*The Cooper Union for the Advancement of Science and Art*

# Charity Navigator: Using Text Processing to Understand Ranking

Ross Kaplan, Anushree Sreedhar, Yvette Wang
Data Science for Social Good
Professor Sam Keene and Will Shapiro

## 1. Objective

Charity Navigator is a non-profit organization that seeks to quantitatively rank charities based on financial, and transparency metrics. Charity Navigator's ultimate goal is to distributing knowledge about the philanthropic marketplace in an easy and efficient manner to overcome national and global challenges. By 2011, the company has gained traction and has been named as "nation's largest and most-utilized evaluator of charities" (1). Given categorical data, tabular quantitative data, text impact data, and ranking of charities, a larger picture can be drawn by using LDA () to glean information from text data to inform Charity Navigator's ranking system. By using a text approach, Charity Navigator can understand what charities think about themselves, rather than what they can financially provide. When used in conjunction with impactful financial data, text analytics will provide an additional layer to hone the current implemented ranking process.

## 2. Background

Charity Navigator receives financial data about charities through their IRS Form 990 that provides information on a charities tax return. These financial statistics are used to evaluate the charity in seven key performing areas, and deliver an overall Financial Health statistic. Most of the data delivered by Charity Navigator reflected financial areas, including working capital, total liabilities, total net assets, administration expenses, program expenses etc. In addition to financial data, Charity Navigator provided text data on behalf of GuideStar, an organization that asks charities five essential questions in order to measure the success of the charities and how they view themselves. Those five questions are as follows:

1. What is the organization aiming to accomplish?
2. What are the organizations key strategies to make this happen?
3. What are the organizations' capabilities to do this?
4. How will they know if they are making progress?
5. What have and haven't they accomplished so far?

Currently, Charity Navigator has not involved these text responses in how they rank charities. However, with further insight as to how charities view themselves, it is possible to include non-financial developments as another factor in the ranking process.

## 3. Results and Discussion

### a. Financial Data vs. Overall Rating

A preliminary analysis on the financial data was performed in order to see ourselves what the kind of correlations existed between financial data and the overall rating. A confusion matrix was made in order to linear correlate all the data, and then the results were processed in a visually appealing manner. The confusion matrix is able to relate given data to describe the performance of a classification model. Figure 1 below is representative of this information. The increasing gradient of blue represents the degree of positivity from 0 to 1 and the increasing gradient of yellow represents the degree of negativity from 0 to -0.074. The more positive the number, the more linearly correlated they are. Hence, across the diagonal there is a 1:1 correlation as represented by a very dark blue. All the other financial features tested against each other was somehow related to one another, as seen by the relatively dark shades of blue. However, when comparing the column and row of the overall rating, there is almost no correlation, as all the boxes are very light. Hence, the financial data was clearly not linearly correlated to the overall rating.

