Classy Data Analysis Project
Final Report

Advisors: Ben Cipollini (Classy) and Ilkay Altintas
Project Team:

- **Budget Manager:** Jeet Nagda | **Project Manager:** Howard Tai | **Project Coordinator:** Erin Hansen | **Report Manager:** Carlos Pimentel |**Record Keeper:** Juan Reyes

Abstract

Comparing and understanding nonprofit organizations is a task typically done by humans and is notoriously time-consuming. While incredibly useful and crucial, it is difficult to systematically gather evidence and data about each organization for the purposes of objectively evaluating them. By filtering on broad categorical labels, a nonprofit researcher might find that organizations such as the NFL and the Women's March are located in the same bucket, which does not provide any meaningful value. To address this shortcoming, we aim to create a new nonprofit organization latent space that can help compare and contrast organizations with greater efficacy and efficiency. By using hierarchical agglomerative clustering, we created a latent space composed of three major axes of comparison: financial standing, organizational purpose, and geographic location. We found that organizations are best understood and compared along these three pillars. This space allows for both searching and similarity ranking using key metrics unique to each pillar. By creating independent and unique clustering models for each axis we were able to successfully encode entirely different feature spaces into a common space. We also present a visual depiction and web-app tool of this new space.

Introduction and Question Formulation

The social sector is typically viewed in terms of nonprofit organizations and the cause categories they belong to. It's clear, however, that while younger generations are active in social causes, they think more in terms of current events and social causes organizations - so much so, that new donor churn now peaks at over 80%.

It is not clear, however, how to lay out a common "social space", where the organizations that drive social change and potential donors could connect, find organizations, get cause-based recommendations, and where discovery of nonprofits could be facilitated. A new social space can also help local organizations find and engage partner organizations in alliances.

In this project, we aimed to build this social space from the ground up. We began with public data generated by non-profit organizations and used these text and financial data to create a low-dimensional map of the social sector, visualizing each non-profit organization's location within this space. This new space describes organizations in terms of its clusters' positions within this space.

The foundational guiding questions that fuel our investigation are:

1. *How can similar organizations be associated with each other on the basis of purpose and mission along with their balance sheet?*
2. *How can a multi-faceted problem involving semantic understanding, financial awareness, and geographical proximity be condensed into an insightful and actionable representation?*
3. *What is the best way for our end users (nonprofit researchers, potential donors) to interact with our model and algorithms?*

Related Work

Prior to our capstone project, previous work in the nonprofit clustering space primarily focused on utilizing the existing numerical-valued fields on the Form 990 or 990EZ to navigate the nonprofit latent space. However, this sort of effort is heavily dependent on an expert-level knowledge on the idiosyncrasies of the Form 990s and non-profit tendencies in practice. Since tax codes change periodically, fields within the 990s change names as well resulting in an undesirable situation where fields that represent the exact same information are given entirely different names. As such, creating a single unified representation of 990 forms across several years is actually a herculean task by itself. To further complicate things, fields within the 990s can share nested relationships where one attribute references another attribute, or sometimes another tax form entirely. Additionally without experience understanding non-profit behaviors and patterns, it is impossible to determine if the financial numbers are subjectively "good" or similar.

To our knowledge, our efforts to tightly integrate semantic meaning into the clustering challenge is a first in this domain. While previous efforts focused primarily on a single descriptive axis (i.e. financial profile), we hope that introducing 2 additional axes will result in a latent space that is more conducive to clear separation amongst nonprofit groups. By including rich textual meaning, we are also able to reduce our numerical features to only the essential and easily-interpretable attributes, thereby simplifying the unified representation across 990 forms.

Team Roles and Responsibilities

- **Budget Manager:** Jeet Nagda
  - Management of AWS funds and cloud-related activity
- **Project Manager:** Howard Tai
  - High-level project planning and assignment of responsibilities
- **Project Coordinator:** Erin Hansen
  - Stakeholder correspondence and record-keeping during meetings
- **Report Manager:** Carlos Pimentel
  - Tracking and coordinating report deliverables

- **Record Keeper:** Juan Reyes
  - Management of Github repo and Docker environment

Data Acquisition

Our core dataset will be the IRS Form 990, a massive tax form that is publicly available in electronic form for many nonprofit organizations. These tax data forms include both financial balance sheet information and text describing the organizations purpose. The mandate of the 990 Form is to claim exemption from income tax for non-profit organizations. We also explored supplementing these tax forms with organization website data. There are a total of 1.5 million non-profit organization of varying sizes and purposes that file the form 990 every year.

The IRS Form 990 data was acquired using an IRSX tool publicly available through GitHub. The IRSX tool invokes an API call to download the XML data payload for a particular organization's form 990. We have the capability of acquiring supplemental web data through web scraping of the non-profits' websites.

Data Preparation

Our dataset was extremely large (with over 1 million samples) which presents a daunting challenge in creating a clustering algorithm that can work for all types of organizations. However, the data samples are relatively clean since the data submitted to the IRS needed to be clean for IRS acceptance. Additionally based on the size and gross receipts of the organization, organizations could elect to file a Form 990EZ which has fewer fields than the generic Form 990. The form 990EZ also had different XML names and information for the financial balance sheet making it necessary to handle both form types.

Data Transformation and Integration:

By utilizing the IRSX tool, we were able to invoke a REST API to download an XML document for each organization already neatly tagged by section. We converted the XML to JSON as we stored it into a MongoDB instance. Once the data was persisted, we could use the document attributes to search and query the organizations for our modeling. Our textual information for the organization purpose was also stored in the same JSON document in this way and available for querying.

Feature Engineering:

Since we created clustering models for each axis of comparison we created different sets of features for each.

Financial Standing

We used key metrics on the balance sheet that are easily compared against other organizations. Some of the most crucial features were age of the organization, and how much money was coming in versus going out. By comparing the gross receipts and the expense amounts year over year, we were able to embed the financial standing of the organization in our clustering algorithm.

By additionally creating ratios of financial numbers we were further able to combine correlation of two variables into one to guide the cluster model towards identifying key trends correctly.

Feature Ratio	Significance
Age	Age of organization indicates maturity
Contribution Revenue %	Able to determine how much of the organizations revenue is from donations
Program Services Expense %	How much the organization is spending on programs to fulfill purpose
Mgmt & General Expense %	How much of the expenses is being spent on management and general expenses, this could indicate bad leadership
Fundraising Expense %	How much of the expenses are spent on generating new revenue
Assets Y/Y	How much the assets changed from the beginning of the year to the end
Salary Y/Y	How much the salary amount changed from the beginning of the year to the end
Expense Y/Y	How much the expense amount changed from the beginning of the year to the end
Fundraising Amount Y/Y	How much the amount spent on generating fundraising changed from the beginning of the year to the end
Revenue Y/Y	How much the amount spent on generating revenue changed from the beginning of the year to the end

The amount and type of data is inconsistent between organizations; however, by conducting the feature engineering described above and finding meaningful relationships between certain fields, we were able to reduce the number of numerical fields from upwards of one hundred down to thirty. Dimensionality reduction through principal component analysis further narrowed our data from thirty features down to five explaining over 90% of the variance.

