Clearing the Skies: Data Science Propels Drone Research

ABSTRACT

As drone technologies rapidly evolve, the National Science Foundation (NSF) has emerged as a critical funding agency, having sponsored over $35 million in drone research across over 300 projects in the past decade alone. However, assessing portfolio wide trends is challenged by fragmented records, with collaborative grants dispersed across programs and duplicate entries for joint projects. This research tackles this fragmentation by leveraging latent semantic analysis (LSA) and latent Dirichlet allocation (LDA) – two powerful textual analysis techniques – to consolidate listings and illuminate funding patterns.

This project puts forward an extensive framework to leverage data from the National Science Foundation (NSF) to progress drone and unmanned aerial vehicle (UAV) research. A core component is consolidating duplicate records of NSF funding for drone/UAV projects. With the funding data unified, two text analysis techniques – Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) – can then be applied to assess prevalent topics and concepts receiving NSF backing. Mapping out current high interest subjects via LSA and LDA delivers strategic insights researchers can utilize to position proposals and frame projects around trends that show funding potential.

The first technical contribution is an inference engine that programmatically matches duplicate drone project entries in the corpus of NSF grant abstracts and titles. Exact string-matching algorithms identify identical records, while fuzzy matching captures near duplicates. Named entity recognition and rule based logic consolidate these variant listings into unified profiles reflecting aggregate drone funding. Cluster analysis groups related projects based on shared vocabulary and concepts. In total, over 150 consolidated project profiles are generated, capturing funding flows across siloed records.

Additionally, using the consolidated records, funded endeavors delving into similar technical areas or ethical considerations can be tracked over time. This allows the NSF and other government agencies to provide research monies to monitor drone/UAV project portfolios at a macrolevel. Identifying clusters of past and present initiatives exploring parallel themes or technologies facilitates improved program planning and allocation of funds to crucial spaces requiring attention.

With a unified corpus, LSA and LDA modeling quantitatively assess semantic relationships and extract interpretive topics. LSA compares drone concept vocabulary across grants to quantify contextual similarity. Projects are then mapped in a vector space to visualize research domains with funding traction. LDA complements this analysis by clustering grants into topics defined by specialized keywords. Examining prevalent topics and granular vocabulary provides a holistic overview of active funding areas – from precision agriculture drones to resilient swarm systems. Tracking grants over time also illuminates rising trajectories and seminal projects spurring new offshoots.

At its core, this framework systematically organizes disjointed NSF funding data that would otherwise be challenging for stakeholders to compile and make sense of independently. Structuring copious records into a cohesive resource removes barriers and opens opportunities to guide responsible drone/UAV innovation. Researchers gain bigger picture visibility that educates grant applications. Funding agencies obtain a consolidated landscape to strategize support dollars. Collectively, transparency and understanding are radically amplified, enabling progress to accelerate projects tackling pressing drone/UAV technical difficulties and ethical dilemmas holding back real world deployment.

Finally, tailored assessments can filter analysis to specific divisions, elucidating trends by key funding agencies. For example, LSA comparisons and LDA topics specialized to the Division of Information and Intelligent Systems unveil surging priorities around autonomous navigation, swarm coordination, and beyond line-of-sight controls. Together, these integrated techniques transform fragmented records into a unified resource for identifying extensions and gaps in portfolios.

A close-up of words

Description automatically generated  
**Wordcloud is based on Award amount for research.**

INTRODUCTION

The advent of drone technology has unlocked game changing potential across far-ranging domains, from infrastructure monitoring to precision agriculture. However, this vast promise is threatened by fragmented funding landscapes that obstruct a comprehensive perspective. Grants related to drones remain scattered across databases, distributing records inefficiently. This siloed arrangement hampers strategic investment guided by holistic insights into the technological and societal capabilities of emerging drone systems.

By consolidating grants through artificial intelligence powered semantic analysis, this research overcomes limited viewpoints enforced by fractured data. Employing techniques like Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), the project transforms the once dispersed corpus into an integrated map illuminating relationships and trajectories essential for continued advancement of drone technology.

LSA uncovers latent semantic structures woven through grants from diverse agencies and specialty areas. The analysis exposes unanticipated conceptual connections challenging conventional notions of fragmentation. LDA then clusters grants into underlying topics, with each topic centered on frequently cooccurring terms that enable intuitive interpretation.