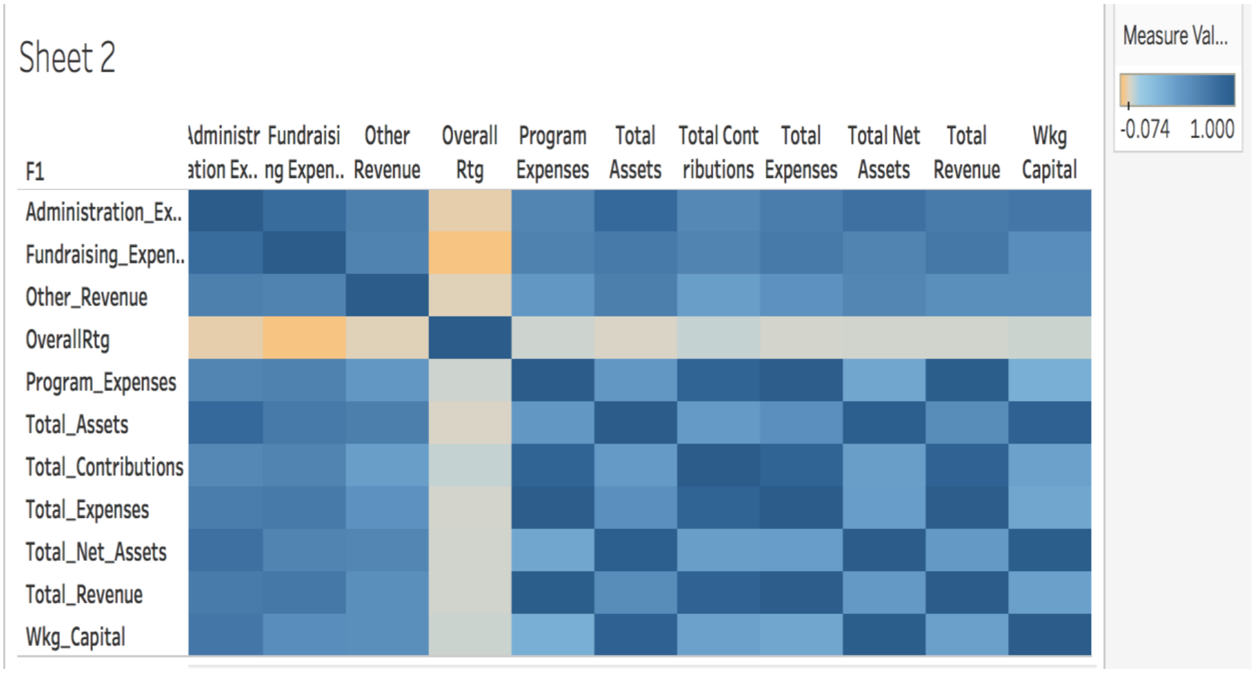


Figure 1: Visual result of financial data vs. overall rating

### b. Initial Text Processing

Using the GuideStar information and removing all the common words in the English language (‘the’, ‘a’, ‘of’, ‘it’), a word cloud was made to initially see if the data is viable and uses words that depict a philanthropic topic. Figure 2 below is the result of the word cloud, with some of the more frequently used words being programs, program, community, work, children, health, and education. Since these words are associated with ultimate goals of certain charities, additional analysis was performed.

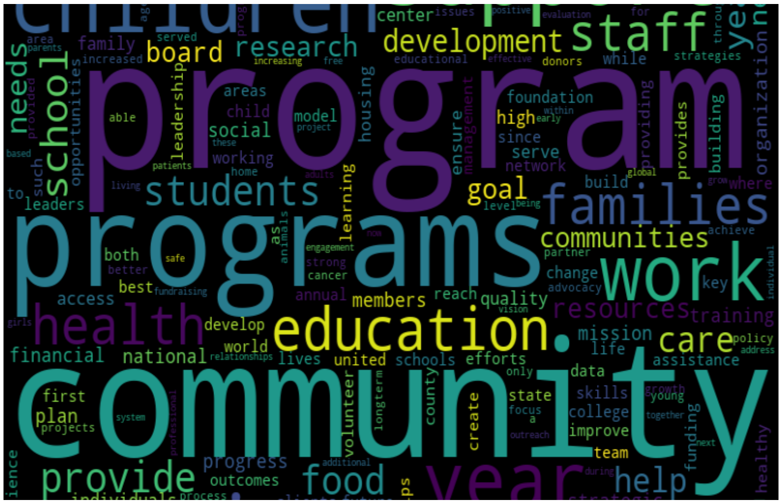
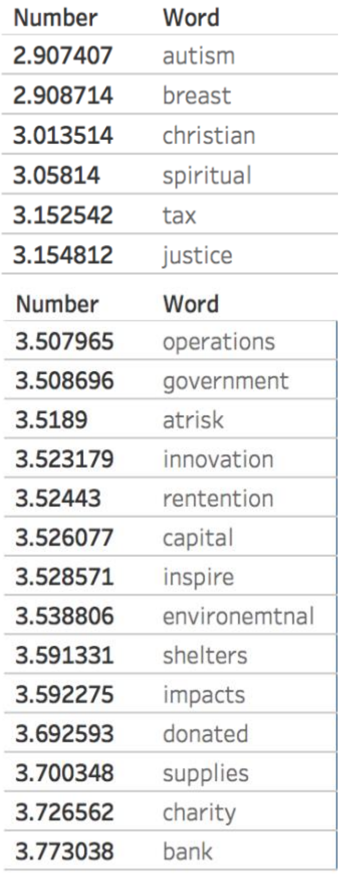


