CIRI | CRITICAL INFRASTRUCTURE RESILIENCE INSTITUTE

A DEPARTMENT OF HOMELAND SECURITY CENTER OF EXCELLENCE

Ana Robinson
Jarod Costello
Dipakkumar (Dipak) P Pravin, PhD (UNT)
CIRI Mentors:
Jose Alejandra Medina Cruz,
Andrea Whitesell

Evaluating Ontologies for Cybersecurity Workforce Development Applications

Agenda

- Context
- The Problem
- The Tool: CyberTalent Bridge (CTB)
- Frameworks & Standards
- Methods
- Results
- Current Conclusion
- Future Work

Context

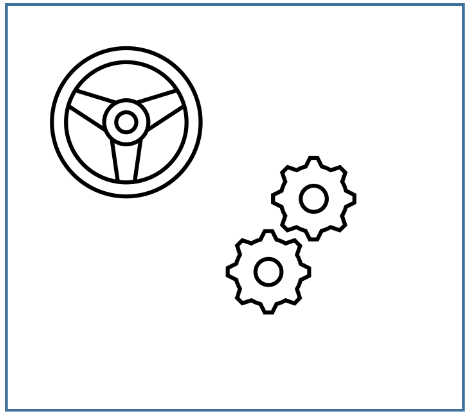
- **Cybersecurity is a complex discipline** requiring knowledge and skills in areas such as computer science, psychology, social science, project management and law to name a few major ones.
- According to one statistics [1], in **USA alone there are close to 500,000 open cybersecurity jobs compared** to an existing workforce of about 1 millions practitioner.
- According to another report [2], "*Cybersecurity Talent Crunch To Create **3.5 Million Unfilled Jobs <u>Globally</u> By 2021***"
- This is a major **challenge for the entire workforce supply chain** starting from training/teaching to recruiting to hiring and then executing projects.

Context (continued)

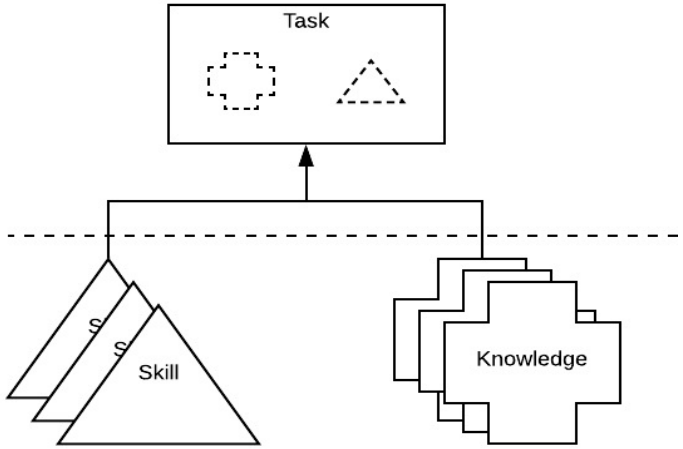
- According to NICE Framework* Executive Summary [3], *"…as information and technology, including many evolving types of operational technology, grow increasingly complex and interconnected <u>it can be difficult to clearly describe the work that is being performed</u> or that we desire to accomplish…"*
- So yes, we need an easy (but more importantly) a common, well-understood way of describing the "work".

*the National Initiative for Cybersecurity Education (NICE) Workforce Framework for Cybersecurity (NICE Framework)

The Problem:



Match



NIST 800-53 Controls Descriptions

NICE Frameworks: **T**ask, **K**nowledge and **S**kills Statements

Caveat: NICE Framework does **not** bundle a given Task with any Knowledge or Skill statement.

The Problem: (an example)

800-53 Controls Description

AC-2: Account Management

The organization; Identifies and selects the following types of information system accounts to support organizational missions/business functions:

- Assigns account managers for information system accounts;
- Establishes conditions for group & role membership;
- Specifies authorized users of the information system, group and role membership, and access authorizations (i.e., privileges) and other attributes (as required) for each account; ……(cont'd)….

NICE Frameworks: **T**ask, **K**nowledge and **S**kills Statements

T0109: Identify and prioritize essential system functions or sub-systems required to support essential capabilities or business functions for restoration or recovery after a system failure or during a system recovery event based on overall system requirements for continuity and availability.

T0830: Track status of information requests, including those processed as collection requests and production requirements, using established procedures. ……(cont'd)….

The Tool: CyberTalent Bridge (CTB)

- **A Quick Demo:** https://cybertalentbridge.com/
- CIRI [6] has developed CTB.
- *"The CyberTalent Bridge* ***empowers businesses to map the knowledge, skills, and abilities (KSAs) required to perform cybersecurity tasks*** *with the cybersecurity KSAs of the personnel within their own workforce as well as candidates for new-hire positions."* [7]
- CTB uses ontologies from NIST 800-53 and NICE Workforce Framework.

Frameworks & Standards

- NIST 800-53: provides a list of *controls* that support the secure operation of a (federal) information system.
- NIST 800-181: The Workforce Framework for Cybersecurity (NICE Framework), provides a set of building blocks for describing the Tasks(T), Knowledge(K), and Skills(S) that are needed to perform cybersecurity work performed by individuals and teams.

*National Institute for Standards and Technology (NIST)

Frameworks & Standards: There are many!

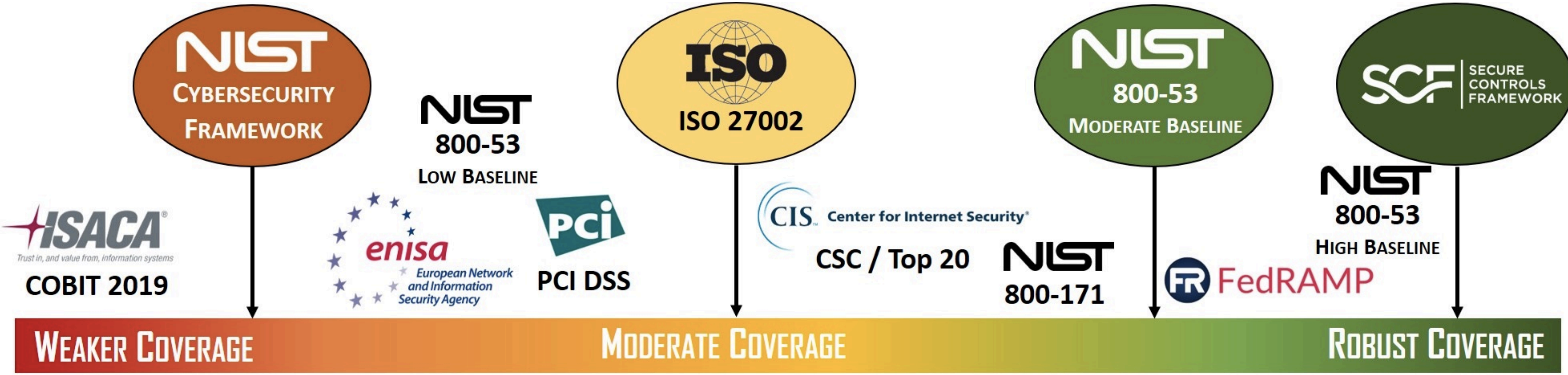
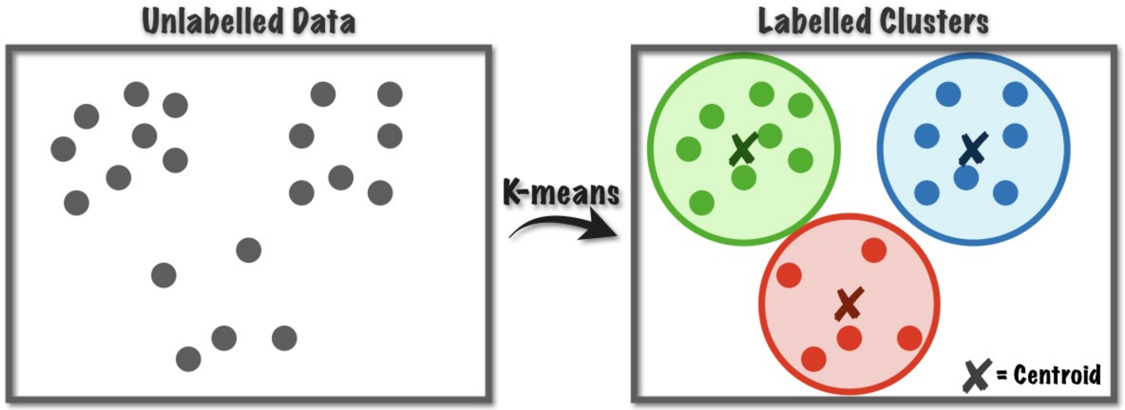


