

The Sum of Us

Visualizing the role of crowdfunding in the United States Healthcare System

Colleen McCaffrey

M.S., Data Visualization candidate, Parsons School of Design

B.A., Literary and Cultural Studies, concentration in Film Studies, College of William and Mary

Thesis Advisors

Daniel Sauter

Aaron Hill

Submitted in partial fulfillment of the requirements for the degree of Master Science in Data Visualization at Parsons School of Design

May, 2019

Abstract

Medical costs are the number 1 cause of bankruptcy¹ in the United States and a federal study conducted in December 2018 shows that 40% of American households can't afford a \$400 financial emergency². The majority of costs in healthcare are incurred by a fraction of the population, but with the current system nearly 1 in 10 Americans³ resort to crowdfunding for medical costs. This project sets to help untangle some of the obscurity behind what systemic issues are intersecting when people need to crowdfund healthcare costs and also look for patterns of behavior and bias within that. The National Conference of State Legislatures released data in December 2018 indicating that health insurance premiums are rising at a rate faster than salaries⁴. When GoFundMe launched in 2010 it was a platform for wedding gifts and travel funds, today it is the the leading medical fundraising platform, with 1 in 3 campaigns being for medical needs and raises some \$650 million annually⁵. Using a hybrid approach of structured and unstructured learning, and the people-first philosophy of human-centered design, this project aims to reveal new data and meaningful insights that can be used by healthcare policymakers and civic-minded innovators to create a more integrative healthcare system and to help close the gap.

¹ ("Bankruptcy Statistics.")

² (Board of Governors of the Federal Reserve System. 2018)

³ ("GoFundMe Medical Fundraising."2019)

⁴ (National Conference of State Legislatures, 2018)

⁵ ("GoFundMe Medical Fundraising.")

Abstract	2
Introduction	3
Treatment	4
Background	4
Methodology	9
Data Collection	10
Data Analysis	13
Design	16
Findings	18
Conclusion	21
Considerations and Future Directions	22
Bibliography	24

Introduction

As of April 2019, health expenses account for 17.9%⁶ of the U.S. GDP, the highest of developed countries in the world. Even with insurance, many people end up paying for out-of-pocket treatments, medication, and care that is not covered by their insurance plan. Since 2013 average national salaries have stagnated while the out of-pocket costs per individual with employer supported healthcare has risen 16.7%, or around \$5,893⁷. For those with a chronic disease the costs have risen \$20,000 on average per individual⁸, and trends over time show that this gap is only widening. As of 2018, Americans pay 3.5 trillion dollars annually in premiums alone⁹.

With these gaps in the system more people are turning to GoFundMe to supplement insurance costs and the indirect costs of being out of work due to an illness or caring for a loved one, which often in the political conversations around healthcare, fall behind prescription drug costs and surgery. At the 2017 Techonomy Health Convention data-driven healthcare panel, Andrew Thompson, President and CEO, Proteus Digital Health made the following statement

"We don't have a healthcare system. We have a sick care system. And it's important to note that it was built in the last century to do a very important job, which was to deal with acute disease and trauma....Today, we have very different challenges, 75% or 85% of what we need to deal with is chronic disease that's dealt with in community settings, not in hospitals. So, we need to supplement, and in many ways magnify the power of this magnificent sickcare system with a healthcare system."

That concept is part of the inspiration for this thesis project. When we consider the implications that our healthcare system, in its current state, covers primarily hospital and medical professional related costs, but neglects to support many of the associated costs of access, at-home recovery, or ongoing treatment. The fractures in the system are hard to ignore.

This data visualization project aims to be a tool from which other analysts, healthcare professionals, and policymakers can gain new insights and involve the user in the decision-making process to crowdsource - instead of money - solutions for a broken healthcare

⁶ (World Health Organization Global Health Expenditure database. 2017.)

⁷ (Healthcare Cost Institute. 2019)

⁸ (Healthcare Cost Institute. 2019)

⁹ (Centers for Medicare and Medicaid Services. 2017)

system. The exploratory nature of the visualization is designed to encourage political discourse, policy reform, and civic engagement by presenting new data insights and helping to uncover opportunities for systemic change.

Treatment

Background

With the dismantling of the Affordable Care Act¹⁰ and rising costs of prescriptions drugs, people are doing what they do in times of need- turning to their communities. They have adopted a technology to meet their needs for something that Americans pay 3.5 trillion dollars annually in premiums alone.¹¹ So while GoFundMe never had the goal of being a medical funding platform, it is now the nation's top¹². Sometimes unintended consequences become a feature not a fault, which happened with crowdfunding.