Figure 2: Word Cloud on GuideStar Information

Another way to view the importance of this text data was to average the ranking of the charities that included that particular word. In essence, this analysis would depict the weight, or importance, of this word with regards to how it is viewed in the philanthropic world. Figure 3 is a sample of some words and their associated average ranking. For example, for every charity that contained the word “autism”, the average ranking of these charities faired at roughly a 2.9 out of 5. This information can be useful to Charity Navigator so that they can see how charities that write about themselves in a certain way are actually fairing in the philanthropic world. In addition, this type of data would help in formulating different lists about what types of charities are “trendig” versus some that require more help (donation wise, volunteers, etc.)



| Number | Word |
|---|---|
| 2.907407 | autism |
| 2.908714 | breast |
| 3.013514 | christian |
| 3.05814 | spiritual |
| 3.152542 | tax |
| 3.154812 | justice |

| Number | Word |
|---|---|
| 3.507965 | operations |
| 3.508696 | government |
| 3.5189 | atrisk |
| 3.523179 | innovation |
| 3.52443 | rentention |
| 3.526077 | capital |
| 3.528571 | inspire |
| 3.538806 | environemtnal |
| 3.591331 | shelters |
| 3.592275 | impacts |
| 3.692593 | donated |
| 3.700348 | supplies |
| 3.726562 | charity |
| 3.773038 | bank |

Figure 3: Important Words and Associated Average Rating

### c. LDA

As part of natural language processing, Latent Dirichlet Allocation (LDA) is a way to project words into a very high-dimensional space and find clusters of data in this space. LDA is a type of topic modelling analysis, or a statistical model that processes text in order to discover topics that exist within that set.

---

Figure 4 shows the general high level flow of what happens to data for LDA. Essentially all the impact data (text) is taken in and processed into a bag of words. This means only the count of each word is known, and sentence structure and order are intentionally ignored. We used a form of TF-IDF to remove words that are common in English.

To explain TF-IDF, essentially, a word is given a value for both TF and IDF, and the ratio is computed to find out how important the word is to a document. The TF term is the term frequency, or how many times a word appears in a document.  The IDF term is how many times that word appears in English.

If that was too complicated, we basically just removed words that are overly common in English (like “the”, “is”, etc.) that will not be an important topic to include. Once those words are removed and the most important terms are kept, we cluster the remaining words and produced the results seen in Figure 7 and Figure 8.