Image Credits: https://www.complianceforge.com/faq/nist-800-53-vs-iso-27002-vs-nist-csf.html

Methods:

- **The mapping process should be automated** to accommodate for revisions in the standards to reflect new cybersecurity areas as well as different standards that currently exist.
- **Natural Language Processing** techniques are needed as the base standards are developed by the subject matter experts (SMEs).
    - **Supervised Machine Learning** would be ideal but the appropriate training data does not exist currently.
    - **Unsupervised Learning is appropriate** since the natural language inputs carry inherent meanings that can be used for matching:
        - **K-means Clustering** is a good starting point to match Controls to TKS statements**.**

K-Means Clustering:



- We used Python, **Scikit-Learn and Pandas** frameworks [8]
- Experimented and Evaluated **several cluster sizes**
- **Results were surprising and not useful**.
  - Different cluster sizes produced different bundles with the same control.
  - Experiments with different controls had the same unstable results.
  - Skewed(imbalanced) nature of data input is the most likely the cause[9].

Image Credits to: https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c

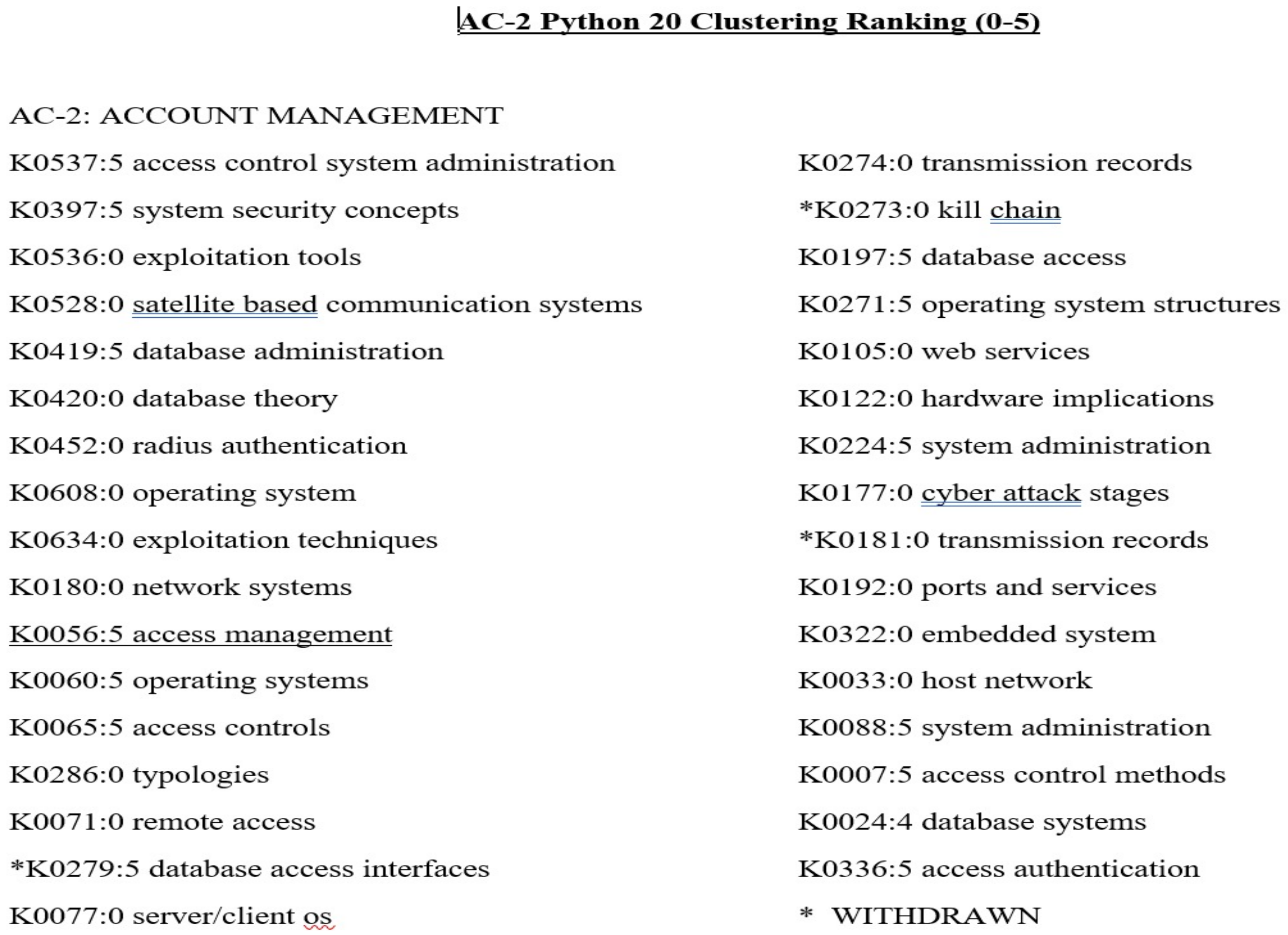
K0537: Knowledge of system administration concepts for the Unix/Linux and Windows operating systems (e.g., process management, directory structure, installed applications, Access Controls)

K0397 Knowledge of security concepts in operating systems (e.g., Linux, Unix.)

K0536 Knowledge of structure, approach, and strategy of exploitation tools (e.g., sniffers, keyloggers) and techniques (e.g., gaining backdoor access, collecting/exfiltrating data, conducting vulnerability analysis of other systems in the network).

K0528 Knowledge of satellite-based communication systems.

K0419 Knowledge of database administration and maintenance.



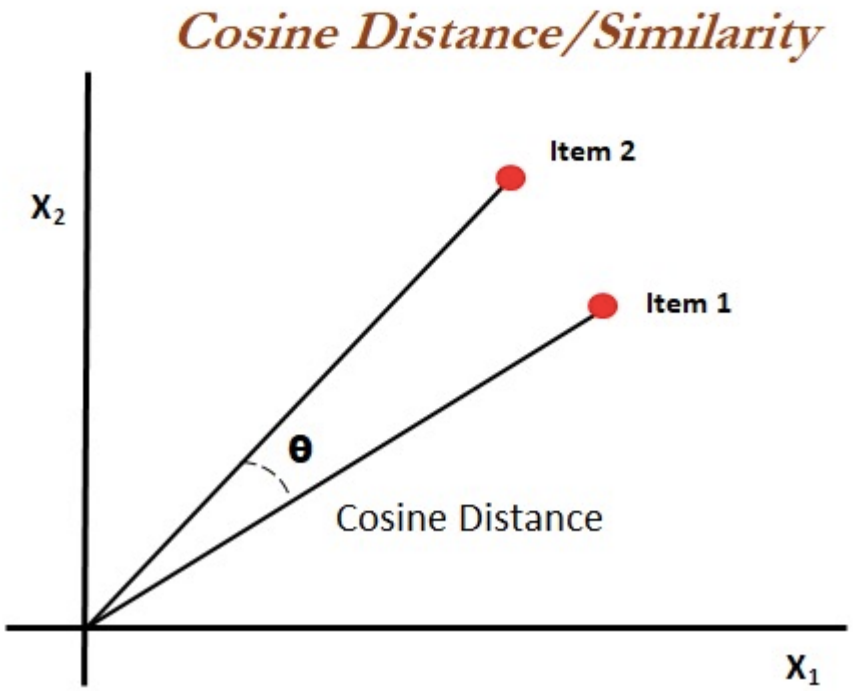
AC-2 Python 20 Clustering Ranking (0-5)

