National Science Foundation Awards:

Tracing the journey and Impact of Research Projects

Francisco Herrera

The University of Texas at San Antonio

Dr. Ritu Arora

CS-4243-001 Large Scale Data Management

Abstract

The "Award Tracing" initiative seeks to track the lifecycle and evaluate the impact of federally funded programmers. This research endeavor involves a number of stages, including data collection, topic modelling, database creation, I b interface development, catalogue compilation, and the extraction of insights and statistics. The primary purpose is to provide a thorough analysis of the funded initiatives, casting light on their development and societal impact. The initiative entails collecting award data from a number of federal agencies, including the NSF, NIH, DOE, and DOJ. Keywords that capture the essence of each project are derived from abstracts and titles using topic modelling techniques. The collected data is stored in a database for efficient retrieval and administration. A user-friendly I b interface is devised to facilitate database access and exploration. In addition, a catalogue presenting project information in a standard format is created. Through rigorous analysis of the collected data, insights and statistics are derived to illustrate the social impact of the initiative. This initiative seeks to increase understanding of federally funded programmers, their development, and their contribution to society.

Introduction

Federal agencies play a crucial role in financing a vast array of projects aimed at advancing scientific research, technological innovation, and societal I lfare. Tracing the lifecycle and assessing the
impact of these funded initiatives can be a difficult and time-consuming endeavour. The "Award Tracing" initiative aims to resolve this difficulty by implementing a comprehensive strategy to monitor the development and evaluate the societal impact of federally funded projects.

The primary objective of this research is to provide a comprehensive analysis of funded projects, casting light on their voyage from inception to completion and gaining an understanding of the
results they generate. Award Tracing employs a systematic methodology comprising data collection, topic modelling, database creation, I b interface development, catalogue generation, and the
derivation of insights in an effort to uncover the narrative behind each project and quantify its societal impact.

To accomplish this objective, a number of federal agencies, including the National Science Foundation (NSF), National Institutes of Health (NIH), Department of Energy (DOE), and Department of Justice (DOJ), will be investigated. These organizations fund initiatives in a variety of disciplines,
including scientific research, technology development, healthcare, and law enforcement. Award Tracing seeks to provide a comprehensive view of funded projects by collecting award data from these agencies and integrating it into a centralized database.

In addition, the project employs sophisticated methods such as topic modelling to extract keywords that capture the essence of each project. This allows for a greater comprehension of the project's focal areas, which facilitates efficient categorization and analysis. The collected data and derived.
keywords serve as the basis for the subsequent phases of the project, which include the construction of a user-friendly I b interface and a standardized catalogue that presents project information in an accessible format.

Award Tracing's ultimate objective is to contribute to the body of knowledge surrounding federally funded projects and their impact on society. This research endeavor seeks to provide valuable.
information to stakeholders, policymakers, and researchers interested in understanding the outcomes and societal value generated by federal funding by analyzing collected data and deriving meaningful insights.

Methods

Collecting and Processing Information

This project's data acquisition process involved retrieving information about awarded initiatives from the I bsites of various federal agencies. I concentrated on collecting information from the National Science Foundation (NSF) website and other agencies, including the National Institutes of Health (NIH), Department of Energy (DOE), and Department of Justice (DOJ). I conducted queries on the NSF website using keywords associated with the funding programmers of interest, such as CSSI, SI2, DIBBS, CICI, MRI, OAC, and CCF.

Regarding the remaining agencies, I visited their respective websites and investigated the accessible datasets associated with the awarded projects. In instances where datasets were not readily available, I reached out to the departmental contact via email to acquire the necessary data.

Subject Modelling

I used topic modelling techniques to extract relevant keywords and obtain insight from the project abstracts and titles. The abstracts and titles of the awarded initiatives were the primary focus of our analysis. I utilized the Gensim library, which offers efficient topic modelling algorithm implementations. I removed stop words and converted the text to lowercase before pre-processing the text data. The text was then tokenized into individual words. The resulting corpus was used to construct a dictionary of terms by assigning a unique integer id to each distinct word. The corpus was then transformed into a Bag-of-Words representation, with each document represented as a sparse vector of word frequencies.

On the pre-processed corpus, a Latent Dirichlet Allocation (LDA) model was trained. The LDA model is a probabilistic generative model that associates documents with topics and words with topics. I utilised the LdaModel class from the Gensim module, specifying five topics and executing ten training iterations. After training the LDA model, the best words associated with each topic were extracted. These words represented the most significant and representative terms within each topic, illuminating the primary themes and areas of emphasis of the awarded projects.

Data Storage and Internet Access

I created a database to organize and archive the collated project information. The database was designed to contain project-specific information including titles, abstracts, keywords, and pertinent metadata. I employed suitable database management systems and frameworks to ensure the efficient storage, retrieval, and administration of data. I devised a web-based interface to facilitate user access to and exploration of the project catalogue. Modern web development technologies, such as HTML, CSS, and JavaScript, were utilized to design the interface. It allowed users to search and browse the catalogue of projects, view detailed project information, and access insights and statistics that were generated. The web interface provided a seamless and interactive user.
experience, allowing users to readily investigate the social impact of the initiatives.

Our project trace system was founded on a combination of data collection, topic modelling, database storage, and web interface development. These techniques enabled us to efficiently record, analyze, and present the lifecycle and progression path of federally funded initiatives.

Results

Implementation of the Award Tracing system produced significant results in terms of documenting and analyzing federally funded awarded projects. The following essential results were accomplished:

Data Collection and Processing:

- I effectively retrieved an exhaustive list of awarded projects from the website of the National Science Foundation (NSF) and other agencies, such as the NIH, DOE, and DOJ. I obtained the datasets pertaining to the awarded projects through diligent efforts, either directly from the agency websites or by contacting the departmental point of contact.

