Issues in Information Systems
Volume 23, Issue 4, pp. 70-80, 2022

DOI: https://doi.org/10.48009/4_iis_2022_107

Critical mineral trend analysis using text mining

Selassie Adiwokor, *Missouri University of Science and Technology, skakbb@umsystem.edu*
Eugene Gyawu, *Missouri University of Science and Technology, eagnvh@umsystem.edu*
Kyle Johnson, *Missouri University of Science and Technology, kdjdk7@umsystem.edu*
Allan Kimurgor Kosgey, *Missouri University of Science and Technology,akk5mf@umsystem.edu*
Wen-Bin Yu, *Missouri University of Science and Technology, yuwen@umsystem.edu*

Abstract

The supply of minerals is important to the world's major and emerging industrial economies. Over time, the supply of these minerals has been a concern to the US (United States) government. This led to Executive Order 13817 of 2017 which outline minerals that are critical to the US economy and national security. The US Geological Survey (USGS) executes the order to ensure a constant supply of critical minerals. Despite the US government and other state agencies' steps to define and review the critical mineral list yearly, there has not been any published list by academia on what they describe as critical minerals. The objectives of this study are to identify the trends in critical minerals using research publications and reveal the knowledge gap between academia and the government-published literature. Text mining will be used as a tool to achieve the set objectives in this publication. Thirty-one articles were obtained for this research which was sampled at random from publications from 2008 to 2020 using "Critical minerals" as the keyword. Our results show Helium and Iron Ore carried no supply risk, and unsurprisingly were not found in the research articles. Gold, however, also carried no risk, but still carried weight in the research discussions.

Keywords: Critical minerals, Text mining

Introduction

Critical minerals are defined as mineral materials essential to the economic and national security of the United States and are subject to the risk of disruption of supply (NRC, 2008). There are about 54 minerals published by the US government Critical in 2018. These minerals are deemed critical because of their associated supply risk which may be due to trade policies, geopolitical issues, geological scarcity, and other factors. The U.S (United States) (United States). government and other state agencies define and review the critical mineral list yearly. Despite numerous research conducted by the government which is focused on critical minerals, there has not been any published list by academia on what they describe as critical minerals. This work is an attempt to reveal the knowledge gap between academic publishing literature and government research on critical minerals using the text mining approach.

Fortier et al. 2019 conducted an analysis of the US critical minerals list according to USGS. They employed quantitative criteria to measure the country's concentration of production, and net import reliance. From the data, they produced a list of minerals and their supply risk. The list of minerals and their supply risk form the basis of our study.

Text mining is defined as the automatic retrieval of information from various written resources by a computer of new and or previously unknown information (Hearst, 2003). The text mining approach has been used by Russian scientists to analyze trends in Russia's extractive industry (Gokhberg et. al 2020). The authors used a combination of expert-based foresight activities as well as statistical analyses with text mining techniques (machine learning and artificial intelligence) to report on large-scale technologies. Emphasis was placed on the technological advancement of industries that exploit natural resources (rare earth metals, phosphate) and hydrocarbons. (Tugrul, et al., 2009) reiterate that the widely used foresight method with expert panels is the Delphi method. However, as difficulties arise with information accessibilities, complex methods specifically used for scanning strategic intelligence are being developed.

A similar study conducted by (Kim et al., 2018) used text mining to analyze trends in the field of medical informatics. The authors extracted articles from the PubMed archives. They performed cluster analysis on the papers to group commonly occurring, nontrivial words, based upon common topics. They were able to create word clouds showing terms in each cluster with size varying by weight of the term. They also created line graphs showing the number of times words surfaced in literature for each cluster over time as the articles were published.

The objective of this research is to: (i) determine the current trend of critical minerals from peer-reviewed journal articles, (ii) cluster these identified critical minerals using the supply risk weight stipulated by the USGS, and (iii) compare the weights to a term frequency weight.

Data used in this research was acquired from articles from scholarly literature repositories such as Scopus, Google Scholar, and Google search engine. "Critical minerals" was used as the keyword to retrieve the articles for text mining. Thirty-one articles were obtained for the research. These articles were sampled at random but constricted to publications from 2008 to 2020. The acquired data were pre-processed using the term frequencies (TF) of the minerals. The minerals in the papers were extracted and a word cloud was produced from them. The 54 minerals that are ranked by the USGS used in this paper from the highest risk to the lowest risk are:
"Gallium, Niobium, Cobalt, Neodymium, Ruthenium, Rhodium, Dysprosium, Aluminum, Fluorspar, Platinum, Iridium, Praseodymium, Cerium, Lanthanum, Bismuth, Yttrium, Antimony, Tantalum, Hafnium, Tungsten, Vanadium, Tin, Magnesium, Germanium, Palladium, Titanium, Zinc, Graphite, Chromium, Arsenic, Barite, Indium, Samarium, Manganese, Lithium, Tellurium, Lead, Potash, Strontium, Rhenium, Nickel, Copper, Beryllium, Feldspar, Phosphate, Silver, Mica, Selenium, Cadmium, Zirconium, Molybdenum, Gold, Helium, Iron Ore".

