



INSTITUTE FOR DEFENSE ANALYSES

Multi-Use Text Analytics

Arun S. Maiya, Project Leader

Margaret E. Zientek

Dale Visser

October 2023

Distribution Statement A.
Approved for public release:
distribution is unlimited.

IDA Product 3000530



The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication

This work was conducted by the IDA Systems and Analyses Center under contract HQ0034-19-D-0001, Project AI-5-5214, "DAMO Triage R&D Support," for the Maintaining Technology Advantage (MTA), OUSD (R&E). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgements

For More Information

Arun S. Maiya, Project Leader
amaiya@ida.org, 703-578-2790

Margaret E. Myers, Director, Information Technology and Systems Division
mmyers@ida.org, 703-578-2782

Copyright Notice

© 2023 Institute for Defense Analyses
730 East Glebe Road, Alexandria, Virginia 22305 • (703) 845-2000.

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 (Feb. 2014).

Executive Summary

How does one effectively explore and extract insights from large collections of highly heterogeneous documents across a diverse set of use cases? The Institute for Defense Analyses (IDA), a nonprofit corporation operating three federally funded research and development centers, helps government sponsors answer challenging questions that lie at the intersection of national security, science, technology, and policy. These questions often involve making sense of large numbers of documents and files. Moreover, factors important to one question may not be important for others. Most existing solutions to document analysis and review employ a one-size-fits-all approach that cannot be easily adapted to different use cases. IDA has developed in-house text analytics capabilities to facilitate search, exploration, and analysis of large collections of files in ways that enable greater flexibility in answering research questions. An extensible set of machine learning and natural language processing methods can be targeted and applied in real time to subsets of documents of interest through a no-code, point-and-click, web-based user interface, which allows for more focused insight extractions. In addition, programmatic APIs are locally and remotely accessible, which allow users to customize document processing and analyses for a wide range of different use cases without requiring source code modifications. For instance, researchers have previously used the API to ingest documents from diverse data sources and train custom machine learning models to auto-label documents to make them findable or filterable. In this talk, I provide a high-level overview of this work with illustrative examples.



Multi-Use Text Analytics

Margaret Zientek, Dale Visser, Arun Maiya (Project Leader)

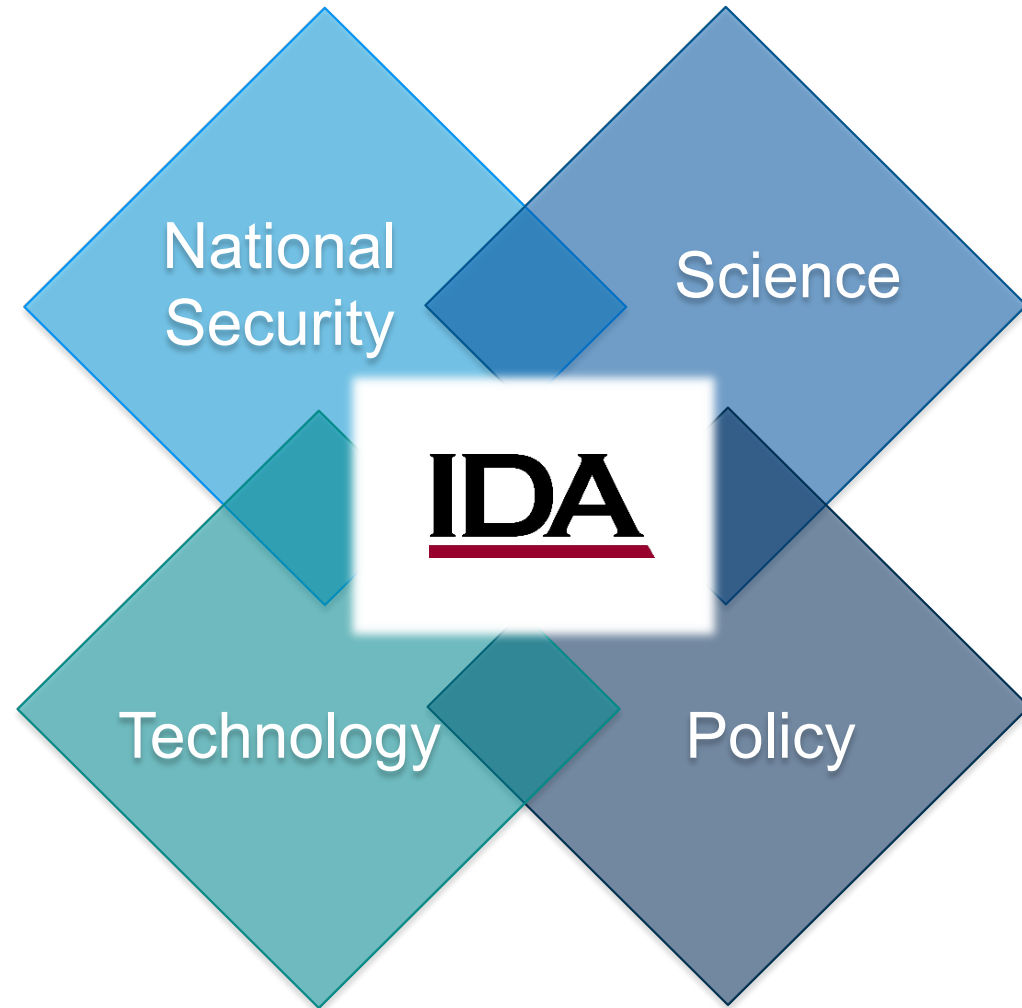
9 November 2023

Institute for Defense Analyses
730 East Glebe Road • Alexandria, Virginia 22305

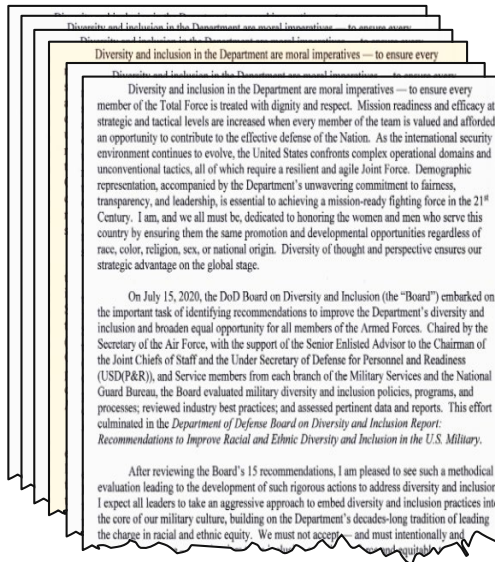
IDA is a Research Organization with a Wide Range of Use-Cases for Text Analytics



Image from <https://www.ida.org>



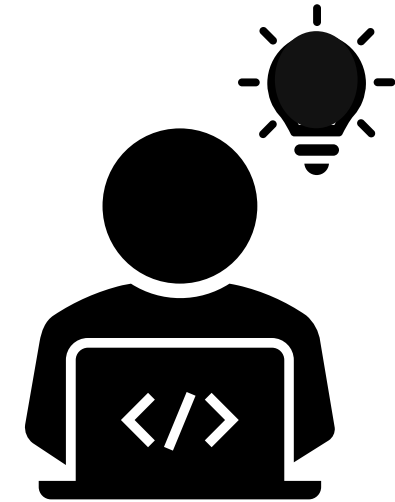
Agenda



Common Tasks for Text Analytics at IDA



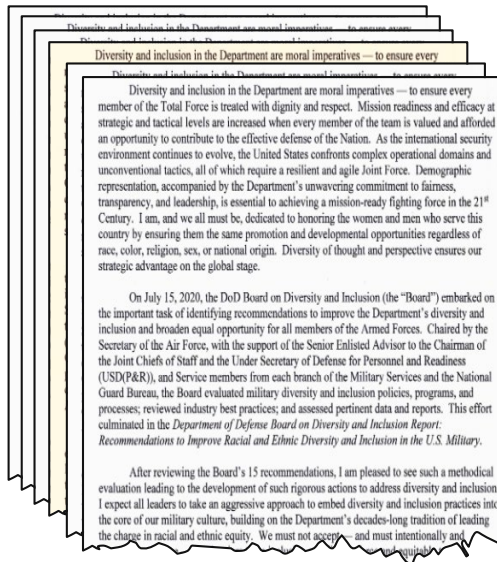
Customizable Software Can Be Tailored to the User's Needs



Real-time Analysis Features Enable Greater Insight into Data

Task 1: Search – Find Important Documents

Large Corpus



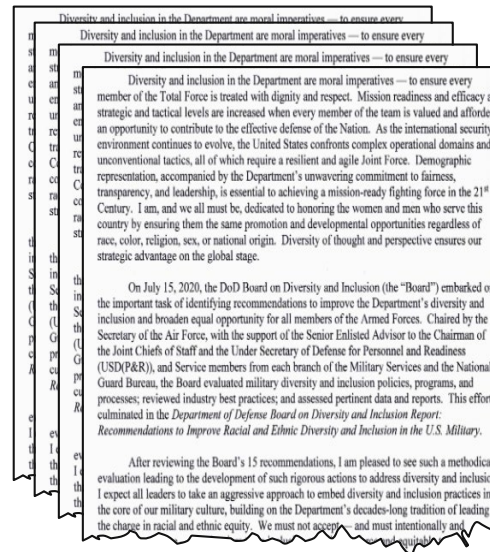
Diversity and inclusion in the Department are moral imperatives — to ensure every member of the Total Force is treated with dignity and respect. Mission readiness and efficacy at strategic and tactical levels are increased when every member of the team is valued and afforded an opportunity to contribute to the effective defense of the Nation. As the international security environment continues to evolve, the United States confronts complex operational domains and unconventional tactics, all of which require a resilient and agile Joint Force. Demographic representation, accompanied by the Department's unwavering commitment to fairness, transparency, and leadership, is essential to achieving a mission-ready fighting force in the 21st Century. I am, and we all must be, dedicated to honoring the women and men who serve this country by ensuring them the same promotion and developmental opportunities regardless of race, color, religion, sex, or national origin. Diversity of thought and perspective ensures our strategic advantage on the global stage.

On July 15, 2020, the DoD Board on Diversity and Inclusion (the "Board") embarked on the important task of identifying recommendations to improve the Department's diversity and inclusion and broaden equal opportunity for all members of the Armed Forces. Chaired by the Secretary of the Air Force, with the support of the Senior Enlisted Advisor to the Chairman of the Joint Chiefs of Staff and the Under Secretary of Defense for Personnel and Readiness (USD(P&R)), and Service members from each branch of the Military Services and the National Guard Bureau, the Board evaluated military diversity and inclusion policies, programs, and processes; reviewed industry best practices; and assessed pertinent data and reports. This effort culminated in the *Department of Defense Board on Diversity and Inclusion Report: Recommendations to Improve Racial and Ethnic Diversity and Inclusion in the U.S. Military*.

