiency and enhanced readiness, both indispensable attributes in navigating the intricacies of a dynamic and multifaceted global security landscape.

## Literature Review

### Automated Talent Management

NLP constitutes a branch of AI dedicated to endowing computers with the ability to comprehend, interpret, and generate human language meaningfully. Within talent management and human resources, NLP emerges as a transformative technology, presenting innovative solutions to enduring challenges and encompassing a spectrum of HR functions, from recruitment and talent acquisition to employee engagement and development.

One prominent application of NLP in talent management involves the automation of job descriptions and candidate resume analysis during recruitment processes. NLP-driven tools adeptly sift through job requirements and applicant qualifications, enabling HR professionals to swiftly pinpoint the most suitable candidates for specific roles (Vanetik & Kogan, 2023). Moreover, NLP-enabled sentiment analysis of job postings and social media activity offers invaluable insights into employer branding and aids organizations in gauging their perception of the job market (Allioui & Mourdi, 2023).

Furthermore, NLP plays a pivotal role in fostering employee engagement and retention. Through analyzing employee feedback, encompassing survey responses, performance evaluations, and informal communication channels like emails and chat logs, NLP identifies patterns and sentiment trends indicative of potential areas of concern or dissatisfaction. Timely detection of employee disengagement empowers HR teams to intervene proactively, enhancing retention rates and bolstering workplace satisfaction (Gomathi et al., 2023). Additionally, NLP facilitates the development of personalized learning and growth strategies by identifying individual skill gaps and recommending pertinent training resources.

NLP promises to revolutionize talent management and HR practices by automating mundane tasks, providing insights into employee sentiment, and facilitating data-driven decision-making. From streamlining recruitment procedures to enhancing employee engagement and development initiatives, the multifaceted applications of NLP contribute to realizing more efficient and strategic HR operations.

### Skill Identification

As previously discussed, the process of skill identification entails analyzing extensive volumes of textual data to pinpoint specific skills possessed by individuals or required for particular roles. In this regard, NLP algorithms demonstrate exceptional efficacy, leveraging techniques such as Named Entity Recognition (NER) to extract skill-related keywords and phrases from unstructured text automatically. These algorithms adeptly discern subtle variations of skills, including synonyms or related terms, ensuring a comprehensive comprehension of an individual’s or role’s skill repertoire. Furthermore, NLP can contextualize these skills, distinguishing between incidental mentions and those integral to an individual’s proficiency or role requirements.

Following skill identification, NLP offers the potential to streamline skill mapping processes by establishing connections between identified skills and specific job roles, career trajectories, or developmental pathways (Mohanty et al., 2023). Through analyzing skill-role relationships within extensive datasets, NLP algorithms uncover patterns and correlations, thereby generating skill-to-role mappings that are both data-driven and adaptable. This automated approach enhances talent management by furnishing decision-makers with precise insights into skill requirements across diverse roles and career trajectories. Additionally, NLP-powered decision support systems proffer actionable recommendations from skill analyses, empowering HR professionals and organizational leaders to make informed decisions about talent acquisition, training initiatives, and skill gap remediation. For instance, by scrutinizing skill data, NLP can propose tailored learning trajectories for employees, enabling them to cultivate critical skills aligned with their career aspirations and organizational imperatives (Caratozzolo et al., 2023). Overall, NLP’s automation of talent management processes enhances operational efficiency, mitigates bias, and facilitates data-driven decision-making within the complexities of contemporary workforce environments.

## Framework

Efficiently handling extensive textual data sets poses a significant contemporary challenge in information management. This section introduces a novel framework designed to streamline content extraction, text summarization, and the creation of executive summaries. These tasks hold critical importance across diverse domains, ranging from academic research to corporate decision-making, facilitating rapid and informed information retrieval and decision support. As illustrated in Figure 1, the framework comprises two key phases: Phase 1 encompasses Text Extraction and Summarization, while Phase 2 focuses on Skill Identification and Analytics. Notably, all framework processes have been developed utilizing AI-assisted technologies.

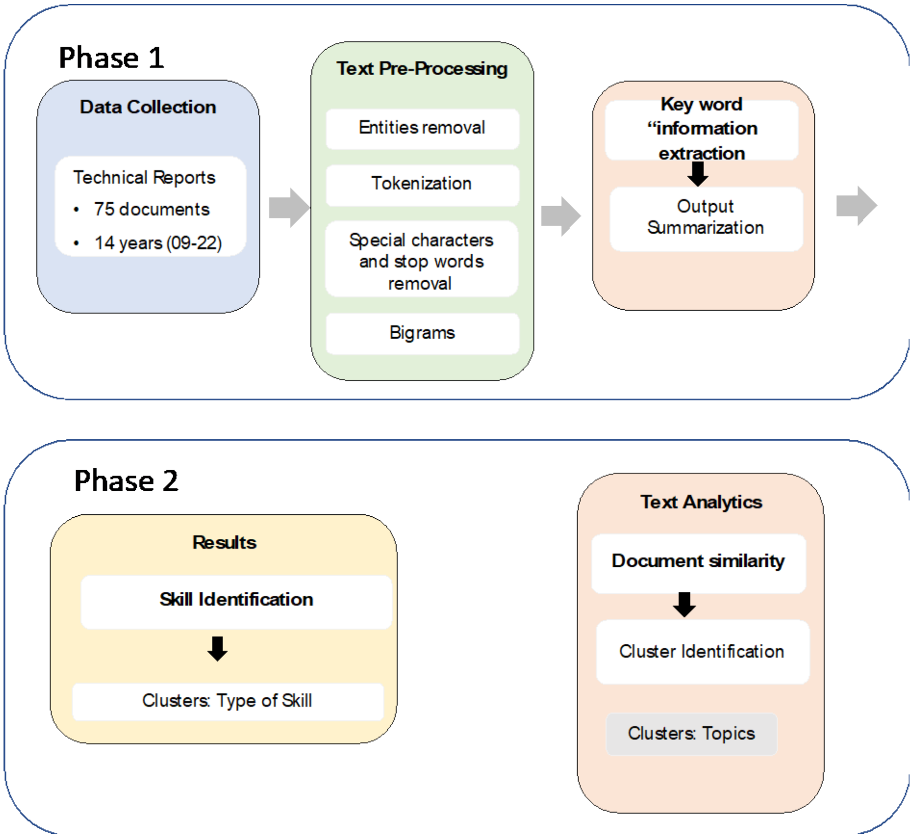


**Figure 1. Skill Identification Framework Phases 1 and 2**

The following sections describe the critical components of Phases 1 and 2, showcasing their functionality and applications.