Together, the quantitative modeling coalesces grants plagued by scattering into a unified landscape. Funding priorities become quantifiable, with specialized agency needs rising to the surface. The research charts rising trajectories related to critical drone application spheres displaying particular promise, like infrastructure analysis, autonomous navigation, and precision agriculture.

For researchers, this integrated aerial map reveals invaluable insights into themes and keywords that resonate with funding bodies. By targeting proposals to align with evolving agency priorities related to drone systems, chances of securing coveted grants improve. The work thus serves as a beacon for traversing complex funding environments strategically.

Meanwhile, policymakers receive a wellspring of intelligence enabling strategic decision making. Understanding granular funding patterns and trends facilitates optimized resource allocation toward the most promising drone technology arenas. The research empowers a panoramic perspective necessary for shaping investments holistically rather than through myopic silos.

Industry stakeholders also gain awareness of emerging trajectories at the intersection of academic inquiry and commercial potential – a collaborative sweet spot primed for innovation. As drone applications bridge theory and practice across settings like crop analysis, surveillance services, and mapping, corporate leaders can anchor strategic plans based on the cutting edges of drone technology revealed by the funding insights.

Overall, by decoding the complex language of funding dynamics underlying a domain challenged by fragmentation, this project catalyzes transformation. The commitment goes beyond mere data exploration to impart actionable clarity for diverse stakeholders navigating the future of drones. Researchers receive illuminated pathways through competitive grant environments. Policymakers wield an optimization tool for responsible resource allocation. Corporate strategists access signposts to horizons of innovation.

Through this multifaceted impact, the research enables drone technology to continue ascending in service of society’s greatest challenges, from environmental sustainability to infrastructure safety. The promise of this burgeoning field threatened by the pitfall of fragmented insights now sees a route to realization.

Methods and Materials

Grant data preprocessing for NSF funded drone research involved crucial steps to ensure clean text and optimal conditions for unbiased semantic and topic modeling. Starting with raw HTML and XML files, basic cleaning entailed eliminating tags, normalizing Unicode characters, converting text to lowercase, removing punctuation and numbers not conveying meaningful content, and lemmatizing words to base forms. Removing extremely common words unlikely to distinguish topics such as articles, pronouns, and prepositions proved necessary as well, utilizing a customized stop words list for the corpus. Unusual words potentially skewing models were replaced with relevant terminology or excluded if they could not be sufficiently interpreted and appropriately fitted in the broader context.

With a refined text corpus, baseline calculations provided preliminary insights into word frequencies across all grants. Histogram and bar chart visualizations of term frequencies revealed word occurrences, enabling analysis of prevalent language.

Term frequency–inverse document frequency (TFIDF) widened the understanding of key words defining individual grants. By adjusting values based on how selective a term is across the full collection of grants, distinctive features for each project proposal could emerge. Results revealed both specialized terminology exclusive to niche proposals as well as foundational language binding concepts spanning drone research as a whole. Clustering grants based on similarities and differences of TFIDF values surfaced concentrated areas of distinct activity.

Latent Semantic Analysis (LSA), an unsupervised machine learning technique, extracted relationships between concepts contained in the funding dataset. By statistically associating grants sharing vocabulary and revealing connections between documents and terms, LSA delivered a multidimensional semantic space capturing both granular word connotations as well as broader subject patterns. Conceptual trajectories traced funding trends focused on critical themes like drone endurance, autonomy, and security. LSA integration and projection matrices compressed complex textual information into representative vectors ripe for further analysis.

Complementing LSA, Latent Dirichlet Allocation (LDA) performed unsupervised clustering to categorize grants into topics. LDA probabilistically assigned documents to groups based on cooccurring language indicative of common concepts. Qualitative assessment accompanied the quantitative outputs, manually reviewing and interpreting the most significant terms in each topic to assign descriptive labels. For enhanced coherence, LDA models were iteratively refined by tweaking parameters and testing against gold standard human judgement until a model with distinct, cohesive topics emerged. The final topic model provided a high-level taxonomy of prominent activity revealed across thousands of NSFfunded drone research proposals.