After reviewing the Board's 15 recommendations, I am pleased to see such a methodical evaluation leading to the development of such rigorous actions to address diversity and inclusion. I expect all leaders to take an aggressive approach to embed diversity and inclusion practices into the core of our military culture, building on the Department's decades-long tradition of leading the charge in racial and ethnic equity. We must not accept... and must intentionally and...



Uninteresting

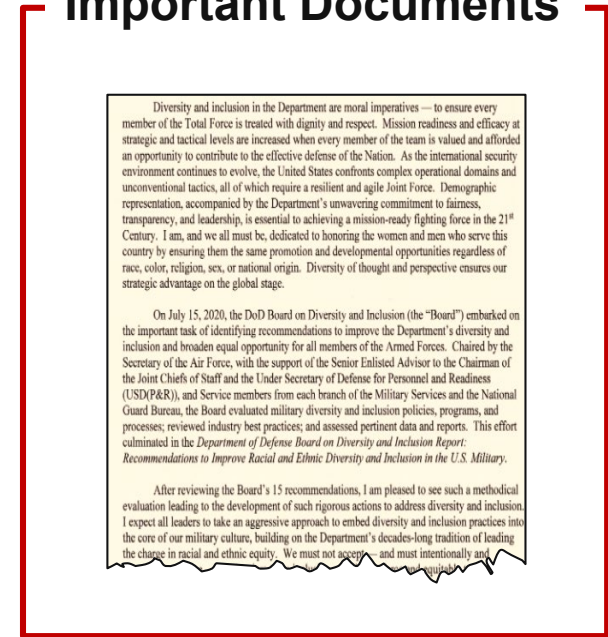


Diversity and inclusion in the Department are moral imperatives — to ensure every member of the Total Force is treated with dignity and respect. Mission readiness and efficacy at strategic and tactical levels are increased when every member of the team is valued and afforded an opportunity to contribute to the effective defense of the Nation. As the international security environment continues to evolve, the United States confronts complex operational domains and unconventional tactics, all of which require a resilient and agile Joint Force. Demographic representation, accompanied by the Department's unwavering commitment to fairness, transparency, and leadership, is essential to achieving a mission-ready fighting force in the 21st Century. I am, and we all must be, dedicated to honoring the women and men who serve this country by ensuring them the same promotion and developmental opportunities regardless of race, color, religion, sex, or national origin. Diversity of thought and perspective ensures our strategic advantage on the global stage.

On July 15, 2020, the DoD Board on Diversity and Inclusion (the "Board") embarked on the important task of identifying recommendations to improve the Department's diversity and inclusion and broaden equal opportunity for all members of the Armed Forces. Chaired by the Secretary of the Air Force, with the support of the Senior Enlisted Advisor to the Chairman of the Joint Chiefs of Staff and the Under Secretary of Defense for Personnel and Readiness (USD(P&R)), and Service members from each branch of the Military Services and the National Guard Bureau, the Board evaluated military diversity and inclusion policies, programs, and processes; reviewed industry best practices; and assessed pertinent data and reports. This effort culminated in the *Department of Defense Board on Diversity and Inclusion Report: Recommendations to Improve Racial and Ethnic Diversity and Inclusion in the U.S. Military*.

After reviewing the Board's 15 recommendations, I am pleased to see such a methodical evaluation leading to the development of such rigorous actions to address diversity and inclusion. I expect all leaders to take an aggressive approach to embed diversity and inclusion practices into the core of our military culture, building on the Department's decades-long tradition of leading the charge in racial and ethnic equity. We must not accept... and must intentionally and...

Important Documents

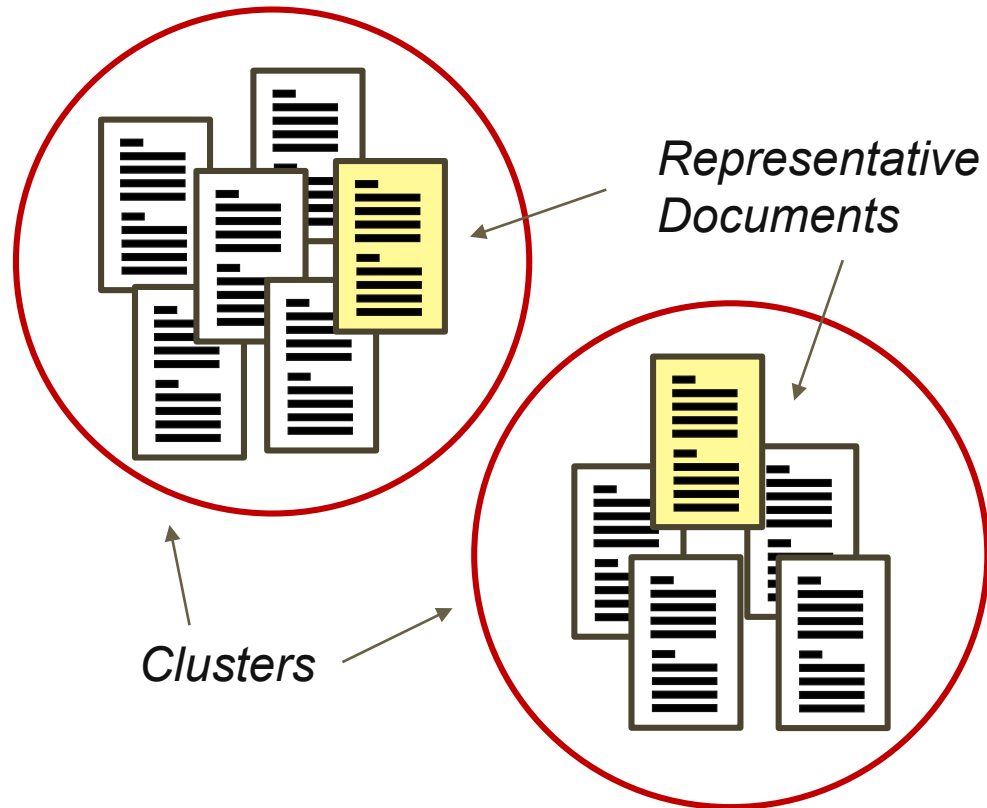


Diversity and inclusion in the Department are moral imperatives — to ensure every member of the Total Force is treated with dignity and respect. Mission readiness and efficacy at strategic and tactical levels are increased when every member of the team is valued and afforded an opportunity to contribute to the effective defense of the Nation. As the international security environment continues to evolve, the United States confronts complex operational domains and unconventional tactics, all of which require a resilient and agile Joint Force. Demographic representation, accompanied by the Department's unwavering commitment to fairness, transparency, and leadership, is essential to achieving a mission-ready fighting force in the 21st Century. I am, and we all must be, dedicated to honoring the women and men who serve this country by ensuring them the same promotion and developmental opportunities regardless of race, color, religion, sex, or national origin. Diversity of thought and perspective ensures our strategic advantage on the global stage.

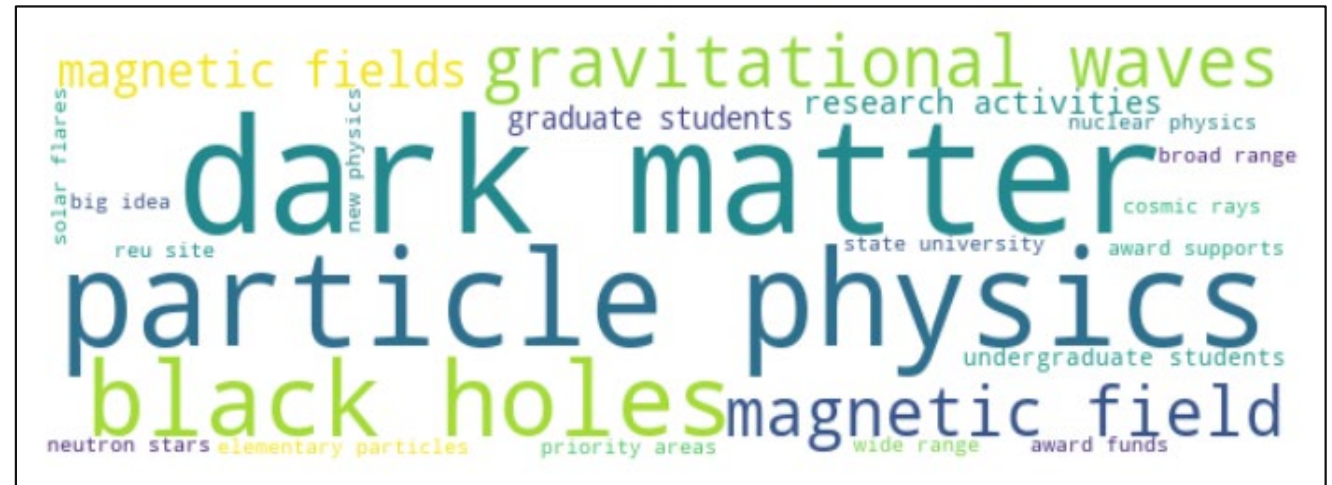
On July 15, 2020, the DoD Board on Diversity and Inclusion (the "Board") embarked on the important task of identifying recommendations to improve the Department's diversity and inclusion and broaden equal opportunity for all members of the Armed Forces. Chaired by the Secretary of the Air Force, with the support of the Senior Enlisted Advisor to the Chairman of the Joint Chiefs of Staff and the Under Secretary of Defense for Personnel and Readiness (USD(P&R)), and Service members from each branch of the Military Services and the National Guard Bureau, the Board evaluated military diversity and inclusion policies, programs, and processes; reviewed industry best practices; and assessed pertinent data and reports. This effort culminated in the *Department of Defense Board on Diversity and Inclusion Report: Recommendations to Improve Racial and Ethnic Diversity and Inclusion in the U.S. Military*.

After reviewing the Board's 15 recommendations, I am pleased to see such a methodical evaluation leading to the development of such rigorous actions to address diversity and inclusion. I expect all leaders to take an aggressive approach to embed diversity and inclusion practices into the core of our military culture, building on the Department's decades-long tradition of leading the charge in racial and ethnic equity. We must not accept... and must intentionally and...