AC-2: ACCOUNT MANAGEMENT

K0537:5 access control system administration
K0397:5 system security concepts
K0536:0 exploitation tools
K0528:0 satellite based communication systems
K0419:5 database administration
K0420:0 database theory
K0452:0 radius authentication
K0608:0 operating system
K0634:0 exploitation techniques
K0180:0 network systems
K0056:5 access management
K0060:5 operating systems
K0065:5 access controls
K0286:0 typologies
K0071:0 remote access
*K0279:5 database access interfaces
K0077:0 server/client os

K0274:0 transmission records
*K0273:0 kill chain
K0197:5 database access
K0271:5 operating system structures
K0105:0 web services
K0122:0 hardware implications
K0224:5 system administration
K0177:0 cyber attack stages
*K0181:0 transmission records
K0192:0 ports and services
K0322:0 embedded system
K0033:0 host network
K0088:5 system administration
K0007:5 access control methods
K0024:4 database systems
K0336:5 access authentication
* WITHDRAWN

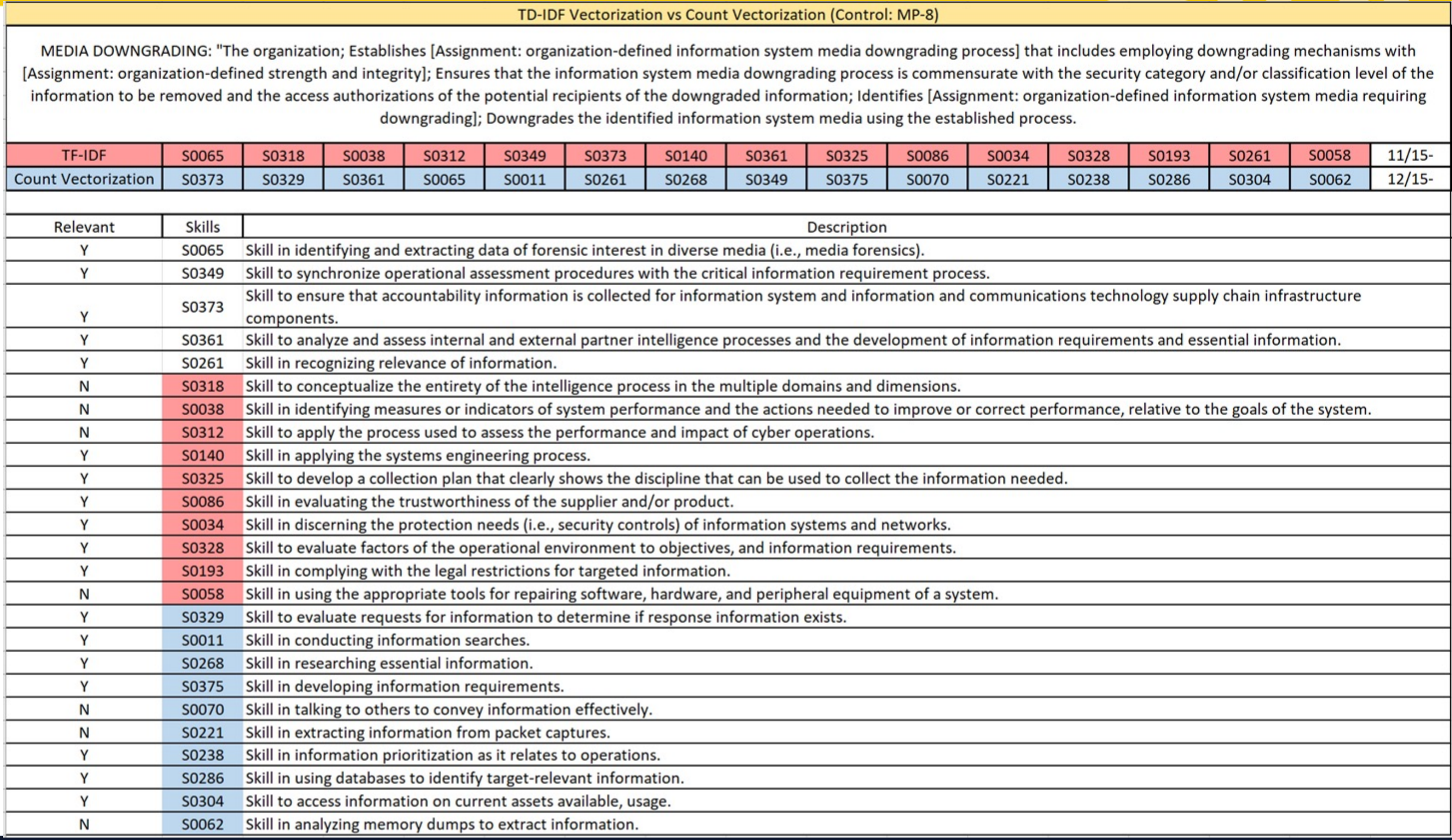
Cosine Similarity:



- Each document (either control description or a TKS statement) is vectorized using Bag of Words (BOW) or Term-Frequency/Inverse Document Frequency (TF-IDF).
- Vectorization converts text to numbers.
- **Results were understandable & useful**.

Image Credits https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/eb9cd609-e44a-40a2-9c3a-f16fc4f5289a.xhtml

Results:



TD-IDF Vectorization vs Count Vectorization (Control: MP-8)

MEDIA DOWNGRADING: "The organization; Establishes [Assignment: organization-defined information system media downgrading process] that includes employing downgrading mechanisms with [Assignment: organization-defined strength and integrity]; Ensures that the information system media downgrading process is commensurate with the security category and/or classification level of the information to be removed and the access authorizations of the potential recipients of the downgraded information; Identifies [Assignment: organization-defined information system media requiring downgrading]; Downgrades the identified information system media using the established process.

TF-IDF	S0065	S0318	S0038	S0312	S0349	S0373	S0140	S0361	S0325	S0086	S0034	S0328	S0193	S0261	S0058	11/15-
Count Vectorization	S0373	S0329	S0361	S0065	S0011	S0261	S0268	S0349	S0375	S0070	S0221	S0238	S0286	S0304	S0062	12/15-

Relevant	Skills	Description
Y	S0065	Skill in identifying and extracting data of forensic interest in diverse media (i.e., media forensics).
Y	S0349	Skill to synchronize operational assessment procedures with the critical information requirement process.
Y	S0373	Skill to ensure that accountability information is collected for information system and information and communications technology supply chain infrastructure components.
Y	S0361	Skill to analyze and assess internal and external partner intelligence processes and the development of information requirements and essential information.
Y	S0261	Skill in recognizing relevance of information.
N	S0318	Skill to conceptualize the entirety of the intelligence process in the multiple domains and dimensions.
N	S0038	Skill in identifying measures or indicators of system performance and the actions needed to improve or correct performance, relative to the goals of the system.
N	S0312	Skill to apply the process used to assess the performance and impact of cyber operations.
Y	S0140	Skill in applying the systems engineering process.
Y	S0325	Skill to develop a collection plan that clearly shows the discipline that can be used to collect the information needed.
Y	S0086	Skill in evaluating the trustworthiness of the supplier and/or product.
Y	S0034	Skill in discerning the protection needs (i.e., security controls) of information systems and networks.
Y	S0328	Skill to evaluate factors of the operational environment to objectives, and information requirements.
Y	S0193	Skill in complying with the legal restrictions for targeted information.
N	S0058	Skill in using the appropriate tools for repairing software, hardware, and peripheral equipment of a system.
Y	S0329	Skill to evaluate requests for information to determine if response information exists.
Y	S0011	Skill in conducting information searches.
Y	S0268	Skill in researching essential information.
Y	S0375	Skill in developing information requirements.
N	S0070	Skill in talking to others to convey information effectively.
N	S0221	Skill in extracting information from packet captures.
Y	S0238	Skill in information prioritization as it relates to operations.
Y	S0286	Skill in using databases to identify target-relevant information.
Y	S0304	Skill to access information on current assets available, usage.
N	S0062	Skill in analyzing memory dumps to extract information.