Organizational Purpose

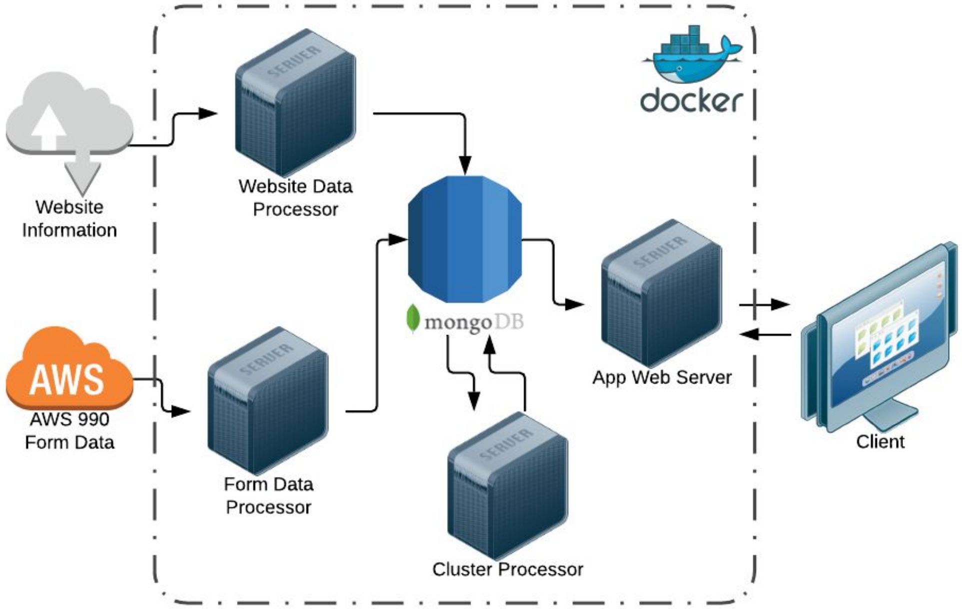
The Form 990 has a required subsection where filers must include a small description of the organization’s description and self-declared purpose. These subsections allowed the IRS to categorize and lookup information about the organization if needed. We used these sources of text in the Universal Sentence Encoder (USE) for our clustering. Pre-processing of this text includes making all the text lower case, clearing all numerical digits, and removing punctuation marks. We leave the stop words for the sentence encoder to allow encoding of meaning and order. The Universal Sentence Encoder outputs embeddings in a 512 feature space which is subsequently reduced down to a 200-dimensional space using principal component analysis (PCA). This reduction operation was able to retain 95% variance while eliminating more than half of the original features.

In our development process, we also intended on integrating web-scrapped text as part of the “Organization Purpose” data. The idea behind this decision was rooted in an observation that some nonprofits have relatively nondescript mission statements in the Form 990s. Since semantic meaning contributes significantly to our cluster fidelity, we desired to have an additional pool of data to bolster the richness of nonprofits’ textual description. Though our final product does not utilize this functionality due to the general uncleanliness of web-scrapped text, we believe that future efforts in this area can potentially yield even more powerful text-based features.

Geographic Location

For encoding geographic location we simply use the center of the zip code region provided on the Form 990. This allowed us to compute the Haversine distance to indicate if two organizations were nearby.

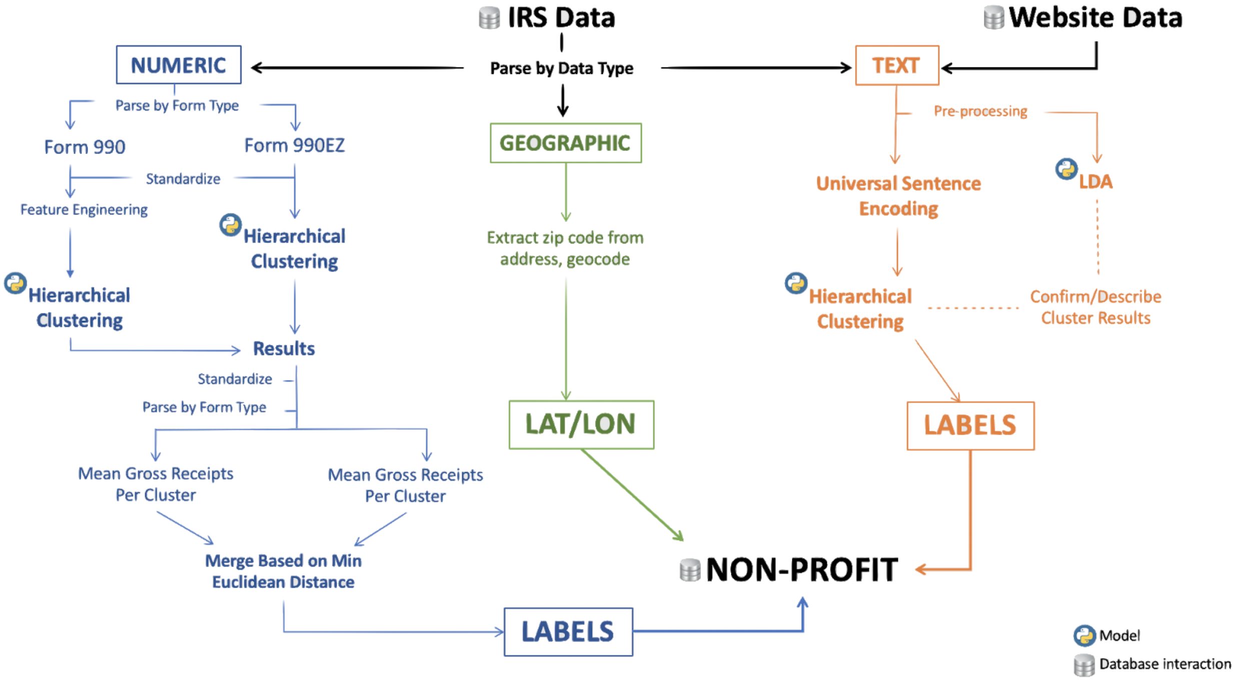
Data Pipeline and Architecture:



Our data pipeline was divided into different modules to accomplish discrete tasks. In our final product, we used Python and Docker throughout as our common language and platform. The purpose of our first module (Form Data Processor) was to gather and fetch information from the AWS repository using the IRSX tool for each of the organizations in a giant manifest. We had an additional experimental module (Website Data Processor) which additionally scraped the HTML text of the organization's website listed on the form 990. The XML payload was parsed and stored as a document in a MongoDB instance. We then had a clustering module (Cluster Processor) which read data samples from the MongoDB instance to create three labels or cluster IDs for each organization: one for each axis of comparison. These labels were loaded back into the MongoDb document for persistence. Finally our last module (App Web Server) reads from the database to create visualizations for high level queries, or a given input organization. This architecture is diagrammed in the preceding figure.

Analysis Methods

In order to create independent axes of similarity within our data without throwing away information, we created multiple branches of our pipeline to adhere to different data types (text, spatial, numeric) and different data volumes (form type).