Task 2: Discovery – Find Themes in Documents



Example Cluster



Task 3: Comparison – Analyze Similar Documents

Enacted Law

PUBLIC LAW 117-263—DEC. 23, 2022 136 STAT. 2395

Public Law 117-263
117th Congress

An Act

To authorize appropriations for fiscal year 2023 for military activities of the Department of Defense, for military construction, and for defense activities of the Department of Energy, to prescribe military personnel strengths for such fiscal year, and for other purposes.

Dec. 23, 2022
[H.R. 7776]

James M. Inhofe National Defense Authorization Act for Fiscal Year 2023.
10 USC 101 note.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

(a) IN GENERAL.—This Act may be cited as the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023.”

(b) REFERENCES.—Any reference in this or any other Act to the “National Defense Authorization Act for Fiscal Year 2023” shall be deemed to be a reference to the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023.”

SEC. 2. ORGANIZATION OF ACT INTO DIVISIONS; TABLE OF CONTENTS.

(a) DIVISIONS.—This Act is organized into 11 divisions as follows:

- (1) Division A—Department of Defense Authorizations.
- (2) Division B—Military Construction Authorizations.
- (3) Division C—Department of Energy National Security Authorizations and Other Authorizations.
- (4) Division D—Funding Tables.
- (5) Division E—Non-Department of Defense Matters.
- (6) Division F—Intelligence Authorization Act for Fiscal Year 2023.
- (7) Division G—Homeland Security.
- (8) Division H—Water Resources.
- (9) Division I—Department of State Authorizations.
- (10) Division J—Oceans and Atmosphere.
- (11) Division K—Don Young Coast Guard Authorization Act of 2022.

(b) TABLE OF CONTENTS.—The table of contents for this Act is as follows:

Sec. 1. Short title.
Sec. 2. Organization of Act into divisions; table of contents.
Sec. 3. Congressional defense committees.
Sec. 4. Budgetary effects of this Act.
Sec. 5. Explanatory statement.

DIVISION A—DEPARTMENT OF DEFENSE AUTHORIZATIONS

TITLE I—PROCUREMENT

Subtitle A—Authorization of Appropriations

Sec. 101. Authorization of appropriations.

House Version

PUBLIC LAW 117-263—DEC. 23, 2022 136 STAT. 2395

Public Law 117-263
117th Congress

An Act

To authorize appropriations for fiscal year 2023 for military activities of the Department of Defense, for military construction, and for defense activities of the Department of Energy, to prescribe military personnel strengths for such fiscal year, and for other purposes.

Dec. 23, 2022
[H.R. 7776]

James M. Inhofe National Defense Authorization Act for Fiscal Year 2023.
10 USC 101 note.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

(a) IN GENERAL.—This Act may be cited as the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023.”

(b) REFERENCES.—Any reference in this or any other Act to the “National Defense Authorization Act for Fiscal Year 2023” shall be deemed to be a reference to the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023.”

SEC. 2. ORGANIZATION OF ACT INTO DIVISIONS; TABLE OF CONTENTS.

(a) DIVISIONS.—This Act is organized into 11 divisions as follows:

- (1) Division A—Department of Defense Authorizations.
- (2) Division B—Military Construction Authorizations.
- (3) Division C—Department of Energy National Security Authorizations and Other Authorizations.
- (4) Division D—Funding Tables.
- (5) Division E—Non-Department of Defense Matters.
- (6) Division F—Intelligence Authorization Act for Fiscal Year 2023.
- (7) Division G—Homeland Security.
- (8) Division H—Water Resources.
- (9) Division I—Department of State Authorizations.
- (10) Division J—Oceans and Atmosphere.
- (11) Division K—Don Young Coast Guard Authorization Act of 2022.

(b) TABLE OF CONTENTS.—The table of contents for this Act is as follows:

Sec. 1. Short title.
Sec. 2. Organization of Act into divisions; table of contents.
Sec. 3. Congressional defense committees.
Sec. 4. Budgetary effects of this Act.
Sec. 5. Explanatory statement.

DIVISION A—DEPARTMENT OF DEFENSE AUTHORIZATIONS

TITLE I—PROCUREMENT

Subtitle A—Authorization of Appropriations

Sec. 101. Authorization of appropriations.

Senate Version

PUBLIC LAW 117-263—DEC. 23, 2022 136 STAT. 2395

Public Law 117-263
117th Congress

An Act

To authorize appropriations for fiscal year 2023 for military activities of the Department of Defense, for military construction, and for defense activities of the Department of Energy, to prescribe military personnel strengths for such fiscal year, and for other purposes.

Dec. 23, 2022
[H.R. 7776]

James M. Inhofe National Defense Authorization Act for Fiscal Year 2023.
10 USC 101 note.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

(a) IN GENERAL.—This Act may be cited as the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023.”

(b) REFERENCES.—Any reference in this or any other Act to the “National Defense Authorization Act for Fiscal Year 2023” shall be deemed to be a reference to the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023.”

SEC. 2. ORGANIZATION OF ACT INTO DIVISIONS; TABLE OF CONTENTS.

(a) DIVISIONS.—This Act is organized into 11 divisions as follows:

- (1) Division A—Department of Defense Authorizations.
- (2) Division B—Military Construction Authorizations.
- (3) Division C—Department of Energy National Security Authorizations and Other Authorizations.
- (4) Division D—Funding Tables.
- (5) Division E—Non-Department of Defense Matters.
- (6) Division F—Intelligence Authorization Act for Fiscal Year 2023.
- (7) Division G—Homeland Security.
- (8) Division H—Water Resources.
- (9) Division I—Department of State Authorizations.
- (10) Division J—Oceans and Atmosphere.
- (11) Division K—Don Young Coast Guard Authorization Act of 2022.

(b) TABLE OF CONTENTS.—The table of contents for this Act is as follows:

Sec. 1. Short title.
Sec. 2. Organization of Act into divisions; table of contents.
Sec. 3. Congressional defense committees.
Sec. 4. Budgetary effects of this Act.
Sec. 5. Explanatory statement.

DIVISION A—DEPARTMENT OF DEFENSE AUTHORIZATIONS

TITLE I—PROCUREMENT

Subtitle A—Authorization of Appropriations

Sec. 101. Authorization of appropriations.

IDATA is a highly customizable capability for searching and exploring large collections of documents with real-time analysis features



IDATA

Raw Documents

Diversity and inclusion in the Department are moral imperatives — to ensure every member of the Total Force is treated with dignity and respect. Mission readiness and efficacy at strategic and tactical levels are increased when every member of the team is valued and afforded an opportunity to contribute to the effective defense of the Nation. As the international security environment continues to evolve, the United States confronts complex operational domains and unconventional tactics, all of which require a resilient and agile Joint Force. Demographic representation, accompanied by the Department's unwavering commitment to fairness, transparency, and leadership, is essential to achieving a mission-ready fighting force in the 21st Century. I am, and we all must be, dedicated to honoring the women and men who serve this country by ensuring them the same promotion and developmental opportunities regardless of race, color, religion, sex, or national origin. Diversity of thought and perspective ensures our strategic advantage on the global stage.

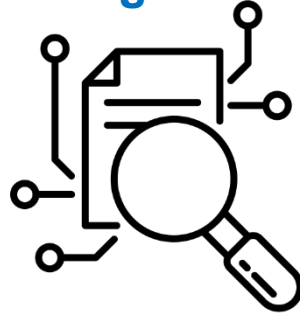
Diversity and inclusion in the Department are moral imperatives — to ensure every member of the Total Force is treated with dignity and respect. Mission readiness and efficacy at strategic and tactical levels are increased when every member of the team is valued and afforded an opportunity to contribute to the effective defense of the Nation. As the international security environment continues to evolve, the United States confronts complex operational domains and unconventional tactics, all of which require a resilient and agile Joint Force. Demographic representation, accompanied by the Department's unwavering commitment to fairness, transparency, and leadership, is essential to achieving a mission-ready fighting force in the 21st Century. I am, and we all must be, dedicated to honoring the women and men who serve this country by ensuring them the same promotion and developmental opportunities regardless of race, color, religion, sex, or national origin. Diversity of thought and perspective ensures our strategic advantage on the global stage.

On July 15, 2020, the DoD Board on Diversity and Inclusion (the "Board") embarked on the important task of identifying recommendations to improve the Department's diversity and inclusion and broaden equal opportunity for all members of the Armed Forces. Chaired by the Secretary of the Air Force, with the support of the Senior Enlisted Advisor to the Chairman of the Joint Chiefs of Staff and the Under Secretary of Defense for Personnel and Readiness (USD(P&R)), and Service members from each branch of the Military Services and the National Guard Bureau, the Board evaluated military diversity and inclusion policies, programs, and processes; reviewed industry best practices; and assessed pertinent data and reports. This effort culminated in the *Department of Defense Board on Diversity and Inclusion Report: Recommendations to Improve Racial and Ethnic Diversity and Inclusion in the U.S. Military*.

After reviewing the Board's 15 recommendations, I am pleased to see such a methodical evaluation leading to the development of such rigorous actions to address diversity and inclusion. I expect all leaders to take an aggressive approach to embed diversity and inclusion practices into the core of our military culture, building on the Department's decades-long tradition of leading the change in racial and ethnic equity. We must not accept... and must intentionally and...



Ingest



- **OCR**
- **Text extraction**
- **Document metadata**
- **Keyword discovery**
- ...



Indexed Documents

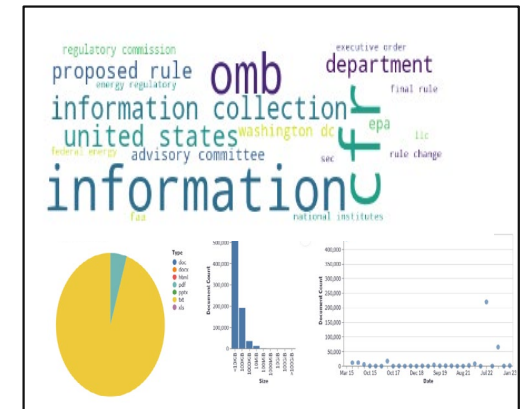
Diversity and inclusion in the Department are moral imperatives — to ensure every member of the Total Force is treated with dignity and respect. Mission readiness and efficacy at strategic and tactical levels are increased when every member of the team is valued and afforded an opportunity to contribute to the effective defense of the Nation. As the international security environment continues to evolve, the United States confronts complex operational domains and unconventional tactics, all of which require a resilient and agile Joint Force. Demographic representation, accompanied by the Department's unwavering commitment to fairness, transparency, and leadership, is essential to achieving a mission-ready fighting force in the 21st Century. I am, and we all must be, dedicated to honoring the women and men who serve this country by ensuring them the same promotion and developmental opportunities regardless of race, color, religion, sex, or national origin. Diversity of thought and perspective ensures our strategic advantage on the global stage.