Together, meticulous preprocessing paired with LSA's ability to expose semantic connections and LDA's topic clustering capacities delivered both an intricate grant level perspective alongside a simplified landscape overview of critical developments, obstacles, goals, and directions guiding the trajectory of drone research progress. The multifaceted analysis framework supplies stakeholders ranging from principal investigators to funding agencies an information foundation to strategize proposals and policies guiding responsible drone innovation.

**Methodology**

The first step in our methodology focuses on preparing the dataset for robust analysis. This vital process, termed data preprocessing, executes crucial refinements to the corpus, enabling insightful exploratory potential.

**1. Data Selection**

The dataset hones in on recently funded NSF grants from niche divisions closely tied to the study's pursuits. This targeted focus retains maximum relevance to the research questions at hand while uncovering insights potentially obscured in larger corpora. Tailoring the textual corpus solidifies analytic traction, allowing revelations to emerge within purpose-built boundaries.

**2. Encoding Standardization**

Explanation:

Text encoding defines the digital formatting of character representations within strings. By uniformly translating the corpus to adopt the universal UTF-8 encoding system, the rendered characters consistently and accurately symbolize the intended semantic assignments. Enforcing this encoding consistency streamlines downstream textual processing and analytic procedures by guaranteeing symbols map to meanings free of distortion.

Significance:

Non-standardized encodings can lead individual character encodings to discordantly digitize, resulting in defective rendering of information. For example, accented alphabetic characters common in many global languages may fail to encode properly under outdated or fragmented encoding systems. Thus UTF-8 adoption critically circumvents such issues by applying a unified framework, ensuring text integrity through cross-platform symbolic fidelity.

**3. Corpus Compilation**

Explanation:

A textual corpus constitutes a collected body of textual documents targeted for analysis. This research compiles its core analytical corpus from the extracted grant abstract texts provided in the broader dataset. Isolating these textual summaries as the informative locus enables concentrated scrutiny while retaining the essence of the broader funding narratives from which they derive.

Significance:

The distillation of full grant documentation down to salient abstract excerpts condenses otherwise lengthy and diffuse datasets into directly scrutinizable narrative selections carrying rich potential utility for both computational and qualitative examination. Focusing analytic efforts on these informationally dense textual loci allows revelations to surface amidst an ultimately more wieldy corpus.

**4. Stopword Filtering**

Explanation:

Stopwords constitute ubiquitous terms like "and," "is," or "the" carrying little unique semantic information in isolation. By filtering these terms from the corpus prior to analysis, the remaining linguistic components more clearly signal key semantic content by highlighting only the most informative contributing terms.

Significance:

Though stopwords feature prevalently across most texts, overrepresentation tends to dilute the information density of corpuses as these terms contribute limited distinct meaning. Removing them therefore serves to de-noise the dataset, prioritizing only the most content-bearing and value-added language to downstream processes. This distillation allows analytic focus on narrative-driving semantics.

**5. Abstract Refinement**

Explanation:

The full abstract refinement process applies a series of data cleansing steps including stripping numeric digits or punctuation elements and harmonizing capitalization formats. Together these refinements pare down competing signals to hone in on the purely linguistic core content.

Significance:

Abstract cleansing through stripping extraneous elements and homogenizing case equalizes the textual data, removing potentially confusing artifacts unrelated to core semantics. It serves to emphasize the language itself by erasing dissonant signals in formatting. Standardizing case additionally normalizes variance that could skew analysis.

**6. Creating a Tibble**

Explanation

We are structuring the scraped grants data into a tibble data frame within R, consolidating key variables like unique award ID, funding abstract narratives, and awarded dollar amounts into an organized table format optimized for text analysis. Methodical organization of heterogeneous web data transforms an unstructured corpus into an analytical foundation.

Significance

Constructing a tibble filters noise, aligning unsystematic web data into standardized columns containing the most informative fields. This curation precipitates deeper investigation by assembling the building blocks for textual and statistical modeling.

**7. Bigram Creation**

Explanation

Leveraging the quanteda R package, we generated adjacent pairwise word combinations known as bigrams from the abstract column of our tibble. Bigrams capture proximate terms frequently occurring together within the grant funding narratives, surfacing potential conceptual connections.