Image Credits:

IR-4: K-Means vs. Cosine Similarity

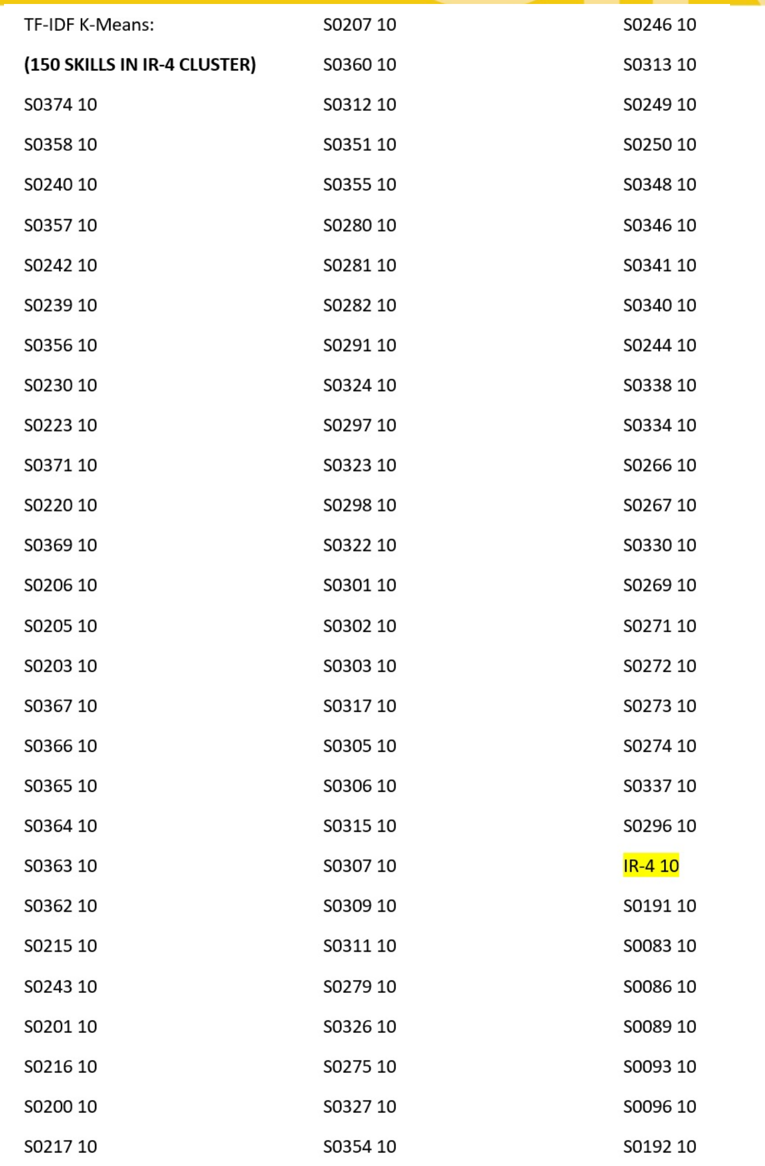
🟪 SKILL = TF-IDF & COUNT

🟥 SKILL = TF-IDF

🟦 SKILL = COUNT

COSINE	K-MEANS
🟪 S0054	ABSENT
🟪 S0365	PRESENT
🟪 S0176	PRESENT
🟪 S0350	ABSENT
🟪 S0337	PRESENT
🟪 S0150	ABSENT
🟪 S0032	ABSENT
🟪 S0200	PRESENT
🟪 S0377	ABSENT
🟪 S0186	ABSENT
🟪 S0329	ABSENT
🟪 S0371	PRESENT
🟥 S0282	ABSENT
🟥 S0374	ABSENT
🟥 S0064	ABSENT
🟦 S0100	PRESENT
🟦 S0309	PRESENT
🟦 S0197	ABSENT

PRESENT: 7/18 ABSENT: 11/18

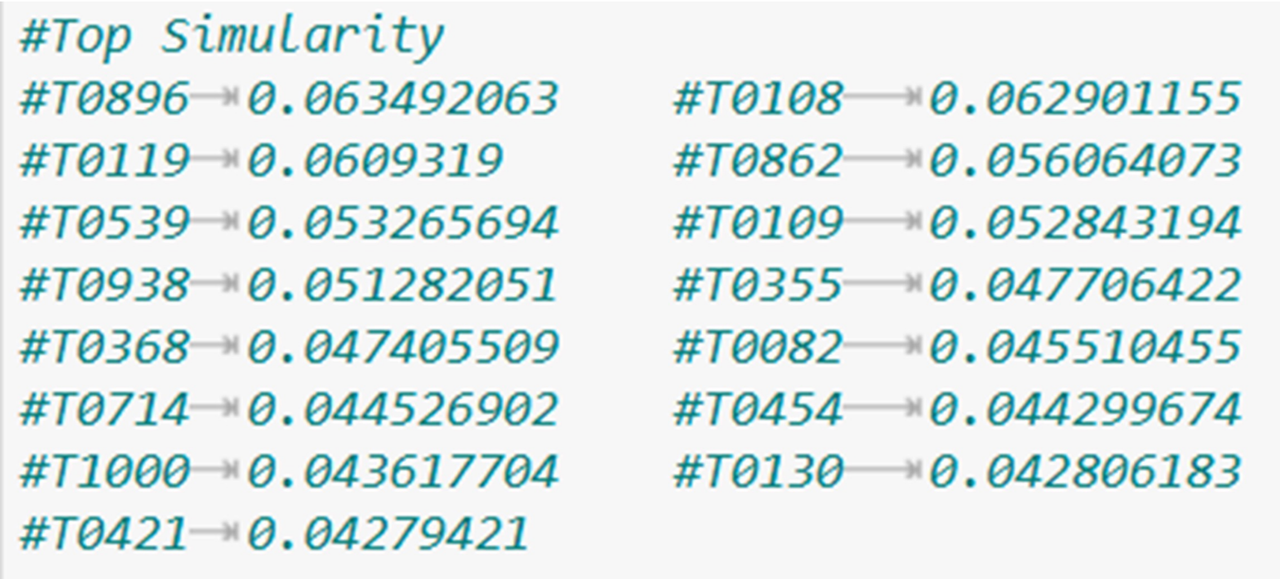


TF-IDF K-Means:
(150 SKILLS IN IR-4 CLUSTER)
S0374 10
S0358 10
S0240 10
S0357 10
S0242 10
S0239 10
S0356 10
S0230 10
S0223 10
S0371 10
S0220 10
S0369 10
S0206 10
S0205 10
S0203 10
S0367 10
S0366 10
S0365 10
S0364 10
S0363 10
S0362 10
S0215 10
S0243 10
S0201 10
S0216 10
S0200 10
S0217 10

S0207 10
S0360 10
S0312 10
S0351 10
S0355 10
S0280 10
S0281 10
S0282 10
S0291 10
S0324 10
S0297 10
S0323 10
S0298 10
S0322 10
S0301 10
S0302 10
S0303 10
S0317 10
S0305 10
S0306 10
S0315 10
S0307 10
S0309 10
S0311 10
S0279 10
S0326 10
S0275 10
S0327 10
S0354 10