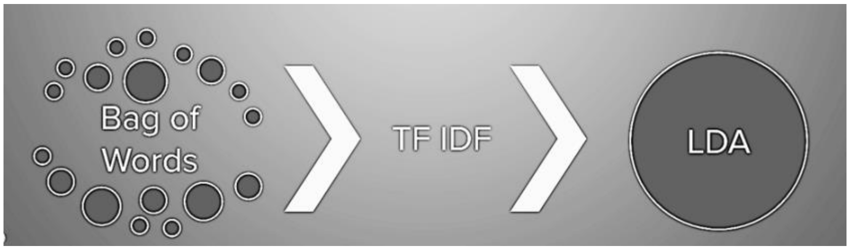
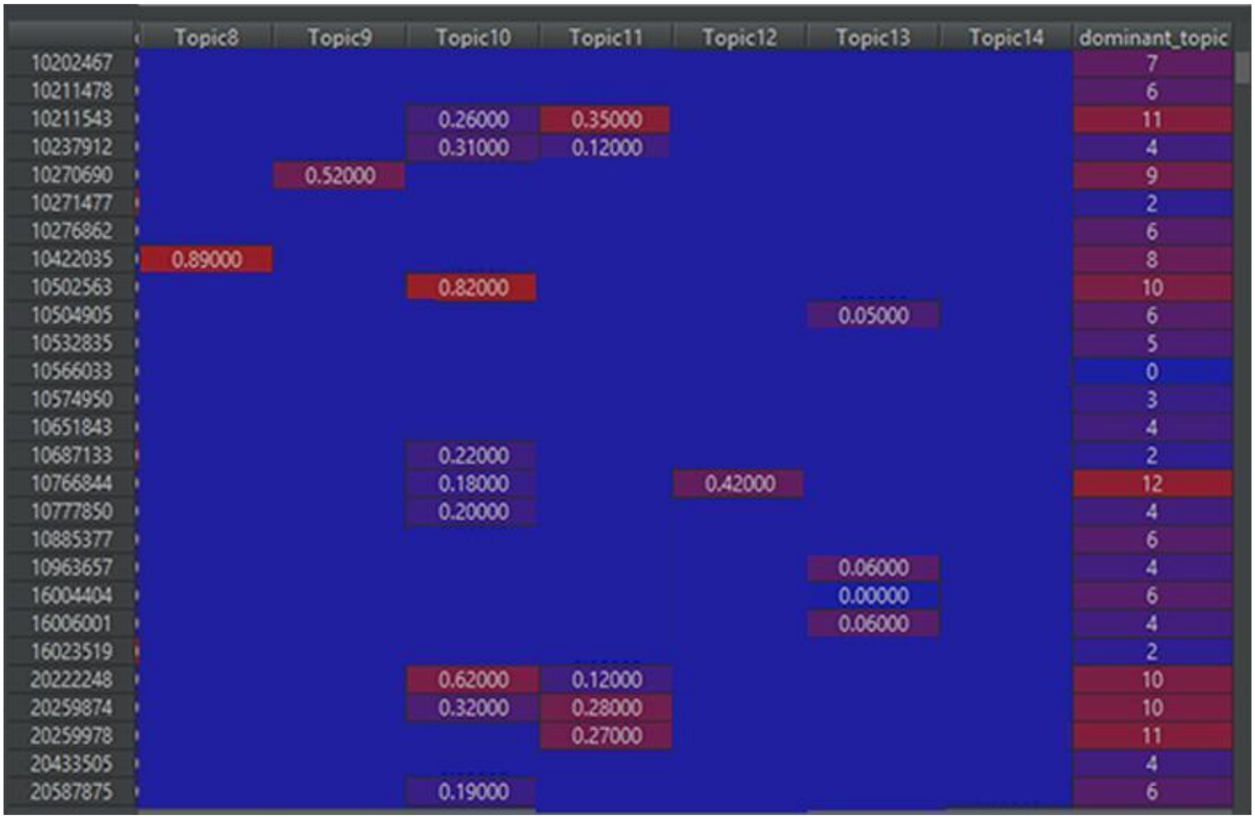