Significance

Elucidating recurrent word pairs within abstracts supplements singular term frequency analysis with higher order lexical relationships. Bigrams reveal semantic links between vocabulary through colocation patterns. By honing in on phraseology, we move closer towards conveyed meaning within proposals.

**8. Stopword Removal from Bigrams**

Explanation

We refined the bigram dataset by filtering out common stopwords using the quanteda stopwords function. Stopwords are high frequency terms like articles, prepositions, or pronouns that contribute little topical meaning.

Significance

Eliminating stopwords sharpens our focus to most salient word combinations. Excluding superficial syntactic pairings leaves meaningful terminology untouched for downstream utility. It reduces noisy edges to sharpen overall resolution.

**9. Tokens Unnesting**

Explanation

Unnesting transitions our analysis from strict word pairs back to individual words. Applying R’s unnest\_tokens function splits apart bigrams into their constituent oneword tokens.

Significance

Fragmenting bigrams reverts to a singular token basis while retaining bigram insights. This flexibility permits both granular and phrase level exploration. Individual tokens become inputs for subsequent techniques like tfidf weighting and semantic analysis.

**10. Lemmatization**

Explanation

We leveraged the textstem R package to reduce inflected token variants to common lemmas. Lemmatization strips affixes to retrieve root dictionary forms, consolidating word families.

Significance

Lemmatization combats vocabulary variation by collapsing derivatives into shared stems. It counters scattering concepts across superficial syntactic shifts. Consolidating related terms amplifies signals within textual data by aggregating dispersed usage. Together these steps curate a refined corpus.

**11. Corpus Update**

Explanation:

Following the lemmatization process that reduces words to their root forms, the aggregated textual corpus undergoes a comprehensive update. This update globally aligns the corpus to reflect the newly lemmatized term representations. Applying this crucial step pre-analysis ensures all downstream processes build upon a refined dataset utilizing standardized terminology.

Significance:

Updating the master corpus with lemmatized variants proves vital for safeguarding accuracy as the analysis trajectory unfolds. It synchronizes the corpus vocabulary, encapsulating all word variations under unified lexical stems. This consistency forms an essential foundation upholding the structural integrity of subsequent analytic processes.

**12. Document Term Matrix Creation**

Explanation:

Leveraging the foundational bag-of-words textual modeling approach, a Document Term Matrix (DTM) is constructed to represent the corpus. This matrix provides a structured visualization of term frequencies, denoting the occurrence rates of words by document. The output formats the textual landscape into a quantifiable matrix, elucidating the variable importance of linguistic components based on measurable frequency.

Significance:

The DTM holds an integral role in textual analytics by quantifying nebulous language into definitizable term frequencies. It lends tangibility to abstract language through numerical transformation. The matrix view illuminates the intersections between corpus documents and constituent terms, crystallizing patterns otherwise obscured. In aggregating, structuring and visualizing term repetition rates by source content, DTM construction constitutes a lynchpin utility for unlocking insights into the textual terrain.

A close-up of words

Description automatically generated

Wordcloud for dataset from 2015 -2022

**Natural Language Processing (NLP) Techniques:**

Now that the dataset has undergone thorough preprocessing, the methodology progresses to advanced NLP techniques aimed at extracting meaningful insights from the grant abstracts.

1. Optimal Topic Number Selection

Explanation:

With the refined corpus now ready for interrogation, a series of precise tuning procedures identify the optimal number of latent topics to maximize thematic clarity when applying probabilistic topic modeling. Employing a suite of quantitative evaluation metrics coupled with ldatuning methods, the intricate process balancing under and overspecification through a harmonic assessment of model likelihood determines the pivotal number of topics.

Significance:

Pinpointing the ideal topical breadth proves foundational to the integrity of subsequent semantic mining operations. Defining too few topics risks merging distinct themes while too many invites fragmentation or incoherence. Enlisting mathematically-driven tuning protocols injects rigor into the notoriously delicate process of scoping an interpretable topical architecture for fruitful modeling. By optimizing for harmonic coherence, the method hunts the goldilocks balance between concision and diffusion to support both granular investigation and higher-order thematic cohesion. The determination of scope grants focus to the forthcoming analytical odyssey.