Crowdsourcing healthcare funds now affects 1 in 10 people annually¹³ and will only continue to grow. But finding a solution for the fractured healthcare system will take more than neighborly goodwill. Both donors and campaign owners alike suffer the fatigue from the never-ending cycle of donation requests¹⁴. Crowdfunding healthcare, can be useful for some isolated instances but to truly find a sustainable solution a holistic approach must be taken. Some politicians support a single-payer healthcare system or medicare for all, but crowdsourcing healthcare needs is not an American only problem, it is prevalent in Canada and other countries who already have healthcare systems like that in place¹⁵.

Medical issues are the number one cause of bankruptcy, with more than 2 out of 3 bankruptcies being due to medical costs.¹⁶ there are many studies out there combating the "Myth of Medical Bankruptcies"¹⁷ and just as much data confirming it¹⁸. If the federal reserve study concluded that 40% of families don't have \$400 to cover emergency expenses¹⁹, then we can also assume tens of thousands of medical bills for either deductibles, premiums, or emergency services would

¹⁰ (Payne, Emily and Chris Nicholls. 2019)

¹¹ (Centers for Medicare and Medicaid Services. 2017)

¹² (Bluth, Rachel. 2019)

¹³ ("GoFundMe Medical Fundraising." 2019)

¹⁴ (Alkon, Cheryl. 2018)

¹⁵ (Bluth, Rachel. 2019)

¹⁶ (Tozzi, John and Zachary Tracer. 2018)

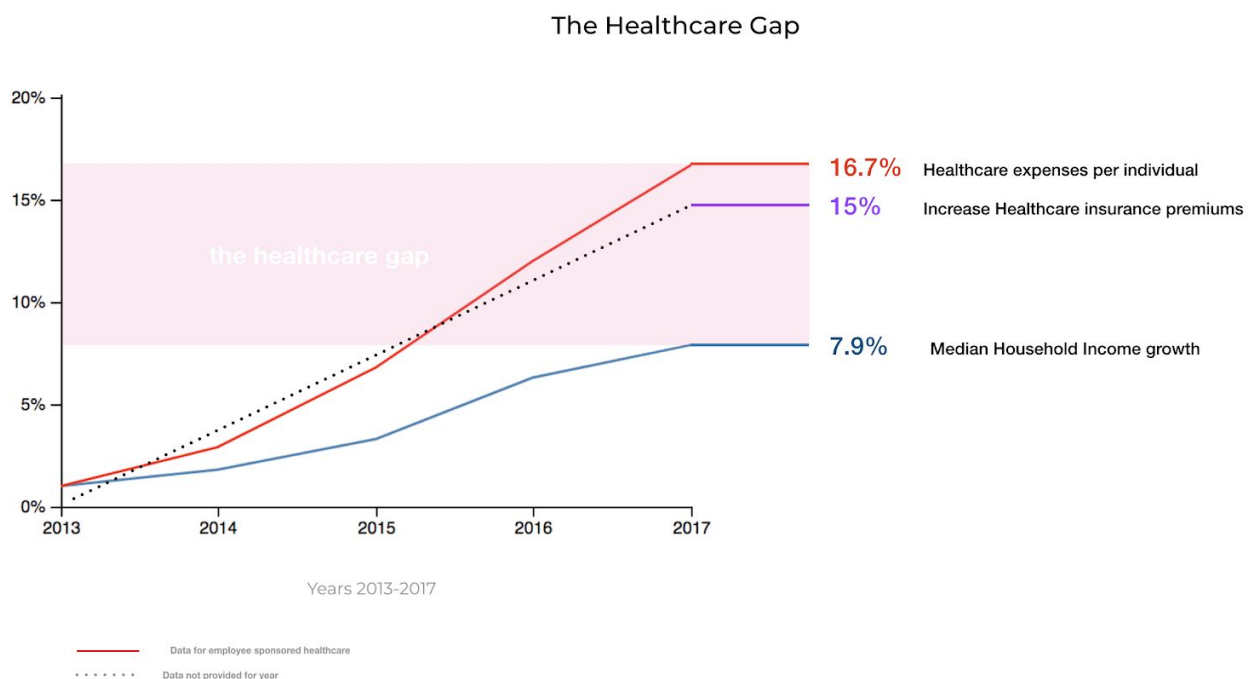
¹⁷ ("People are Raising \$650 Million on GoFundMe each Year to Attack Rising Healthcare Costs." 2019)

¹⁸ "Bankruptcy Statistics."

