Topic Analysis for NSF (National Science Foundation) grant programs using the LDA model.

Capstone Project

Animesh Kumar Jha
Student ID -23505138
Fall 2023

Supervised by: Prof. Bryan Heidorn

Abstract

In the dynamic and rapidly evolving landscape of Big Data research, securing funding is a pivotal step for advancing scientific inquiry and technological innovation. This project undertakes an extensive investigation into National Science Foundation (NSF) funded research topics within the domain of Big Data, utilizing advanced Latent Dirichlet Allocation (LDA) topic analysis. By delving into the abstracts of research papers that have successfully secured NSF funding, this study aims not only to identify prevalent themes but also to discern the keywords within those topics that attract the highest financial support.

Motivation for this research is deeply rooted in the transformative influence of Big Data across diverse industries, ranging from healthcare and finance to cybersecurity and beyond. As the NSF continues to be a key player in funding groundbreaking research, understanding the intricate patterns in funding allocation becomes imperative for researchers, policymakers, and industry stakeholders alike.

The methodological approach adopted for this study is comprehensive and multi-faceted. Initial steps involve the compilation of a comprehensive dataset encompassing NSF-funded research papers pertaining to Big Data, with a specific focus on abstracts as the primary source of textual information. The LDA model, known for its effectiveness in uncovering hidden topics in large datasets, is then applied to unveil latent topics within these abstracts. Each identified topic is characterized by a set of keywords, and the prevalence of these topics is quantified, providing a detailed map of overarching research themes within NSF-funded Big Data research. Furthermore, the analysis extends to discerning specific keywords within these topics that exhibit a correlation with higher funding levels, offering nuanced insights into the factors influencing financial support for Big Data research.

The anticipated findings of this study promise to provide a nuanced and holistic understanding of the research landscape, equipping researchers with valuable insights into the most promising avenues for securing funding in Big Data-related domains. By discerning keywords associated with higher funding, scholars can tailor their research proposals to align with funding priorities, thereby increasing the likelihood of securing coveted NSF grants.

Beyond the immediate benefits to researchers, the results will offer crucial insights for policymakers and funding agencies, guiding informed decisions on resource allocation and strategic planning. An understanding of the evolving dynamics in Big Data research

and funding patterns is fundamental for fostering innovation and propelling advancements in Big Data technologies.

The implications of this research extend far beyond the confines of academia, reaching into industry sectors that heavily leverage Big Data technologies. Businesses stand to gain valuable insights into emerging trends and focal points, aligning their strategies with the forefront of research and innovation. This collaboration between academia and industry ensures that funded research translates into tangible real-world applications, driving societal progress and economic development.

This project leverages the power of LDA topic analysis to unravel the intricate funding patterns within NSF-supported Big Data research. By dissecting the abstracts of funded papers, the study provides a comprehensive roadmap for researchers, policymakers, and industry stakeholders to navigate the expansive and dynamic landscape of Big Data research. This, in turn, fosters collaboration, drives innovation, and contributes significantly to the ongoing evolution of this critical and influential domain.

Introduction

In the epoch of information, where the digital footprint of humanity grows exponentially, the field of Big Data stands as a beacon of innovation and discovery. As organizations, industries, and scientific communities grapple with an ever-expanding volume of data, the impetus to decipher patterns, extract insights, and harness the transformative potential of Big Data has never been more critical. At the forefront of this dynamic landscape, the National Science Foundation (NSF) serves as a key catalyst, providing crucial financial support to researchers who endeavor to unravel the complexities of Big Data.

This project embarks on a comprehensive journey into the funding dynamics of Big Data research, focusing specifically on NSF grants. The overarching goal is to dissect and understand the intricate interplay between research topics, keywords, and funding patterns within the realm of Big Data. Employing the powerful analytical tool of Latent Dirichlet Allocation (LDA), this study aims to shed light on not only the prevalent themes in NSF-funded research but also the specific keywords that wield influence in securing financial support.

Big Data, characterized [1] by the 3Vs—Volume, Velocity, and Variety—represents a paradigm shift in how information is processed and leveraged. The interdisciplinary nature of Big Data research spans fields as diverse as computer science, statistics, machine learning, and domain-specific applications. From predicting disease outbreaks to optimizing supply chains, Big Data has become the linchpin of decision-making processes across various sectors.

The significance of Big Data is underscored by its pervasive impact on modern society. Healthcare systems leverage it for personalized medicine, financial institutions deploy it for risk management, and smart cities harness its capabilities for urban planning. As the applications burgeon, the need for cutting-edge research to address the challenges and harness the opportunities presented by Big Data becomes increasingly apparent.

As a leading federal agency dedicated to advancing scientific research across diverse domains, the NSF plays a pivotal role in nurturing innovation and pushing the boundaries of knowledge. The foundation's investments in Big Data research underscore a commitment to fostering breakthroughs that transcend disciplinary boundaries and address the multifaceted challenges posed by the era of Big Data.

NSF grants serve as a lifeline for researchers, providing the financial resources needed to conduct in-depth studies, develop novel methodologies, and push the boundaries of existing knowledge. In the context of Big Data, where the scale and complexity of research often demand substantial resources, NSF funding becomes a cornerstone for driving impactful discoveries.

In the intricate landscape of Big Data research, where the torrent of information flows ceaselessly, the role of funding, particularly from the National Science Foundation (NSF), is a linchpin for propelling scientific inquiry and technological innovation. While the significance of NSF funding in advancing research within the expansive domain of Big Data is indisputable, a nuanced understanding of the intricate funding dynamics remains elusive. It is this enigma that our project seeks to unravel—a quest driven by the imperative to provide profound insights into the patterns governing the allocation of financial resources within the vast and dynamic realm of Big Data research.

Researchers, policymakers, and industry stakeholders find themselves at the nexus of a burgeoning field where the potential for groundbreaking discoveries and technological advancements is immense. However, the journey from conceptualization to realization often hinges on securing the necessary financial support. It is in this context that our project emerges, propelled by the conviction that a comprehensive exploration of NSF funding dynamics holds the key to unlocking the full potential of Big Data research.

Motivated by the imperative to bridge the gap between funding allocation and research priorities, our project takes a holistic approach, delving into the intersection of research topics, keywords, and funding within the specific context of NSF-supported Big Data projects. The endeavor is not merely an academic exercise; it is a pursuit grounded in the practical implications that the findings hold for the diverse stakeholders invested in the trajectory of Big Data research.

