Paper 5203-2020

Unsupervised Contextual Clustering of Abstracts

Jacob Noble; Himanshu Gamit;
University of St Thomas, St Paul

ABSTRACT

This study utilizes publicly available data from the National Science Foundation (NSF) Web Application Programming Interface (API). In this paper, various machine learning techniques are demonstrated to explore, analyze and recommend similar proposal abstracts to aid the NSF or Awardee with the Merit Review Process. These techniques extract textual context and group it with similar context. The goal of the analysis was to utilize the Doc2Vec unsupervised learning algorithms to embed NSF funding proposal abstracts text into vector space. Once vectorized, the abstracts were grouped together using K-means clustering. These techniques together proved to be successful at grouping similar proposals together and could be used to find similar proposals to newly submitted NSF funding proposals.

To perform text analysis, SAS® University Edition is used which supports SASPy, SAS® Studio and Python JupyterLab. Gensim Doc2vec is used to generate document vectors for proposal abstracts. Afterwards, document vectors were used to cluster similar abstracts using SAS® Studio KMeans Clustering Module. For visualization, the abstract embeddings were reduced to two dimensions using Principal Component Analysis (PCA) within SAS® Studio. This was then compared to a t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction technique as part of the Scikit-learn machine learning toolkit for Python.

Conclusively, NSF proposal abstract text analysis can help an awardee read and improve their proposal model by identifying similar proposal abstracts from the last 24 years. It could also help NSF evaluators identify similar existing proposals that indirectly provides insights on whether a new proposal is going to be fruitful or not.

INTRODUCTION

The National Science Foundation is an independent federal agency created by Congress in 1950 "to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense"[1]. The NSF is vital because it supports basic research and people to create knowledge that transforms the future. The NSF and awardees go through a long application submission and merit review process that involves submitting, reviewing and evaluating about 50,000 proposals a year. Of that, determining research proposals which have the greatest potential and would be the most fruitful investment of taxpayer dollars is a big challenge [1].

The aim of this paper is to demonstrate how text insights can help awardees improve their proposals and better contribute to existing research. This also allows the NSF team to optimize processes and channel appropriate funding to a variety of projects.

PROBLEM STATEMENT

Every year the NSF receives about 50,000 research proposal, which is evaluated by a diverse mix of NSF and university faculty members, who assess the significance and the quality of the funding proposal. A large amount of manual effort is required to screen the proposals before they are approved for funding using taxpayer money. This creates challenges related

to scalability, consistency of project vetting across awardees, and identification of projects which require special assistance.

For awardees, it is highly beneficial to be able to identify similar funding proposals. It can often be a challenge to find collaborators on projects. Isaac Newton said, "If I have seen further than others, it is by standing upon the shoulders of giants." One impediment of proposing research is ensuring that the research will build upon the research of others.

DATASET AND DATA PREPARATION

The dataset was attained from the NSF Award Search Web API (ASWA), which is made publicly available by the NSF. This web REST API provides an interface to the Research Spending and Results (RS&R) functionality available through the NSF's Research.gov system. The award search data demonstrates how federal research dollars are being spent, what research is being performed, and how the outcomes of research are benefiting society.

The dataset was chosen for its unique ability to describe current and past research trends within the academic community. The dataset contains rich information surrounding each proposal such as the funding department, awardee and principal investigator information, and funding amount. There are many opportunities for study and research around this dataset aside from the abstract embedding and clustering technique discussed in this paper.

The dataset of project proposals includes data from 1985-09-06 to 2019-09-23. The dataset has 52 columns and 329321 rows. Of that, only the abstract text was used. In preparing the large data set, instances with missing abstract texts were excluded from the dataset. This did not have any adverse effects on the model as empty abstract text data would have not altered the embedding process. No duplicates were ensured by enforcing uniqueness on the award ID provided by the ASWA.

The Doc2Vec model cannot take raw text. Each abstract text is split into individual words or tokens while removing punctuation and numbers. Each token is then lowercased, and any accents are removed. This is a relatively simple approach to text processing and tokenization. After all the data preparation was complete, a total of 309661 abstract texts were considered to perform Doc2Vec embedding.

ANALYSIS METHODS

DOCUMENT VECTORS

Most machine learning and statistical models require numeric and often fixed length inputs. In this paper, tokenized abstract text from the NSF funding proposals have been embedded into vectors with 300 dimensions.