On July 15, 2020, the DoD Board on Diversity and Inclusion (the "Board") embarked on the important task of identifying recommendations to improve the Department's diversity and inclusion and broaden equal opportunity for all members of the Armed Forces. Chaired by the Secretary of the Air Force, with the support of the Senior Enlisted Advisor to the Chairman of the Joint Chiefs of Staff and the Under Secretary of Defense for Personnel and Readiness (USD(P&R)), and Service members from each branch of the Military Services and the National Guard Bureau, the Board evaluated military diversity and inclusion policies, programs, and processes; reviewed industry best practices; and assessed pertinent data and reports. This effort culminated in the *Department of Defense Board on Diversity and Inclusion Report: Recommendations to Improve Racial and Ethnic Diversity and Inclusion in the U.S. Military*.

After reviewing the Board's 15 recommendations, I am pleased to see such a methodical evaluation leading to the development of such rigorous actions to address diversity and inclusion. I expect all leaders to take an aggressive approach to embed diversity and inclusion practices into the core of our military culture, building on the Department's decades-long tradition of leading the change in racial and ethnic equity. We must not accept... and must intentionally and...

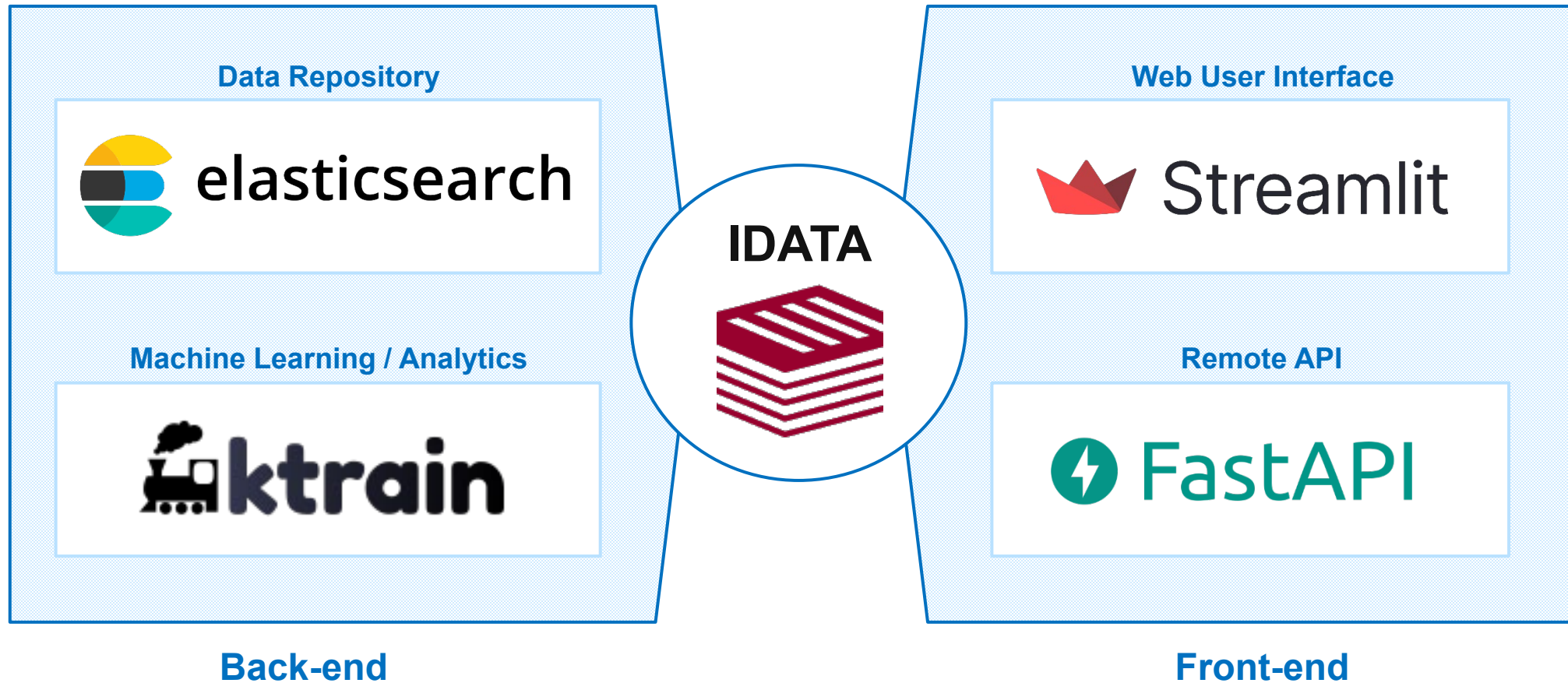


Real-time Analysis



- **Keyword searches**
- **Theme discovery**
- **Language translation**
- ...

IDATA is a python package powered by open-source tools



Multiple Interfaces for Users to Interact with Their Data

Web User Interface

IDATA: Search

Search Help

4337 results (0.40 seconds)

Show Word Cloud

You can select keywords of interest using the Filters/stop_keywords menu on the left.

national institute ikt 000000 gao
 research service
 nist final rule
 public law proposed rule dod
 po 000000 united states
 fiscal year available free
 national security congressional research

check all

No.	Filename	MD5	Preview	View
<input type="checkbox"/> 1.	Understanding AI Technology.pdf	4786a0c81c3d8741c028f000	[Preview]	Download (1.2M)

Highlights:
 Machine Learning is closely related to statistics and allows machines to learn from data... Machine Learning: At the key difference between a handwritten knowledge system and a Machine Learning system is in where it acquires its knowledge...
 Top Keywords: number of patents artificial intelligence data point handwritten knowledge intelligent technology joint ai machine learning training data understanding artificial
 Topic Cluster: data use reporting big report support analysis reported cases used
 Potential Offices: OIGM

Show File Metadata

No.	Filename	MD5	Preview	View
<input type="checkbox"/> 2.	...ing Behavior of Selected Appt	ab2205b404a10b120400004a10	[Preview]	Download (4.2M)

Highlights:
 Instructions can be conducted by demonstrating the machine simulation software. Additional instructions and demonstrations are obtained Learn, see disclaimer on page three... Figure 7 gives a brief overview of the learning module. Module 1 solicitation contains software for performing three-dimensional simulation of a 3-axis milling machine... The following are a few possible examples of application-based learning: One of the major concerns of machine tool metrology is the heat generated during machining... Learning about the physics of metrology could be included in training module on machine tool metrology and coordinate measuring machine (CMM) metrology... During the project-based activities, the students learn some of the basics of machining processes and machine control programming.
 Top Keywords: elementary elementary school learning module manufacturing technology math class milling machine mhm mhm project nia technology learning
 Topic Cluster: project project contribution award transportation development funds environmental planning funding
 Potential Offices: IGOT

Show File Metadata

Programmatic API

Adding Documents to the Search Engine

The `ingest` method can be used to index documents sequentially.

```
resp, success, fail, docs_success, docs_fail = es.ingest([document1.to_
print(success) # Tip: If you run this cell twice, it will show 0, sinc
```

Searching and Tagging

```
resp = es.search('"ktrain" AND "machine learning"',
filters=[{'extension': ['pdf']}],
project_ids=[project_id], size=1)
```

The returned response is the raw response as described in the [Elasticsearch documentation](#)

```
resp.keys()
dict_keys(['took', 'timed_out', '_shards', 'hits', 'aggregations'])
```

```
resp['hits']['hits'][0]['_source']['filepath']
['some/subfolder/2004.10703-6aa1799.pdf']
```

```
resp['hits']['hits'][0]['_source']['md5']
'6aa1799b36f35ee2684a91d977a6fcb4'
```

Documents can also be tagged.

```
md5 = resp['hits']['hits'][0]['_source']['md5']
es.add_tag('important', md5s=[md5])
```

Remote-Accessible API

search

POST `/api/search/` Search

GET `/api/search/{md5}` Document From Md5

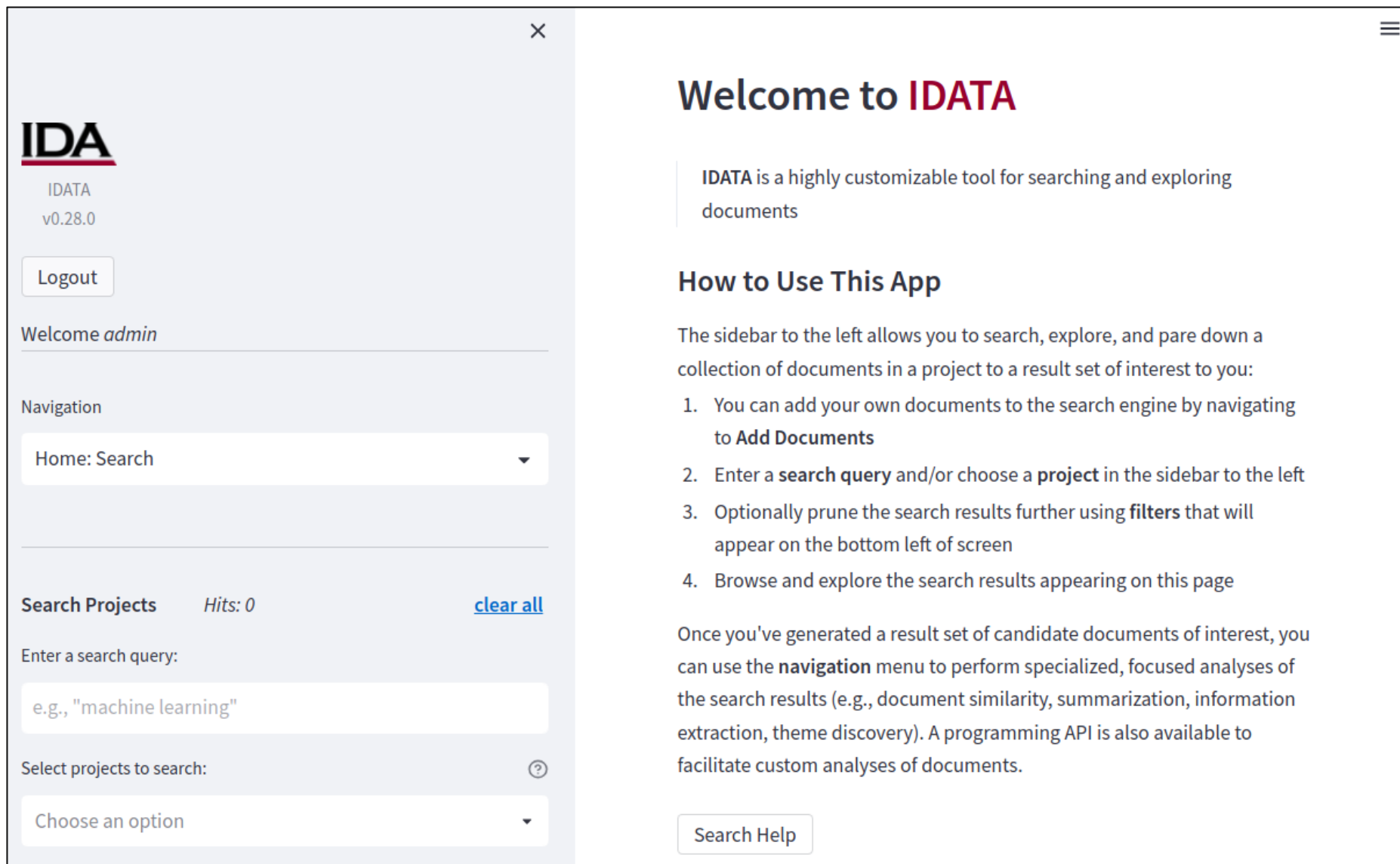
POST `/api/search/check` Check Query Syntax