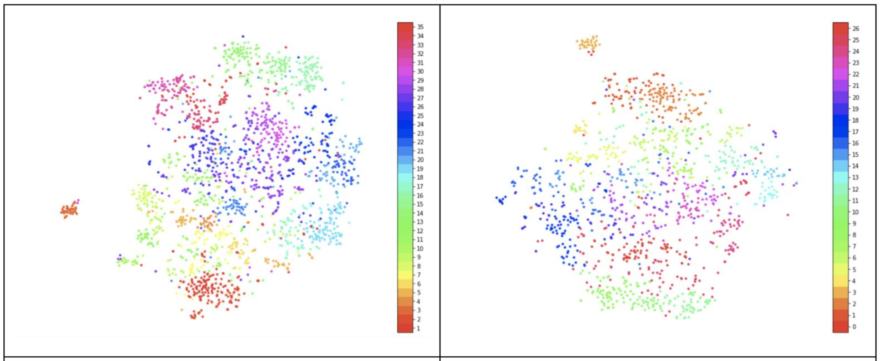
The result is three modeling algorithms plus geographic distance calculation. Our algorithms are hierarchical clustering models utilizing the gap statistic to find the most natural number of clusters for each group. Because we must split by form type, two of the three models are within the financial pipeline, though we wish the merge the results and have only one financial space in the end. Since each form type was first determined by the organization's gross receipts, we can calculate the mean Gross Receipts Amount for each cluster in both form types, and merge the 990EZ cluster with the 990 cluster with the minimum Euclidean distance for this metric.

The final database describes each organization in just four attributes: cluster labels for the text and financial spaces, as well as latitude and longitude for the geographic space.

Organizational Purpose (Text):

The organizational purpose text model performance is best explained using a visual medium. To provide a brief overview, the text clustering model was evaluated with the purpose of assessing the quality of clusters. Since this is an unsupervised task, the traditional notions of "accuracy", "precision", or "recall" cannot be applied. Due to this limitation, for our first method we decided to utilize a t-SNE decomposition to obtain a visual validation of cluster behavior along with a latent Dirichlet allocation (LDA) based method for validating cluster cohesiveness.

The intuition behind our LDA validation approach is rooted in the fact that the Universal Sentence Encoder (USE) embeddings are completely blind to the LDA transformation. By using LDA to independently model the topics within the text corpus and subsequently use these topics to assess the USE-embedded clusters, we are effectively benchmarking USE, hierarchical clustering, and the gap statistic against LDA. Our expectations for this validation is that each USE cluster should have a predominant LDA topic (or at most, a few topics). As a validation of USE embeddings' robustness, we found that a PCA decomposition of the embeddings from 512 to 200 retained 95% variance while a reduction down to 70 dimensions kept 84%.

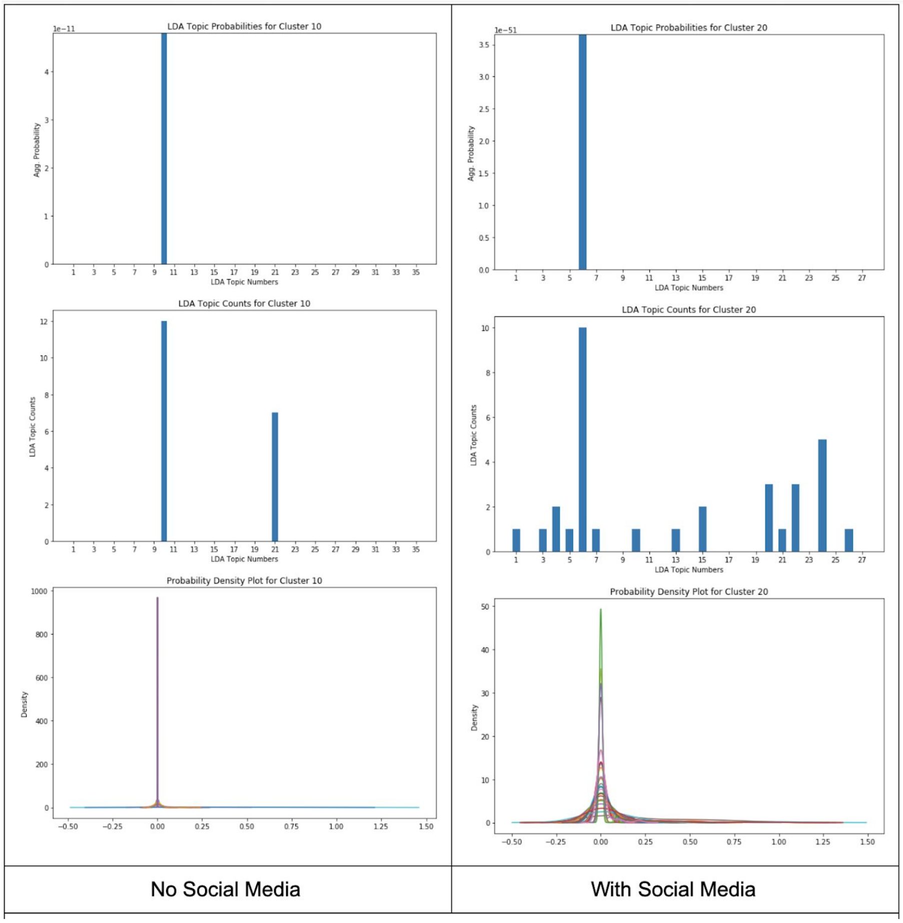


No Social Media

With Social Media

The figures above and below contain visualization results for the USE-embedding text model. In both figures, side-by-side comparisons of vanilla documents and website-enhanced documents are shown.

For Figure 1, t-SNE decomposed representations of the 512-dimensional USE embeddings are depicted. From a single glance, it can be seen that in both plots, clear clusters can be readily identified. Interestingly, it appears that the vanilla corpus yielded tighter clusters than the website-augmented version. However, this observation is well-aligned with the quality of the web-scraped text which is not consistently reliable. This finding prompted us to proceed with the capstone project without the inclusion of scrapped text since the development of a preprocessing strategy that can successfully clean the wide spectrum of web-scrapped text would’ve been a significant investment.

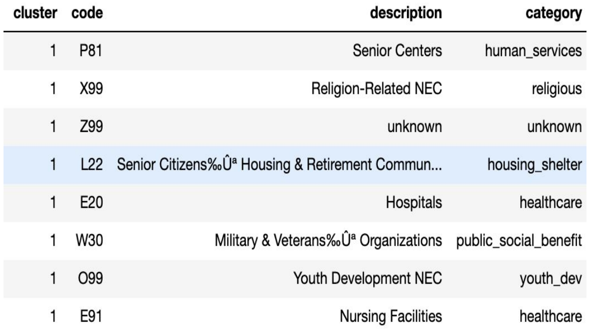


This figure tells a more interesting tale. The two sets of plots shown in Figure 2 describes our cluster-vetting procedure for both corpus types. The top-most plot is a bar plot depicting aggregate topic probabilities for all documents within a certain USE cluster (cluster 10 for the left-hand figure). The middle plot is a bar plot summarizing the counts of the most-likely LDA topics to have occurred within a USE cluster. For the left-hand figure showing results for Cluster 10, it can be seen that LDA topics 10 and 21 are the most-popular. Finally, the bottom-most plot depicts the probability density plots for every LDA topic within a specified USE cluster.