Using the term-frequency matrix data, four clusters were identified using the weights assigned to critical minerals by the USGS which are minerals with high supply risk.

Methodology

This part of the paper provides a detailed account of data acquisition, preprocessing of data, data exploration, and application of text mining techniques.

Data acquisition and preprocessing

Articles from scholarly literature repositories such as Scopus, Google Scholar, and google search engine were manually collated. Keywords such as "critical minerals," "mining," and "United States" were used as search parameters to retrieve the articles for text mining. Thirty-one articles were obtained for the research. These articles were sampled at random but constricted to publications from 2008 to 2020. The year 2008

was chosen as the beginning year since the United State Geological Survey started recognizing critical minerals around the same year.

One of the prime objectives of this project was to identify the trends in critical minerals using research publications. Initially, the researchers decided to use only the introductory part of the articles identified. But upon consultation, they decided to use three main parts of the articles which were: (i) abstract, (ii) introduction, and (iii) result and discussion. These parts of the article were transferred from these repositories into an Excel file to be analyzed in python and SAS Enterprise Miner.

Text mining method

The act of discovering knowledge from vast amounts of unstructured data sources using a semi-automated pattern extracting process is deemed as text mining (Delen, 2014). Text mining is intently associated with data mining making use of identical processes, however, the input to a text mining system is a series of unstructured textual content files like Extensible Markup Language files, and Portable Document Format files, just to mention a few. The advantages of text mining are in fields that engender big data. Typical areas include marketing, social media, engineering, and health.

Text mining process

A standardized process, such as CRoss Industry Standard Process for Data Mining (CRISP-DM), used by many data scientists was adopted for this research. CRISP-DM possesses six phases that make up the life cycle of data science. These phases provide answers to typically asked questions to produce a good scientific result. Typical questions asked under each phase are listed below:

- Business understanding – What does the business need?
- Data understanding – What data do we have or need? How clean is that data?
- Data Preparation – How do we organize the data for modeling?
- Evaluation – What model best meets the business objectives?
- Deployment – How do stakeholders access the result?

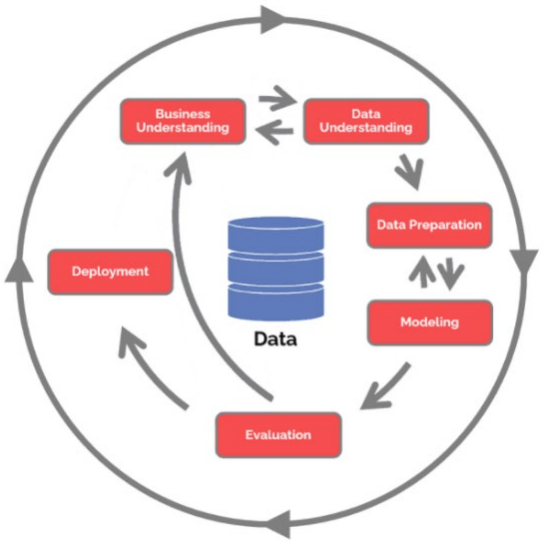


Figure 1: Diagram of CRISP-DM used as data science life cycle (Alliance, 2021)

Figure 1 represents the six phases of data mining. However, in text-mining, these six phases could be simplified into three main phases which are: (i) Corpus creation, (ii) data pre-processing, and, (iii) Knowledge recognition. These phases were adopted from Kim and Delen's research (Kim & Delen, 2018) as shown in Figure 2.