POST `/api/collect/` Collect

```
app_url = "http://127.0.0.1:8000/"
# Set up the session and login
#sess = Session(app_url, token=token)
sess = Session(app_url)
# Check that you are logged in
sess.get("/user")
```

```
Type username for app: test
Type test password to login to app: .....
{'username': 'test', 'user_id': 3}
```

Streamlit-Based User Interface



The screenshot displays a Streamlit-based user interface for IDATA. On the left is a sidebar with a light gray background, containing the IDATA logo, version information (v0.28.0), a Logout button, a Welcome message for 'admin', a Navigation dropdown menu currently set to 'Home: Search', a search status section showing 'Search Projects' with 'Hits: 0' and a 'clear all' link, a search query input field with the placeholder text 'e.g., "machine learning"', and a project selection dropdown menu with the placeholder text 'Choose an option'. The main content area on the right has a white background and features a 'Welcome to IDATA' heading, a brief description of the tool, a 'How to Use This App' section with a numbered list of four steps, a paragraph explaining the navigation menu, and a 'Search Help' button at the bottom.

IDA
IDATA
v0.28.0

Logout

Welcome *admin*

Navigation

Home: Search

Search Projects Hits: 0 [clear all](#)

Enter a search query:

e.g., "machine learning"

Select projects to search: ?

Choose an option

Welcome to IDATA

IDATA is a highly customizable tool for searching and exploring documents

How to Use This App

The sidebar to the left allows you to search, explore, and pare down a collection of documents in a project to a result set of interest to you:

1. You can add your own documents to the search engine by navigating to **Add Documents**
2. Enter a **search query** and/or choose a **project** in the sidebar to the left
3. Optionally prune the search results further using **filters** that will appear on the bottom left of screen
4. Browse and explore the search results appearing on this page

Once you've generated a result set of candidate documents of interest, you can use the **navigation** menu to perform specialized, focused analyses of the search results (e.g., document similarity, summarization, information extraction, theme discovery). A programming API is also available to facilitate custom analyses of documents.

Search Help

Standard Search for “Machine Learning”

Key file metadata

No.	Filename	MD5	Preview	View
<input type="checkbox"/> 1.	2004.10703.pdf	6aa1799b36f35ee2684a91d977a6fcb4	[hover]	Download File (273.7 KiB)

demo_project (1) : demo_files_extracted/others/2004.10703.pdf

Highlights: *← Search highlight*

ktrain: A Low-Code Library for Augmented **Machine Learning** ktrain: A Low-Code Library for Augmented **Machine Learning** Arun S....Moreover, each of the three steps above can be accomplished in as little as three or four lines of code, which we refer to as “low-code” **machine learning** . ktrain can be used with any **machine learning** model...Non-Supervised ML Tasks All the examples covered thus far involve supervised **machine learning** **Learning** ence to build sophisticated **machine learning** models with minimal coding....Scikit-learn: **Machine learning** in Python. Journal of **Machine Learning** Research, 12:2825–2830, 2011. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”

Tag File *← User labelling*

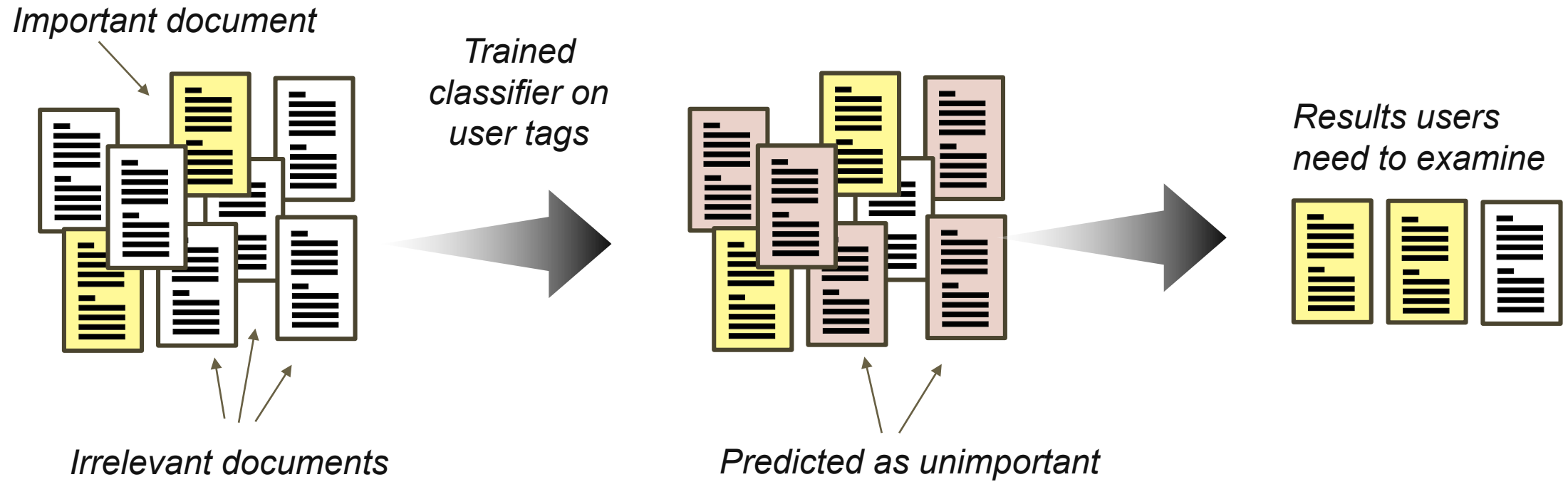
Show File Metadata

filetype: pdf|report

[View Inline](#) *← Easy access to file*

← User view marker

Machine Learning Pipeline to Improve Search Results



Low Code Text Analytics from ktrain

Example: Building a Chinese-Language Sentiment Analyzer Using BERT (a Deep Learning Model by Google)



```
import ktrain
from ktrain import text as txt

# STEP 1: Load and preprocess data
trn, val, preproc = txt.texts_from_folder('data/ChnSentiCorp',
                                          maxlen=75,
                                          preprocess_mode='bert')

# STEP 2: Load model
model = txt.text_classifier('bert', trn, preproc=preproc)
learner = ktrain.get_learner(model, train_data=trn, val_data=val, batch_size=32)

# STEP 3: Estimate learning rate (a neural network hyperparameter)
learner.lr_find(show_plot=True)

# STEP 4: Train model using 1cycle learning rate schedule for 4 epochs
learner.fit_onecycle(2e-5, 4)
```

ktrain is available on GitHub: <https://github.com/amaiya/ktrain>

Features in Users' Hands

For Users

- Standard search
 - Lucene query syntax
 - Field filtering
- Document tagging
- Key phrase extraction
- Dynamic faceting
- Export results to spreadsheets
- View hits in documents
- Find similar documents
- Paragraph-level comparison of similar documents
- Word clouds
- Batch search
- Corpus statistics
- Theme discovery
- Named entity extraction
- Sentiment analysis
- Document summarization
- Question-answering
- Folder browser
- Image browser
- Language translation
- Report bugs and feedback

For Admins

- Customizable document-processing and user interface
- Ability to auto-tag documents with custom ML classifiers
- Custom report generation
- CLI management scripts
 - Bulk ingest of files
 - Export data, update tags ...
- Automatic document deduplication
- Manage users/projects
- Add/delete documents
- Automatic user activity logging

Example Use Cases

NSF Grants

 Award Abstract # 1912813
Particle Physics at CMS in the High Luminosity Era

NSF Org:	PHY Division Of Physics
Recipient:	CORNELL UNIVERSITY
Initial Amendment Date:	July 9, 2019
Latest Amendment Date:	August 4, 2021
Award Number:	1912813
Award Instrument:	Continuing Grant
Program Manager:	James Shank jshank@nsf.gov (703)292-4516 PHY Division Of Physics MPS Direct For Mathematical & Physical Scien
Start Date:	August 1, 2019
End Date:	July 31, 2023 (Estimated)
Total Intended Award Amount:	\$3,075,000.00

- ✓ Search
- ✓ Discover
- Compare

NDAA

PUBLIC LAW 117-263—DEC. 23, 2022 136 STAT. 2395

Public Law 117-263
117th Congress

An Act

To authorize appropriations for fiscal year 2023 for military activities of the Department of Defense, for military construction, and for defense activities of the Department of Energy, to prescribe military personnel strengths for such fiscal year, and for other purposes. Dec. 23, 2022
[H.R. 7776]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.
James M. Inhofe National Defense Authorization Act for Fiscal Year 2023. 10 USC 101 note.