¹⁹ (Board of Governors of the Federal Reserve System. 2018)

cause many families to suffer financially, and many would never fully recover from that kind of setback.

To answer the question, “What can crowdfunding tell us about the United States Healthcare Gap?” we must first define the gap. Since 2013, medical expenditure per individual has increased 16.7%, or around \$5,893, and insurance premiums have increased 15%, all while salaries have stagnated. For those with a chronic disease? The costs have risen \$20,000 on average per individual.²⁰ As of April 2019, health expenses account for almost 18% of the U.S. GDP, which is the highest of developed countries in the world. And trends over time show that this gap is only widening.



In December 2018, the Commonwealth Fund published a brief²¹ using ‘data from the federal Medical Expenditure Panel Survey–Insurance Component (MEPS–IC) to examine trends in employer premiums at the state level to see how much workers and their families are paying for their employer coverage in terms of premium contributions and deductibles. [They] examine the

²⁰ (Healthcare Cost Institute. 2019)

²¹(Healthcare Cost Institute. 2019)

size of these costs relative to income for those at the midrange of income distribution. The MEPS–IC is the most comprehensive national survey of U.S. employer health plans. It surveyed more than 40,000 business establishments in 2017, with an overall response rate of 65.8 percent.” The brief noted that “while the Affordable Care Act’s marketplaces receive a lot of media and political attention, the truth is that far more Americans get their coverage through employers. In 2017, more than half (56%) of people under age 65 — about 152 million people — had insurance through an employer, either their own or a family member’s. In contrast, only 9 percent had a plan purchased on the individual market, including the marketplaces.”²²

Further highlights from the report:²³

- After climbing modestly between 2011 and 2016, average premiums for employer health plans rose sharply in 2017. Annual single-person premiums climbed above \$7,000 in eight states; family premiums were \$20,000 or higher in seven states and D.C.
- Rising overall employer premiums increased the amount that workers and their families contribute. Average annual premium contributions for single-person plans ranged from \$675 in Hawaii to \$1,747 in Massachusetts; family plans ranged from \$3,646 in Michigan to \$6,533 in Delaware.
- Average employee premium contributions across single and family plans amounted to 6.9 percent of U.S. median income in 2017, up from 5.1 percent in 2008. In 11 states, premium contributions were 8 percent of median income or more, with a high of 10.2 percent in Louisiana.
- The average annual deductible for single-person policies rose to \$1,808 in 2017, ranging from a low of \$863 in Hawaii to a high of about \$2,300 in Maine and New Hampshire. Average deductibles across single and family plans amounted to 4.8 percent of median income in 2017, up from 2.7 percent in 2008. In three states (Florida, Mississippi, and Tennessee), average deductibles comprised more than 6 percent of median income.
- Combined, average employee premium contributions and potential out-of-pocket spending to meet deductibles across single and family policies rose to \$7,240 in 2017 and was \$8,000 or more in eight states. Nationally, this potential spending amounted to 11.7 percent of median income in 2017, up from 7.8 percent a decade earlier. In Louisiana and Mississippi, these combined costs rose to 15 percent or more of median income.

²² (Healthcare Cost Institute. 2019)

²³(Healthcare Cost Institute. 2019)

A news release, published by the Bureau of Labor Statistics from the U.S. Department of Labor, estimated that even for those employed, medical benefits and sick days and paid holidays are not available to everyone, only 69% of private sector and 89% of public sector, and the average percent the employee is responsible for is higher for those in the bottom half of salary ranges.²⁴ It further estimated that

“Medical care benefits were available to 69 percent of private industry workers and 89 percent of state and local government workers in March 2018, the U.S. Bureau of Labor Statistics reported today. In private industry, access to employer-sponsored medical care benefits varied by establishment size. Fifty-five percent of private industry workers in small establishments (those with fewer than 100 employees) were offered medical care benefits. These benefits were offered to 83 percent of workers in medium-size establishments (those employing between 100 and 499 workers) and 88 percent of workers in large establishments (those with 500 employees or more). In state and local government, medical care benefits were available to 85 percent of workers in small establishments, 86 percent of workers in medium-size establishments, and 92 percent of workers in large establishments.”