2. Topic Model Optimization

Explanation:

Building atop the identified optimal architecture, customized topic modeling procedures further tailor model parameters to the idiosyncrasies of the NSF corpus. Model optimizations account for corpus linguistic qualities and inherent structural patterns such as term frequencies and co-occurrence tendencies. This adaption curates model fidelity to the dataset at hand.

Significance:

While topic models promise multidimensional insights, their utility relies firmly on customization to the underlying data universe. One-size-fits all approaches falter in extracting meaningful themes without accounting for corpus linguistic qualities. Adaptive parameter tuning aligns the model to innate dataset characteristics, bolstering the probability of surfacing poignant observations in the final output. Optimization therefore constitutes a vital installment supporting impactful modeling.

A graph of a number of lines

Description automatically generated with medium confidence

Optimal Topic Graph using ldatuning package

A graph showing a line

Description automatically generated

LDA model is run using 21 topics and then saved and tidied. The top 10 words are extracted for each topic and concatenated.

**Latent Dirichlet Allocation Implementation**

Explanation:

With optimal parameters secured, Latent Dirichlet Allocation (LDA) launches over the corpus. LDA operates by statistically distributing terms across inferred latent topics, leveraging the patterns in language structure and term co-occurrence to learn hidden thematic structures. Powered by 11 precisely tuned topics, the Bayesian computational model dissects linguistic layers to expose underlying semantic dimensions.

Significance:

As a pillar probabilistic topic modeling technique, LDA holds profound power to extract insights from the complex textual dataset. By detecting shared term patterns and clustering documents accordingly, LDA reveals systemic concept associations otherwise obscured within the word maelstrom. It imparts structure and dimensionality to nebulous language, illuminating multifaceted research dimensions buried below the surface.

**Topic Term Extraction**

Explanation:

With topics structurally molded, the 10 most influential words within each thematic cluster surface. These terms, ranked by explanatory power, offer a pointed representation of the latent concepts uncovered from the depths of the corpus. They signify semantic gravity centers orienting topics around crucial language.

Significance:

Extracting the foremost terms of each topic provides an efficient semantic handle for reasoning about the intricate modeling outputs. These words crystallize topics into their most central linguistic elements, providing a straightforward vehicle for conceptual interpretation and labeling. They distill key aspects of emergent themes for ready examination while retaining fidelity to data-derived discoveries.



TI-IDF Wordcloud

**Bag of Words**

Explanation:

The Bag of Words technique conceptualizes text as an unordered collection of words devoid of grammar, syntax or word order. By focusing singly on individual word frequencies while disregarding linguistics conventions, documents mathematically distill to proportional word counts.

Significance:

This atomic approach to language grants computational simplicity for modeling the corpus lexicon. Stripping complex linguistic nuances to enumerate term prevalence constructs an intuitive foundation for quantifying document differences and similarities. Bag of Words powers myriad text analysis building blocks.

**2. Bigrams**

Explanation:

Bigrams constitute word pairings within textual sequences, capturing adjacent terms for unlocking contextual insights. As atomic units, individual words struggle to encapsulate latent meaning. Bigrams retain vital sequencing to expose phrasal semantics.

Significance:

Atomic words fail to relay inter-term conceptual connections central to interpretation. Bigrams reveal key term associations, elucidating meaning otherwise obscured. They provide base units receiving the balance of linguistic context with computational simplicity.

**3. Stopwords**

Explanation:

Stopwords represent ubiquitous terms like “and” or “the” that offer minimal distinct semantic value. As meaningless linguistic components, they remain while carrying little information. Stopword filtering removes this noisy debris.

Significance:

By filtering stopwords, analyses concentrate computational resources on the meatiest linguistic chunks. Eliminating ubiquitous but uninformative terms enhances the signal received from critical keywords and grants procedures acute lexical focus.

**4. Tokens**

Explanation:

Tokenization bisects text into atomic word units via parsing procedures for targeting individual terms. This dissection focuses the analytical microscope to linguistic granularity critical for precision.

Significance:

By cracking sentences into discrete lexemes primed for scrutiny, tokenization institutes textual analysis while unpacking meaning buried within bewildering lexical complexity. Tokens form the basis of penetrating computational examinations.