Figure 4: High Level Overview of LDA

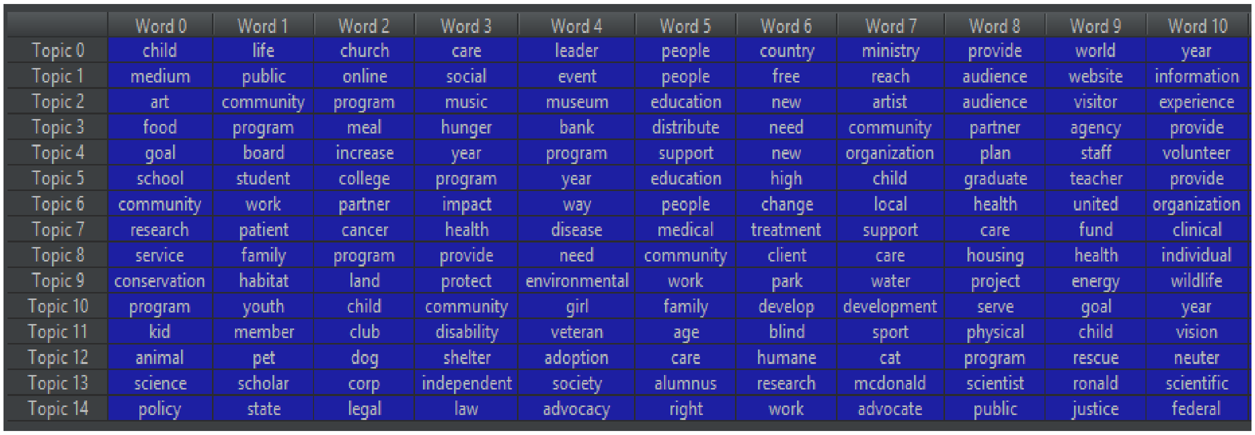


| | Topic8 | Topic9 | Topic10 | Topic11 | Topic12 | Topic13 | Topic14 | dominant_topic |
|---|---|---|---|---|---|---|---|---|
| 10202467 | | | | | | | | 7 |
| 10211478 | | | | | | | | 6 |
| 10211543 | | | 0.26000 | 0.35000 | | | | 11 |
| 10237912 | | | 0.31000 | 0.12000 | | | | 4 |
| 10270690 | | 0.52000 | | | | | | 9 |
| 10271477 | | | | | | | | 2 |
| 10276862 | | | | | | | | 6 |
| 10422035 | 0.89000 | | | | | | | 8 |
| 10502563 | | | 0.82000 | | | | | 10 |
| 10504905 | | | | | | 0.05000 | | 6 |
| 10532835 | | | | | | | | 5 |
| 10566033 | | | | | | | | 0 |
| 10574950 | | | | | | | | 3 |
| 10651843 | | | | | | | | 4 |
| 10687133 | | | 0.22000 | | | | | 2 |
| 10766844 | | | 0.18000 | | 0.42000 | | | 12 |
| 10777850 | | | 0.20000 | | | | | 4 |
| 10885377 | | | | | | | | 6 |
| 10963657 | | | | | | 0.06000 | | 4 |
| 16004404 | | | | | | 0.00000 | | 6 |
| 16006001 | | | | | | 0.06000 | | 4 |
| 16023519 | | | | | | | | 2 |
| 20222248 | | | 0.62000 | 0.12000 | | | | 10 |
| 20259874 | | | 0.32000 | 0.28000 | | | | 10 |
| 20259978 | | | | 0.27000 | | | | 11 |
| 20433505 | | | | | | | | 4 |
| 20587875 | | | 0.19000 | | | | | 6 |

Figure 7: Sample of topics (horizontal axis) and list of charities (vertical axis left) with the weightage of each category

When building the model, a grid search was used to decide the best model by changing two parameters, the number of topics and learning decay. The best model found had 15 topics with a log likelihood score of -2504928.86406, and perplexity of 1109.64665427.

LDA gave a result of 15 categories as the sample shown in Figure 7. The top 11 key words of each category was listed in Figure 8. From the list, we can see that the charities are well categorized by the model, as there are words like art, food, school, health, etc. from which we can easily distinguish one category from another. For example, as in Figure 9, the word contributes to the topic most is "student" which has about 5,000 occurrences in Topic 6 and in total. Hence, Topic 6 is a category on education, and other key words like college, career and scholarship implies Topic 6 has its major charities focusing on college education.



| | Word 0 | Word 1 | Word 2 | Word 3 | Word 4 | Word 5 | Word 6 | Word 7 | Word 8 | Word 9 | Word 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Topic 0 | child | life | church | care | leader | people | country | ministry | provide | world | year |
| Topic 1 | medium | public | online | social | event | people | free | reach | audience | website | information |
| Topic 2 | art | community | program | music | museum | education | new | artist | audience | visitor | experience |
| Topic 3 | food | program | meal | hunger | bank | distribute | need | community | partner | agency | provide |
| Topic 4 | goal | board | increase | year | program | support | new | organization | plan | staff | volunteer |
| Topic 5 | school | student | college | program | year | education | high | child | graduate | teacher | provide |
| Topic 6 | community | work | partner | impact | way | people | change | local | health | united | organization |
| Topic 7 | research | patient | cancer | health | disease | medical | treatment | support | care | fund | clinical |
| Topic 8 | service | family | program | provide | need | community | client | care | housing | health | individual |
| Topic 9 | conservation | habitat | land | protect | environmental | work | park | water | project | energy | wildlife |
| Topic 10 | program | youth | child | community | girl | family | develop | development | serve | goal | year |
| Topic 11 | kid | member | club | disability | veteran | age | blind | sport | physical | child | vision |
| Topic 12 | animal | pet | dog | shelter | adoption | care | humane | cat | program | rescue | neuter |
| Topic 13 | science | scholar | corp | independent | society | alumnus | research | mcdonald | scientist | ronald | scientific |
| Topic 14 | policy | state | legal | law | advocacy | right | work | advocate | public | justice | federal |

Figure 8: Tabular list of topics with words

These results can be viewed at the following site, here.