### Phase 1: Text Extraction and Summarization

In Phase 1, a robust toolset is provided for content extraction, keyword identification, summarization, and executive summary generation. This automation greatly enhances the efficiency and effectiveness of information management, enabling users to grasp crucial insights from extensive textual documents swiftly. The framework simplifies handling large volumes of text-based data, empowering users to make informed decisions, conduct comprehensive research, and produce concise executive summaries for effective communication and knowledge dissemination.

### Code 1: Content Extraction and Identification of Key Information

Code 1 illustrates a Python-based solution (refer to Code 1 in Appendix A) for extracting content from PDF documents while identifying and highlighting specific keywords or phrases of interest. It employs NLP techniques to analyze the extracted text. The code functions as follows:

- Content Extraction: Utilizing the PyPDF2 library, the code extracts text from PDF documents, facilitating the processing of diverse textual content.
- Keyword Identification: Employing NLP, the code identifies keywords, lemmas, and stems of specified search terms within the extracted text and highlights them within the sentences, enabling users to identify relevant information swiftly.

- Output: Extracted sentences containing the specified keywords are presented with highlighted terms, and the code exports these sentences to a CSV file for further analysis or reference.

### Code 2: Text Summarization for Executive Summary Generation

Code 2 presents a Python script (refer to Code 2 in Appendix A) for generating concise executive summaries from lengthy textual documents utilizing the “summarizer” library. The code operates as follows:

- Text Extraction: The code extracts text content from PDF files, preparing it for summarization.
- Text Summarization: Leveraging the “summarizer” library, the code generates an executive summary by selecting the most informative sentences from the document. Users can specify the desired summary length in sentences.
- Output: The executive summary is printed to the console, offering a succinct overview of the document’s main points. This summary is then processed through AI-assisted technology to request an executive summary regarding skills and competencies.

### *Phase 2.a*

### Skill Identification and Analytics

Phase 2 of the framework, comprising Code 3 and Code 4, offers a versatile toolkit for document summarization and bigram extraction, essential for information organization, retrieval, and insight generation. These processes facilitate uncovering hidden relationships between documents, identifying shared bigrams indicative of common themes, and visualizing document similarity for effective content management. Codes 3 and 4 are pivotal components of this framework, demonstrating their functionality and applications in the context of skills, capabilities, and requirements.

### Code 3: Document Similarity Analysis and Visualization

Code 3 (refer to Code 3 in Appendix A) provides a solution for comparing the similarity between PDF documents within a specified directory. The Python script performs the following tasks:

- Text Extraction: Extracts text content from multiple PDF documents in a designated directory.
- Cosine Similarity Calculation: Computes the cosine similarity between these documents using Term Frequency-Inverse Document Frequency (TF-IDF) vectors, quantifying the degree of textual resemblance.
- Heatmap Visualization: Generates a heatmap visually representing the similarity matrix to aid interpretation. The intensity of colors in the heatmap indicates the degree of similarity between pairs of documents, enabling users to identify clusters of related documents.
- Output: Presents results as a heatmap and a CSV file containing the similarity matrix for further analysis.

### Code 4: Bigram Extraction and Document Clustering

Code 4 introduces a Python script for extracting bigrams (pairs of adjacent words) from PDF documents and their subsequent clustering (refer to Code 4 in Appendix A). The code performs the following tasks:

- Text Extraction: Extracts text content from multiple PDF documents in a specified directory.

- Text Preprocessing: Preprocesses the extracted text, including tokenization, removal of stopwords, and stemming using the Porter stemmer.
- Bigram Generation: Generates bigrams representing pairs of significant words from the preprocessed text, capturing meaningful word combinations for context and insights.
- TF-IDF Vectorization: Transforms bigrams into TF-IDF vectors, quantifying their importance in each document numerically.
- K-Means Clustering: Clusters documents with similar bigrams using the K-Means algorithm, grouping them into clusters.
- Output: Prints clusters of shared bigrams and associated documents. Additionally, AI-assisted technology structures unstructured text by providing insights into skills, talent, and capabilities. The AI-assisted technology is used to provide structure by requesting the following: Provide structure in terms of skill, talent, and, capabilities to the following unstructured text.

### *Phase 2.b*

### Network Analytics and Semantic Clustering

In this phase, the Louvain community detection algorithm is used as a pivotal tool in network analysis to identify communities or clusters within a given network based on the modularity of its structure (Puertas et al., 2021). This algorithm iteratively optimizes the network's modularity by dynamically reassigning nodes to different communities, forming cohesive and densely connected groups. In conjunction with Louvain community detection, NLP bigram extraction plays a crucial role in constructing the network. NLP techniques extract meaningful pairs of adjacent words, bigrams, from textual data. These bigrams serve as the nodes in the network, representing key concepts or entities derived from the text. The connections between nodes are established based on the co-occurrence of bigrams within a specified proximity, thereby capturing semantic relationships and associations in the text. By integrating Louvain community detection with NLP bigram extraction, a comprehensive semantic network can be developed, facilitating the exploration and analysis of complex textual data structures.