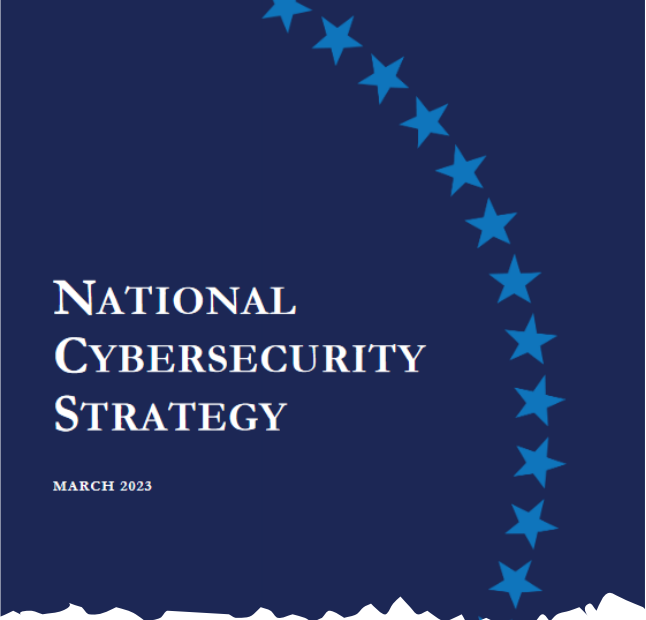
(a) IN GENERAL.—This Act may be cited as the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023”.
(b) REFERENCES.—Any reference in this or any other Act to the “National Defense Authorization Act for Fiscal Year 2023” shall be deemed to be a reference to the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023”.

SEC. 2. ORGANIZATION OF ACT INTO DIVISIONS; TABLE OF CONTENTS.
(a) DIVISIONS.—This Act is organized into 11 divisions as follows:

- (1) Division A—Department of Defense Authorizations.
- (2) Division B—Military Construction Authorizations.
- (3) Division C—Department of Energy National Security Authorizations and Other Authorizations.


- Search
- ✓ Discover
- ✓ Compare

National Strategies



- Search
- Discover
- ✓ Compare

Examining NSF Grants

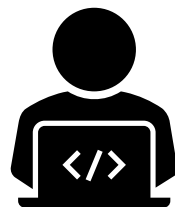
 Award Abstract # 1912813
Particle Physics at CMS in the High Luminosity Era

NSF Org:	PHY Division Of Physics
Recipient:	CORNELL UNIVERSITY
Initial Amendment Date:	July 9, 2019
Latest Amendment Date:	August 4, 2021
Award Number:	1912813
Award Instrument:	Continuing Grant
Program Manager:	James Shank jshank@nsf.gov (703)292-4516 PHY Division Of Physics MPS Direct For Mathematical & Physical Scien
Start Date:	August 1, 2019
End Date:	July 31, 2023 (Estimated)
Total Intended Award Amount:	\$3,075,000.00



IDATA

Standard Ingest



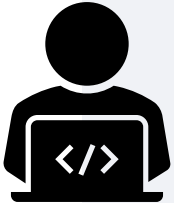
User Pre-Processing
of XML Files

```
<?xml version="1.0" encoding="UTF-8"?>
<rootTag>
<Award>
<AwardTitle>![CDATA[CI CoE: Demo Pilot: Advancing Research Computing
<AGENCY>NSF</AGENCY>
<AwardEffectiveDate>05/15/2021</AwardEffectiveDate>
<AwardExpirationDate>04/30/2024</AwardExpirationDate>
<AwardTotalIntnAmount>1497243.00</AwardTotalIntnAmount>
<AwardAmount>1497243</AwardAmount>
<AwardInstrument>
<Value>Standard Grant</Value>
</AwardInstrument>
<Organization>
<Code>05090000</Code>
<Directorate>
<Abbreviation>CSE</Abbreviation>
<LongName>Direct For Computer & Info Scie & Enginr</LongName>
</Directorate>
```

```
AWARD_DICT = {
    'text' = TITLE + ABSTRACT,
    'investigator' = INVESTIGATOR,
    'institution' = INSTITUTION,
    ... }
```

Flexible API Allows for Creating Custom Analyses

User Input



```
DYNAMIC_FIELDS = ['investigator',  
                  'institution',  
                  'effective_date',  
                  'expiration_date',  
                  'award_amount']
```

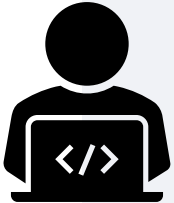
Results

Collaborative Research: An Evidence-Based Approach Towards Increasing Female Participation in STEM Entrepreneurship This University of Wisconsin-Madison as the lead institution and the institution, to explore the entrepreneurial proclivity of undergrad

Investigator: Jonathan T Eckhardt
Institution: University of Wisconsin-Madison
Effective Date: 2021-08-01
Expiration Date: 2024-07-31
Award Amount: \$983,486.00

Search and Filter on User-Defined Dynamic Fields

User Input



```
DYNAMIC_FIELDS = ['investigator',  
                  'institution',  
                  'effective_date',  
                  'expiration_date',  
                  'award_amount']
```

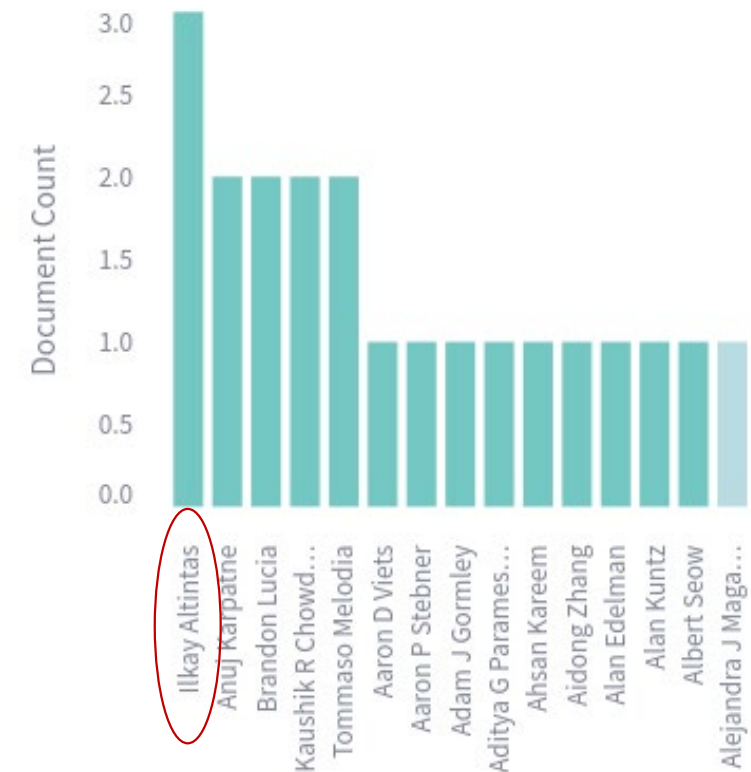
Enter a search query:



```
"machine learning" AND award_amount:[1000000 TO *]
```

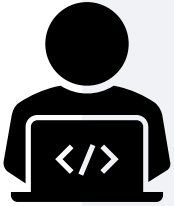
Results

Investigator



Discover Clusters in the Filtered Results

User Input



```
DYNAMIC_FIELDS = ['investigator',  
                  'institution',  
                  'effective_date',  
                  'expiration_date',  
                  'award_amount']
```

Enter a search query:



```
"machine learning" AND award_amount:[1000000 TO *]
```

There are a total of **133** documents in your search results.
Press the **Discover Themes** button below to begin.

Discover Themes



Results

cybersecurity systems program workforce government activities national security service required (7)



CyberCorps Scholarship for Service: Cybersecurity Workforce Preparation in the Age of Artificial Intelligence This project aims to address the growing need for a highly-skilled national cybersecurity workforce capable of responding to rapidly evolving challenges in the age of artificial intelligence (AI). With funding from the NSF CyberCorps(R) Scholarship for Service (SFS) program, Georgia State...

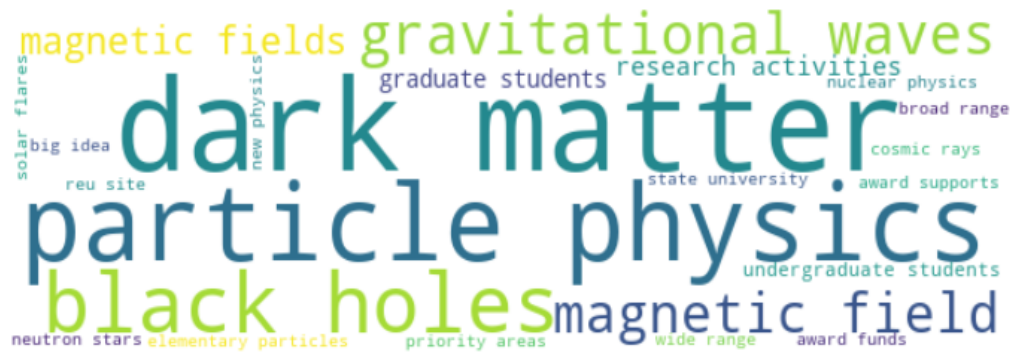
Investigator: Balasubramaniam Ramesh | Daniel Takabi | Dmitry Zhdanov | Zhipeng Cai
Institution: Georgia State University Research Foundation, Inc.
Effective Date: 2022-02-01
Expiration Date: 2027-01-31
Award Amount: \$1,119,210.00

Discover Clusters in Different Types of Documents

NSF Awards

Select a discovered theme

physics matter magnetic dark particles particle field universe nuclear experiment (415) | ▾



Dark matter is an as-yet-identified form of **matter**....The identity and properties of the **dark matter** nevertheless remain unknown....The nature of the **dark matter** and the physical laws governing the interactions of **dark matter** can be investigated in both earth-bound laboratories and indirectly through astronomical observations.... **Dark matter** lies beyond our current understanding of **matter** and forces in nature....In the first component, Professor Zentner will simulate the effects of exotic **dark matter** candidates --- particularly asymmetric **dark matter**, self-interacting **dark matter**, and strongly-interacting **dark**

Investigator: Andrew R Zentner

Institution: University of Pittsburgh

Effective Date: 2021-09-01

Expiration Date: 2024-08-31

Award Amount: \$184,769.00

NDAAs Sections

Select a discovered theme

health global pandemic preparedness prevention fund international countries behavioral strategy ... | ▾

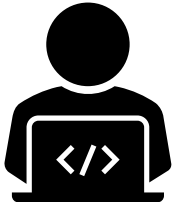


SEC. 5562. INTERNATIONAL PANDEMIC PREVENTION AND PREPAREDNESS. SEC. 5562. INTERNATIONAL PANDEMIC PREVENTION AND PREPAREDNESS. (a) United States International Activities To Advance Global Health Security and Diplomacy Strategy and Report.-- (1) In general.--The President shall develop, update, maintain, and advance a comprehensive strategy for improving United States global ...