**5. Lemmatization**

Explanation:

Lemmatization associates related term variants with common roots, condensing diverse morphological forms like “analyze,” “analyzes,” “analyzed” to a universal base -- typically the lemma.

Significance:

This normalization equalizes semantic treatment of related term variants, boosting model accuracy by merging disparate surface forms to reveal underlying uniformity. Lemmatization confronts vocabulary fluidity through consolidation.

**6. Document Term Matrix**

Explanation:

The Document Term Matrix constitutes a mathematical structure visualizing term frequency across documents. This matrix formats rows to represent individual documents, columns to capture specific vocabulary terms, and cells to embed occurrence counts of a given term within a particular document.

Significance:

Quantitatively outlining corpus term dispersion through a formal matrix architecture enables refined computational text analysis approaches. By translating opaque linguistic features into unambiguous numeric attributes of term prevalence, the DTM structurally reinforces the corpus for examination through statistical learning algorithms requiring numerical inputs. It tames vocabulary chaos through principled design.

**7. Latent Dirichlet Allocation**

Explanation:

Latent Dirichlet Allocation (LDA) operates as a probabilistic modeling approach to decompose document collections into latent topic architectures. It assumes documents manifest as distributions over topics, where each topic forms probability distributions over words. LDA computationally divines these elusive associations.

Significance:

as a preeminent technique for uncovering thematic structures woven through texts, LDA holds profound significance for granting insights into the topical composition of the corpus. By revealing hidden semantic links between frequently co-occurring terms, LDA offers a vehicle for systematically inspecting conceptual connections threading through the domain’s knowledge base. It brings to light driving research themes.

This multi-step methodology enlists a synthesized assembly of data conditioning, quantitative synthesis, and probabilistic topic modeling techniques, traced from raw corpus to refined revelations. The coherent funnel of increasingly sophisticated transforms structurally reinforces the textual evidence base to unlock and visually organize otherwise inaccessible understandings of the grant funding knowledge landscape.

**Result**

This research leveraged text mining techniques to map the landscape of drone and UAV research based on grants awarded by the National Science Foundation (NSF). Meticulous data preprocessing set the stage for accurate modeling, with cleaning steps like stripping HTML tags, converting text to lowercase, removing punctuation and numbers not contributing semantic content. Initial term frequency analysis and bigram associations revealed common vocabulary and concepts under exploration across projects. Latent Semantic Analysis (LSA) then exposed underlying semantic relationships spanning proposals sharing terminology. Meanwhile, Latent Dirichlet Allocation (LDA) clustered documents into coherent topics linked by subject matter based on word co-occurrence tendencies.

The optimal number of topics was determined to be 21 after iterative validation assessments. The top terms in each topic paint a qualitative picture of research directions receiving funding. Clear focuses emerge ranging from developing drone swarm communication protocols and coordination algorithms to improving battery technologies and investigating causes of failure. Mapping connections and distances between topics also grants a holistic landscape overview. Tight clusters indicate subsets of thematic concentration, while spaced points mark relatively distinct areas being funded.

Preprocessing and modeling unlock critical insights unattainable through manual analysis given the volume of textual data. The framework supplies an empirical foundation for tracking activity, informing funding priorities, and strategizing project proposals to align with support trends. Researchers gain visibility into rising subject areas to position compelling grants. Agencies can pinpoint saturated spaces or gaps requiring attention when developing initiatives. Collectively, systematically organizing disjointed records and clarifying relationships converts opaque data into transparent knowledge to responsibly advance drone and UAV research.

The analysis techniques exhibit tremendous potential for continual monitoring as an updated funding dataset could be seamlessly reprocessed to compare emerging activity. This scalable approach generates an evolving navigation aid steering stakeholders towards milestones holding the most promise and peril for drone technologies.

Top 10 Terms in each 21 LDA Topic

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

**A close-up of a diagram

Description automatically generated**

**Challenges Faced**

The foremost challenges stemmed from working with heterogeneous raw web data requiring substantial wrangling. As an unstructured mix of HTML, XML, and text spanning decades, meticulous preprocessing was imperative before analysis. Cleaning irregular formats, parsing relevant fields, handling encoding inconsistencies, and filtering noise posed early obstacles. The absence of labeling or categorization also meant relying solely on contents to ascertain semantic connections.