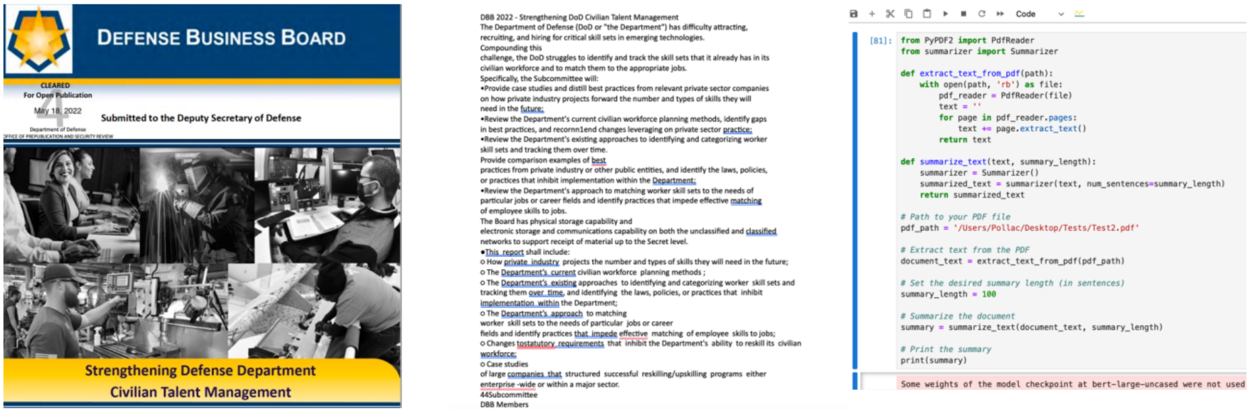
## Framework Implementation and Discussion

### Data Set

The dataset comprises a selection of 75 reports curated from an array of interactions with esteemed experts affiliated with the Systems Engineering Research Center (SERC) and the Acquisition Innovation Research Center (AIRC). These reports provide a rich spectrum of content, spanning sectors and perspectives, including but not limited to governmental insights, industry viewpoints encapsulated in position papers, and scholarly discourse in academic journal articles. A representative sample is included in Appendix A for a glimpse into these reports. If needed, the interested reader may request access to the complete archive.

### Skill Extraction Results

The skill extraction process yielded results from all 75 reports, each of which underwent integration into the semi-automated framework, resulting in the generation of executive reports. Figure 2 illustrates the implementation of Phase 1 of the framework using a specific report as an example, while Table 1 presents the corresponding summary output. Notably, the framework successfully generated 60 executive summaries, although 20 initial documents provided minimal or no information regarding skill sets. These executive summaries can be provided upon request.


DEFENSE BUSINESS BOARD
CLEARED
For Open Publication
May 18, 2022
Department of Defense
OFFICE OF PREPUBLICATION AND SECURITY REVIEW
Submitted to the Deputy Secretary of Defense
Strengthening Defense Department
Civilian Talent Management

DBB 2022 - Strengthening DoD Civilian Talent Management
The Department of Defense (DoD or "the Department") has difficulty attracting,
recruiting, and hiring for critical skill sets in emerging technologies.
Compounding this
challenge, the DoD struggles to identify and track the skill sets that it already has in its
civilian workforce and to match them to the appropriate jobs.
Specifically, the Subcommittee will:
•Provide case studies and distill best practices from relevant private sector companies
on how private industry projects forward the number and types of skills they will
need in the future;
•Review the Department's current civilian workforce planning methods, identify gaps
in best practices, and recornn1end changes leveraging on private sector practice;
•Review the Department's existing approaches to identifying and categorizing worker
skill sets and tracking them over time.
Provide comparison examples of best
practices from private industry or other public entities, and identify the laws, policies,
or practices that inhibit implementation within the Department;
•Review the Department's approach to matching worker skill sets to the needs of
particular jobs or career fields and identify practices that impede effective matching
of employee skills to jobs.
The Board has physical storage capability and
electronic storage and communications capability on both the unclassified and classified
networks to support receipt of material up to the Secret level.
•This report shall include:
o How private industry projects the number and types of skills they will need in the future;
o The Department's current civilian workforce planning methods ;
o The Department's existing approaches to identifying and categorizing worker skill sets and
tracking them over time, and identifying the laws, policies, or practices that inhibit
implementation within the Department;
o The Department's approach to matching
worker skill sets to the needs of particular jobs or career
fields and identify practices that impede effective matching of employee skills to jobs;
o Changes tostatutory requirements that inhibit the Department's ability to reskill its civilian
workforce;
o Case studies
of large companies that structured successful reskilling/upskilling programs either
enterprise -wide or within a major sector.
44Subcommittee
DBB Members

```
[81]: from PyPDF2 import PdfReader
      from summarizer import Summarizer

      def extract_text_from_pdf(path):
          with open(path, 'rb') as file:
              pdf_reader = PdfReader(file)
              text = ''
              for page in pdf_reader.pages:
                  text += page.extract_text()
              return text

      def summarize_text(text, summary_length):
          summarizer = Summarizer()
          summarized_text = summarizer(text, num_sentences=summary_length)
          return summarized_text

      # Path to your PDF file
      pdf_path = '/Users/Pollac/Desktop/Tests/Test2.pdf'

      # Extract text from the PDF
      document_text = extract_text_from_pdf(pdf_path)

      # Set the desired summary length (in sentences)
      summary_length = 100

      # Summarize the document
      summary = summarize_text(document_text, summary_length)

      # Print the summary
      print(summary)

      Some weights of the model checkpoint at bert-large-uncased were not used
```

Input: Report (95 structured pages)

Output: Keyword information Extraction (7 unstructured pages)

Summarize Output (1 page)

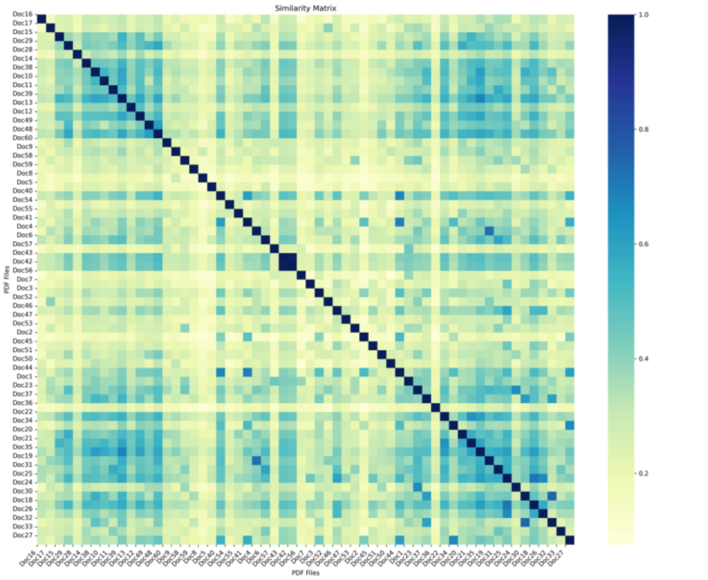
**Figure 2. Defense Business Board Document Phase 1 Example**