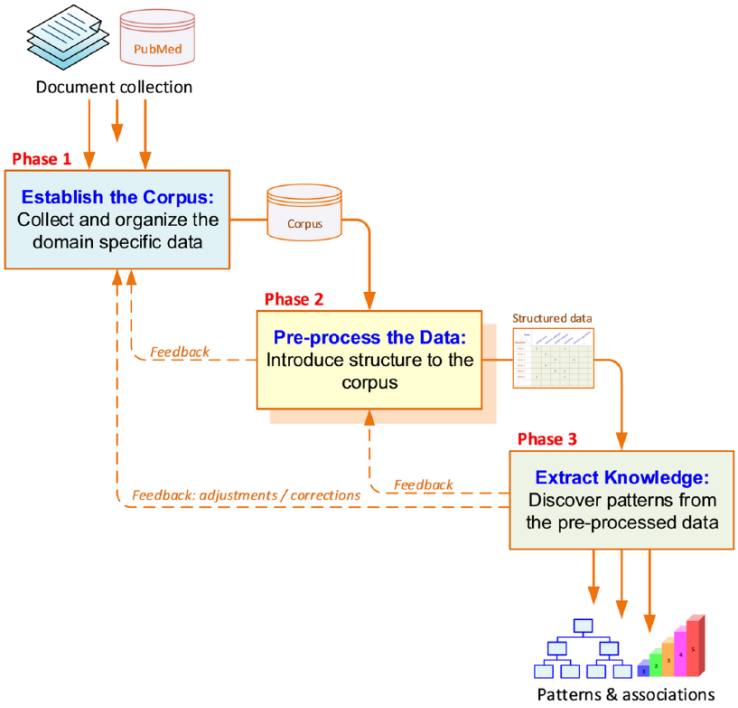


Figure 2: Modified CRISP-DM for text mining showing three phases (Kim & Delen, 2018)

Corpus creation

In this phase, documents found to be related to the research context were gathered. This was done by retrieving the abstract, introduction, and the result portions of all identified articles into a spreadsheet (shown in figure 3). In the spreadsheet, five columns represent the authors, year of publication, abstract of the article, the introduction of the article, and finally, the results and discussion section of the articles. The rows, however, represent each collected article. The format of textual data was ASCII for ease of processing.

Data pre-processing

In this phase, the final output was the term-frequency (TF) of minerals. There were preliminary procedures executed to arrive at the output mentioned above. The first task was to eliminate the stop words. These words are normal everyday English words that do not have any significance in the documents. The second task was to explore the data. This was done to establish the context of the data. Especially, knowing if the corpus created has the right information needed for the research. A word cloud of all abstracts, introduction, and result columns in the corpus was created for that purpose. The third task was lemmatization or stemming. This task was for the purposed of reducing the terms or words to a simplified form. The process was executed using a python script. To identify all the key terms relevant for this research, researchers

performed filtration using what they called a critical minerals list table. This table contains the list of minerals (terms) needed for the research. After filtration, the frequency of terms (minerals) was obtained and exported in a spreadsheet. Figure 4 illustrates the word cloud generated to validate the context of the articles and step 'A' in figure 5 illustrated the python script used.

	A	B	C	D	E
1	Author	Year	Abstract	Introduction	Results and discusions
2	Coulomb, R., Dietz, S., God	2015	Raw materials are es	Resource Council of the National Academies,	Most of the analysis is carried out for the OECD economy on
3	Hayes, S. M., & McCullough	2018	Mineral criticality is	One consequence of incre	General trends Despite differing frames of reference, there are
4	McCullough, E., & Nassar, N	2017	Increasing reliance on non-renewable	Introduction Modern technology is in large	The normalized R indicator values for each of the minerals examined
5	McLellan, B. C., Yamasue, E	2016	Abstract: The nexus	1. Introduction As global	Results and Discussion The combina
6	Viebahn, P., Soukup, O., Sa	2015	The German governr	IntroductionMajor redu	4. Assessment of critical minerals from a resourceperspectiveFinally,
7	Humphreys, D. (2014)	2014	The geology and tecl	supporting infrastructure	Recycling and reuse of metals Mode
8	Humphries, M. (2015)	2015	China is the world's	China is in a new era of de	The NRC produced a criticality matr
9	Stensgaard, B. M., Stendal,	2017	Minerals are essenti	2. Critical raw materials F	4. Potential companion-metal comr
10	LASLEY, S. (2018)	2018	Following a U.S. Gec	The terms critical minerals and strategic	Critical minerals strategy ordered Tr
11	Sarapää, O., Lauri, L. S., Al	2015	The economy of Finl	Hi-tech metals and critica	2. RARE EARTH ELEMENTS (REE) Rare
12	Sievers, H., & Tercero, L. (2	2012	Even if geological co	19.1Historical backgrour	19.2Criticality assessment for the Eu
13	Mclellan, C. B. (2015)	2015	In order to address	The minerals-energy nexu	CRITICAL MINERALS IN CLEAN ENERG
14	Pongrácz, E. (2014)	2014	Critical materials sh	Future energy strategies r	While the scientific community has
15	Evrard, M., & Pirard, E. (201	2013	The Chelopech gold-	Europe is more and more	Pixels number has been counted in e
16	Radwanek-Bak, B. (2014)	2014	The assurance of futi	The problem of increasing	SWOT ANALYSIS OF CRM RECOVERY D
17	Parthemore, C. (2011)	2011	Reliable access to	Minerals are a subject of	Recommendations for U.S. Policyma
18	Yellishetty, M., Ranjith, P. G	2009	Background, aim an	Background, aim and sco	LCA is increasingly used in the miner
19	Peens, A. M. (2009)	2009	Mining is an importa	2.1. The Minera/and Petr	4.3 Summary It is thus clear that in t
20	Rogich, D. G., & Matos, G. F	2008	This paper provides	The current metals and m	Outputs (releases) to the environme
21	Graedel, T. E., & Cao, J. (20	2010	We have assembled	I t is without any doubt tl	Results Metal use index (ψj;k) values
22	Wang, J., Yang, L., Lin, J., &	2020	In the context of dep	organized by the United Nations Framework	DISCUSSION Importance of Investme
23	Toledano, P., Brauch, M., K	2020	The green energy tra	The green energy transitic	Anticipated demand for minerals is f
24	Hammarstrom, J. M., & Dicke	2019	Rare earth elements	The U.S. Geological Surve	The focus areas outlined throughout
25	Whittle, D., Yellishetty, M., W	2020	Critical minerals ass	A project titled "Evaluati	Science, 2019). • Reducing administ
26	Komal Habib and Henrik Wer	2014	The dependency on	1. Introduction Motivate	3.1. The scale of demand vs. supply l
27	Junbeum Kim, Bertrand Guill	2014	Critical materials as	There has been a dramatic	In Tables 4 and 5 and Fig. 4, the calcu
28	Michele L. Bustamante, Gabr	2014	Transitioning to a su	The global energy supply	To learn from the above methodolog
29	Ayman Elshkaki, T.E. Graedel	2015	Several scenarios ha	Several scenarios for the f	3.1. The demand for the metals requi
30	Ayman Elshkaki, T.E. Graedel	2014	Wind power techno	Wind power is one of the	3.1. Electricity generation by wind p
31	Komal Habib, Henrik Wenzel	2015	The recent debate o	A transition from the cur	This section presents the methodolo
32	Ayman Elshkaki, T.E. Graedel	2013	The demand for ene	The demand for energy ar	The total global electricity demand b