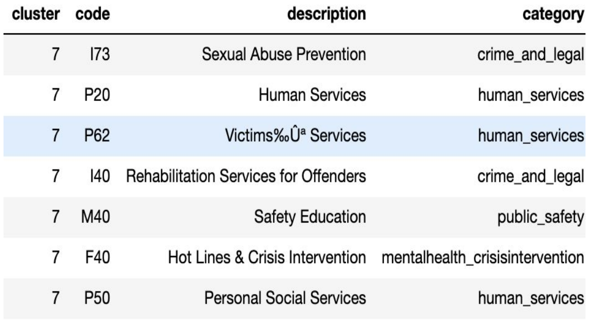
This arrangement of graphics enabled us to rapidly assess the quality of each USE cluster by cross-referencing the three plots simultaneously. It should be expected that the spike identified in the top-most plot will be mirrored in the middle plot. In the case of Cluster 10 (shown on the left), it can be seen that LDA topic 10 is indeed reflected in both top and center. The absence of topic 21 in the top plot gives us additional insights that while LDA topic 10 is the dominant LDA topic of Cluster 10, LDA topic 21 can be considered a sub-topic. The massive spike in the bottom-most probability density plot is indicative of USE Cluster 10 being very compact.

For our second method of model validation, we incorporated the National Taxonomy of Exempt Entities (NTEE) system used by the IRS to classify non-profit organizations by their activity. This is the current standard of practice and could be viewed for our purposes as the baseline model. The NTEE codes are acquired from the IRS Exempt Organization Business Master File Extract (EO BMF). This data is acquired by US region and then aggregated. We then combine these codes with their descriptions from the IRS website. Below we compare our clusters generated by our text model and how the IRS classifies these records. It can be seen by the makeup of NTEE codes within each cluster that the results from our clustering algorithm are logical. It is also evident from the diversity of NTEE codes within these clusters that limiting an organization’s peers to a single code is often too strict of a system. It may be appropriate -- and perhaps even more accurate -- to group organizations that operate in the “crime and legal” space with “public safety” organizations, for example. Grouping “mental health crisis intervention” with “healthcare” organizations also seems logical in some cases, but these peers would be overlooked with the given NTEE code system.

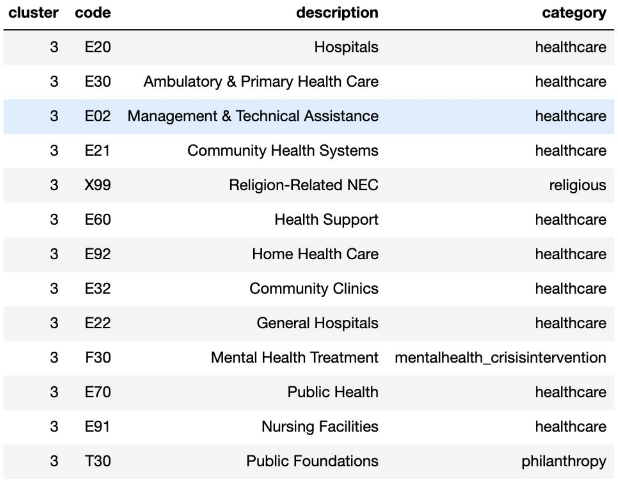
Unique NTEE codes by Generated Text Clusters:



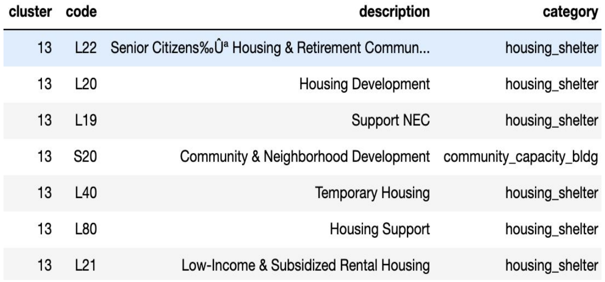
cluster	code	description	category
1	P81	Senior Centers	human_services
1	X99	Religion-Related NEC	religious
1	Z99	unknown	unknown
1	L22	Senior Citizens‰Ûª Housing & Retirement Commun...	housing_shelter
1	E20	Hospitals	healthcare
1	W30	Military & Veterans‰Ûª Organizations	public_social_benefit
1	O99	Youth Development NEC	youth_dev
1	E91	Nursing Facilities	healthcare



cluster	code	description	category
7	I73	Sexual Abuse Prevention	crime_and_legal
7	P20	Human Services	human_services
7	P62	Victims‰Ûª Services	human_services
7	I40	Rehabilitation Services for Offenders	crime_and_legal
7	M40	Safety Education	public_safety
7	F40	Hot Lines & Crisis Intervention	mentalhealth_crisisintervention
7	P50	Personal Social Services	human_services



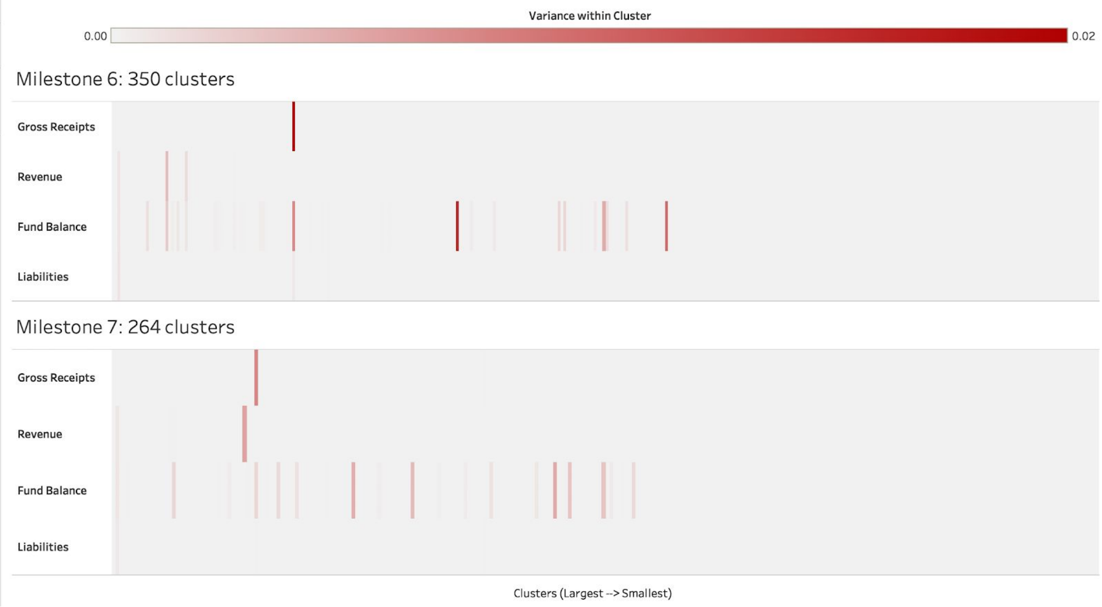
cluster	code	description	category
3	E20	Hospitals	healthcare
3	E30	Ambulatory & Primary Health Care	healthcare
3	E02	Management & Technical Assistance	healthcare
3	E21	Community Health Systems	healthcare
3	X99	Religion-Related NEC	religious
3	E60	Health Support	healthcare
3	E92	Home Health Care	healthcare
3	E32	Community Clinics	healthcare
3	E22	General Hospitals	healthcare
3	F30	Mental Health Treatment	mentalhealth_crisisintervention
3	E70	Public Health	healthcare
3	E91	Nursing Facilities	healthcare
3	T30	Public Foundations	philanthropy