The tapestry of funded research projects, woven through the allocation of NSF grants, forms a rich canvas that encapsulates the pulse of innovation and inquiry within the Big Data landscape. By immersing ourselves in the complexities of this tapestry, we aspire to go beyond mere

observation. Our goal is to construct a comprehensive roadmap—a navigational guide for researchers navigating the terrain of funding acquisition in Big Data.

For researchers, this roadmap becomes an invaluable tool, offering insights into the prevailing themes and critical keywords that resonate with funding agencies. Armed with this knowledge, researchers can tailor their proposals with a nuanced understanding of the funding priorities, increasing their chances of securing coveted NSF grants. The project thus becomes a beacon for those traversing the competitive landscape of grant applications, illuminating the path to successful funding endeavors.

Policymakers, entrusted with the responsibility of strategic decision-making and resource allocation, find in our project a wellspring of insights. Understanding the ebb and flow of funding patterns within Big Data research enables informed decision-making, guiding the formulation of policies that align with the evolving priorities of scientific inquiry. It is a tool for shaping the trajectory of research at a macro level, ensuring that resources are directed toward avenues that hold the promise of significant impact. Industry stakeholders, deeply enmeshed in the tapestry of technological advancements and innovation, stand to gain strategic advantages from the outcomes of our project. The intersection of academic research and industrial applications is a fertile ground for collaboration. Our project acts as a mediator, facilitating a symbiotic relationship between academia and industry. By offering insights into emerging trends and pivotal keywords, industry players can align their strategies with the forefront of research and innovation, fostering a collaborative ecosystem that propels both sectors forward.

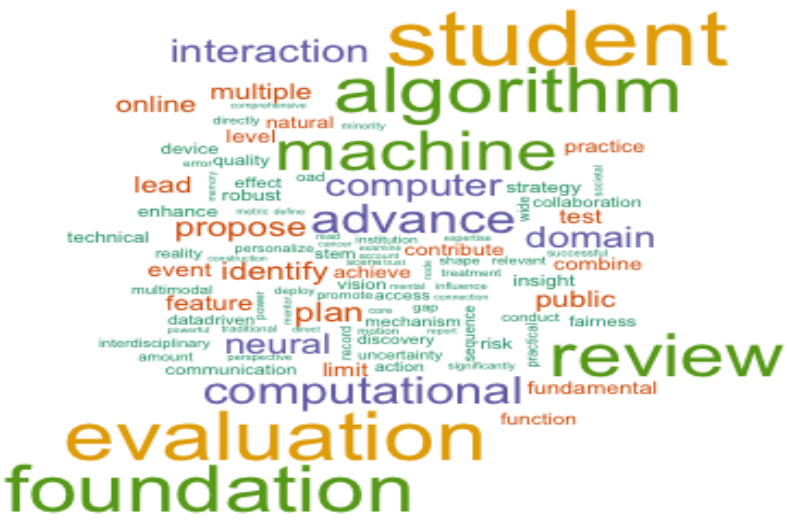


Fig 1. Initial Word Cloud after Data Preprocessing

The project is not a mere exploration of data; it is a commitment to decoding the language of funding dynamics within NSF-supported Big Data research. It is an odyssey into the heart of innovation, guided by the mission to provide tangible benefits for researchers, policymakers, and industry stakeholders. As we embark on this journey, the anticipation is not just for unraveling patterns but for catalyzing a ripple effect that transforms the landscape of Big Data research, making it more accessible, impactful, and collaborative for all.

The methodology employed in this study is designed to be rigorous and comprehensive, leveraging the capabilities of Latent Dirichlet Allocation (LDA) to extract meaningful insights from a vast corpus of NSF-funded Big Data research papers.

Data Compilation:

The initial step involves the compilation of a robust dataset comprising NSF-funded research papers in the field of Big Data. This dataset serves as the foundation for our analysis, with a specific focus on abstracts as they encapsulate the essence of the research and provide a rich source of textual information.

LDA Topic Analysis:

LDA, a probabilistic model [2] for uncovering latent topics in large datasets, serves as the analytical backbone of this project. Applied to the abstracts of NSF-funded research papers, LDA facilitates the identification of latent topics and their associated keywords. This allows us to discern the prevalent themes within funded research projects and quantify the prevalence of each topic in the corpus.

Keyword-Funding Correlation Analysis:

Taking the analysis a step further, our study seeks to identify specific keywords within the identified topics that exhibit a correlation with higher funding levels. This granular exploration aims to uncover the factors that influence funding allocation, providing researchers with valuable insights into tailoring their proposals to align with funding priorities.

The anticipated findings of this study are poised to unfurl a tapestry of insights with far-reaching implications across the spectrum of Big Data research stakeholders. For researchers navigating the competitive landscape of grant applications, the outcomes promise a treasure trove of knowledge—a nuanced understanding of prevalent topics and influential keywords that can intricately shape the crafting of research proposals. Armed with this knowledge, researchers can strategically align their endeavors with the priorities of NSF funding, enhancing the likelihood of securing crucial financial support.

Policymakers and funding agencies, entrusted with the formidable task of steering the course of scientific inquiry, find in these anticipated findings a compass for informed decision-making. Insights into the evolving research priorities within the Big Data domain become pivotal tools in guiding strategic decisions on resource allocation and program development. The study acts as

a dynamic lens, allowing policymakers to adjust their focus in tandem with the ever-shifting currents of innovation, ensuring that funding aligns with the pulse of cutting-edge research.

The industry, as a significant beneficiary of Big Data innovations, stands at the threshold of strategic advantages that the study's outcomes offer. The dynamics of funded research projects unfold as a roadmap, providing a panoramic view of emerging trends and influential keywords. Businesses, by deciphering this map, can strategically position themselves at the forefront of innovation. The study becomes a compass for industry players, guiding them to align their strategies with the ever-evolving landscape of research, fostering collaborations with academia that are mutually beneficial.

As this project embarks on a comprehensive exploration of NSF grants within the domain of Big Data research[3], the objective is not merely to observe but to contribute valuable insights that act as catalysts for change. The intricate interplay between research topics, keywords, and funding patterns becomes the focal point of our analysis. Delving into the funding dynamics is not an isolated endeavor; it is a commitment to shaping the trajectory of Big Data research.

Our goal extends beyond the confines of academia. It is a commitment to fostering collaboration, driving innovation, and catalyzing advancements in this critical domain. By unraveling the threads of funding dynamics, we seek to create a narrative that inspires collaboration between researchers, policymakers, and industry stakeholders. This collaboration is the crucible where ideas transform into solutions, and theoretical concepts find real-world applications.