Figure 3: A snippet of articles compiled in excel

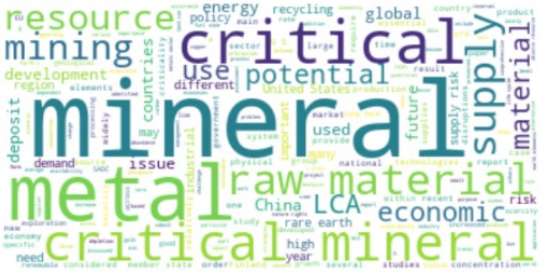
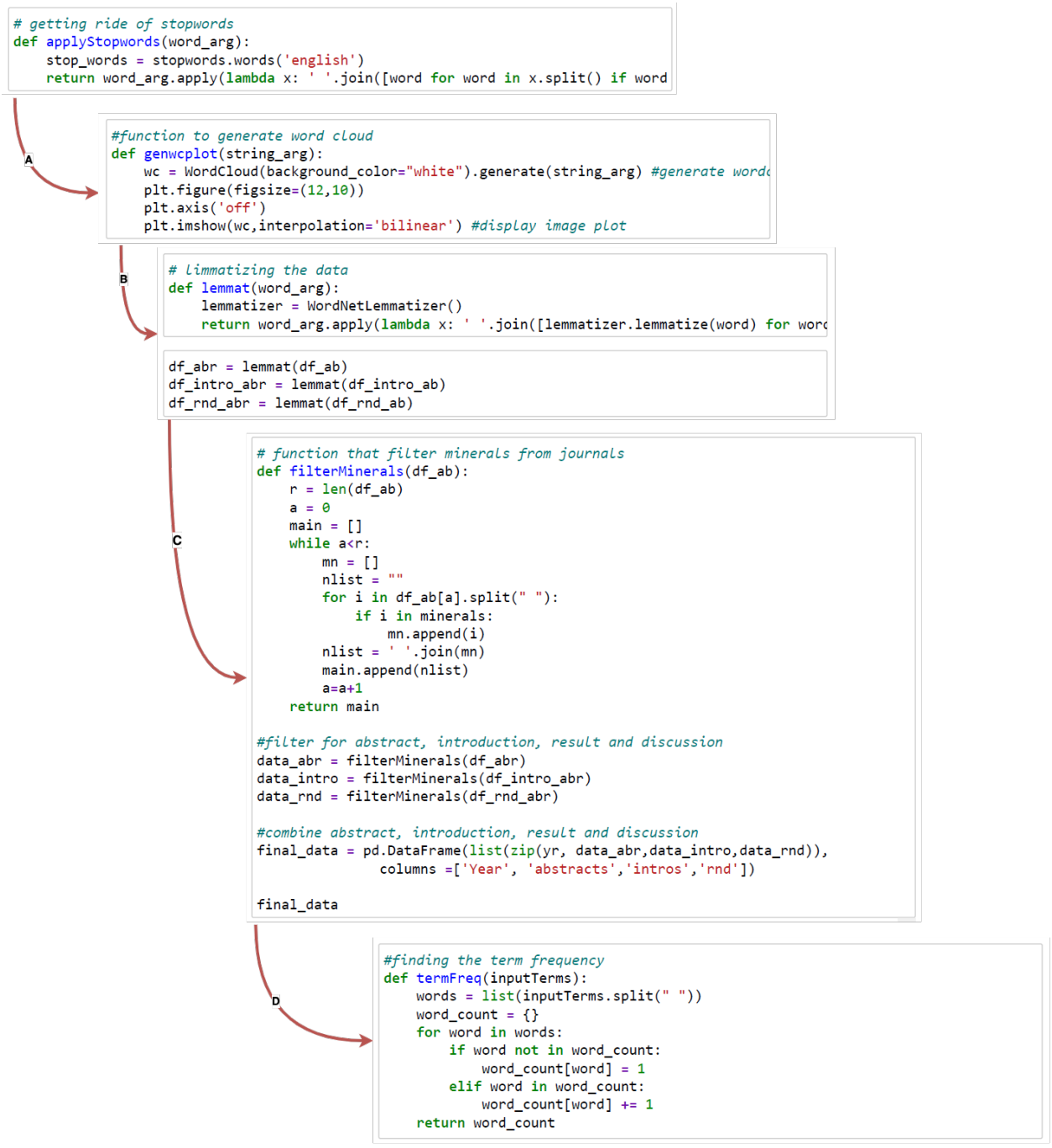