Topic Modelling: I used topic modelling techniques, specifically Latent Dirichlet Allocation (LDA), to extract keywords and obtain insight from the abstracts and titles of the projects.

- By training the LDA model on the pre-processed corpus, I were able to identify five distinct topics that represent the primary themes and focal points of the awarded projects. The top terms
associated with each topic provided valuable information about the significant terms and concepts within the projects, thereby facilitating comprehension of their scope and impact.

Database Creation and Web Interface:

- I effectively created a database to hold and organize the collected project data, including titles, abstracts, keywords, and pertinent metadata. The Award Tracing system's web-based interface
enabled interactive searching, browsing, and exploration of the project catalogue.

- Users could gain access to comprehensive project information, observe standardized project listings, and derive insights and statistics demonstrating the societal impact of the projects.

Catalogue of Products:

- The collected data served as the basis for the compilation of an exhaustive catalogue of the awarded projects.

- Each catalogue page displayed comprehensive project information, including the title, abstract, keywords, and additional metadata and the standard format assured consistency and made comparisons between initiatives straightforward.

Deriving Insights and Statistics:

- The collected data allowed us to derive insightful insights and statistics that demonstrated the impact of the funded initiatives on society.

- By analyzing the project data, I I re able to identify patterns, trends, and significant contributions in numerous research and innovation fields. These insights provided policymakers, researchers, and other interested parties with valuable information regarding the outcomes and societal implications of the initiatives.

Overall, the Award Tracing system effectively tracked the lifecycle and progression of federally funded initiatives. It enabled efficient data collection, topic modelling analysis, database storage, and I b-
based exploration, yielding valuable insights into the social impact of the initiatives. The user-

friendly interface of the system increased accessibility and provided a comprehensive catalogue of projects for further research and decision-making.

Discussion

The Award Tracing system has demonstrated its ability to track the lifecycle and progression of federally funded initiatives. The implementation of multiple components, such as data collection, topic modelling, database creation, I b interface development, and insight extraction, has
contributed to a comprehensive comprehension of the awarded projects and their societal impact.

The ability of the Award Tracing system to collect project data from multiple federal agencies is one of its most significant strengths. This broadens the scope of the analysis and provides an all-
encompassing perspective of the funded initiatives across various domains and disciplines. The incorporation of diverse agencies ensures that the Award Tracing system encompasses a vast array of research fields, thereby fostering inter-disciplinary collaborations and the exchange of knowledge.

Utilizing topic modelling, specifically Latent Dirichlet Allocation (LDA), has been instrumental in revealing the underlying themes and concepts within the abstracts and titles of projects. By identifying key topics and extracting relevant keywords, the system provides researchers and stakeholders with fast insights into the funded projects' primary areas of focus. This facilitates comprehension of the project landscape, identification of emergent research trends, and
identification of potential areas for further study.

The construction of a comprehensive database for organising and storing the collected project data has proved to be beneficial. In addition to facilitating the efficient search and retrieval of project information, the database also enables the system to generate standard project listings. The standardized format ensures consistency and uniformity in the presentation of project details, making it simpler for users to compare and evaluate various projects. Moreover, the I b-based interface improves accessibility and user experience, allowing stakeholders to interact with the database and intuitively investigate project information.

The Award Tracing system generates a catalogue of products that provides a consolidated view of funded initiatives and a I alth of information for various stakeholders. Researchers, policymakers, and funding agencies can use the catalogue to identify projects aligned with their areas of interest, evaluate the impact of past projects, and make informed decisions about future funding allocations. The standardized layout of the catalogue pages facilitates comprehension and analysis, enabling users to efficiently derive pertinent information.

The Award Tracing system enables stakeholders to evaluate the societal impact of funded projects by deriving insights and statistics from the collected project data. The analysis of project outcomes, contributions, and societal consequences can inform policy decisions, influence research orientations, and direct resource allocation. The system's insights can also be used to identify successful project models, foster collaborations, and identify areas where additional investments may be required.

Despite the fact that the Award Tracing system has demonstrated significant strengths, certain. limitations must be acknowledged. The system is dependent upon the availability of data from federal agencies. In some instances, gaining access to and acquiring the required data may be difficult, necessitating additional effort in contacting agencies and obtaining permissions. Second, the quality of the project abstracts and titles determines the precision

and dependability of the extracted.

keywords and topics. Incomplete or ambiguous descriptions may result in topic modelling outcomes that are less precise. The availability and accessibility of project updates and final reports may
influence the system's efficacy in capturing the complete lifecycle and progression path of projects.

The Award Tracing system has provided valuable insights into the lifecycle, impact, and progression of federally funded initiatives. By incorporating data collection, topic modelling, database creation, and I b interface development, the system has made analysis, cataloguing, and exploration of awarded projects more efficient. The system's insights contribute to evidence-based decision making, nurture cross-disciplinary collaborations, and promote research funding transparency and accountability. The Award Tracing system has the potential to improve federal funding agency research assessment, project evaluation, and knowledge dissemination.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Gensim. (n.d.). Gensim: *Topic modeling for humans*. Retrieved from https://radimrehurek.com/gensim/

National Science Foundation. (n.d.). NSF Award Search: *Simple search result*. Retrieved from https://www.nsf.gov/awardsearch/simpleSearchResult?queryText=CSSI

Ritchie, A. (2020). *Beautiful Soup Documentation*. Retrieved from https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Witten, I. H., Frank, E., & Hall, M. A. (2016). Data mining: *Practical machine learning tools and techniques* (4th