S0246 10
S0313 10
S0249 10
S0250 10
S0348 10
S0346 10
S0341 10
S0340 10
S0244 10
S0338 10
S0334 10
S0266 10
S0267 10
S0330 10
S0269 10
S0271 10
S0272 10
S0273 10
S0274 10
S0337 10
S0296 10
IR-4 10
S0191 10
S0083 10
S0086 10
S0089 10
S0093 10
S0096 10
S0192 10

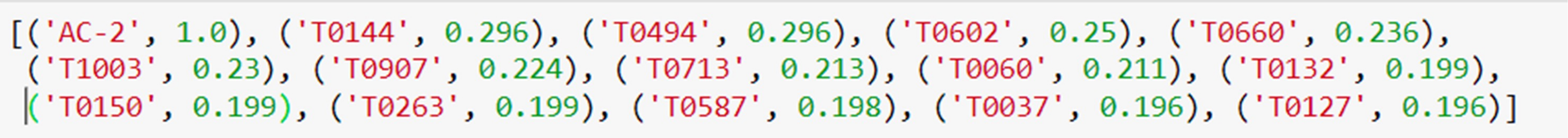
Similarity: Sequence Matcher



```
#Top Simularity
#T0896	0.063492063		#T0108	0.062901155
#T0119	0.0609319		#T0862	0.056064073
#T0539	0.053265694		#T0109	0.052843194
#T0938	0.051282051		#T0355	0.047706422
#T0368	0.047405509		#T0082	0.045510455
#T0714	0.044526902		#T0454	0.044299674
#T1000	0.043617704		#T0130	0.042806183
#T0421	0.04279421
```

- Weightage numbers are based on sequence matching.
- Tasks are being compared to AC-2 using a similarity sequence matcher
- 7/10 accuracy for AC-2 v/s 9/10 for Cosine Similarity
- **Results were ok in a few cases,** but not as accurate as Cosine similarity.

Cosine Similarity



```
[('AC-2', 1.0), ('T0144', 0.296), ('T0494', 0.296), ('T0602', 0.25), ('T0660', 0.236),
 ('T1003', 0.23), ('T0907', 0.224), ('T0713', 0.213), ('T0060', 0.211), ('T0132', 0.199),
 ('T0150', 0.199), ('T0263', 0.199), ('T0587', 0.198), ('T0037', 0.196), ('T0127', 0.196)]
```

Current Conclusion:

- **The mapping process could be automated.**
- **Cosine Similarity based on TF-IDF vectorization** gives useful results.
- Cosine Similarity with Count Vectorization is close as well in the small sample we have tried.

Future work:

- **Expand the experiment to cover more controls from 800-53** .
- Experiment with different parameters of TF-IDF Vectorization to give more weightage to a few key terms versus equal weightage [13]
- Understand how to deal with imbalanced document corpus [9]

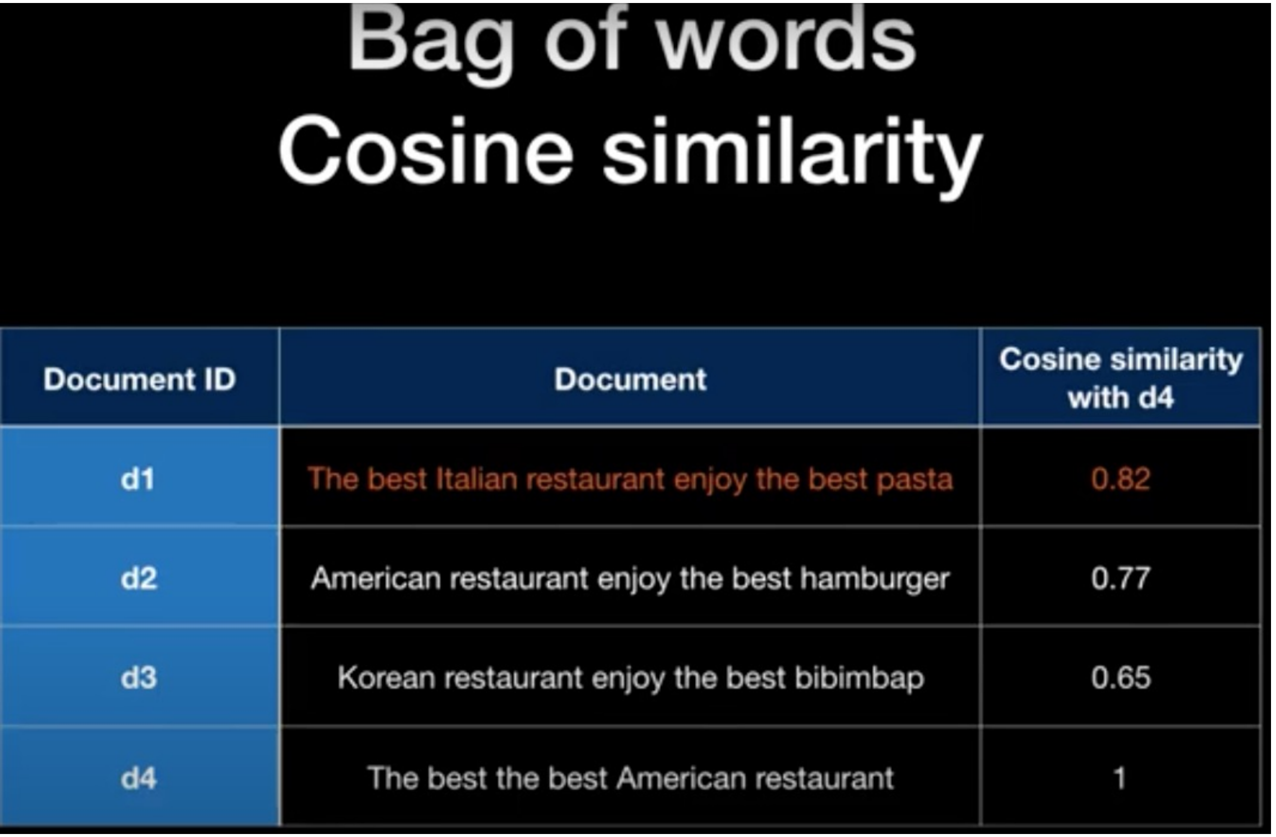
Acknowledgement

This research was performed under an appointment to the U.S. Department of Homeland Security (DHS) Science & Technology (S&T) Directorate Office of University Programs Summer Research Team Program for Minority Serving Institutions, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and DHS. ORISE is managed by ORAU under DOE contract number DE-SC0014664. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of DHS, DOE or ORAU/ORISE.