The subsequent sections of this study will be a deep dive into the detailed analysis and findings, peeling back the layers to offer a nuanced understanding of the funding landscape within NSF-supported Big Data research. Each section is a step closer to demystifying the complex relationships between research topics, keywords, and funding patterns. It is an exploration that transcends data points and statistical analyses—it is a journey into the heart of innovation, where the implications of our findings resonate not just in academic corridors but reverberate across industries and shape the future landscape of Big Data research.

Methodology

This intricate process was undertaken to decode the funding dynamics within the National Science Foundation (NSF) grant data from the Division of Information and Intelligent Systems Organization. Spanning the years 2018 to 2022, this methodology employs a meticulous approach, encompassing data preprocessing and advanced Natural Language Processing (NLP) techniques. Each step is accompanied by an explanation of relevant NLP terms, providing a thorough understanding of their significance in the research context.

Data Preprocessing:

The initial phase of the methodology is dedicated to preparing the dataset for in-depth analysis. Data preprocessing involves several key steps aimed at cleaning and structuring the data to facilitate meaningful exploration.

1. Data Selection:
The dataset under scrutiny comprises NSF grant fund data from the Division of Information and Intelligent Systems Organization for the years 2018 to 2022. This specific temporal and organizational focus ensures relevance to the domain of interest.

2. Encoding Conversion:
Explanation:
Encoding refers to the character encoding of text data. Converting the dataset encoding to UTF-8 ensures uniformity in character representation and facilitates consistent processing of textual information.

Significance:
Diverse characters may exist in textual data, and different encodings can represent these characters differently. Standardizing to UTF-8 mitigates potential issues arising from varied character representations.

3. Creating a Corpus:
Explanation:
A corpus is a collection of textual data. In this context, the corpus is created from the Abstract column of the dataset, serving as the primary source of textual information for subsequent NLP analyses.

Significance:
Building a corpus is a fundamental step in NLP. It consolidates text data into a format suitable for analysis, enabling the extraction of meaningful patterns and insights.

4. Removing Stopwords:
Explanation:
Stopwords are common words with limited semantic value (e.g., "and," "the," "is"). Removing stopwords from the corpus enhances the relevance of subsequent analyses by focusing on content-carrying words.

Significance:
Stopwords are ubiquitous but often lack meaningful content. Removing them streamlines the dataset, allowing analyses to concentrate on terms carrying more substantive information.

5. Cleaning Abstracts:
Explanation:

Cleaning involves removing special characters, numbers, and extra white spaces from the Abstract column. Additionally, all character values are converted to lowercase for uniformity.

Significance:
Cleaning abstracts standardizes the text data, eliminating noise and ensuring consistency. Converting characters to lowercase avoids discrepancies in subsequent analyses due to case variations.

6. Creating a Tibble:
Explanation:
A tibble is constructed, including essential components such as award ID, abstract, and awarded amount. The Abstract column of this tibble serves as the primary text corpus for subsequent analysis.

Significance:
Structuring the data into a tibble facilitates organized and efficient handling of relevant components. It prepares the dataset for further exploration while maintaining essential information.

7. Bigram Creation:
Explanation:
Bigrams represent pairs of adjacent words in a sequence. Generating bigrams from the Abstract column provides insights into co-occurring terms, revealing potential associations within the research abstracts.

Significance:
Bigrams offer a more nuanced understanding of language usage by capturing word pairs. This enhances the context of the analysis, allowing for a deeper exploration of semantic relationships.

8. Stopword Removal from Bigrams:
Explanation:
Removing stopwords from the generated bigrams refines the dataset further, enhancing the relevance of subsequent analyses.

Significance:
Extending stopwords removal to bigrams maintains focus on meaningful content. It is a continuation of the effort to reduce noise and emphasize content-carrying terms.

9. Tokens Unnesting:
Explanation:
Unnesting involves splitting the bigrams and extracting tokens from the Abstract column. This process facilitates a granular exploration of individual terms within the abstracts.

Significance:
Tokens represent the basic units of text. Unnesting bigrams into tokens allows for a more detailed examination of individual terms, contributing to a comprehensive analysis.

10. Lemmatization:
Explanation:
Lemmatization transforms words to their base or root form. This step reduces variations of words to a common base, aiding in more accurate analysis.

Significance:
Lemmatization enhances the efficiency of subsequent analyses by consolidating different forms of words. It ensures that variations of a word are treated as a single entity.

11. Updating Text Corpus:
Explanation:
Post-lemmatization, the text corpus is updated to reflect the lemmatized representation of the textual data. This step ensures subsequent analyses are based on a refined and standardized dataset.

Significance:
Updating the text corpus with lemmatized terms ensures consistency and accuracy in downstream analyses. It aligns the dataset with the standardized form of words.

12. Document Term Matrix (DTM) Creation:
Explanation:
A Document Term Matrix (DTM) is constructed using the bag of words method. This matrix provides a structured representation of the frequency and importance of terms within the dataset.

Significance:
DTM is a foundational element in text analysis. It quantifies the occurrences of terms, facilitating the exploration of patterns and relationships between documents and terms.

Fig 2. Word Cloud after Bag of Words calculations

Natural Language Processing (NLP) Techniques:

With the dataset appropriately preprocessed, the methodology transitions to advanced NLP techniques to extract meaningful insights from the grant abstracts.

1. Determining Optimal Number of Topics:
Explanation:
Ldatuning methods are employed to ascertain the optimal number of topics based on different evaluation metrics. The harmonic mean of log-likelihoods is calculated to identify the number of topics that maximizes coherence across the dataset.

Significance:
Determining the optimal number of topics is crucial for a focused analysis. Ldatuning methods use statistical metrics to guide the selection, ensuring meaningful and coherent topics are identified.