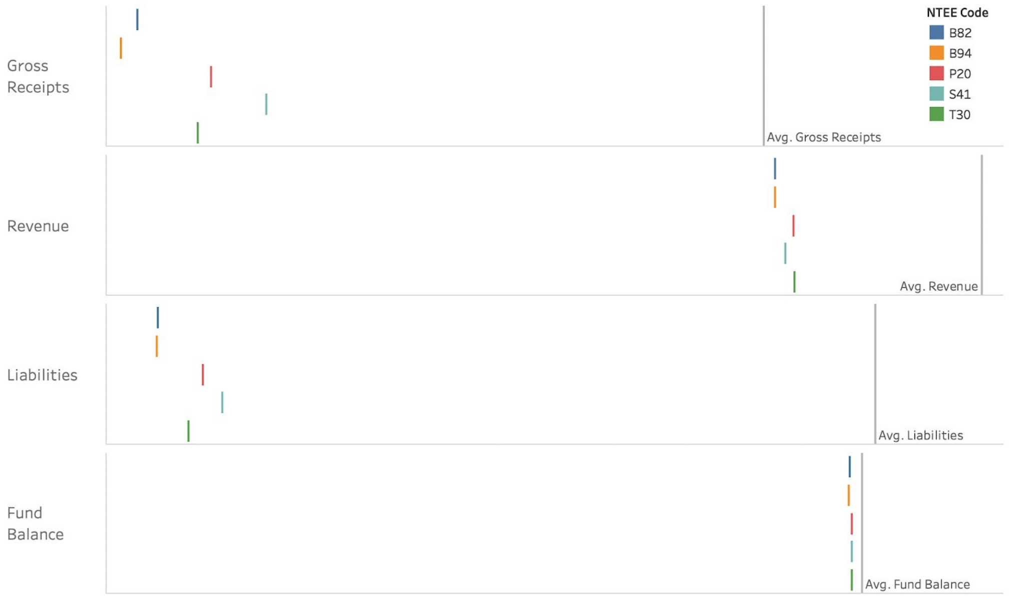
cluster	code	description	category
13	L22	Senior Citizens‰Ûª Housing & Retirement Commun...	housing_shelter
13	L20	Housing Development	housing_shelter
13	L19	Support NEC	housing_shelter
13	S20	Community & Neighborhood Development	community_capacity_bldg
13	L40	Temporary Housing	housing_shelter
13	L80	Housing Support	housing_shelter
13	L21	Low-Income & Subsidized Rental Housing	housing_shelter

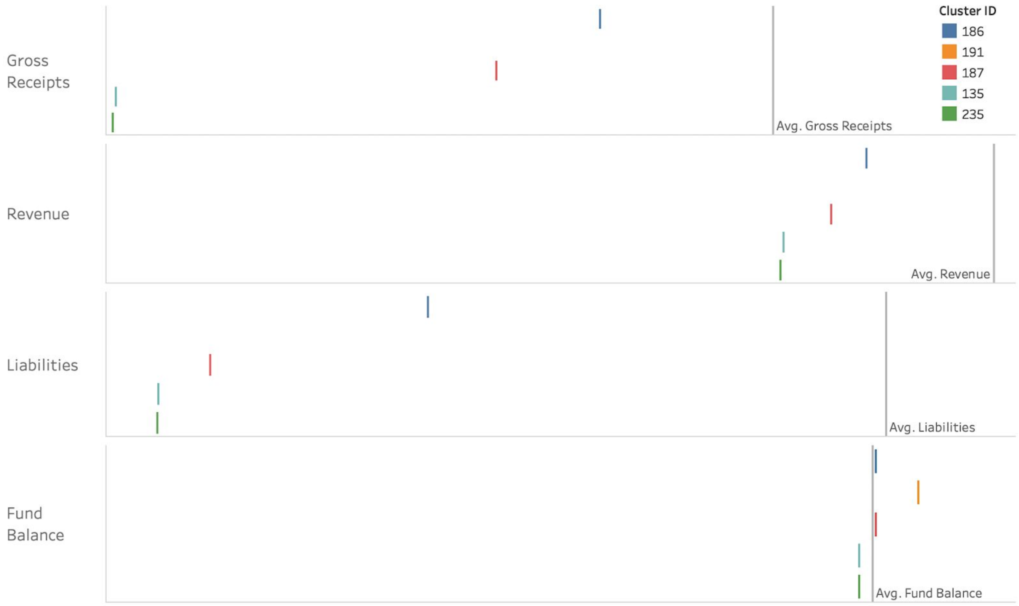
Financial Standing (Numerical) Model:

Two key metrics we will use to evaluate our financial clusters are *variances of core financial metrics within clusters* (ideally the in-cluster variance will be low for all metrics and clusters) and the *cluster distribution*; i.e. number of clusters and cluster sizes. Over a handful of iterations in our modeling process, we were able to reduce the number of clusters outputted by 38%. The figure below is a heatmap of in-cluster variance, with the darker red squares indicating a higher variance. It is evident that even though the number of clusters reduced by 84 between these two model iterations, we did not introduce much in-cluster variance. In fact, in many cases we were even able to reduce it. This proved that the new implementations during our Evaluation and Interpretation phase were effective.



Similar to the text cluster validation, we wished to compare our clusters against those formed by grouping NTEE codes to show we are able to beat the baseline model. Seeing as the NTEE codes are self-reported and have little to do with financial metrics, we expected our model to have much better results. The two figures below depict the mean values of the four core financial metrics for the five largest groupings in each methodology.