Figure 4: Word cloud used to validate the context of the article in the corpus




```
# getting ride of stopwords
def applyStopwords(word_arg):
    stop_words = stopwords.words('english')
    return word_arg.apply(lambda x: ' '.join([word for word in x.split() if word
```

```
#function to generate word cloud
def genwcplot(string_arg):
    wc = WordCloud(background_color="white").generate(string_arg) #generate wordc
    plt.figure(figsize=(12,10))
    plt.axis('off')
    plt.imshow(wc,interpolation='bilinear') #display image plot
```

```
# limmatizing the data
def lemmat(word_arg):
    lemmatizer = WordNetLemmatizer()
    return word_arg.apply(lambda x: ' '.join([lemmatizer.lemmatize(word) for worc
```

```
df_abr = lemmat(df_ab)
df_intro_abr = lemmat(df_intro_ab)
df_rnd_abr = lemmat(df_rnd_ab)
```

```
# function that filter minerals from journals
def filterMinerals(df_ab):
    r = len(df_ab)
    a = 0
    main = []
    while a<r:
        mn = []
        nlist = ""
        for i in df_ab[a].split(" "):
            if i in minerals:
                mn.append(i)
        nlist = ' '.join(mn)
        main.append(nlist)
        a=a+1
    return main

#filter for abstract, introduction, result and discussion
data_abr = filterMinerals(df_abr)
data_intro = filterMinerals(df_intro_abr)
data_rnd = filterMinerals(df_rnd_abr)

#combine abstract, introduction, result and discussion
final_data = pd.DataFrame(list(zip(yr, data_abr,data_intro,data_rnd)),
               columns =['Year', 'abstracts','intros','rnd'])

final_data
```

```
#finding the term frequency
def termFreq(inputTerms):
    words = list(inputTerms.split(" "))
    word_count = {}
    for word in words:
        if word not in word_count:
            word_count[word] = 1
        elif word in word_count:
            word_count[word] += 1
    return word_count
```

Figure 5: Illustrate python scripts used in performing the analysis in chronological order

Knowledge recognition

This phase was the critical part of the research. Being able to identify patterns in data needs an adequate data/text mining skill set. Using the term-frequency matrix data, four clusters were identified using the weights assigned to critical minerals by the USGS which are minerals with high supply risk. These weights

were separated to have an upper and lower bound making up four groups shown in table 1 below. A python script was generated based on these bound to cluster the minerals.

Table 1: Clusters generated from critical minerals weights for high-risk supply

Cluster	Lower Bound	Upper Bound	Total CML
Extremely High	0.56	0.67	14
High	0.48	0.55	13
Moderate	0.36	0.47	14
Low	0	0.34	13

```python
# grouping minerals into high - low
def groupElements(str_element,minerals):
    ext_high = minerals[0:14]
    high = minerals[14:27]
    moderate = minerals[27:41]
    low = minerals[41:54]
    risk_cat = ['extremely high','high','moderate','low','NaN']
    risk=[]
    for i in str_element:
        if i in ext_high:
            risk.append(risk_cat[0])
        elif i in high:
            risk.append(risk_cat[1])
        elif i in moderate:
            risk.append(risk_cat[2])
        elif i in low:
            risk.append(risk_cat[3])
        else:
            risk.append(risk_cat[4])
    return risk
```

Figure 6: Illustrate the Python code snippet used in clustering minerals based on table 1.

Results

As the United States Geological Survey (USGS) has created a list of minerals that are considered "critical," it was natural to note the appearance of each of those minerals within the corpus.