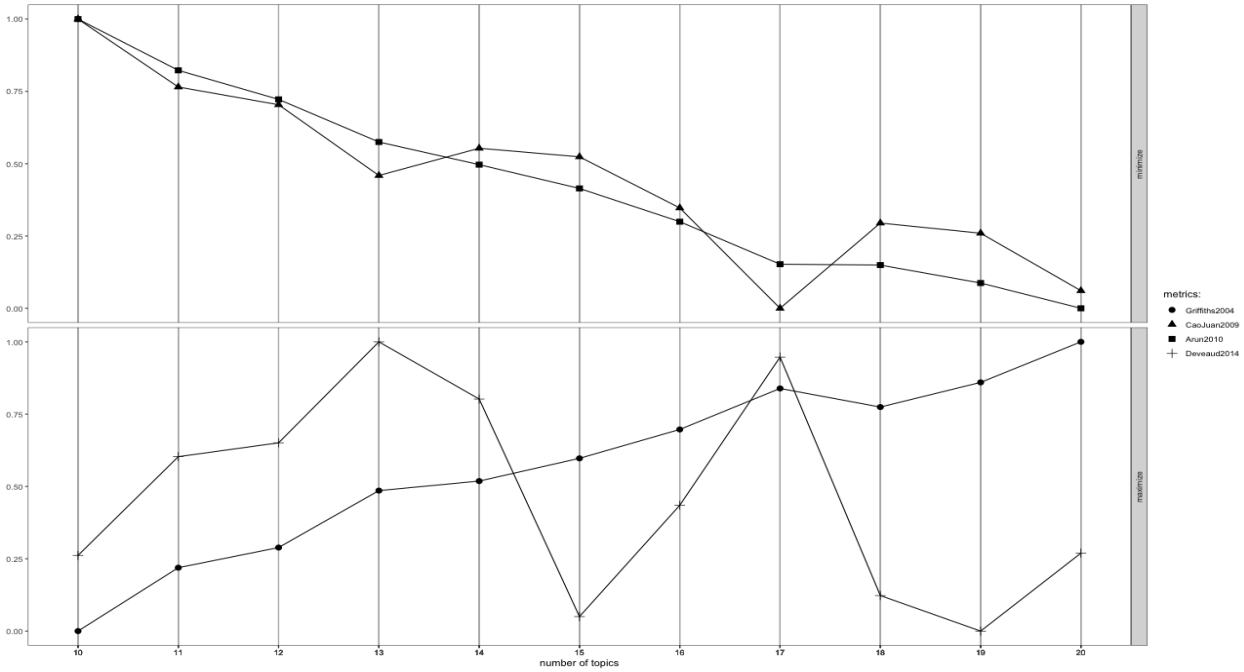


Fig 3. Optimal Topic Graph using ldatuning package

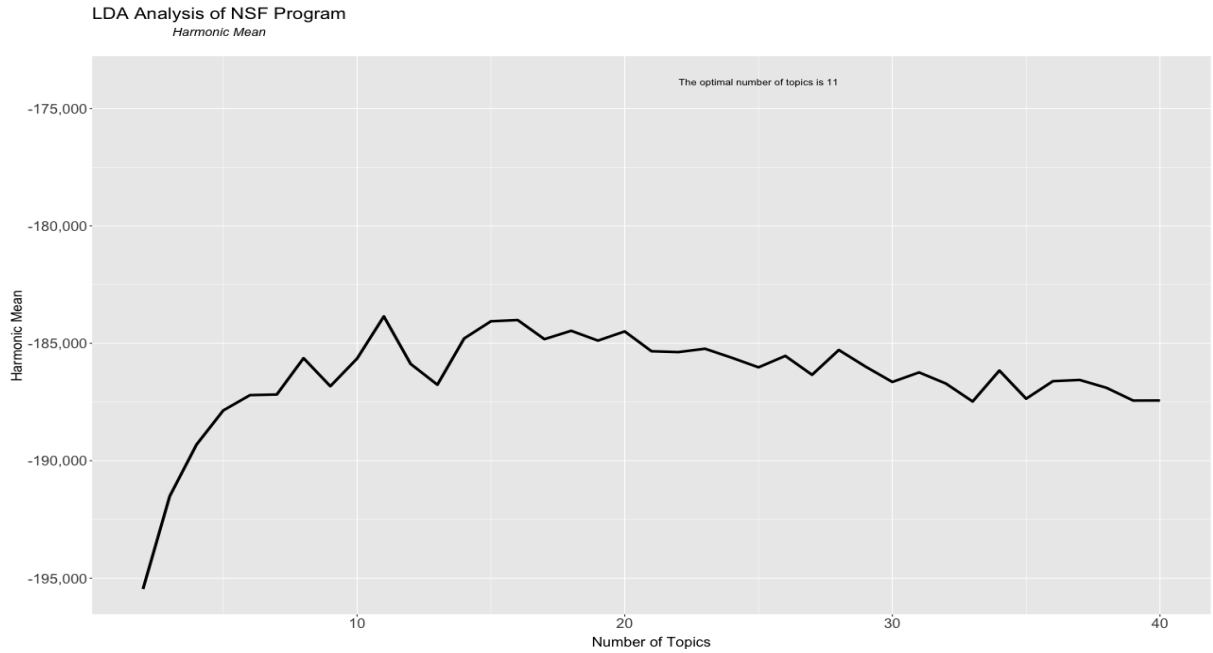


Fig 4. Optimal Topic Analysis using Harmonic Mean

2. LDA Model Implementation:

Explanation:

The Latent Dirichlet Allocation (LDA) model is implemented with the determined optimal number of topics, set at 11 in this case. LDA is a generative probabilistic model that uncovers latent topics within a collection of documents.

Significance:
LDA is a powerful technique for topic modeling. It identifies underlying themes within documents, providing a structured way to understand the distribution of topics across the dataset.

3. Top Words Extraction:
Explanation:
The top 10 words for each of the 11 topics are extracted from the LDA model. These words represent the most influential terms within each topic and are concatenated to provide a concise representation of the thematic elements.

Significance:
Extracting top words from each topic condenses the information, offering a snapshot of the key terms defining each thematic area. It aids in the interpretation and labeling of topics.

Fig 5. Word Cloud after Bag of Words calculations

Explanation of NLP Terms:

Now, let's delve into the detailed explanation of each NLP term employed in the methodology:

1. Bag of Words (BoW):
Explanation:
Bag of Words[4] is a common method in NLP that represents text as an unordered set of words, disregarding grammar and word order but focusing on the frequency of individual words.

Significance:
BoW simplifies the complexity of language into a structured format, allowing the analysis of word frequencies across documents. It serves as the foundation for creating a Document Term Matrix.