Title: SEC. 5562. INTERNATIONAL PANDEMIC PREVENTION AND PREPAREDNESS.

Tags: TITLE_LV-FOREIGN_AFFAIRS_MATTERS--Subtitle_D-International_Pandemic_Preparedness

NDAAs Processing



```
DYNAMIC_FIELDS = ['title.keyword']
```

PUBLIC LAW 117-263—DEC. 23, 2022 136 STAT. 2395

Public Law 117-263
117th Congress
An Act

To authorize appropriations for fiscal year 2023 for military activities of the Department of Defense, for military construction, and for defense activities of the Department of Energy, to prescribe military personnel strengths for such fiscal year, and for other purposes.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

(a) IN GENERAL.—This Act may be cited as the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023”.

(b) REFERENCES.—Any reference in this or any other Act to the “National Defense Authorization Act for Fiscal Year 2023” shall be deemed to be a reference to the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023”.

SEC. 2. ORGANIZATION OF ACT INTO DIVISIONS; TABLE OF CONTENTS.

(a) DIVISIONS.—This Act is organized into 11 divisions as follows:

- (1) Division A—Department of Defense Authorizations.
- (2) Division B—Military Construction Authorizations.
- (3) Division C—Department of Energy National Security Authorizations and Other Authorizations.

James M. Inhofe National Defense Authorization Act for Fiscal Year 2023.
10 USC 101 note.

Results

NDAAs(1) : DIVISION A--DEPARTMENT OF DEFENSE AUTHORIZATIONS/TITLE I--PROCUREMENT/Subtitle A--Authorization of Appropriations/101.txt

Preview:

SEC. 101. AUTHORIZATION OF APPROPRIATIONS. SEC. 101. AUTHORIZATION OF APPROPRIATIONS. Funds are hereby authorized to be appropriated for fiscal year 2023 for procurement for the Army, the Navy and the Marine Corps, the Air Force and the Space Force, and Defense-wide activities, as specified in the funding table in section 4101.

Title: SEC. 101. AUTHORIZATION OF APPROPRIATIONS.

Tags: TITLE_I--PROCUREMENT--Subtitle_A--Authorization_of_Appropriations

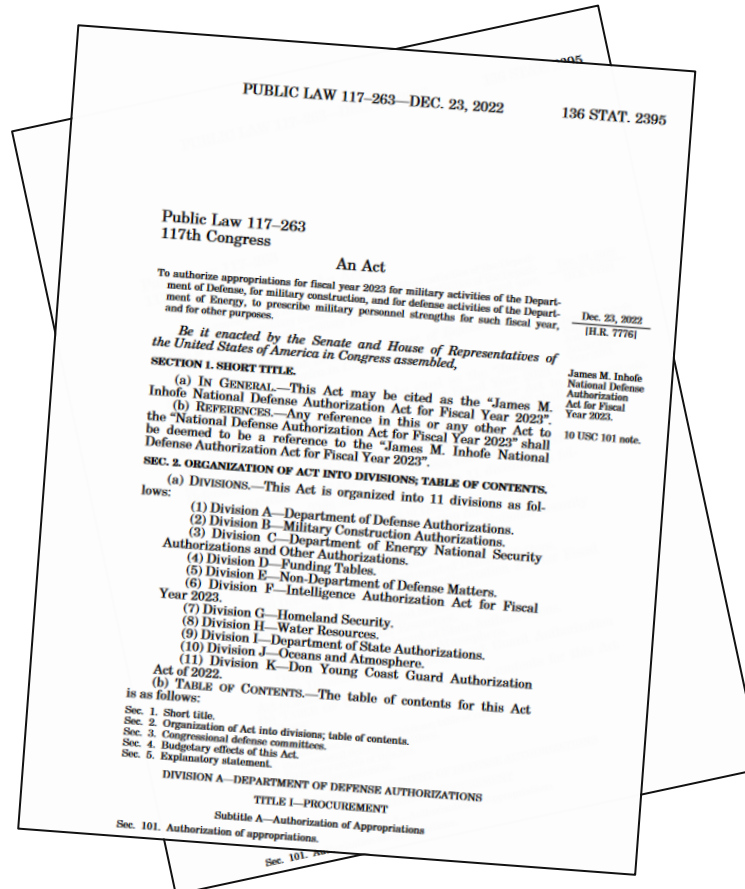
Compare Different Versions of NDAA

Search Result #1: NDAA/NDAA_2023_House.txt ▾

Similar File #1: NDAA/NDAA_2023_Senate.txt ▾

Press the **Compare** button after selecting the files to begin.

Compare



Compare Different Versions of NDAA

Search Result #1: NDAA/NDAA_2023_House.txt ▼

Similar File #1: NDAA/NDAA_2023_Senate.txt ▼

Press the **Compare** button after selecting the files to begin.

Compare

Similar Paragraphs Identical Paragraphs Unmatched Paragraphs Side-by-Side of Text

The Armed Forces are authorized strengths for active duty personnel as of September 30, 2023, as follows: (1) The Army, 473,000. (2) The Navy, **348,220**. (3) The Marine Corps, 177,000. (4) The Air Force, **323,400**. (5) The Space Force, 8,600.

The Armed Forces are authorized strengths for active duty personnel as of September 30, 2023, as follows: (1) The Army, 473,000. (2) The Navy, **354,000**. (3) The Marine Corps, 177,000. (4) The Air Force, **325,344**. (5) The Space Force, 8,600.

Compare Different Versions of NDAA

Search Result #1: NDAA/NDAA_2023_House.txt ▼

Similar File #1: NDAA/NDAA_2023_Senate.txt ▼

Press the **Compare** button after selecting the files to begin.

Compare

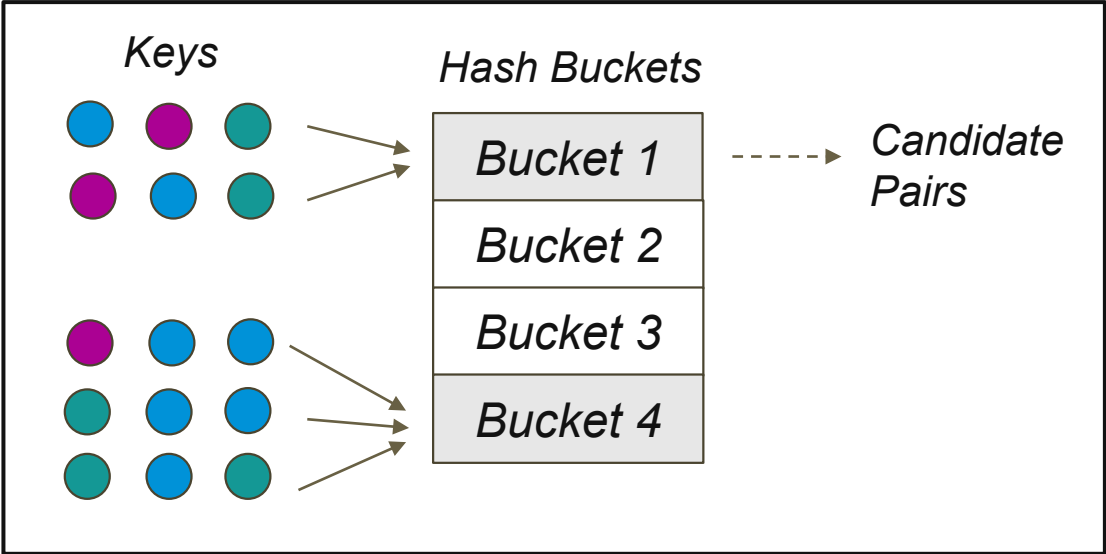
Similar Paragraphs Identical Paragraphs Unmatched Paragraphs Side-by-Side of Text

(a) Establishment.--The Secretary of Defense shall establish a **voluntary** pilot program to promote the safe storage of personally owned firearms. (b) Elements.--Under the pilot program under subsection (a), the Secretary of Defense shall furnish to members of the Armed Forces **described in subsection (c) secure gun storage or safety** devices for the purpose of securing personally owned firearms when not in use (including by directly providing, subsidizing, or

(a) Establishment.--The Secretary of Defense shall establish a pilot program to promote the safe storage of personally owned firearms. (b) Elements.--Under the pilot program under subsection (a), the Secretary of Defense shall furnish to members of the Armed Forces **who are participating in the pilot program at military installations selected under subsection (e) locking devices or firearm safes, or both,** for the purpose of securing personally owned firearms when

Document Comparison

Locality Sensitive Hashing Algorithm

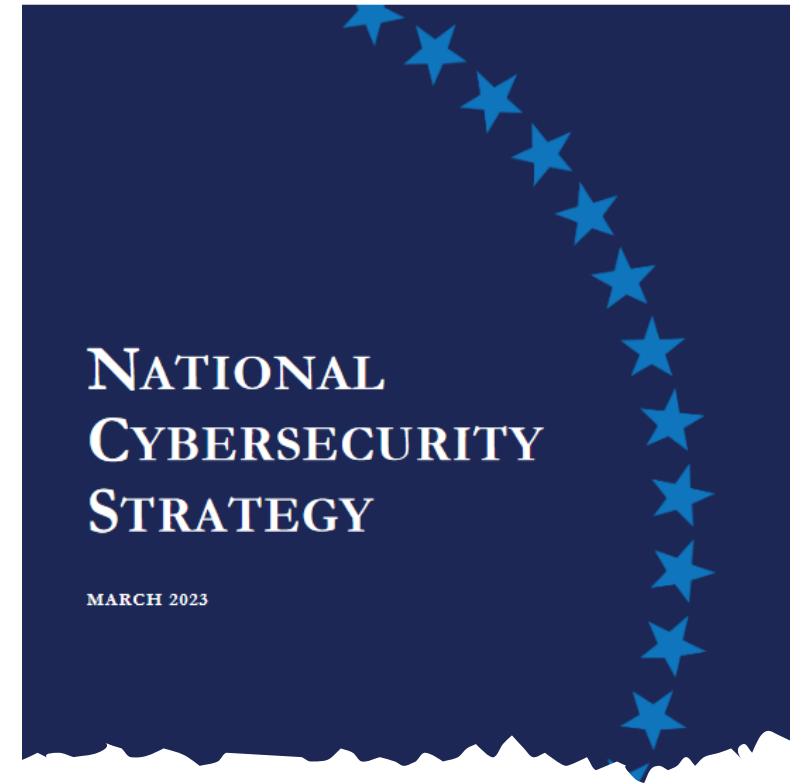


Gestalt Pattern Matching



- ✓ Near-duplicate documents
- ✗ Documents with similar content

Comparison of Cybersecurity Strategies



Semantic Comparison of Cybersecurity Strategies



```
NCS_DICT = { "Sec 1.1": Text, "Sec 1.2": Text, ... }
```

```
EXEC_ORDER_DICT = { "Sec 01": Text, ... }
```

```
Results = SemanticMatcher.execute(EXEC_ORDER_DICT, NCS_DICT)
```

Executive Order 14028

Sec. 4. Enhancing Software Supply Chain Security

Sec. 4. Enhancing Software Supply Chain Security. (a) The security of software used by the Federal Government is vital to the Federal Government's ability to perform its critical functions. The development of commercial software often lacks transparency, sufficient focus on the ability of the software to resist attack, and adequate controls to prevent tampering by malicious actors. There is a pressing need to implement more rigorous and predictable mechanisms for ensuring that products function securely, and as intended. The security and integrity of "critical software"—software that performs functions critical to trust (such as affording or requiring elevated system privileges or direct access to networking and computing resources)—is a particular concern. Accordingly, the Federal Government must take action to rapidly improve the security and integrity of the software supply chain, with a priority on addressing critical software.