The frequency of each mineral was collected across each section and combined. Once the frequency was established, the USGS supply risk weights were applied to each mineral to create a "score." These scores were derived by dividing individual frequencies by the total frequency and multiplying by ten. These scores represent a quantitative figure for the level of research discussion around a particular mineral. A visualization of these metrics by risk category can be seen in figures 7 and 8 as well as prominent minerals in each category or cluster.

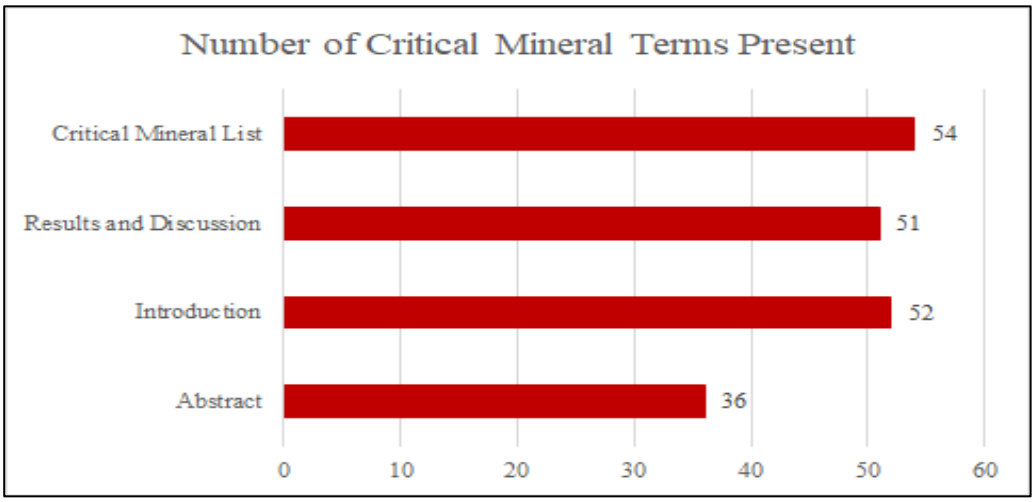


Figure 7: Number of critical minerals mentioned in each section of the corpus

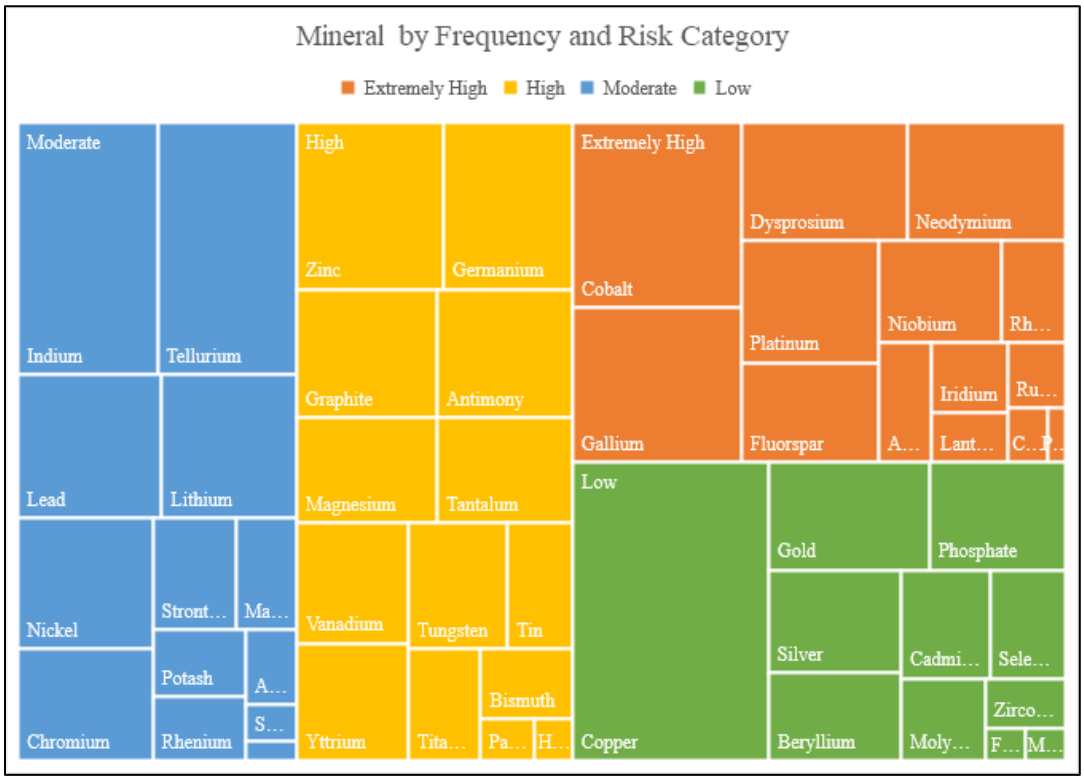


Figure 8: Frequency by Category

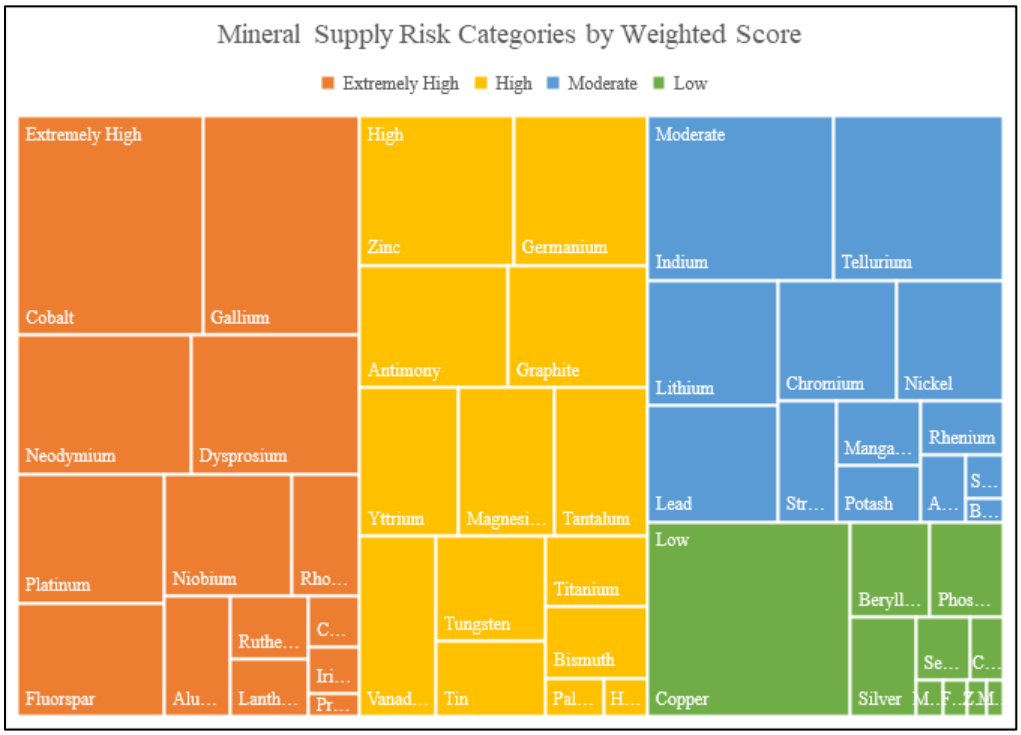


Figure 9: Weighted Score by Category

From this, the research discussion associated with each risk category can be approximated. The next step in comparison was to find the weight of each mineral as it appears in the corpus. This is illustrated in Figure 10 alongside the supply risk scores assigned by USGS.

Discussion

Based on Figure 10, the level of discussion surrounding each mineral on the critical mineral list varies. Iridium, Praseodymium, and cerium fall under the extremely high-risk category according to the USGS ranking but they are not discussed as much as Copper and Gold which fell under the low-risk category. Figure 10 also shows that academic research and economic research on critical minerals do not correlate based on the importance of the mineral commodity. This might be because the minerals have not been Primary Commodities and that is not a primary driver for developing a mine. Gold and copper are primary commodities that usually occur with other by-products such as primary sulphide mineralization (Fortier et al. 2019). Although Critical minerals can be either primary commodities, coproducts, or byproduct commodities based on the commodity and the mineral deposit type. Critical minerals such as rare earth elements, beryllium, antimony, and lithium are usually produced as primary commodities (Fortier et al. 2019). Our results confirm that these minerals are been talked about in academic literature as well. Helium and Iron Ore carried no supply risk and unsurprisingly were minimal in the research articles surrounding.

Conclusion

This study serves as a starting point for further research into the field of critical minerals. Minerals which are part of higher risk categories are the most pressing to research from an economic and strategic perspective. Additionally, minerals that have a lower presence in research literature compared to their

supply risk may also warrant further study. To conduct further exploration of the corpus which has been gathered in this study. The goal is to identify underlying relationships between concepts found in the research. In a nutshell, this research brings to light the current research focus in academia on the USGS-established critical minerals. This validates the need for more probing into critical minerals that have more significance based on their weight allotted by USGS.