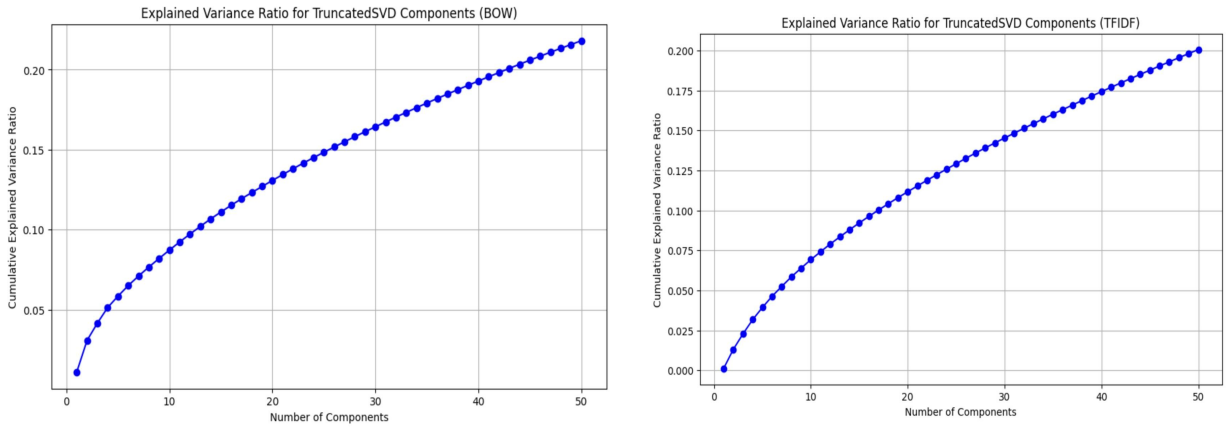


Fig 6. Comparison of Bag of Words and TF-IDF methods for information retention

2. Bigrams:
Explanation:
Bigrams are pairs of adjacent words in a sequence. Analyzing bigrams helps understand the relationships and co-occurrences between words, providing more context than analyzing individual words.

Significance:
Bigrams capture contextual information by considering pairs of words. This enhances the understanding of how terms interact within the abstracts, contributing to a more nuanced analysis.

3. Stopwords:
Explanation:
Stopwords are common words that are often filtered out during NLP analyses as they carry little semantic meaning. Examples include articles, prepositions, and conjunctions.

Significance:
Filtering out stopwords streamlines analyses by focusing on content-bearing words. It improves the accuracy of subsequent analyses by removing frequently occurring but less informative terms.

4. Tokens:
Explanation:
Tokens are individual units of text, typically words. During NLP analyses, text is often tokenized, breaking it down into these basic units for further analysis.

Significance:
Tokenization is a fundamental step that enables the analysis of individual words. It forms the basis for various NLP techniques, allowing a detailed examination of text data.

5. Lemmatization:
Explanation:
Lemmatization is the process of reducing words to their base or root form. It helps in reducing variations of words to a common base, aiding in more accurate analysis.

Significance:
Lemmatization ensures consistency in the treatment of words by reducing them to their essential form. It addresses variations in word forms, enhancing the accuracy of subsequent analyses.

6. Document Term Matrix (DTM):
Explanation:
A Document Term Matrix is a mathematical matrix that represents the frequency of terms in a collection of documents. Rows represent documents, and columns represent terms, with each cell containing the frequency of a term in a document.

Significance:
DTM quantifies the occurrence of terms across documents, enabling the exploration of patterns and relationships. It provides a structured representation for further analysis.

7. Latent Dirichlet Allocation (LDA):
Explanation:
LDA is a probabilistic model used for topic modeling. It assumes that each document is a mixture of a small number of topics, and each word's presence is attributable to one of the document's topics.

Significance:
LDA uncovers latent topics within a collection of documents, providing a structured way to understand the distribution of topics across the dataset. It is pivotal for thematic analysis.

This comprehensive methodology outlines a meticulous and multi-step approach, from data preprocessing to advanced NLP techniques, in analyzing NSF grant fund data. Each step is carefully explained to provide a deeper understanding of its significance in the research context. The subsequent sections of this research paper will delve into the detailed analysis and findings, offering a nuanced understanding of the funding landscape within NSF-supported Big Data research.

Results

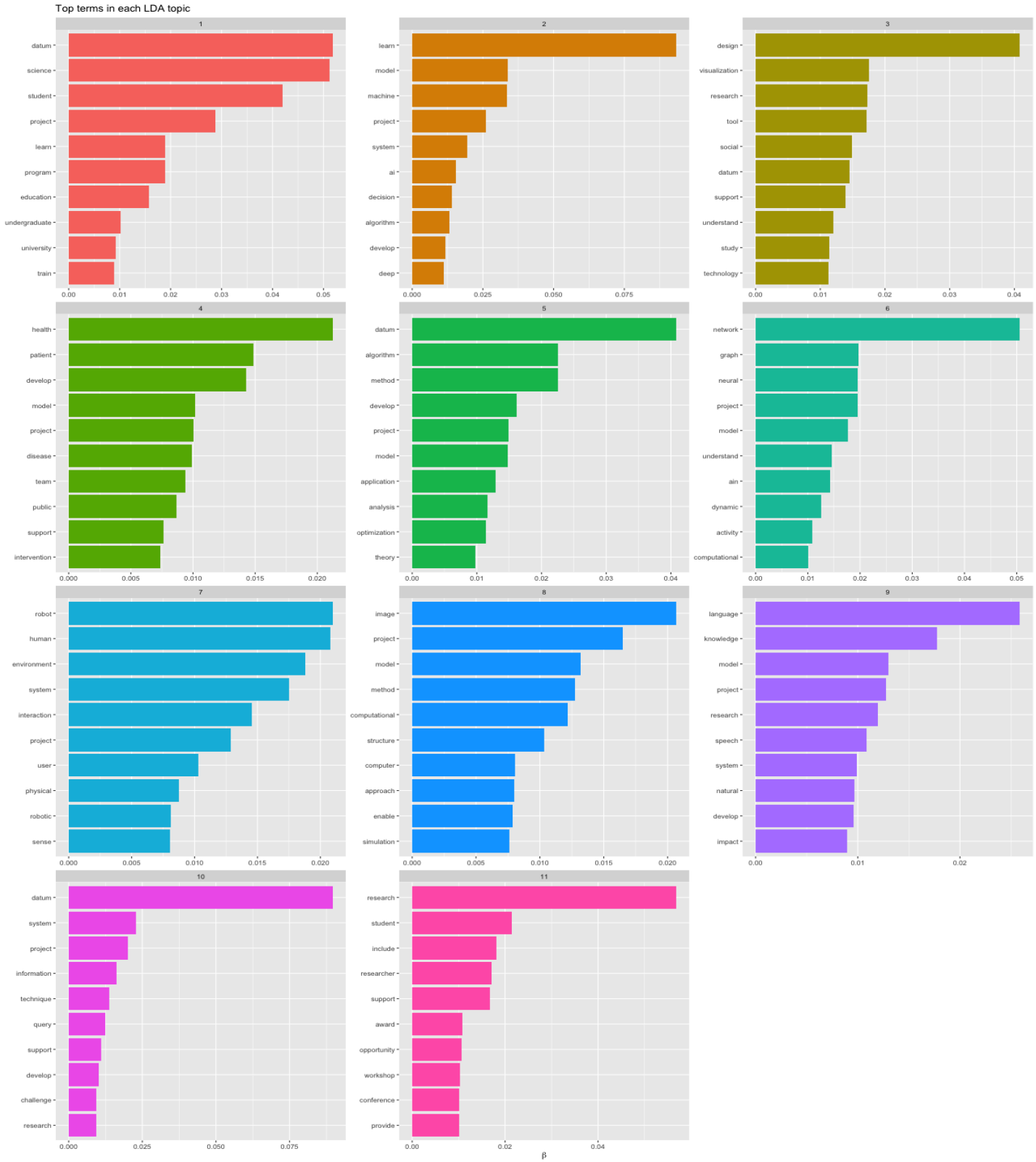
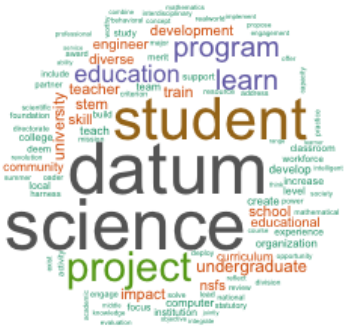
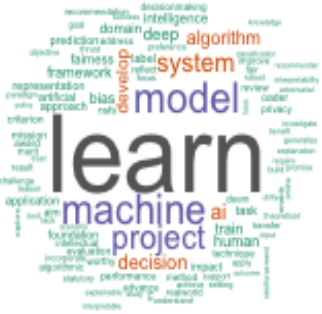