With grants containing distinct vocabularies and writing styles, sparseness and variability threatened to obscure overarching conceptual links. Creators flexibly interchanged terminology like drones, UAVs, quadrotors, and UAS. Fortunately, steps like lemmatization, removing custom stop words, and using bigrams alongside TF-IDF vectors counteracted dispersion across synonyms and phrasing variations. However, subtle complexities around evolving acronyms still risked distraction.

Since proposals spanned technical specifications to ethics, balancing and representing niche keywords with universal terms during modeling iterations remained an art – not pure science. Custom tuning addressed distinctiveness while seeking sufficient term overlap to surface cogent topics. Further, with LDA yielding numerous viable clustering combinations, definitive maps stayed elusive. Tradeoffs between granularity and aggregation test even human judgement on optimality.

Quantitatively measuring semantic interpretability also posed challenges. While metrics like cosine distance and topic coherence scores offered evaluation support, subjective scrutiny was indispensable for contextual coherence. No single metric universally dictates model fidelity; rather, combinations of quantitative signals and qualitative assessments allow more informed tuning. Even after refinement though, edge cases likely persist given intricacies of language.

With grants separated across decades, changes in formatting conventions and mainstream terminology complicated identifying temporal trajectories. Techniques like dynamic topic modeling accommodate evolutions but require chronological metadata frequently absent. Connecting topics over time warrants deeper integration of funding details like award dates. Associated analysis of investigator backgrounds using named entity recognition may also clarify critical influences.

Overall, balancing noisiness, variability, complexity, and continuity inherent with multifaceted textual data creates ripe conditions for misrepresentation. No solution universally reconciles tensions across completeness, precision, concision, and evolution. Yet text mining presents a formidable opportunity when responsibly executed – revealing what no unaided expert could systematically expose at scale. So while challenges persist, adopting best practices mitigates threats on the path to maximizing insights extracted to advance discovery. Ongoing gains surface by fusing automated techniques with human guidance to beat challenges hampering understanding.

**Conclusion**

In concluding the Topic Modeling Analysis employing Latent Dirichlet Allocation (LDA) on the drone and UAV grant dataset, an exploratory expedition through funding patterns comes into focus. The LDA model adeptly distills thousands of unique proposals from the corpus into an optimal set of 21 interpretable topics. This topic count is meticulously determined through quantitative coherence checks and qualitative assessments to ensure statistical significance and semantic alignment with embedded structures within the source text.

The creation of Word Cloud visualizations proves instrumental as an intermediary translation layer, converting abstract statistical outputs into tangible thematic summaries represented by immediately recognizable terminology. The prominence of particular terms within each cloud offers a window into the specific conceptual essence at the core of related proposals tied together by subject matter and frequently co-located language. Taken together, these visualizations unlock an aerial view of the diverse sub-domains garnering NSF backing across over two decades, illuminating both longstanding challenges and emerging fronts.

In constructing the foundational document-term matrix for input into the LDA model, the strategic choice is made to retain the complete Bag-of-Words representation without filtering low frequency terms. This decision stewards nuanced vocabularies within niche specialties that may be erroneously stripped by alternate weighting schemes yet contribute to distinguishing concepts. Preserving textual integrity thereby enhances downstream discovery of latent connections. The document-term matrix subsequently operates as a canvas visualizing thematic brushstrokes rather than a distortion exaggerating particular elements.

The multifaceted insights extend beyond isolated topics and keywords to higher-level interrelationships between subjects depicted through the Intertopic Distance Map. This cartographic visualization orients semantic clusters and proximities, pointing to potential cross-pollination across specializations exploring similar challenges through alternate frames. It also exposes more nascent spaces warranting closer examination to address technological obstacles or ethical uncertainties standing in the way of responsible real-world drone deployment.

Ultimately, by peering through these lexical windows, stakeholders spanning researchers, policy advisors, and industry partners achieve elevated clarity regarding the status, direction, and purpose of innovation trajectories based on areas commanding funding support. The modeled landscape thereby spotlights promising inroads while also keeping peripheral visions attuned to gaps in understanding. Convertible to an interactive dashboard, continual ingestion of new grants provides a steering compass as the topography accelerates across an atmosphere of open curiosity.