**Table 1. Executive Summary for Defense Business Board Document**

| | |
|---|---|
| 1. Talent Acquisition and Recruitment: Skills in attracting and recruiting professionals with critical skill sets in emerging technologies. This includes expertise in sourcing candidates, conducting interviews, assessing qualifications, and employing effective recruitment strategies. | 6. Change Management: Proficiency in managing change within the organization to facilitate the adoption of new talent management practices. This includes communication, stakeholder engagement, and developing strategies to address resistance or challenges related to implementing new approaches to talent management. |
| 2. Workforce Planning: Skills in strategic workforce planning to anticipate and align human capital needs with organizational goals. This involves analyzing current and future skill requirements, identifying gaps, and developing plans to address those gaps through recruitment, training, or other talent management initiatives. | 7. Data Management and Analysis: Competence in data management and analysis to support talent management decisions. This includes using technology platforms and data lakes to track and analyze job-related data, employee skills, and workforce trends, enabling informed decision-making and proactive planning. |
| 3. Skill Set Identification and Tracking: Competence in identifying and categorizing worker skill sets, as well as establishing systems to track and update these skills over time. This includes using technology and data analysis to monitor skill inventories, assess skill gaps, and ensure accurate matching of employee skills to job requirements. | 8. Collaboration and Relationship Building: Skills in building partnerships and collaborative relationships with stakeholders both within and outside the organization. This includes fostering cooperation between different departments, leveraging external expertise, and engaging with private industry partners to exchange knowledge and best practices. |
| 4. Comparative Analysis and Benchmarking: Skills in conducting comparative analysis between the DoD's talent management practices and those of private sector companies or other public entities. This entails identifying best practices, gaps, and areas for improvement, and making recommendations for adapting private industry practices to enhance talent management in the DoD. | 9. Talent Development and Upskilling: Abilities in designing and implementing talent development programs that enhance employees' skills and promote lifelong learning. This involves creating learning opportunities, providing access to training resources, and encouraging employees to expand their knowledge and expertise in line with organizational needs. |
| 5. Skill Matching and Job Alignment: Abilities in matching worker skill sets to the needs of specific jobs or career fields within the DoD. This requires understanding the knowledge, experience, and competencies required for different roles and effectively aligning employees' skills to maximize their contributions and job satisfaction. | 10. Knowledge of Emerging Technologies: Understanding and awareness of emerging technologies relevant to the defense sector. This includes staying abreast of advancements in areas such as cybersecurity, artificial intelligence, data analytics, robotics, and other emerging fields that impact DoD operations and require specialized skill sets. |

## Skill Identification

After acquiring the executive summaries, Phase 2.a involved creating a similarity matrix for all documents, as demonstrated in Figure 3, and extracting pertinent bigrams from document clusters. The task of identifying skill sets was completed utilizing AI-assisted technology, with ChatGPT being the designated tool for this task, as outlined in Table 2. The final output encompasses the comprehensive compilation of skill sets from all 15 clusters (available upon request, given its size). It is essential to emphasize that labeling these skill sets is AI-generated and may require further refinement based on a thorough understanding of DoD requirements.



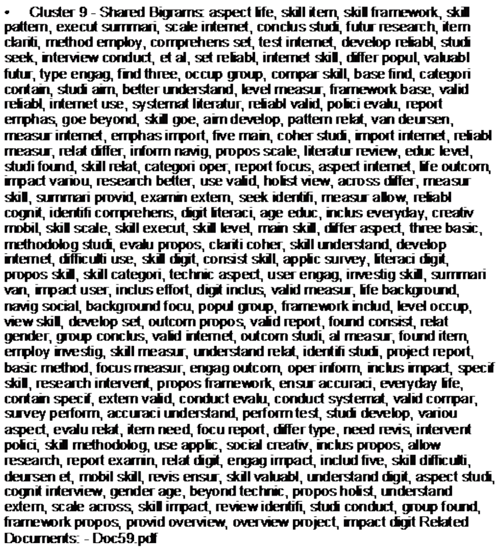

- Cluster 9 - Shared Bigrams: aspect life, skill item, skill framework, skill pattern, execut summari, scale internet, conclus studi, futur research, item clariti, method employ, comprehens set, test internet, develop reliabl, studi seek, interview conduct, et al, set reliabl, internet skill, differ popul, valuabl futur, type engag, find three, occup group, compar skill, base find, categori contain, studi aim, better understand, level measur, framework base, valid reliabl, internet use, systemat literatur, reliabl valid, polici evalu, report emphas, goe beyond, skill goe, aim develop, pattern relat, van deursen, measur internet, emphas import, five main, coher studi, import internet, reliabl measur, relat differ, inform navig, propos scale, literatur review, educ level, studi found, skill relat, categori oper, report focus, aspect internet, life outcom, impact variou, research better, use valid, holist view, across differ, measur skill, summari provid, examin extern, seek identifi, measur allow, reliabl cognit, identifi comprehens, digit literaci, age educ, inclus everyday, creativ mobil, skill scale, skill execut, skill level, main skill, differ aspect, three basic, methodolog studi, evalu propos, clariti coher, skill understand, develop internet, difficulti use, skill digit, consist skill, applic survey, literaci digit, propos skill, skill categori, technic aspect, user engag, investig skill, summari van, impact user, inclus effort, digit inclus, valid measur, life background, navig social, background focu, popul group, framework includ, level occup, view skill, develop set, outcom propos, valid report, found consist, relat gender, group conclus, valid internet, outcom studi, al measur, found item, employ investig, skill measur, understand relat, identifi studi, project report, basic method, focus measur, engag outcom, oper inform, inclus impact, specif skill, research intervent, propos framework, ensur accuraci, everyday life, contain specif, extern valid, conduct evalu, conduct systemat, valid compar, survey perform, accuraci understand, perform test, studi develop, variou aspect, evalu relat, item need, focu report, differ type, need revis, intervent polici, skill methodolog, use applic, social creativ, inclus propos, allow research, report examin, relat digit, engag impact, includ five, skill difficulti, deursen et, mobil skill, revis ensur, skill valuabl, understand digit, aspect studi, cognit interview, gender age, beyond technic, propos holist, understand extern, scale across, skill impact, review identifi, studi conduct, group found, framework propos, provid overview, overview project, impact digit Related Documents: - Doc59.pdf