Fig 7. Top Terms in each LDA Topic

In the pursuit of unraveling the funding dynamics within NSF grant data from the Division of Information and Intelligent Systems Organization (2018-2022), the analysis extends beyond numerical summaries. Visual representations play a pivotal role in interpreting the nuances of the latent topics identified through the Topic Modeling process. This section delves into the insights provided by Word Clouds for each of the 11 topics and an Intertopic Distance Map of the 30 most salient terms.





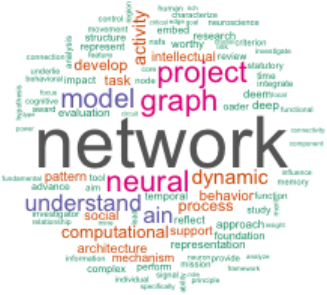


Fig 8. Word Cloud From Topic 1 to Topic 11

Word Clouds serve as visual summaries, portraying the prominence of words within each topic based on their frequency. Larger and bolder words indicate higher occurrence, offering an intuitive glimpse into the thematic essence of each topic.

Topic 1:

Words like "Science," "datum," and "Student" dominate the cloud.
Interpretation: This topic revolves around the core concepts of education on Big Data.

Topic 2:

Noteworthy terms include "learning," "machine," and "algorithm."
Interpretation: Machine learning algorithms and methodologies are central to the theme of this topic.

Topic 3:

Words like "algorithm," "design," and "model" dominate the cloud.

Interpretation: This topic revolves around the core concepts of data modeling and algorithm development.

Topic 4:

Keywords like "health," "care," and "data" are emphasized.

Interpretation: This topic is oriented towards research at the intersection of healthcare and data analytics.

Topic 5:

Prominent terms include "optimization," "algorithm," and "network."

Interpretation: Optimization algorithms and network-related research are central themes within this topic.

Topic 6:

Dominated by terms like "sensor," "network," and "data."

Interpretation: The topic emphasizes research related to sensor networks and data collection.

Topic 7:

Keywords like "Environment," "Robot," and "human" stand out.

Interpretation: This topic centers on research related to Robotics.

Topic 8:

Prominent words include "language," "processing," and "natural."

Interpretation: Natural language processing (NLP) appears to be a key theme within this topic.

Topic 9:

Key terms like "image," "processing," and "algorithm" feature prominently.

Interpretation: Image processing algorithms and methodologies are focal points within this topic.

Topic 10:

Notable words include "knowledge," "representation," and "semantic."

Interpretation: The topic is centered on the representation and semantic understanding of knowledge.

Topic 11:

Keywords like "research," "Development," and "opportunity" stand out.

Interpretation: This topic centers on research related to Big Data Research.

The Intertopic Distance Map[7] illustrates the proximity or similarity between different topics based on the 30 most salient terms. Closer positioning indicates higher semantic similarity.

Proximity Insights:

- Topics 8, 9, and 2 are closely clustered, suggesting shared thematic elements. This proximity aligns with the prominence of terms like "algorithm" and "data" in these topics, indicating a common focus on algorithmic and data-centric research.
- Topics 5 and 10 also exhibit proximity, indicating potential connections between research on security/privacy systems and knowledge representation/semantic understanding.
- Topic 7, centered on robotics, is relatively isolated, emphasizing its unique thematic characteristics.

Keyword Overlaps:

The overlap of terms across multiple topics, such as "algorithm" and "data," underscores the interdisciplinary nature of Big Data research.

Distinctive Topics:

The clear separation of Topic 9 (Image processing) and Topic 7 (Robotics) from other clusters highlights their distinct thematic focuses.

Strategic Collaboration:

Proximity in the Intertopic Distance Map suggests potential interdisciplinary collaboration opportunities. Researchers exploring algorithmic and data-centric themes may find common ground for collaboration.

Identifying Research Niches:

The isolation of Topic 7 (Robotics) and clear separation of Topics 9 and 7 indicate specialized research areas. This insight can guide researchers and stakeholders in identifying niche domains for further exploration.