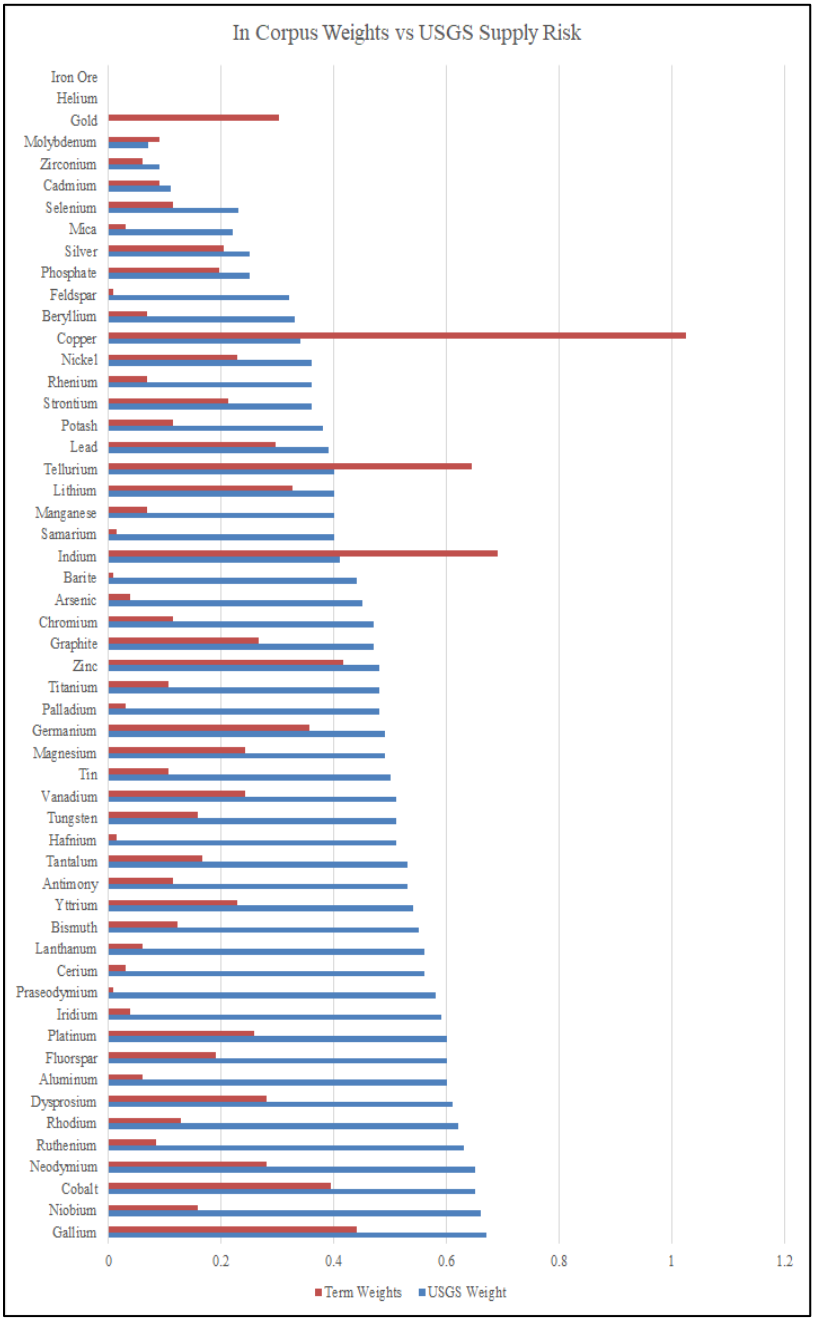


Figure 10: Mineral term weight in corpus compared to USGS supply risk weight

References

Data Science Process Alliance. (n.d.). Retrieved December 14, 2021, from https://www.datascience-pm.com/crisp-dm-2/

Delen, D. (2014). Real-world data mining: applied business analytics and decision making. FT Press.

Fortier, S. M., Hammarstrom, J. H., Ryker, S. J., Day, W. C., Seal, R. R., & U.S. Geological Survey. (2019). USGS Critical Minerals Review. USGS Critical Minerals Review, 35–36. https://www.researchgate.net/publication/339438964

Fortier, S. M., Hammarstrom, J. H., Ryker, S. J., Day, W. C., & Seal, R. R. (2019). USGS critical minerals review. Mining Engineering, 71(5), 35-47.

Hearst, M. (2003). What is text mining. SIMS, UC Berkeley, 5

Kim, Yong-Mi, and Dursun Delen. "Medical Informatics Research Trend Analysis: A Text Mining Approach." Health Informatics Journal, (December 2018), 432-52. https://doi.org/10.1177/1460458216678443.

Leonid, G., Kuzminov, I., Khabirova, E., & Thurner, T. (2020). Advanced text-mining for trend analysis of Russia's extractive industries. Futures 115, 102476.

National Research Council. (2008). Minerals, critical minerals, and the US economy. National Academies Press.

Tugrul, D., Justice, J., Krampits, M., Letts, M., Subramanian, G., & Thirumalai, M. (2009). Data center metrics: An energy efficiency model for information technology managers. Management of Environmental Quality: An International Journal.