References

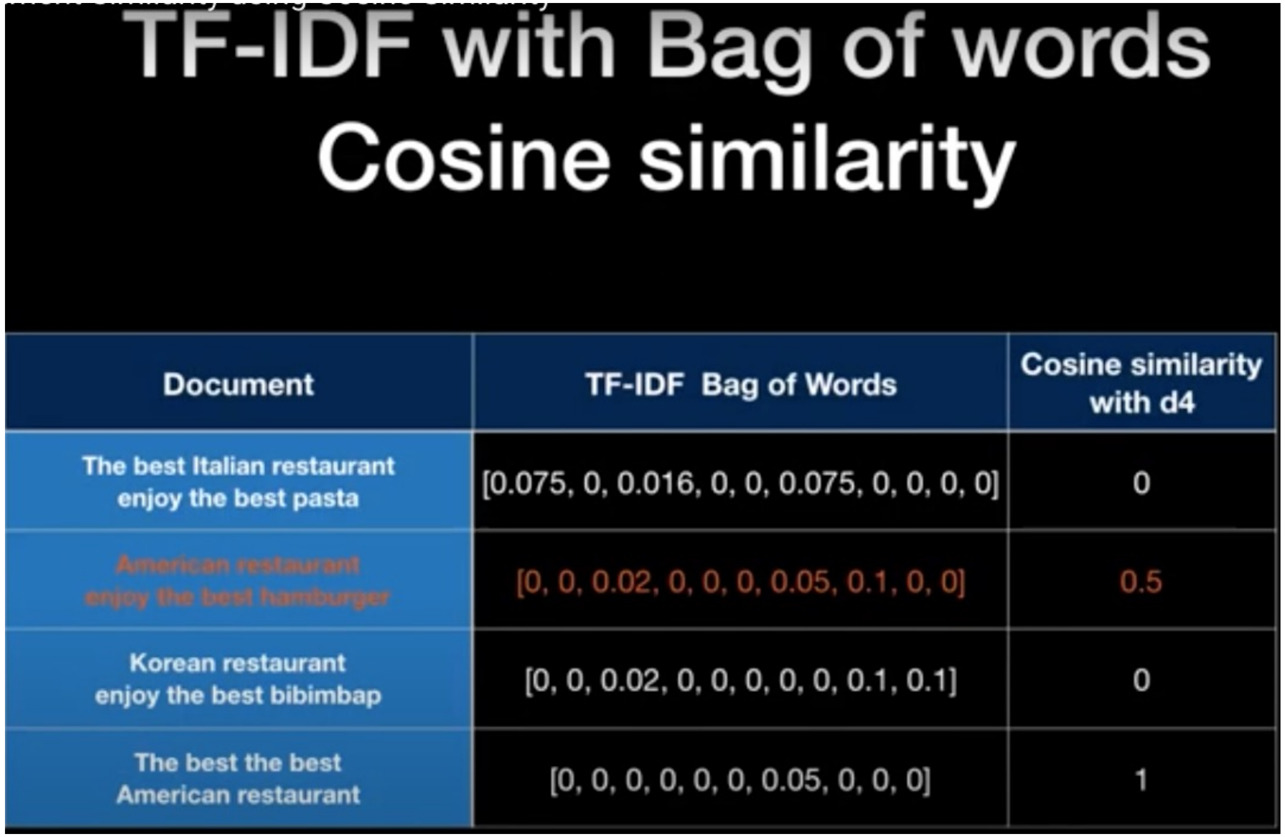
- [1] https://www.cyberseek.org/heatmap.html
- [2] https://cybersecurityventures.com/jobs/
- [3] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-181r1.pdf National Initiative for Cybersecurity Education (NICE) Workforce Framework for Cybersecurity (NICE Framework)
- [4] https://www.complianceforge.com/faq/nist-800-53-vs-iso-27002-vs-nist-csf.html Comparison of frameworks
- [5] https://pages.nist.gov/OSCAL/ Open Security Controls Assessment Language
- [6] CIRI developed CTB matches project needs to current talents. https://cybertalentbridge.com/projects.
- [7] CIRI announcement of CTB and its context. https://ciri.illinois.edu/news/21767
- [8] https://towardsdatascience.com/clustering-documents-with-python-97314ad6a78d
- [9] Santosh Kumar, et al, 2015, “*Subset K-Means Approach for Handling Imbalanced Distributed Data*“ in S.C. Satapathy et al. (eds.), Emerging ICT for Bridging the Future − Volume 2
- [10] Cosine Similarity: https://www.youtube.com/watch?v=hc3DCn8viWs
- [11] Topic Modelling: https://www.youtube.com/watch?v=i74DVqMsRWY
- [12] Choosing the right estimator: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
- [13] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Backup



Bag of words Cosine similarity

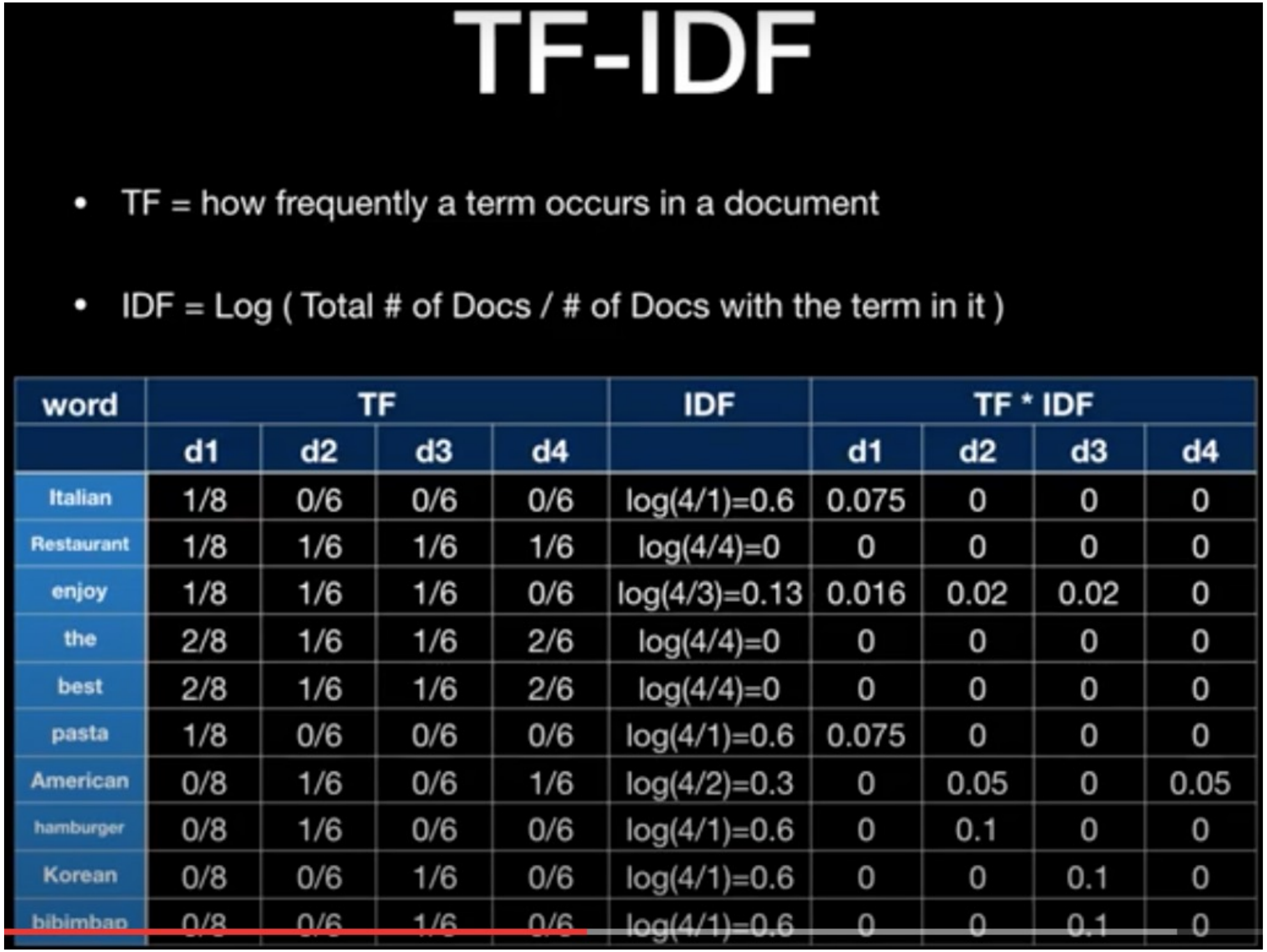
Document ID	Document	Cosine similarity with d4
d1	The best Italian restaurant enjoy the best pasta	0.82
d2	American restaurant enjoy the best hamburger	0.77
d3	Korean restaurant enjoy the best bibimbap	0.65
d4	The best the best American restaurant	1