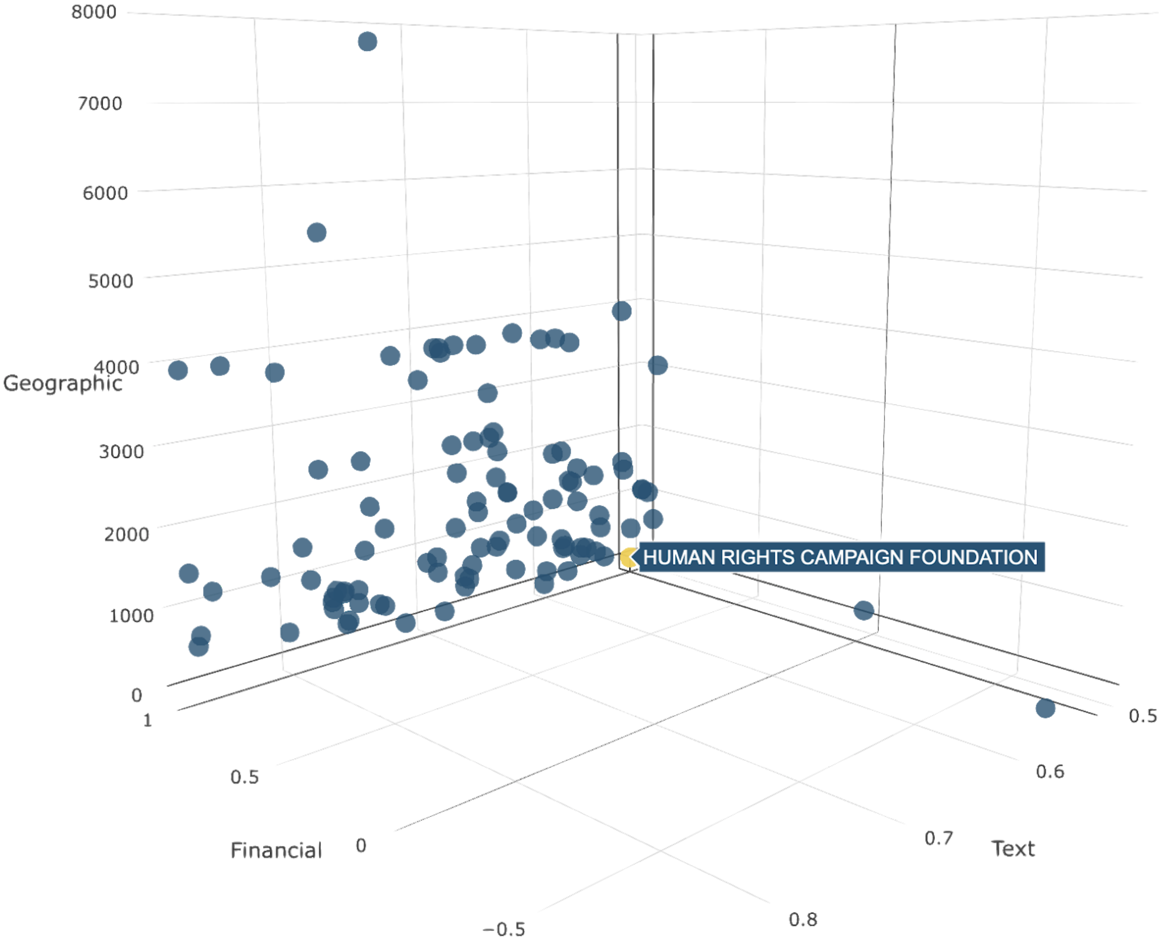




The intention behind this visualization was to test the hypothesis that grouping by NTEE codes would be similar to taking a random sample of organizations. If the means of these groups hovered around the overall averages, are hypothesis would be confirmed; however, as seen in the top figure, this turned out not to be the case. The five largest groups formed by NTEE codes have averages much smaller than the overall averages, indicating there may be certain NTEE codes where nonprofits are larger and drive the averages up for these metrics. What we can glean from these visualizations is that the groups created by the NTEE codes are very similar to one another, proving there is not much financial diversity between these clusters. The clusters created by our algorithm (shown in the bottom figure) are capturing more of the financial qualities of the organizations in our sample, which is exactly what we were hoping to achieve.

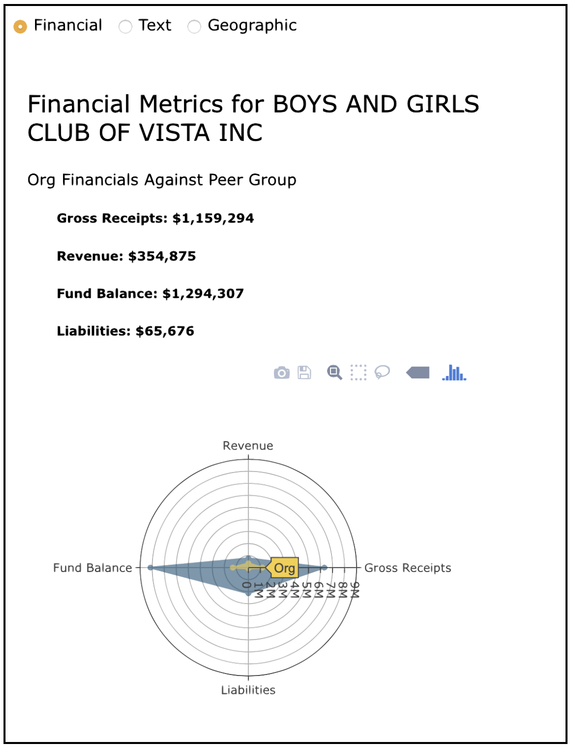
Findings and Reporting

Perhaps our most significant finding is that the social space we sought to define is not one-dimensional. By utilizing all three of our axes of comparison, we can create a three-dimensional social space that is both comprehensive and interpretable. In this way, very strong (or very weak) associations in one space do not have an overpowering influence on the final similarity measure. Instead, we are returning that power to the end-user, allowing them to see results in all three spaces and decide which definition of “similarity” holds more weight for their purpose. The figure below is a visual example of this space, depicting the Human Rights Campaign Foundation and its peer group.



To communicate the results of similarity measures in three spaces, we created a web application using Dash by Plotly. The main area of focus in our dashboard is a three-dimensional scatter plot similar to the one above. Because there are hundred of thousands of nonprofits in the database, we allow the user to select an organization of interest and depict a single peer group at a time. Each of the three data spaces originally contained a multitude of vectors that the clustering analysis used to calculate relationships. Now we intend to collapse those into a single dimension -- a projection of those values onto a single axis -- by taking into account the distance on the respective space of the primary organization and its neighbors. This is done by calculating the cosine similarity between the vector representing each organization and the "centroid" vector of the cluster hosting the organization of interest. The centroid vector is simply the mean of the cluster.

Once we project each one of the three data types (numerical, text, and geographical) towards an individual axis we have an ideal situation to consolidate those metrics into a simple X,Y,Z plot rendered by our graphical interface and easily interpreted by the users visually. Any collection of points in a 3D volume could be visually analyzed and all relationships between different objects compared quickly. It is intuitive and natural for humans to extract knowledge from such visualization.



● Financial ○ Text ○ Geographic

Financial Metrics for BOYS AND GIRLS CLUB OF VISTA INC

Org Financials Against Peer Group

Gross Receipts: $1,159,294

Revenue: $354,875

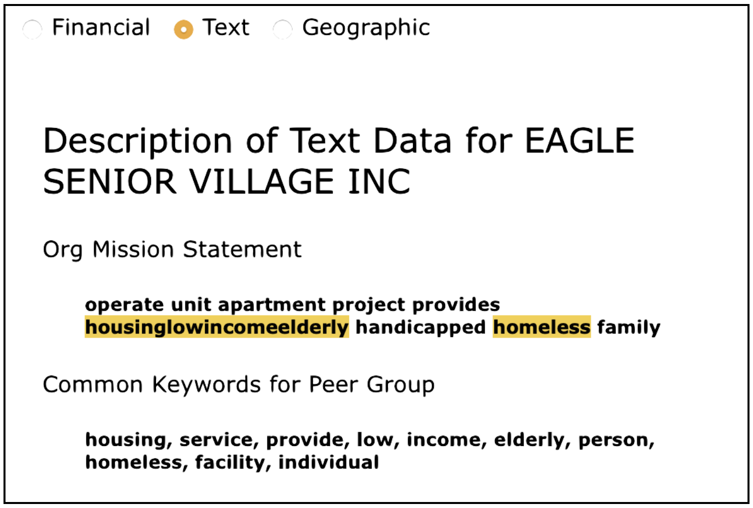
Fund Balance: $1,294,307

Liabilities: $65,676

The distance between points on each of the axes in the plot is directly proportional to the spatial distance our clustering analysis produced for each of the data types. Since all three axis are orthogonal to each other we eliminate any possible distortion or visual dependencies that one data type could exert on another data type. The distances on each axis might not be on the same scale but within a particular axis all distances will be accurately represented and correctly encoded without unintended artifacts or visual limitations. The visual distance between any two points in the volume will be proportional to the mathematical distance that we calculated based on clustering. This visual distance and relativity is the idea we wish to capture in the visualization.