**Figure 3. Similarity Matrix for All Executive Summaries of Phase 1 and Cluster 9**

## Network Analytics

After completing Phase 1 (see Figure 4) of the framework and to construct a semantic network suitable community detection, the following steps involve defining the nodes of the network, which are determined as the most frequent bigrams within the corpus. Following node selection, the network's links are established based on the co-occurrence of bigrams within a proximity window of size 10. These links are then weighted according to the frequency of co-occurrences. Once the semantic network is established, Louvain community detection is applied to identify semantic clusters within the network. The Louvain algorithm operates iteratively, beginning with small communities and gradually adjusting the modularity by adding or removing nodes from communities. This process continues until an optimal modularity score is achieved. Nodes with higher modularity scores, indicative of denser clusters, are grouped together, thus forming distinct semantic clusters within the network.

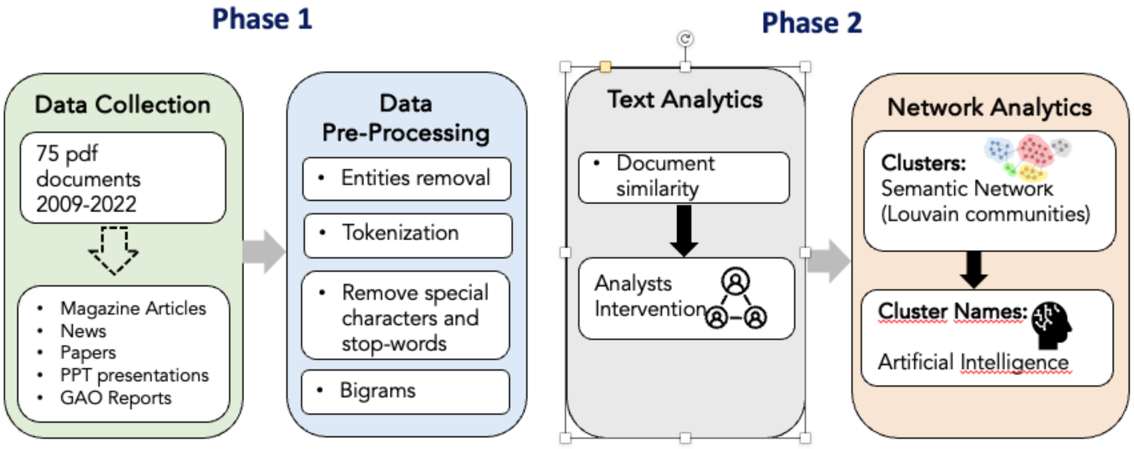


**Figure 4. Network Analytics Stage within Framework Phases 1 and 2**

The network analysis of the implemented Louvain algorithm on the series of documents generated a network consisting of 318 nodes and 4,743 edges, indicating a density of 0.09410. The network exhibited a transitivity of 0.37795 and an average clustering coefficient of 0.5030, suggesting a moderate level of clustering within the network. Furthermore, nodes with higher centrality, such as “United_States,” “Department_defense,” and “National_security,” emerged as prominent entities, highlighting their significance in the network structure. Additionally, nodes with higher closeness, including “Artificial_Intelligence,” “Big_data,” and “Data_science,” were identified, underscoring their pivotal role in facilitating efficient information flow within the network. These findings provide valuable insights into the key entities and their interconnectedness within the document network, shedding light on the underlying themes and relationships present in the dataset. Figure 5 illustrates the final network obtained along with the communities identified.