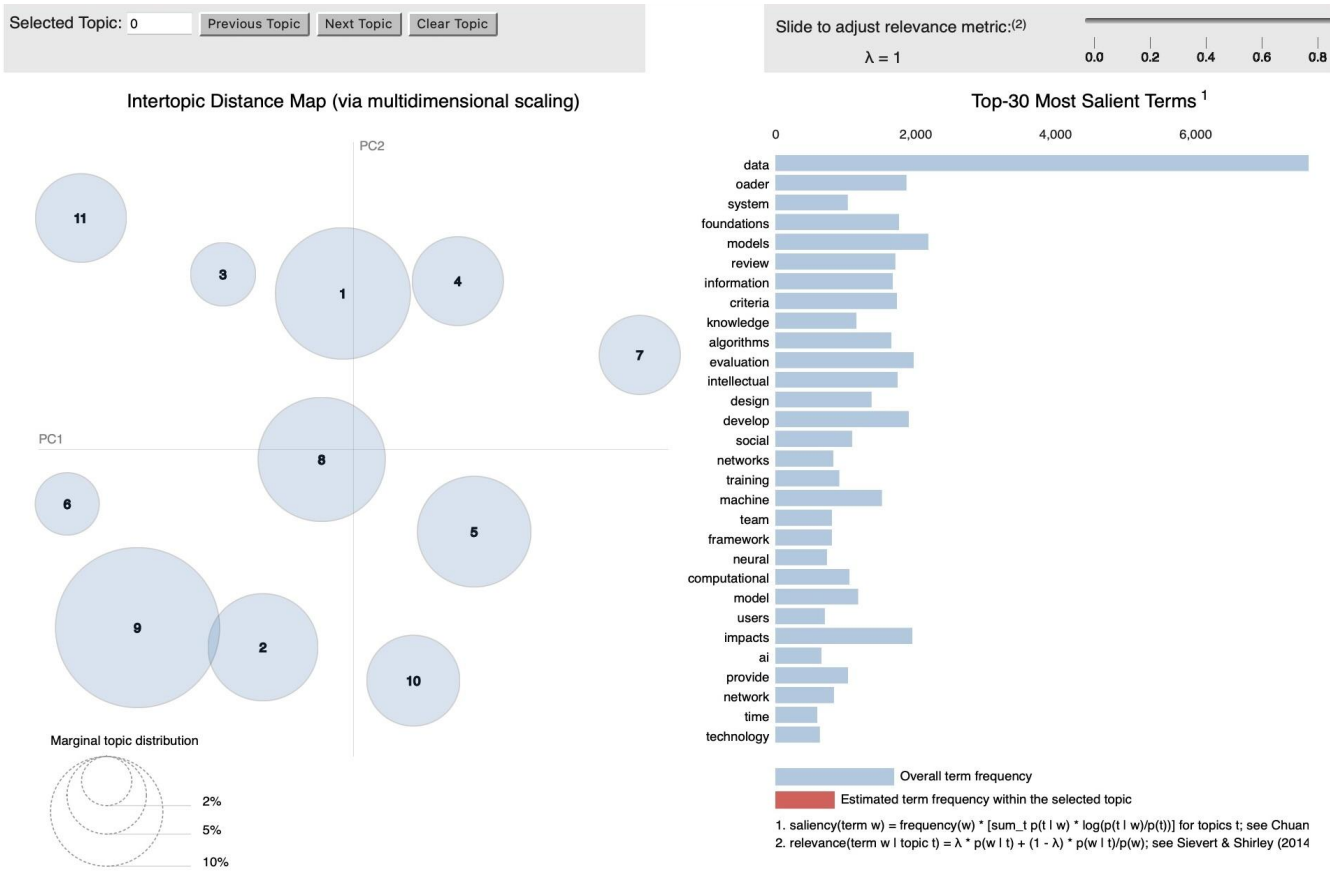


Fig 9. LDA Intertopic Distance Map

Conclusion

In concluding the Topic Modeling Analysis employing Latent Dirichlet Allocation (LDA), a journey through the intricate landscape of NSF grant data from the Division of Information and Intelligent Systems Organization unfolds. The LDA model, a powerful tool in uncovering latent topics within a collection of documents, has been adeptly utilized to distill the richness of this dataset into distinct thematic realms.

The heart of the analysis lies in the optimal determination of the number of topics, a crucial aspect that significantly influences the interpretability and coherence of the findings. Leveraging the ldatuning package methods and harmonic mean calculations, a meticulous exploration led to the pinpointing of the optimal number of topics. This precision ensures that the identified topics are not only statistically robust but also align with the inherent structure and content of the grant abstracts. The choice of an optimal number of topics serves as a cornerstone, enabling meaningful insights and actionable inferences.

The creation of keyword visualizations, specifically Word Clouds, emerges as a pivotal aspect in translating the abstract topics into tangible, interpretable narratives. These visual summaries encapsulate the essence of each topic, with the prominence of specific terms conveying the thematic focus. The nuanced nature of Big Data research is brought to the forefront as these Word Clouds unveil the core concepts and recurring themes within the funded projects. It is through these visual representations that researchers, policymakers, and industry stakeholders gain a bird's eye view of the diverse domains encapsulated in the grant abstracts.

The methodological choice of the Bag of Words method for constructing the document term matrix underscores a strategic decision aimed at optimizing information retention. This choice is rooted in the understanding that the Bag of Words method outperforms TF-IDF in this specific context. The document term matrix, a foundational element in the LDA analysis, serves as the canvas on which the topics are painted. The meticulous construction of this matrix ensures that the subsequent analysis is grounded in a robust representation of the textual data, allowing for nuanced exploration.

The exploration of topics extends beyond individual themes to their relationships and interconnectedness. The Intertopic Distance Map emerges as a cartographer, sketching the landscape of semantic proximity and thematic overlap. The visualization of topic clusters and their spatial relationships offers a holistic understanding of the broader research ecosystem. Proximity hints at potential collaborative opportunities and interdisciplinary intersections, guiding researchers and stakeholders in navigating the complex terrain of Big Data research.

In the grand tapestry of NSF-supported Big Data research, these visualizations act as portals, inviting stakeholders to peer into the diverse realms explored by funded projects. The insights derived from Word Clouds and the Intertopic Distance Map transcend mere visual appeal; they serve as beacons illuminating strategic paths for researchers seeking collaboration, policymakers making informed decisions, and industry stakeholders aligning their strategies with the cutting edge of innovation.

As the analysis concludes, the kaleidoscope of topics, keywords, and their relationships within NSF grant data reflects the vibrant and ever-evolving nature of Big Data research. The journey through the labyrinth of abstracts, guided by the meticulous interplay of methodologies and visualizations, sets the stage for a deeper understanding of funding dynamics, fostering collaborative endeavors, and propelling the trajectory of innovation in this critical domain.

Challenges Faced

Embarking on the data cleaning phase, we encounter a formidable challenge, one that resonates with the inherent complexity of dealing with long and intricate texts within abstracts. The inherent difficulty lies in navigating through extensive textual information, where nuances and patterns are deeply embedded. As we grapple with the expansive nature of abstracts, extracting precise requirements that seamlessly integrate with the Latent Dirichlet Allocation (LDA) model becomes a demanding task. The sheer length of these abstracts adds layers of intricacy, making it imperative to delve into methods that not only explore but also cleanse the data with heightened accuracy.