TF-IDF with Bag of words Cosine similarity

Document	TF-IDF Bag of Words	Cosine similarity with d4
The best Italian restaurant enjoy the best pasta	[0.075, 0, 0.016, 0, 0, 0.075, 0, 0, 0, 0]	0
American restaurant enjoy the best hamburger	[0, 0, 0.02, 0, 0, 0, 0.05, 0.1, 0, 0]	0.5
Korean restaurant enjoy the best bibimbap	[0, 0, 0.02, 0, 0, 0, 0, 0, 0.1, 0.1]	0
The best the best American restaurant	[0, 0, 0, 0, 0, 0, 0.05, 0, 0, 0]	1

Image Credits: [10]



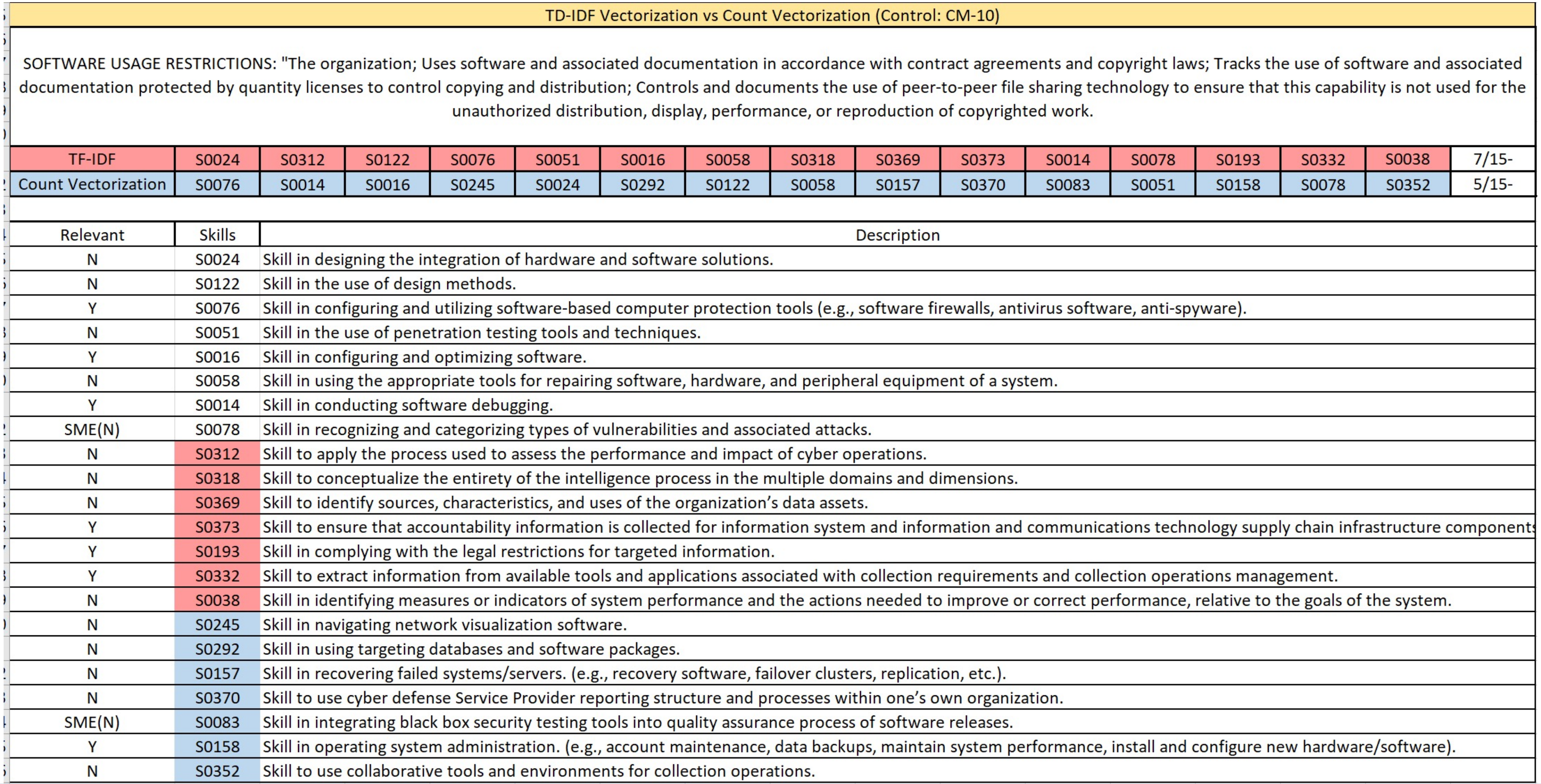
TF-IDF

- TF = how frequently a term occurs in a document
- IDF = Log (Total # of Docs / # of Docs with the term in it)

word	TF				IDF	TF * IDF			
	d1	d2	d3	d4		d1	d2	d3	d4
Italian	1/8	0/6	0/6	0/6	log(4/1)=0.6	0.075	0	0	0
Restaurant	1/8	1/6	1/6	1/6	log(4/4)=0	0	0	0	0
enjoy	1/8	1/6	1/6	0/6	log(4/3)=0.13	0.016	0.02	0.02	0
the	2/8	1/6	1/6	2/6	log(4/4)=0	0	0	0	0
best	2/8	1/6	1/6	2/6	log(4/4)=0	0	0	0	0
pasta	1/8	0/6	0/6	0/6	log(4/1)=0.6	0.075	0	0	0
American	0/8	1/6	0/6	1/6	log(4/2)=0.3	0	0.05	0	0.05
hamburger	0/8	1/6	0/6	0/6	log(4/1)=0.6	0	0.1	0	0
Korean	0/8	0/6	1/6	0/6	log(4/1)=0.6	0	0	0.1	0
bibimbap	0/8	0/6	1/6	0/6	log(4/1)=0.6	0	0	0.1	0

Image Credits: [10]