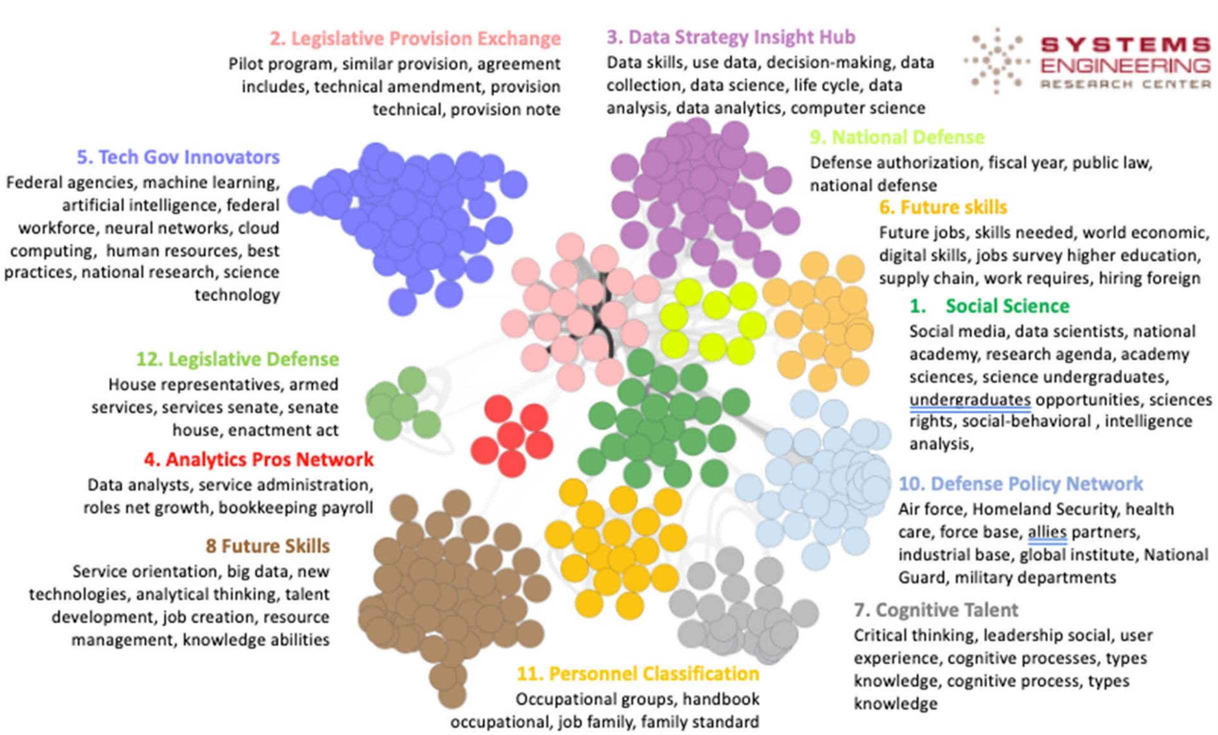

2. Legislative Provision Exchange
Pilot program, similar provision, agreement includes, technical amendment, provision technical, provision note
3. Data Strategy Insight Hub
Data skills, use data, decision-making, data collection, data science, life cycle, data analysis, data analytics, computer science
SYSTEMS ENGINEERING RESEARCH CENTER
5. Tech Gov Innovators
Federal agencies, machine learning, artificial intelligence, federal workforce, neural networks, cloud computing, human resources, best practices, national research, science technology
9. National Defense
Defense authorization, fiscal year, public law, national defense
6. Future skills
Future jobs, skills needed, world economic, digital skills, jobs survey higher education, supply chain, work requires, hiring foreign
1. Social Science
Social media, data scientists, national academy, research agenda, academy sciences, science undergraduates, undergraduates opportunities, sciences rights, social-behavioral , intelligence analysis,
12. Legislative Defense
House representatives, armed services, services senate, senate house, enactment act
4. Analytics Pros Network
Data analysts, service administration, roles net growth, bookkeeping payroll
10. Defense Policy Network
Air force, Homeland Security, health care, force base, allies partners, industrial base, global institute, National Guard, military departments
8 Future Skills
Service orientation, big data, new technologies, analytical thinking, talent development, job creation, resource management, knowledge abilities
7. Cognitive Talent
Critical thinking, leadership social, user experience, cognitive processes, types knowledge, cognitive process, types knowledge
11. Personnel Classification
Occupational groups, handbook occupational, job family, family standard

**Figure 5. Skill Communities Identified**

## Conclusions and Future Work

The proposed framework has adeptly addressed the core research inquiries guiding this study. Initially, it confronts the intricate task of skill identification within the DoD workforce by harnessing the capabilities of NLP algorithms. Through meticulous analysis of a diverse corpus comprising over 80 documents from industry, government, and academic reports, the framework seamlessly automates the detection of pivotal skills. This approach furnishes valuable insights into the indispensable skill sets for tackling global challenges and proficiently managing the DoD workforce, bolstering the organization’s preparedness and robustness.

Moreover, the framework propels the realm of decision support within the DoD by capitalizing on NLP analysis to empower decision-makers. By distilling actionable insights and recommendations from the extensive document analysis, it equips DoD leadership with the requisite knowledge to make judicious decisions concerning talent acquisition, training initiatives, and the mitigation of skill disparities. In an epoch characterized by dynamic global challenges, the framework's contributions transcend mere automation; it emerges as a strategic apparatus for augmenting the agility and efficacy of the DoD, thereby enhancing the organization's ability to navigate evolving complexities with acumen and foresight.

Promising opportunities for further research and development are evident in several domains, with Skill Mapping emerging as a crucial area for advancement within the DoD. Future endeavors in this field aim to develop an intelligent system capable of mapping identified skills to specific job roles and career pathways. This innovative approach has the potential to revolutionize talent allocation, enabling the DoD to match personnel skills with roles and development trajectories precisely, thereby enhancing organizational agility and preparedness.

Moreover, the scope of NLP-driven talent management extends to Predictive Analysis, offering fertile ground for exploration. Future initiatives can focus on forecasting forthcoming skill requirements by harnessing the power of NLP-driven predictive models. These models would facilitate proactive workforce planning and readiness by integrating insights from emerging trends, technological advancements, and evolving operational needs. Such predictive capabilities are poised to significantly impact the DoD's ability to anticipate skill demands, stay ahead of evolving challenges, and cultivate a workforce equipped to navigate the dynamic demands of a rapidly changing global landscape. In summary, the realm of NLP-based talent management holds immense promise, and these avenues for further research and development serve as guiding lights toward fostering a more agile and effective DoD.

## References

Allioui, H., & Mourdi, Y. (2023). Unleashing the potential of AI: Investigating cutting-edge technologies that are transforming businesses. *International Journal of Computer Engineering and Data Science (IJCEDS)*, *3*(2), 1–12. https://ijceds.com/ijceds/article/view/59

Caratozzolo, P., Alvarez-Delgado, A., & Rodriguez-Ruiz, J. (2023). Applications of natural language processing for industry 4.0 skills development. *2023 Future of Educational Innovation-Workshop Series Data in Action*, 1–9. https://doi.org/10.1109/IEEECONF56852.2023.10104796

Gomathi, S., Rajeswari, A., & Kadry, S. (2023). Emerging HR practices—digital upskilling: A strategic way of talent management and engagement. In *Disruptive artificial intelligence and sustainable human resource management* (1st ed.). River Publishers. https://www.taylorfrancis.com/books/edit/10.1201/9781032622743/disruptive-artificial-intelligence-sustainable-human-resource-management-anamika-pandey-balamurugan-balusamy-naveen-chilamkurti?refId=85c9b79f-adb5-461d-a002-77c14fe3d195&context=ubx

Mohanty, S., Behera, A., Mishra, S., Alkhayyat, A., Gupta, D., & Sharma, V. (2023). Resumate: A prototype to enhance recruitment process with NLP based resume parsing. *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*, 1–6. https://doi.org/10.1109/ICIEM59379.2023.10166169

Puertas, E., Moreno-Sandoval, L. G., Redondo, J., Alvarado-Valencia, J., & Pomares-Quimbaya, A. (2021). Detection of sociolinguistic features in digital social networks for the detection of communities. *Cognitive Computation*, *13*, 518–537. https://doi.org/10.1007/s12559-021-09818-9

Vanetik, N., & Kogan, G. (2023). Job vacancy ranking with sentence embeddings, keywords, and named entities. *Information*, *14*(8), 468. https://doi.org/10.3390/info14080468

## Appendix A. Selected Reports

Presidents Management Agenda, Federal Data Strategy, Curated Data Skills Catalog, November 2020. Accessed: https://strategy.data.gov/action-plan/

World Economic Forum, The future of Jobs Report, October 2020,

Accessed: https://www.weforum.org/publications/the-future-of-jobs-report-2020/

Jacobson, S. Maximizing the data Literacy of the Airforce Contracting Workforce, MBA Professional Project, Naval Postgraduate School, December 2021.