The scatter plot is supplemented with further information regarding each space. We wished to show the user how the organization of interest compared to the rest of its peer group and perhaps demonstrate why it was placed in the peer group it ended up in.

For the financial standing of organizations, we chose a radar plot to depict four key financial metrics and how an organization is performing relative to the average of its peer group. After feature engineering and PCA, these four metrics (gross receipts amount, revenue, liabilities, and fund balance) were not direct inputs of the model. They are required fields for every organization and form type, and, therefore, were considered to be “core” financial metrics to describe an organization. In most cases, the shape of the radar plots for the organization were similar to that of its cluster, possibly indicating that ratios between these metrics were considered more telling of financial status by our algorithm than the dollar amounts themselves.



○ Financial ● Text ○ Geographic

Description of Text Data for EAGLE SENIOR VILLAGE INC

Org Mission Statement

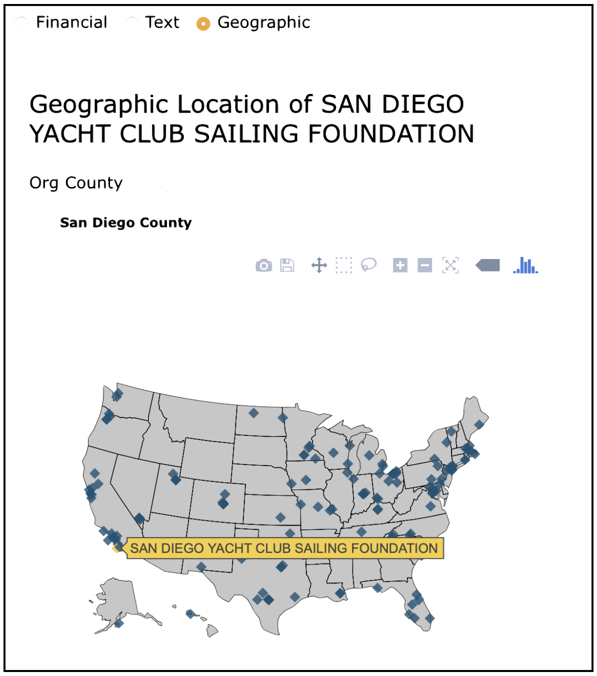
operate unit apartment project provides housinglowincomeelderly handicapped homeless family

Common Keywords for Peer Group

housing, service, provide, low, income, elderly, person, homeless, facility, individual

Our dashboard contains both the mission statement of the organization of interest and the keywords defining its peer group in the text space. The keywords are extracted via LDA topic-modeling on the entirety of the text data within each cluster. They are highlighted in the mission statement after lemmatizing and removing punctuation.

We found that organizations with similar purposes quite frequently used the same or similar key words. This allowed us to explore and visualize the topics and key words for a given cluster as a way to summarize a peer group of organizations. Each topic reveals a latent topic and purpose that may not have existed, similar to the specific NTEE codes. Because each cluster has a strong dominant topic that describes the group, users can read and understand if the topic of the peer group aligns with their search parameters.



○ Financial ○ Text ● Geographic

Geographic Location of SAN DIEGO YACHT CLUB SAILING FOUNDATION

Org County

San Diego County

Finally, the geographic distance is depicted in the most logical way possible: a simple geographic map, where the organization of interest is highlighted, and its surrounding peers can be identified through viewing the name of the organization in the tooltip. Our map is focused on the United States region since the tax data is also for the United States.

Through this dashboard, end users can query organizations of interest by inserting its name or the unique identifier given on the Form 990 (EIN). From here, the user can further explore the peer group by researching the other organizations identified by the tooltips in the visualizations.

Solution Architecture, Performance and Evaluation

As Figure 3 above demonstrates, our solution architecture centers on a single MongoDB instance that serves as the communication backbone between our various clustering models. This framework is subsequently wrapped in a Docker environment that is meant to be easily-deployed on a single computational cloud instance. The use-case that we imagine (from a deployment perspective) is that an interested party can quickly download our Docker image, load it onto an instance, and immediately perform nonprofit clustering.

However, to enable this vision, we need to have high confidence and evaluate the robustness of our clusters. Financial clustering (which operates on strictly numerical values) is interpretable at face-value. On the other hand, the results of text clustering are not. Though the analysis section above details two validation procedures (one using a global LDA approach and another using NTEE codes to cross-validate our results), both methods suffer from an inability to ascertain (on a local, per-cluster level) topic purity.

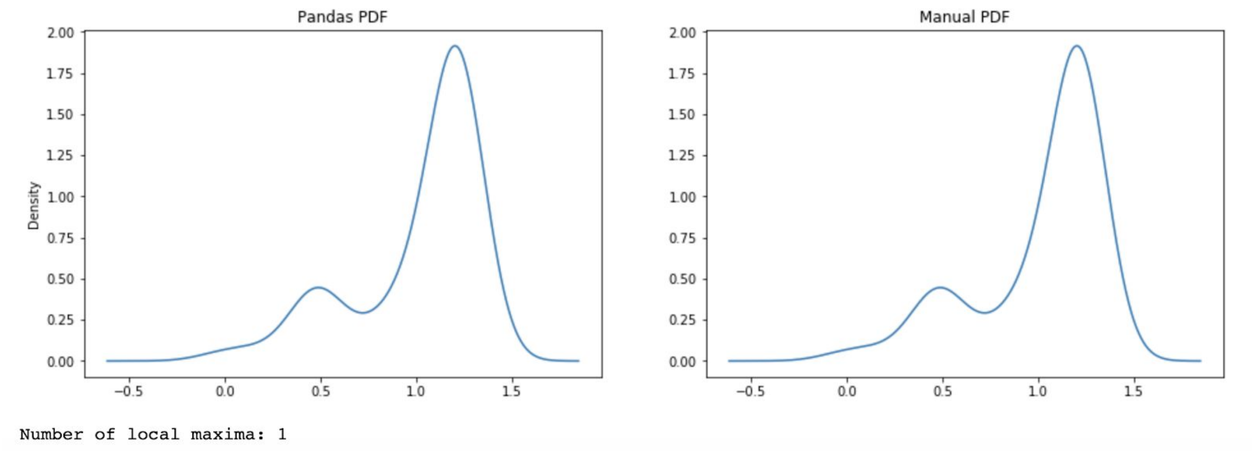
To address this shortcoming, we chose to focus the brunt of our validation efforts into performing a final and exhaustive validation into the quality of the clusters that we have. To accomplish this task, we modified our previous “global” LDA strategy to a “local” form that is superior for the reason that the global strategy is heavily reliant upon an assumption yielded from the gap statistic procedure: the optimal number of clusters.