Long-text data poses unique challenges in terms of exploration and pattern identification. The conventional approaches to data cleaning may need augmentation to accommodate the intricacies of extensive textual information. The quest for methodologies to scrutinize and cleanse such lengthy data emerges as a critical pursuit, aiming not only to streamline the data for analysis but also to enhance the precision and relevance of the subsequent LDA model.

In navigating the labyrinth of long-text data, considerations extend beyond conventional cleaning techniques. The intricacies inherent in the language of abstracts, often laden with specialized terminology and diverse structures, demand a nuanced approach. Methods to distill and refine this textual richness without sacrificing key information become the linchpin of effective data cleaning.

Recognizing the pivotal role of data cleaning as a differentiator in achieving superior results and outputs in any data project, there is a call for innovation in this phase. The efficacy of the LDA model, which hinges on the quality and relevance of the input data, is inherently tied to the success of the data cleaning process. An elevated level of precision in data cleaning becomes a catalyst for unlocking the latent insights hidden within the abstracts.

The challenges posed by long-text data are not impediments but opportunities for refinement and innovation. Exploring advanced methods that can unravel the intricacies of language, identifying and preserving critical information, is pivotal. Techniques such as natural language processing (NLP) and advanced text analysis may offer valuable insights and solutions to the multifaceted challenges presented by lengthy abstracts.

Data cleaning, in the context of extensive textual information, emerges not only as a preparatory step but as a strategic differentiator in the trajectory of a data project. It transcends the conventional notions of cleaning by becoming a conduit for enriching the data, ensuring that the subsequent analysis is not only rigorous but also reflective of the true nuances embedded within the abstracts.

In conclusion, the journey through the data cleaning phase is a dynamic exploration, a quest for methodologies that can unravel the complexities of long-text data. It is a testament to the

recognition that precision in data cleaning is a linchpin for success in any data project. The challenges posed by the length and intricacy of abstracts become waypoints guiding us toward innovative solutions that not only clean but also enhance the richness of the textual data, setting the stage for a more impactful and nuanced LDA analysis.

Future Prospects

Positioned as a comprehensive reference guide, this project stands as a beacon for researchers, policymakers, and industry stakeholders seeking to navigate the ever-evolving landscape of Big Data research. Beyond its role as a mere analysis, this endeavor emerges as a valuable resource, offering insights that extend far beyond the confines of abstract topics and keyword distributions. The project's utility transcends the academic realm, presenting itself as a compass for those venturing into the domain of research grant proposals.

As a reference guide, this project becomes an indispensable tool for researchers aiming to craft proposals that align with emerging topics within the National Science Foundation (NSF) funding landscape. By distilling and visualizing the latent themes within funded projects, the project provides a roadmap for exploration. Researchers can leverage the identified topics and keywords as guiding beacons, aligning their proposals with the pulse of ongoing research, and positioning their work at the forefront of innovation.

The implications extend beyond individual research endeavors; the project serves as a strategic resource for shaping funding priorities. Policymakers and funding agencies can glean valuable insights into the evolving landscape of Big Data research. The identified topics become a compass, guiding decisions on resource allocation, program development, and strategic investments. By aligning funding priorities with the identified themes, policymakers ensure that financial resources are directed toward research that not only addresses current challenges but also anticipates future trends.

Collaboration is a cornerstone of research progress, and this project plays a pivotal role in fostering potential collaborations. By highlighting the interconnectedness of certain topics or the proximity of research clusters, the project becomes a matchmaker, bringing together researchers with shared interests and complementary expertise. The identified themes serve as common ground, facilitating interdisciplinary collaboration and the pooling of resources for more impactful and holistic research outcomes.

The project's value extends to industry stakeholders keen on aligning their strategies with the cutting edge of research and innovation. By understanding the funding dynamics and emerging topics, businesses operating in the Big Data domain can position themselves strategically. The project offers a lens into the future, allowing industry stakeholders to anticipate trends, identify areas for potential collaboration with academia, and tailor their strategies to align with the forefront of research.

Furthermore, the project acts as a catalyst for areas of further exploration. Researchers and stakeholders can use the identified topics as a springboard for deeper dives into specific domains, unveiling new avenues for inquiry and discovery. The project, in this sense, is not a static endpoint but a dynamic catalyst for continuous exploration and advancement within the realm of Big Data research.

In conclusion, this project transcends the boundaries of a typical analysis, transforming into a multifaceted reference guide with far-reaching implications. It empowers researchers with insights for crafting impactful grant proposals, guides policymakers in shaping funding priorities, facilitates collaborations among like-minded researchers, and provides industry stakeholders with a strategic roadmap for innovation. As a living resource, the project not only captures the current state of Big Data research but also propels the community toward future breakthroughs and discoveries.

References-

[1] De Mauro, Andrea & Greco, Marco & Grimaldi, Michele. (2016). A formal definition of Big Data based on its essential features. Library Review. 65. 122-135. 10.1108/LR-06-2015-0061.

[2] Stanford University. (2003). Latent Dirichlet Allocation. https://ai.stanford.edu/~ang/papers/jair03-lda.pdf

[3] Jelodar, Hamed & Wang, Yongli & Yuan, Chi & Feng, Xia. (2017). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey.

[4] Qader, Wisam & M. Ameen, Musa & Ahmed, Bilal. (2019). An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges. 200-204. 10.1109/IEC47844.2019.8950616.

[5] Roman Egger, Joanne Yu. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, Volume(Issue), Article Number. https://www.frontiersin.org/articles/10.3389/fsoc.2022.886498/full

[6] Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63-70). Stanford University. https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf

[7] Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 74-77). https://vis.stanford.edu/files/2012-Termite-AVI.pdf

[8] National Science Foundation.Advance Award Search. https://www.nsf.gov/awardsearch/advancedSearch.jsp

[9] Professor Heidorn. Title of the document. Retrieved from URL. https://docs.google.com/document/d/1QfOOzFrDJvpSEJvZLgGp-fj8gzrxlzm7032GM0nwiqg/ed it

[10] Professor Heidorn, B. 2022. TopicAnalysis. GitHub Repository. https://github.com/BryanHeidorn/TopicAnalysis

[11] Professor Heidorn, B. (2022). Graduate_Education-Topic-Analysis. GitHub Repository. https://github.com/BryanHeidorn/Graduate_Education-Topic-Analysis