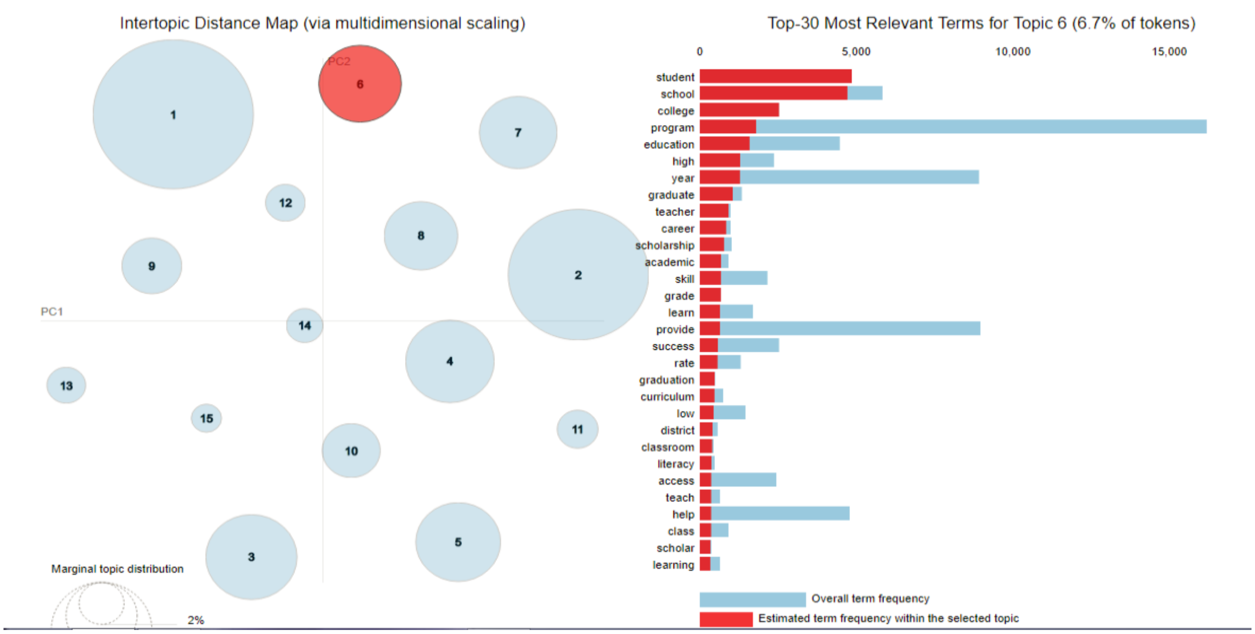


Figure 9: Screenshot of the visualization site

## 4. Conclusion and Recommendation

Seeing as Charity Navigator uses traditional Non-Machine Learning techniques to generate its rankings, we knew that providing Charity Navigator with a machine learning approach to its ranking system wasn't useful or helpful. Similarly, the 990 tax forms also part of the ranking system and are analyzed without machine learning techniques as well. As a result, the only data that's not in use by Charity Navigator is the Impact Data, where charities talk about themselves.

Knowing the sorts of topics that charities talk about is the first step to using the Impact Data in rankings or recommendations for people. When customers visit Amazon's website, they're immediately hit with machine learning generated suggestions on the kinds of products Amazon thinks they'd potentially buy. Similarly, Impact Data (and of course donation history) are the kinds of features that Charity Navigator can employ if they were to attempt to recommend charities to people. In addition to financials and accountability statistics, they can use types of charities they like to give people suggestions.