When handling unstructured text data, there have been several ways that documents have been converted into vectors. Some of these techniques, such as bag-of-words (BOW) or Term Frequency Inverse Document Frequency (TFIDF), involve taking a word count of a document and vectorizing it as an input into a model. The main disadvantage of these techniques is that they do not capture the sequence or semantics of the words within a document. One Machine learning approach to this problem is to use Document Vectors, an unsupervised algorithm that learns continuous distributed vector representations for pieces of texts [3].

The Document Vector framework was initially inspired by the popular Word2Vec model. Word2Vec embeds words into vector space by training a neural network to predict surrounding words or context. The Document Vector model adds to this model by concatenating an additional document vector that is used in combination with the word vector to predict surrounding words[3].

DIMENSIONALITY REDUCTION

Visualizing this type of high dimensionality embedded data can be done by reducing the number of dimensions. Principal Component Analysis (PCA) is a very common dimensionality reduction technique. Principle components are projections of the data, uncorrelated and ordered by variance [4]. The first two or three principal components may maintain most of the variance within the data. Plotting these first principal components can be an effective way to visualize high dimensional data.

A more modern technique for visualizing high dimensionality data is the t-Distributed Stochastic Neighbor Embedding (t-SNE) method. Here, high-dimensional Euclidean distances between data points are converted into conditional probabilities. These probabilities represent the similarity between data points. A t-SNE approach can capture local structures within the high-dimensionality data while also revealing global structures such as clusters. [5]

Both PCA and t-SNE were used to reduce the dimensionality of the embedded abstract vectors to a 2-dimensional space for plotting.

K-MEANS CLUSTERING

K-means clustering is an unsupervised clustering algorithm that groups together N items into K groups or clusters based on the mean Euclidean distance from the center of a group. Clustering word vectors together has shown to correctly group similar words into corresponding clusters[6].

In this paper, K-means clustering was used to cluster similar document vectors together. The latest 100k proposal abstracts were vectorized and grouped together using K-means clustering. The whole dataset was not used to reduce the amount of system resources required to process and cluster the abstract text.

EVALUATION AND VISUALS

ELBOW METHOD

The number of clusters (K) chosen for K-means clustering of the embedded NSF abstract proposals was determined by using the Elbow Method. The Elbow Method involves performing k-means on the data set with a varying number of clusters. The amount of variance within clusters as a function of the number of clusters (K) is then used to determine the optimal number of clusters [7].

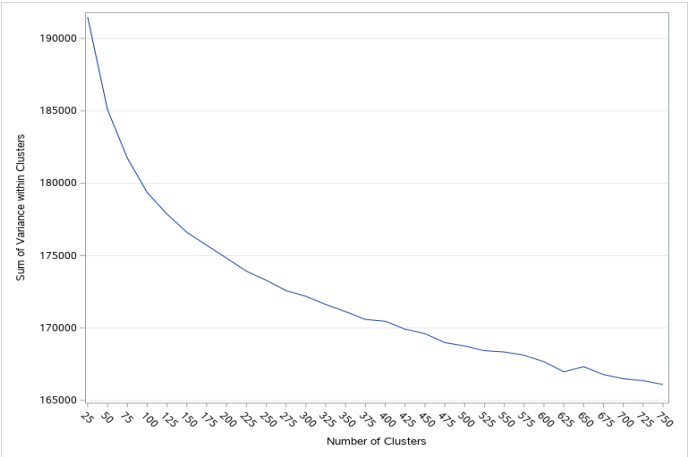

Figure 1. Sum of Variance with Cluster vs. Number of Clusters

This Elbow method does show an initial decreasing in the total variance within clusters as clusters are added. At some number of clusters, adding more clusters may not have a meaningful impact on the variance within a cluster. The number of clusters at this point, the elbow, is then used for the final model.

A K of 200 was used for the final clustering of the abstract embeddings which was determined from Figure 5. The sum of the variance within clusters does continue to decrease past this point. However, this point is in the center of the elbow region. Using a larger number of clusters for this initial project would have produced many clusters with very few abstract embeddings with these clusters.

CLUSTERS INSPECTION

Figures 2 shows the techniques for plotting high dimensionality data with different colors representing different clusters. While both PCA and t-SNE make it possible to visualize high dimensionality data. It is evident that the t-SNE approach produces a visualization that better shows the shape of the clusters of embedded NSF abstracts. Visualizing the first two principal components from PCA fails to show any meaningful structure withing the embedded vector space.