National Cybersecurity Strategy

Strategic Objective 3.3: Shift Liability for Insecure Software Products and Services

Score 70%

Markets impose inadequate costs on—and often reward—those entities that introduce vulnerable products or services into our digital ecosystem. Too many vendors ignore best practices for secure development, ship products with insecure default configurations or known vulnerabilities, and integrate third-party software of unvetted or unknown provenance. Software makers are able to

Strategic Objective 3.5: Leverage Federal Procurement to Improve Accountability

Score 61%

Contracting requirements for vendors that sell to the Federal Government have been an effective tool for improving cybersecurity. EO 14028, "Improving the Nation's Cybersecurity," expands upon this approach, ensuring that contract requirements for cybersecurity are strengthened and standardized across Federal agencies. Continuing to pilot new concepts for setting, enforcing, and

Example Use Cases

NSF Grants

 Award Abstract # 1912813 Particle Physics at CMS in the High Luminosity Era	
NSF Org:	PHY Division Of Physics
Recipient:	CORNELL UNIVERSITY
Initial Amendment Date:	July 9, 2019
Latest Amendment Date:	August 4, 2021
Award Number:	1912813
Award Instrument:	Continuing Grant
Program Manager:	James Shank jshank@nsf.gov (703)292-4516 PHY Division Of Physics MPS Direct For Mathematical & Physical Scien
Start Date:	August 1, 2019
End Date:	July 31, 2023 (Estimated)
Total Intended Award Amount:	\$3,075,000.00

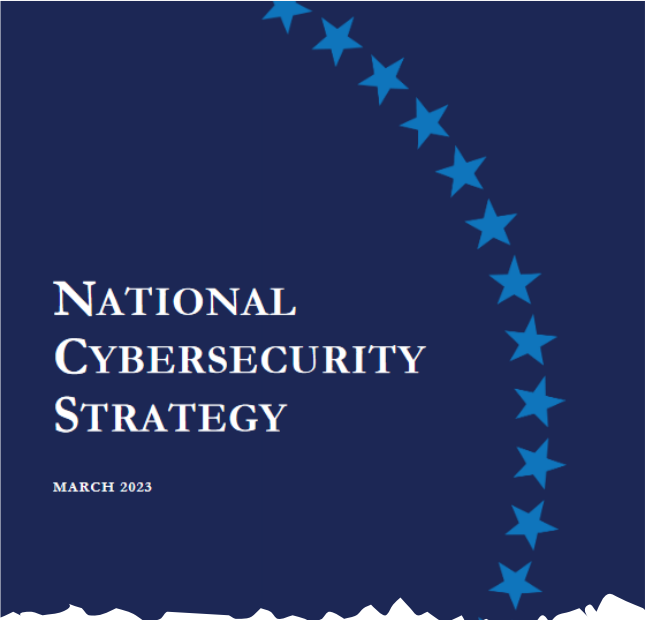
- ✓ Search
- ✓ Discover
- Compare

NDAA

PUBLIC LAW 117-263—DEC. 23, 2022 136 STAT. 2395	
Public Law 117-263 117th Congress An Act	
To authorize appropriations for fiscal year 2023 for military activities of the Department of Defense, for military construction, and for defense activities of the Department of Energy, to prescribe military personnel strengths for such fiscal year, and for other purposes.	
Dec. 23, 2022 [H.R. 7776]	James M. Inhofe National Defense Authorization Act for Fiscal Year 2023.
<i>Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,</i>	
SECTION 1. SHORT TITLE. (a) IN GENERAL.—This Act may be cited as the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023”. (b) REFERENCES.—Any reference in this or any other Act to the “National Defense Authorization Act for Fiscal Year 2023” shall be deemed to be a reference to the “James M. Inhofe National Defense Authorization Act for Fiscal Year 2023”.	
SEC. 2. ORGANIZATION OF ACT INTO DIVISIONS; TABLE OF CONTENTS. (a) DIVISIONS.—This Act is organized into 11 divisions as follows:	
(1) Division A—Department of Defense Authorizations. (2) Division B—Military Construction Authorizations. (3) Division C—Department of Energy National Security Authorizations and Other Authorizations.	

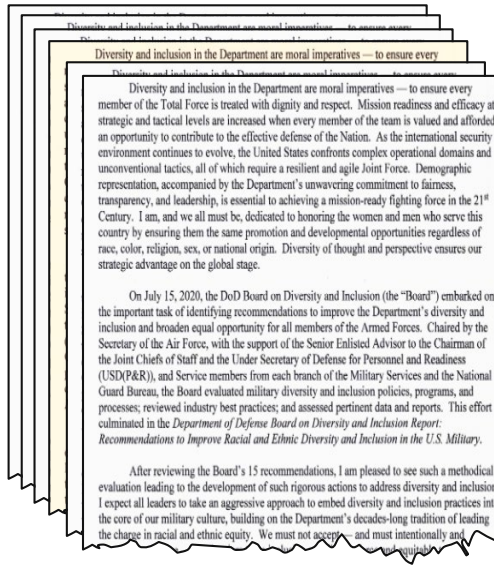
- Search
- ✓ Discover
- ✓ Compare

National Strategies



- Search
- Discover
- ✓ Compare

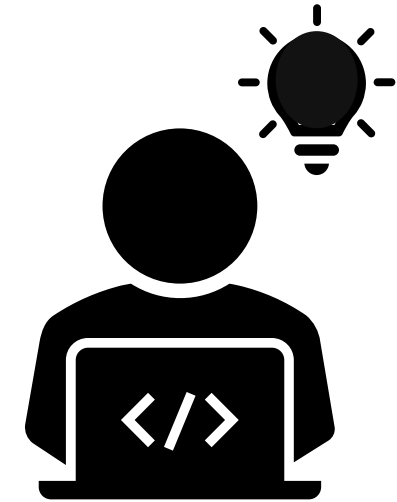
Customizable and Real-Time Analysis Features Enable Users to Tailor Analyses to Their Needs and Generate Better Insights



Common Tasks for Text Analytics at IDA



Customizable Software Can Be Tailored to the User's Needs



Real-time Analysis Features Enable Greater Insight into Data

Questions?

References

- A. S. Maiya et. al. “Exploratory Analysis of Highly Heterogeneous Document Collections” (2013) [<https://arxiv.org/abs/1308.2359>]
- A. S. Maiya. “ktrain: A Low-Code Library for Augmented Machine Learning” (2022) [<https://arxiv.org/abs/2004.10703>]

Our Open-Source Packages:

- ktrain: <https://github.com/amaiya/ktrain>
- OnPrem.LLM: <https://github.com/amaiya/onprem>

Other Questions?

Contact **Margaret Zientek** (mzientek@ida.org) or **Arun Maiya** (amaiya@ida.org)

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188		
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YY) 00-10-23		2. REPORT TYPE Final		3. DATES COVERED (From – To)	
4. TITLE AND SUBTITLE Multi-Use Text Analytics			5a. CONTRACT NUMBER HQ0034-19-D-0001		
			5b. GRANT NUMBER		
			5c. PROGRAM ELEMENT NUMBERS		
6. AUTHOR(S) Margaret E. Zientek, Dale Visser, Arun S. Maiya			5d. PROJECT NUMBER AI-5-5214		
			5e. TASK NUMBER		
			5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSES Institute for Defense Analyses 730 East Glebe Road Alexandria, VA 22305			8. PERFORMING ORGANIZATION REPORT NUMBER IDA Product 3000530		
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Jessica M. Appler, Director, Maintaining Technology Advantage, Office of the Under Secretary of Defense for Research and Engineering 3030 Defense Pentagon Washington, DC 20301-3030			10. SPONSOR'S / MONITOR'S ACRONYM MTA, OUSD (R&E)		
			11. SPONSOR'S / MONITOR'S REPORT NUMBER(S)		
12. DISTRIBUTION / AVAILABILITY STATEMENT Distribution Statement A. Approved for public release: distribution is unlimited.					
13. SUPPLEMENTARY NOTES Project Leader: Arun S. Maiya					
14. ABSTRACT How does one effectively explore and extract insights from large collections of highly heterogeneous documents across a diverse set of use cases? The Institute for Defense Analyses (IDA), a nonprofit corporation operating three federally funded research and development centers, helps government sponsors answer challenging questions that lie at the intersection of national security, science, technology, and policy. These questions often involve making sense of large numbers of documents and files. Moreover, factors important to one question may not be important for others. Most existing solutions to document analysis and review employ a one-size-fits-all approach that cannot be easily adapted to different use cases. IDA has developed in-house text analytics capabilities to facilitate search, exploration, and analysis of large collections of files in ways that enable greater flexibility in answering research questions. An extensible set of machine learning and natural language processing methods can be targeted and applied in real time to subsets of documents of interest through a no-code, point-and-click, web-based user interface, which allows for more focused insight extractions. In addition, programmatic APIs are locally and remotely accessible, which allow users to customize document processing and analyses for a wide range of different use cases without requiring source code modifications. For instance, researchers have previously used the API to ingest documents from diverse data sources and train custom machine learning models to auto-label documents to make them findable or filterable. This presentation provides a high-level overview of the work with illustrative examples.					
15. SUBJECT TERMS text analytics, NLP, machine learning					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Unlimited	18. NUMBER OF PAGES 29	19a. NAME OF RESPONSIBLE PERSON Jessica M. Appler, Director, MTA
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (Include Area Code) 571-256-7008