Since the global LDA model is forced to adhere to a specific number of topics (given as the ideal cluster number), our previous validation procedure was intolerant to small fluctuations in the natural number of topics in a corpus. That is, the idea number of clusters produced from hierarchical clustering does not necessarily have to equal the number of LDA topics. Our answer to this issue is to evaluate topic coherence on a per-cluster basis by projecting each cluster’s samples into a 10-dimensional LDA topic space and subsequently assessing the density of points within this new topology.

The following strategy details this local LDA validation:

1. Cluster documents using hierarchical clustering and the gap statistic
2. For each cluster:
    a. Fit an LDA model on the cluster documents, setting the number of topics to be equal to 10 (arbitrarily)
    b. Transform cluster documents into the 10-dimensional topic space
    c. Construct a pairwise-distance matrix shaped (N, N) where N is the number of documents in a given cluster (various distance metrics can be used: Cosine, Euclidean, Cityblock, Mahalanobis)
    d. For each row (or equivalently, column) in the pairwise distance matrix, construct a probability density plot of distance values
    e. Find the number of local maxima (above a specified threshold) for each probability density plot
    f. Aggregate statistics for number of local maxima (mean, std. dev)

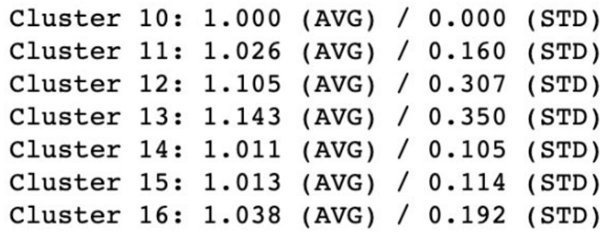
The local-LDA validation procedure outlined above generates a unique probability density function (PDF) for each point in each cluster. In this way, there are as many PDFs corresponding to a single cluster as there are points in that given cluster. Each of these PDFs are generated from a row / column of a pairwise distance matrix depicting the distance from a single point in a cluster to all other points in the cluster. In this way, if there are M points in a cluster, the pairwise distance matrix will be of the shape (M, M).



```
Number of local maxima: 1
```

In the above figure, we can see a comparison of our own PDF-generation function (on the right) versus a Pandas implementation. The first observation is that our implementation is exactly equivalent to the Pandas version (which means we are understanding probability density correctly). The second observation is that this sample PDF has exactly 1 prominent local maxima (even though the overall local maxima is 2). What this means is that in the 10-dimensional topic space that this sample has been projected into, this sample sits in a dense cloud of other samples. This notion of a “dense cloud” in LDA-topic-space can translate equivalently to one dominant LDA topic (which may very-well be a linear combination of the 10 different LDA topics, but one nonetheless).

Using this intuition, we hope that every point in a given cluster has a PDF that shows 1 local maxima and interpret such a finding to mean that all points in a specific cluster are well-clustered. This was our observation across just about every cluster created from our representational dataset (composed of around two thousand samples). This can be observed in the following image snippet which reports on the average number (and standard deviation) of prominent PDF peaks for each cluster.



```
Cluster 10: 1.000 (AVG) / 0.000 (STD)
Cluster 11: 1.026 (AVG) / 0.160 (STD)
Cluster 12: 1.105 (AVG) / 0.307 (STD)
Cluster 13: 1.143 (AVG) / 0.350 (STD)
Cluster 14: 1.011 (AVG) / 0.105 (STD)
Cluster 15: 1.013 (AVG) / 0.114 (STD)
Cluster 16: 1.038 (AVG) / 0.192 (STD)
```

What this means is that our clustering approach is successfully isolating organizations with semantically similar mission statements and descriptions which gives us confidence in our easily-deployable system architecture.

Scalability and Performance

Hierarchical Agglomerative Clustering is a clustering process that requires all of the data points to be held in memory to compute. This means the clustering model is limited by the RAM or memory space of the compute node used if the data volume or features cannot be reduced. We tested our clustering model on 6000 data samples and recorded the times using an AWS EC2 compute instance.

Clustering Model	Compute Specifications (AWS)	Time to Complete (milliseconds)
Financial Standing	122GB RAM, 16 vCPUs (r4.4xlarge)	1216703
Organizational Purpose	122GB RAM, 16 vCPUs (r4.4xlarge)	150673

We believed that with more samples and a larger compute node, the training time will remain in the order of minutes. This means that end users can install a pipeline that can scale with the number of samples with relative ease and cost.

Conclusions

In this capstone project, we've created a tool that nonprofit stakeholders of all kinds can use to quickly gain valuable, actionable insights from the IRS Form 990 database. Our work offers a tremendous value-over-replacement from the previous labor-intensive paradigm of utilizing Form 990 data. This product represents a departure from traditional methods of understanding tax data which involves an extensive and exhaustive effort into unraveling the myriad attributes on Form 990s, many of which change from year to year. Our approach demonstrates that selecting only a few key numerical-valued features in conjunction with utilizing rich text content results in human-interpretable clusters that can be decomposed into raw distance metrics in a 3D representation (each axis embodying one of the three descriptive axes: financial, text, and geography).

In terms of the financial axis, we found that fundamental data pre-processing greatly-improved clustering results. Basic actions such as splitting out dataset along form types and massaging vanilla form attributes into normalized representations such as percent changes and ratios. Removing fields that were highly correlated also aided in denoising the feature space. Our utilization of radar plots to visualize clusters in financial space revealed that in comparing organizations to their clusters, radar shapes were fairly consistent, indicating that the ratios between finance metrics were considered more characteristic of financial status than the dollar amounts themselves. Additionally the clusters may have similar financial ratios but with various thresholds to seperate them.

In terms of the text axis, we discovered that using the Universal Sentence Encoder to obtain text embeddings appeared to be an excellent method for capturing semantic meaning (i.e. the text clusters are very sensible from a human's point of view). Moreover, we validated that the gap statistic method for automatic cluster-scanning is capable of identifying the natural number of clusters and ideal number of topics within our dataset which was subsequently confirmed through a secondary LDA topic modeling.

Taken in conjunction with our tool’s geographic description axis, we’re confident that our project’s efforts have culminated in a robust nonprofit clustering and visualization tool. If future work is to be performed on our baseline algorithms and models, our solution architecture can readily scale to accept new feature sources, preprocessing schemes, and embedding solutions. These additional efforts might look into additional methods of text embedding using state-of-the-art embedding models such as BERT or transformer-based architectures. Additionally it could prove to be helpful to use non-profit website data for both the more rich textual information, as well as page ranking from non-profit links. Another potential avenue of interest would be to attempt a semi-supervised learning approach by utilizing social media connections between nonprofit entities as labels. Though such a graph-based approach would be a deviation from our architecture, the potential gains from including quasi-labels may very well be worth it.