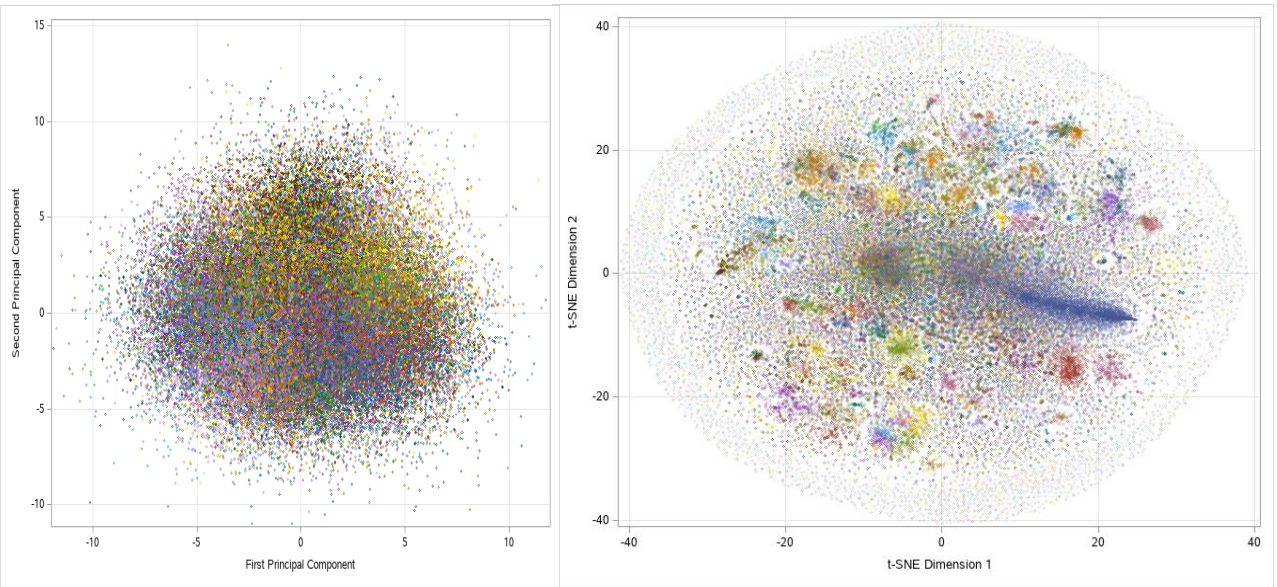

Figure 2. Cluster visualization using PCA (Left) and t-SNE (Right) to reduce dimensionality

Closer inspection of the clusters revealed some interesting insights. One cluster of 99 abstracts grouped together proposals mostly related to infant and childhood development, more specifically, visual, sensory, and language development. These proposals came from several different funding programs such as Developmental Sciences, Linguistics, and Cognitive Neuroscience. There are even some robotics-related proposals that fit in nicely with other proposals in this cluster. Some examples of the proposal titles are:

- Effects of induced maternal stress on the mother, infant, and dad
- Development of Executive Function in Pre-Crawling Infants: The Effect of Robotic-Assisted Locomotor Experience

- The Role of Maternal Sensory Stimulation on Postnatal Development of Language and Communication Skills in Extremely Preterm Infants
- BRAIN EAGER: Robust longitudinal characterization of brain oscillations in the first 3 years of life
- Development and adaptation of active dependency completion mechanisms
- The Role of Maternal Sensory Stimulation on Postnatal Development of Language and Communication Skills in Extremely Preterm Infants

Another cluster with 49 abstracts focused on social research around why people were or were not pursuing STEM-related academic careers, particularly around women in STEM programs. Some of the titles of these abstracts were:

- Diverse Young Women Traveling Pathways to STEM
- A Multi-Method Investigation of the Situational Cues and Contexts Inhibiting Women in STEM Settings
- Patching the STEM Pipeline between College and Work: Investigating Gender Issues in Embeddedness
- Broadening Women's Participation in STEM: The Critical Role of Belonging
- Peer influences on adolescents' self-concept, achievement, and future aspirations in science and mathematics: Does student gender and race matter?
- Decreasing Women's STEM Attrition by Normalizing Ability Concerns

A third cluster of 36 abstracts contained research proposals around metallurgical studies of alloys and superalloys. Some of these titles included:

- Non-Classical Precipitation Mechanisms in Titanium Alloys
- Designing New Economical High-Temperature Aluminum Superalloys
- Fundamental Influences of Grain Size on Oxidation Behavior of Nanocrystalline Alumina-Forming Alloys
- Accelerated Development of Next Generation of Ti Alloys by ICMSE Exploitation of Non-Conventional Transformation Pathways
- Mechanistic and Microstructure-Based Design Approach for Rapid Prototyping of Superalloys

The largest cluster had about 11.5 thousand funding proposals. I brief look at some of these abstracts showed that many of them were to fund individuals to conduct summer internships abroad. There were also a lot of Astronomy related proposals in this cluster.

CLUSTERING ON NEW TEXT

The abstract text from this paper was run through the model to demonstrate the ability of this technique to work with new unseen data. It was vectorized and clustered. Inspecting the assigned cluster found proposals for funding workshops to provide new researchers and investigators with necessary information about the NSF proposal submission process.

GENERALIZATION

This technique for embedding abstract text and clustering can work with any length of text. So, the particular model trained in this paper could find similar funding proposals based on any input text. This could be news articles, a title of a paper, or a few quick sentences of some research topics of interest.

Furthermore, embedding and clustering can be applied to a wide variety of applications and problems. This technique could be used to cluster and classify legal documents or healthcare

claims records. Extending this technique to new applications is only limited by the creativity and imagination required to effectively embed a data object into vector space. Once in vector space, unsupervised clustering techniques can aid in discovering patterns among the embedded objects.

FUTURE WORKS

The merit review process considers several causal factors for a funding proposal. Including more significant causal factors into the embeddings can improve the performance of the embedding and clustering technique. Research Proposal Text data (Summary of Intellectual Merit and Broader Impact) can be leveraged to further enhance the performance of the model. Any available outcomes reports could also be embedded and used for clustering.

Time series analysis techniques could also be performed on the dataset using all the features to predict a funding range for new proposals.

Incorporating recent techniques from natural language processing, like Google's BERT, Pre-trained Word Embedding could improve the quality as well. Other tokenization approaches such as stemming, lemmatization, and removing stop words could also be studied to determine the effect on clustering.

Determining the quality of the cluster model in the paper required a subjective look at the proposals within a group. Future studies could pursue a more qualitative approach to determine the effectiveness of the Doc2Vec and K-means clustering approach performed in this study. This could include identifying known abstract texts that are similar and should be grouped together.

CONCLUSION

In conclusion, this analysis met its goal of identifying similar NSF funding proposals based on proposal abstract text. The models built are precise and reliable on recommending similar abstract. Apart from the review performed by the NSF Committee for a research proposal for approving funding. This analysis attempts to expand the horizons of great research by the use of Natural Language Processing and unsupervised clustering of Similar Documents. The study of NSF Data encourages the awardees to improve their proposal content by suggesting most similar funded proposals.

A combination of insights from SAS® and JupyterLab were used to determine comprehensive and informative insights. Among the various models built and tested, Doc2Vec and K-means provided the best interpretable results of abstracts from the NSF corpus data.

REFERENCES

[1] National Science Foundation, 2019. "About the National Science Foundation". Accessed November 30, 2019. https://www.nsf.gov/about.

[2] National Science Foundation, 2019. "About the National Science Foundation". Accessed November 30, 2019. https://www.nsf.gov/about/glance.jsp.

[3] Le, Quoc and Tomas Mikolov. May 2014. "Distributed Representations of Sentences and Documents." Google Inc.

[4] Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. November 11, 2013. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[5] Van Der Maaten, Laurens and Geoffrey Hinton. November 2008. "Visualizing Data using t-SNE." Journal of Machine Learning Research.

[6] Ma, Long and Yanqing Zhang. 2015. "Using Word2Vec to Process Big Text Data." 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA., pp. 2895-2897.

[7] Bholowalia, Purnima and Arvind Kumar. November 2014. "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN." International Journal of Computer Applications. Vol 105, No. 9.

CONTACT INFORMATION

Jacob Noble jnoble@stthomas.edu Himanshu Gamit himanshu.gamit@stthomas.edu

ACKNOWLEDGMENT

We would like to acknowledge Dr Manjeet Rege for being the advisor on this project. The students and staff at the University of St. Thomas (Graduate Programs in Software) were inspirational and supportive in our intellectual pursuits.