The White House, National Security Strategy, October 2022.

Accessed: https://www.whitehouse.gov/wp-content/uploads/2022/10/Biden-Harris-Administrations-National-Security-Strategy-10.2022.pdf

U.S. Department of Defense, 2022 National Defense Strategy, October 2022.

Accessed: https://www.defense.gov/News/News-Stories/Article/Article/3202438/dod-releases-national-defense-strategy-missile-defense-nuclear-posture-reviews/#:~:text=The%202022%20National%20Defense%20Strategy,and%20partners%20on%20shared%20objectives.

U.S. Department of Defense, National Defense Science and Technology Strategy 2023.

Accessed: https://media.defense.gov/2023/May/09/2003218877/-1/-1/0/NDSTS-FINAL-WEB-VERSION.PDF

Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc. July 2015;103(3):152-3. doi: 10.3163/1536-5050.103.3.010. PMID: 26213509; PMCID: PMC4511057.

Anderson, L. W. and Krathwohl, D. R., et al (Eds..) A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Allyn & Bacon. Boston, MA (Pearson Education Group) 2021

Scherger Group, Future Workforce 2025, September 2019.

Accessed: https://theforge.defence.gov.au/article/future-workforce-2025-scherger-group

McKinsey Global Institute. Skill Shift Automation and the Future of the Workforce. May 2018.

Accessed: https://www.mckinsey.com/featured-insights/future-of-work/skill-shift-automation-and-the-future-of-the-workforce

Future Skills Council. Canada – A Learning Nation: A Skilled, Agile Workforce Ready to Shape the Future. November 2020. ISBN: 978-0-660-35859-8

Deloitte Insights. The skills-based organization: A new operating model for work and the workforce. September 2022.

Accessed: https://www2.deloitte.com/us/en/insights/topics/talent/organizational-skill-based-hiring.html

Gehlhaus, D., Ryseff, J. and Corrigan, J. The Race for U.S. Technical Talent: Can the DOD and DIB Compete? Center for Security and Emerging Technology. August 2023. DOI: 10.51593/20210074

Defense Business Board. Strengthening Defense Department Civilian Talent Management. May 2022

Accessed: https://dbb.defense.gov/Portals/35/Documents/Reports/2022/DBB%20FY22-03%20Talent%20Management%20Study%20Report%2018%20Aug%202022%20-%20CLEARED.pdf

Defense Civilian Personnel Advisory Service. Strategic Workforce Planning Guide. May 2019

Accessed: https://www.dcpas.osd.mil/sites/default/files/DoD%20Strategic%20Workforce%20Planning%20Guide%20-%2030May2019.pdf

## Appendix B. Code

### Code 1

```
import os
import ssl
import certifi
import PyPDF2
import nltk
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.stem import WordNetLemmatizer, PorterStemmer
from termcolor import colored
import csv
# Set the SSL certificate verification
ssl._create_default_https_context = ssl.create_default_context(cafile=certifi.where())
def load_pdf(file_path):
    try:
        with open(file_path, "rb") as file:
            reader = PyPDF2.PdfReader(file)
            text = ""
            for page in reader.pages:
                text += page.extract_text()
            return text
    except FileNotFoundError:
        print("File not found.")
        return None
    except PyPDF2.PdfReadError:
        print("Invalid PDF file.")
        return None
def extract_sentences_with_words(text, words):
    sentences = sent_tokenize(text)
    sentences_with_words = []
    lemmatizer = WordNetLemmatizer()
    stemmer = PorterStemmer()
    # Extract sentences with the specified words
    for sentence in sentences:
        tokens = word_tokenize(sentence)
        lemmas = [lemmatizer.lemmatize(token) for token in tokens]
        stems = [stemmer.stem(token) for token in tokens]

        # Prepare the set of unique word forms (word, lemma, and stem) to search for
        search_words = set()
        for word in words:
            search_words.add(word)
            search_words.add(lemmatizer.lemmatize(word))
            search_words.add(stemmer.stem(word))
        if any(token in search_words for token in tokens) or any(lemma in search_words for lemma
in lemmas) or any(stem in search_words for stem in stems):
            highlighted_sentence = highlight_words(sentence, tokens, lemmas, stems,
search_words)
            sentences_with_words.append(highlighted_sentence)
    return sentences_with_words
def highlight_words(sentence, tokens, lemmas, stems, search_words):
    highlighted_sentence = sentence
    for token in tokens:
        if token in search_words:
            highlighted_sentence = highlighted_sentence.replace(token, colored(token, 'yellow'))
    for lemma in lemmas:
        if lemma in search_words:
            highlighted_sentence = highlighted_sentence.replace(lemma, colored(lemma, 'yellow'))
    for stem in stems:
        if stem in search_words:
            highlighted_sentence = highlighted_sentence.replace(stem, colored(stem, 'green'))
    return highlighted_sentence
# Set the PDF file path
pdf_file = '/Users/ Tests/Test1.pdf'
```

```
# Load the PDF file
text = load_pdf(pdf_file)
if text:
    # Specify the list of words you want to search for
    words = ["skill", "competencies", "capability", "talent"]
    # Extract sentences with the specified words
    sentences = extract_sentences_with_words(text, words)
    # Print the extracted sentences with highlighted words, lemmas, and stems
    if sentences:
        print("Sentences with the specified words:")
        for sentence in sentences:
            print(sentence)
    else:
        print("No sentences found with the specified words.")
    # Save the output to a CSV file
    output_file = 'output.csv'
    with open(output_file, 'w', newline='', encoding='utf-8') as csvfile:
        writer = csv.writer(csvfile)
        writer.writerow(['Sentences with the specified words:'])
        writer.writerows([[sentence] for sentence in sentences])
        writer.writerow(['No sentences found with the specified words.'])
```