Control: CM-10



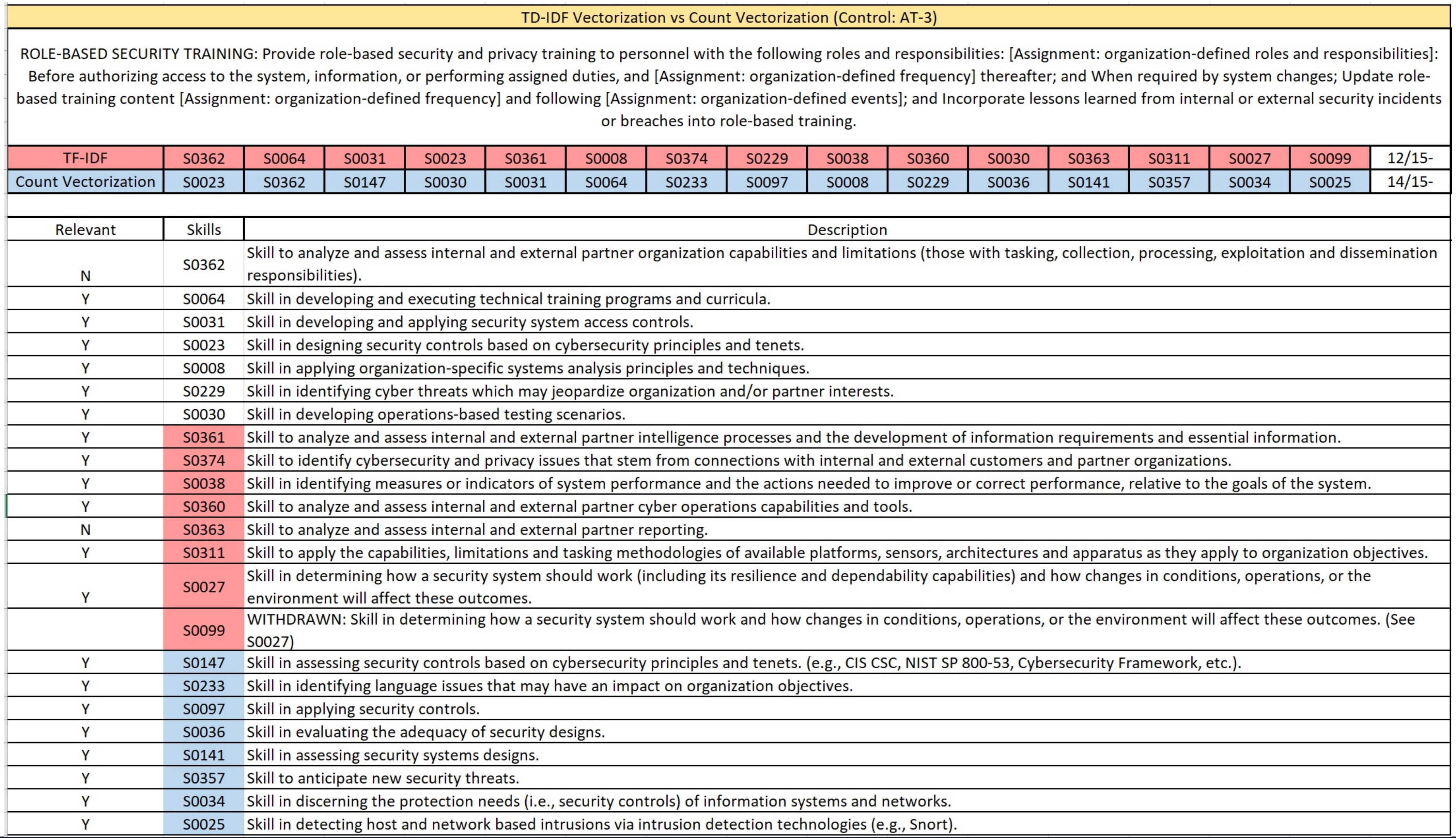
TD-IDF Vectorization vs Count Vectorization (Control: CM-10)

SOFTWARE USAGE RESTRICTIONS: "The organization; Uses software and associated documentation in accordance with contract agreements and copyright laws; Tracks the use of software and associated documentation protected by quantity licenses to control copying and distribution; Controls and documents the use of peer-to-peer file sharing technology to ensure that this capability is not used for the unauthorized distribution, display, performance, or reproduction of copyrighted work.

TF-IDF	S0024	S0312	S0122	S0076	S0051	S0016	S0058	S0318	S0369	S0373	S0014	S0078	S0193	S0332	S0038	7/15-
Count Vectorization	S0076	S0014	S0016	S0245	S0024	S0292	S0122	S0058	S0157	S0370	S0083	S0051	S0158	S0078	S0352	5/15-

Relevant	Skills	Description
N	S0024	Skill in designing the integration of hardware and software solutions.
N	S0122	Skill in the use of design methods.
Y	S0076	Skill in configuring and utilizing software-based computer protection tools (e.g., software firewalls, antivirus software, anti-spyware).
N	S0051	Skill in the use of penetration testing tools and techniques.
Y	S0016	Skill in configuring and optimizing software.
N	S0058	Skill in using the appropriate tools for repairing software, hardware, and peripheral equipment of a system.
Y	S0014	Skill in conducting software debugging.
SME(N)	S0078	Skill in recognizing and categorizing types of vulnerabilities and associated attacks.
N	S0312	Skill to apply the process used to assess the performance and impact of cyber operations.
N	S0318	Skill to conceptualize the entirety of the intelligence process in the multiple domains and dimensions.
N	S0369	Skill to identify sources, characteristics, and uses of the organization's data assets.
Y	S0373	Skill to ensure that accountability information is collected for information system and information and communications technology supply chain infrastructure components
Y	S0193	Skill in complying with the legal restrictions for targeted information.
Y	S0332	Skill to extract information from available tools and applications associated with collection requirements and collection operations management.
N	S0038	Skill in identifying measures or indicators of system performance and the actions needed to improve or correct performance, relative to the goals of the system.
N	S0245	Skill in navigating network visualization software.
N	S0292	Skill in using targeting databases and software packages.
N	S0157	Skill in recovering failed systems/servers. (e.g., recovery software, failover clusters, replication, etc.).
N	S0370	Skill to use cyber defense Service Provider reporting structure and processes within one's own organization.
SME(N)	S0083	Skill in integrating black box security testing tools into quality assurance process of software releases.
Y	S0158	Skill in operating system administration. (e.g., account maintenance, data backups, maintain system performance, install and configure new hardware/software).
N	S0352	Skill to use collaborative tools and environments for collection operations.

Control: AT-3



TD-IDF Vectorization vs Count Vectorization (Control: AT-3)

ROLE-BASED SECURITY TRAINING: Provide role-based security and privacy training to personnel with the following roles and responsibilities: [Assignment: organization-defined roles and responsibilities]: Before authorizing access to the system, information, or performing assigned duties, and [Assignment: organization-defined frequency] thereafter; and When required by system changes; Update role-based training content [Assignment: organization-defined frequency] and following [Assignment: organization-defined events]; and Incorporate lessons learned from internal or external security incidents or breaches into role-based training.

TF-IDF	S0362	S0064	S0031	S0023	S0361	S0008	S0374	S0229	S0038	S0360	S0030	S0363	S0311	S0027	S0099	12/15-
Count Vectorization	S0023	S0362	S0147	S0030	S0031	S0064	S0233	S0097	S0008	S0229	S0036	S0141	S0357	S0034	S0025	14/15-

Relevant	Skills	Description
N	S0362	Skill to analyze and assess internal and external partner organization capabilities and limitations (those with tasking, collection, processing, exploitation and dissemination responsibilities).
Y	S0064	Skill in developing and executing technical training programs and curricula.
Y	S0031	Skill in developing and applying security system access controls.
Y	S0023	Skill in designing security controls based on cybersecurity principles and tenets.
Y	S0008	Skill in applying organization-specific systems analysis principles and techniques.
Y	S0229	Skill in identifying cyber threats which may jeopardize organization and/or partner interests.
Y	S0030	Skill in developing operations-based testing scenarios.
Y	S0361	Skill to analyze and assess internal and external partner intelligence processes and the development of information requirements and essential information.
Y	S0374	Skill to identify cybersecurity and privacy issues that stem from connections with internal and external customers and partner organizations.
Y	S0038	Skill in identifying measures or indicators of system performance and the actions needed to improve or correct performance, relative to the goals of the system.
Y	S0360	Skill to analyze and assess internal and external partner cyber operations capabilities and tools.
N	S0363	Skill to analyze and assess internal and external partner reporting.
Y	S0311	Skill to apply the capabilities, limitations and tasking methodologies of available platforms, sensors, architectures and apparatus as they apply to organization objectives.
Y	S0027	Skill in determining how a security system should work (including its resilience and dependability capabilities) and how changes in conditions, operations, or the environment will affect these outcomes.
	S0099	WITHDRAWN: Skill in determining how a security system should work and how changes in conditions, operations, or the environment will affect these outcomes. (See S0027)
Y	S0147	Skill in assessing security controls based on cybersecurity principles and tenets. (e.g., CIS CSC, NIST SP 800-53, Cybersecurity Framework, etc.).
Y	S0233	Skill in identifying language issues that may have an impact on organization objectives.
Y	S0097	Skill in applying security controls.
Y	S0036	Skill in evaluating the adequacy of security designs.
Y	S0141	Skill in assessing security systems designs.
Y	S0357	Skill to anticipate new security threats.
Y	S0034	Skill in discerning the protection needs (i.e., security controls) of information systems and networks.
Y	S0025	Skill in detecting host and network based intrusions via intrusion detection technologies (e.g., Snort).