**Future Prospects**

Positioned as an exhaustive reference guide, this drone and UAV research mapping endeavor emerges as an indispensable compass for stakeholders navigating the complex funding landscape. The utility stretches beyond isolated statistical outputs to actionable insights guiding proposals, policies, collaborations, and responsible innovation strategies.

For researchers, the illuminated topical structures and associative visualizations supply a strategic roadmap for positioning compelling NSF grants. Applicants can align proposals with rising thematic areas signaled by funding activity, demonstrating continuity with prior supported domains warranting deeper examination to drive progress. The intertopic distance mappings also equip teams to identify potential cross-disciplinary partners studying adjacent challenges from complementary vantage points.

Meanwhile, funding agencies can utilize the bird’s-eye visualization of high-level activity to strategically direct support towards pressing technical obstacles and ethical uncertainties requiring scientific spotlighting. Program managers gain an empirical dashboard highlighting saturated spaces needing no additional capital versus high-potential yet comparatively underfunded frontiers able to absorb and responsibly utilize further backing.

Industry stakeholders equally stand to benefit from observing academic research directions and priority shifts. Aligning internal roadmaps with cutting-edge inquiries on the brink of breakthroughs may expose game-changing innovations worthy of collaboration, investment, or acquisition. The model maps serve as macro-level environmental scans, forecasting emerging technologies and signaled challenges facing real-world integration.

Collectively, transparency into the funding environment precipitates judicious progress across drones and UAVs. The distilled landscape is more than a static analysis – it is a living index, updating alongside database growth. Continuous ingestion of fresh grants will build lexical literacy over time, tracking the evolution and diffusion of seminal concepts. In closing, by extracting signals from data opacity, this reference guide stewards, connects, and orients, driving action by diverse stakeholders united in responsibly advancing drones from abstract concept to ubiquitous capability.

Additionally, statistical models stand to gain enhanced resolution by incorporating awarded dollar amounts alongside abstract texts. Funding levels associated with research topics offer further quantitative signals of relative prioritization. Comparing average or total grants for emerging focuses versus established ones may reveal intensifying activity even if proposal volume remains constant. For example, soaring investment into drone security and privacy concerns could expose escalating attention to ethics despite stagnant case volumes. Meanwhile, downward budgets may point to maturing topics reaching conclusion despite continual interest. Integrating numeric metrics with semantic models thereby supercharges dimensionality, capturing subtler shifts through monetary fluxes. This multilens perspective prevents myopia, balancing textual representations with budget reality checks toward a unified view. In the end, binding linguistic concepts with material backing amplifies insights into the directionality of priorities take flight or stall out. Funding figures also inform higher-level decisions on allocating constrained capital, illuminating cuts with minimized disruptiveness.

Here are 20 research papers that could be used as references for this drone/UAV text mining research:

1. Basole, R.C., Russell, M.G., Huhtamäki, J., Rubens, N., Still, K., Park, H. (2015). Understanding Business Ecosystem Dynamics: A Data-Driven Approach. ACM Transactions on Management Information Systems, 6(2), 6.

2. Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

3. DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). Big Data & Society, 2(2), 2053951715602908.

4. Evans, J.A., Foster, J.G. (2011). Metaknowledge. Science, 331(6018), 721-725.

5. Griffiths, T.L., Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235.

6. Grimmer, J., Stewart, B.M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis, 267-297.

7. Herr II, P.M., Kardes, F.R., Kim, J. (1991). Effects of word-of-mouth and product-attribute information on persuasion: An accessibility-diagnosticity perspective. Journal of consumer research, 17(4), 454-462.

8. Hopp, T., Vargo, C.J., Huang, Y. (2016). Motivational patterns in science funding programs. Public Understanding of Science, 25(8), 932-948.

9. Kanoje, S., Mukhopadhyay, D., Girase, S. (2020). Applications of Drone for Smart Cities: A Review. International Journal of Computer Network and Information Security, 12(3), 20-29.

10. Mimno, D., Blei, D.M. (2011). Bayesian checking for topic models. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 227-237.