**Code 2**

```
from PyPDF2 import PdfReader
from summarizer import Summarizer
def extract_text_from_pdf(path):
    with open(path, 'rb') as file:
        pdf_reader = PdfReader(file)
        text = ''
        for page in pdf_reader.pages:
            text += page.extract_text()
        return text
def summarize_text(text, summary_length):
    summarizer = Summarizer()
    summarized_text = summarizer(text, num_sentences=summary_length)
    return summarized_text
```

```
# Path to your PDF file
pdf_path = '/Users/ Tests/Test2.pdf'
# Extract text from the PDF
document_text = extract_text_from_pdf(pdf_path)
# Set the desired summary length (in sentences)
summary_length = 100
# Summarize the document
summary = summarize_text(document_text, summary_length)
# Print the summary
print(summary)
```

**Code 3**

```
import os
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from PyPDF2 import PdfReader
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
def read_pdf(file_path):
    with open(file_path, 'rb') as file:
        pdf_reader = PdfReader(file)
        text = ""
        for page in pdf_reader.pages:
            text += page.extract_text()
        return text
def compare_similarity(documents_dir):
    # Get a list of PDF files in the given directory
    pdf_files = [file for file in os.listdir(documents_dir) if file.endswith('.pdf')]
    # Read and preprocess the content of each PDF document
    texts = []
    for pdf_file in pdf_files:
        file_path = os.path.join(documents_dir, pdf_file)
        text = read_pdf(file_path)
        texts.append(text)
    # Compute TF-IDF vectors for the documents
    vectorizer = TfidfVectorizer()
    tfidf_matrix = vectorizer.fit_transform(texts)
    # Calculate the cosine similarity matrix
    similarity_matrix = cosine_similarity(tfidf_matrix)
    # Convert similarity matrix to DataFrame for better representation
    similarity_df = pd.DataFrame(similarity_matrix, columns=pdf_files, index=pdf_files)
    # Extract file names without extensions
    pdf_names = [os.path.splitext(pdf_file)[0] for pdf_file in pdf_files]
    # Set up the figure and axis
    plt.figure(figsize=(16, 13))
    ax = sns.heatmap(similarity_df, annot=False, cmap='YlGnBu', fmt=".2f")
    # Customize the plot
    ax.set_title("Similarity Matrix")
    ax.set_xlabel("PDF Files")
    ax.set_ylabel("PDF Files")
    plt.xticks(range(len(pdf_names)), pdf_names, rotation=45, ha='right')
    plt.yticks(range(len(pdf_names)), pdf_names, rotation=0)
    # Save the plot as an image (JPEG or PNG)
    output_image_path = os.path.join(documents_dir, "similarity_matrix.png")
    plt.tight_layout()
    plt.savefig(output_image_path, dpi=300)
    plt.show()
 # Output similarity matrix to a CSV file
    similarity_df.to_csv(os.path.join(documents_dir, "similarity_matrix.csv"))
if __name__ == "__main__":
    # Replace 'path/to/directory' with the path to your directory containing the PDF files
    documents_directory = '/Users /Tests'
    compare_similarity(documents_directory)
```

**Code 4:**

```
import os
import ssl
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
import numpy as np
from PyPDF2 import PdfReader

# Configure SSL context to bypass certificate verification for NLTK downloads
try:
    _create_unverified_https_context = ssl._create_unverified_context
except AttributeError:
    pass
else:
    ssl._create_default_https_context = _create_unverified_https_context

# Download necessary NLTK resources
nltk.download('punkt')
nltk.download('stopwords')

# Function to read text from PDF
def read_pdf(file_path):
    with open(file_path, 'rb') as file:
        pdf_reader = PdfReader(file)
        text = ""
        for page in pdf_reader.pages:
            text += page.extract_text()
        return text

# Function to preprocess the text and generate bigrams
def preprocess_text(text):
    # Tokenize the text into words
    words = nltk.word_tokenize(text.lower())

    # Remove stopwords and non-alphabetic characters from words
    stop_words = set(stopwords.words('english'))
    words = [word for word in words if word.isalpha() and word not in stop_words]

    # Stemming using Porter stemmer
    stemmer = PorterStemmer()
    words = [stemmer.stem(word) for word in words]

    # Generate bigrams
    bigrams = list(nltk.bigrams(words))
    bigrams = [" ".join(bigram) for bigram in bigrams]

    return bigrams

# Specify the directory containing PDF documents
documents_directory = '/Users/Tests'

# Get a list of PDF files in the given directory
pdf_files = [os.path.join(documents_directory, file) for file in os.listdir(documents_directory) if file.endswith('.pdf')]

# Read and preprocess the content of each PDF document
documents = [read_pdf(file_path) for file_path in pdf_files]

# Preprocess the documents and generate bigrams
preprocessed_documents = [preprocess_text(doc) for doc in documents]

# Calculate TF-IDF vectors for bigrams
vectorizer = TfidfVectorizer()
tfidf_vectors = vectorizer.fit_transform([" ".join(bigrams) for bigrams in preprocessed_documents])

# Clustering using K-Means
num_clusters = 15  # Change this value based on the number of clusters you want
kmeans = KMeans(n_clusters=num_clusters, random_state=42)
clusters = kmeans.fit_predict(tfidf_vectors)

# Create clusters of shared bigrams
cluster_bigrams = {}
for i, cluster_label in enumerate(clusters):
    if cluster_label not in cluster_bigrams:
        cluster_bigrams[cluster_label] = set()
    bigrams = preprocessed_documents[i]
    cluster_bigrams[cluster_label].update(bigrams)
# Create a dictionary to store documents corresponding to each cluster
cluster_documents = {}
for i, cluster_label in enumerate(clusters):
    if cluster_label not in cluster_documents:
        cluster_documents[cluster_label] = []
    cluster_documents[cluster_label].append(pdf_files[i])

# Print the clusters and shared bigrams along with associated documents
for cluster_label, bigrams in cluster_bigrams.items():
    print(f"Cluster {cluster_label + 1} - Shared Bigrams:")
    print(", ".join(bigrams))
    print("Related Documents:")
    for document_path in cluster_documents[cluster_label]:
        print(f"- {os.path.basename(document_path)}")
    print("---")
```

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net