A brief published by Kaiser noted that without employee sponsored insurance, people pay more, have access to less adequate services, and are at more of a risk for financial burdens should an emergency arise.²⁵ While the majority of people insured in the United states are so through an employer sponsored health insurance program, a key findings of the Kaiser brief was that the ACA helped provide coverage for those most at risk- the poor and near-poor nonelderly individuals - the individuals are least likely to have a job where health insurance is even offered as a benefit, or offer plans with affordable premiums.

“Coverage gains from 2013 to 2016 were particularly large among groups targeted by the ACA, including adults and poor and low-income individuals. The uninsured rate among nonelderly adults, who are more likely than children to be uninsured, dropped 8.4 percentage points from 20.6% in 2013 to 12.2% in 2016, a 41% decline. In addition, between 2013 and 2016, the uninsured rate declined substantially for poor and near-poor nonelderly individuals (Figure 2). People of color, who had higher uninsured rates than non-Hispanic Whites prior to 2014, had larger coverage gains from 2013 to 2016 than non-Hispanic Whites. Though uninsured rates dropped across all states, they dropped more in states that chose

²⁴ (Bureau of Labor Statistics. 2018)

²⁵ (Henry J. Kaiser Family Foundation. 2018)

to expand Medicaid, decreasing by 7.2 percentage points from 2013 to 2016 compared to a 6.1 percentage point drop in non-expansion states.”

The [report, which can be read in its entirety here](#), goes on to list these key findings about the uninsured population:

How many people are uninsured?

In the past, gaps in the public insurance system and lack of access to affordable private coverage left millions without health insurance. Beginning in 2014, the ACA expanded coverage to millions of previously uninsured people through the expansion of Medicaid and the establishment of Health Insurance Marketplaces. Data show substantial gains in public and private insurance coverage and historic decreases in the number of uninsured people under the ACA, with nearly 20 million gaining coverage. However, for the first time since the implementation of the ACA, the number of uninsured increased by more than half a million in 2017.

Why do people remain uninsured?

Even under the ACA, many uninsured people cite the high cost of insurance as the main reason they lack coverage. In 2017, 45% of uninsured adults said that they remained uninsured because the cost of coverage was too high. Many people do not have access to coverage through a job, and some people, particularly poor adults in states that did not expand Medicaid, remain ineligible for financial assistance for coverage. Some people who are eligible for financial assistance under the ACA may not know they can get help, and undocumented immigrants are ineligible for Medicaid or Marketplace coverage.

Who remains uninsured?

Most uninsured people are in low-income families and have at least one worker in the family. Reflecting the more limited availability of public coverage in some states, adults are more likely to be uninsured than children. People of color are at higher risk of being uninsured than non-Hispanic Whites.

How does not having coverage affect health care access?

People without insurance coverage have worse access to care than people who are insured. One in five uninsured adults in 2017 went without needed medical care due to cost. Studies repeatedly demonstrate that the uninsured are less likely than those with insurance to receive preventive care and services for major health conditions and chronic diseases.

What are the financial implications of being uninsured?

The uninsured often face unaffordable medical bills when they do seek care. In 2017, uninsured nonelderly adults were over twice as likely as their insured counterparts to have had problems paying medical bills in the past 12 months. These bills can quickly translate into medical debt since most of the uninsured have low or moderate incomes and have little, if any, savings.

Methodology

Crowdfunding exposes a lot about the human spirit and the will to survive. One thing we know from the data of crowdfunding is that people are willing to support others in need. Data can tell us a lot about ourselves. Our habits, our beliefs, our motivations and our needs. Crowdfunding has been filling a widening gap created by a tragically broken healthcare system. The visualization part of this thesis project aims to reveal some of those insights and patterns about the types of campaigns and needs the current healthcare system does not meet that drives people to resort to crowdfunding. It will collect and analyze samples of data from the healthcare related categories on GoFundMe.com, “medical”, “emergency”, and use K-means clustering to further evaluate what these campaigns have in common, what makes for a successful healthcare campaigns, and to evaluate if there are commonalities around the types of campaigns more people are willing to contribute to and those that have a higher success rate of meeting their goal.

Using a hybrid approach of structured and unstructured learning, and the philosophy of human-centered design, I collected campaign data dating back to the last election in November 2018 and analyzed the corpus using k-means clustering and tf-idf to find the most relevant unique terms within the clusters.

Why GoFundMe campaign data? GoFundMe has consolidated GiveForward, Generosity, and YouCaring and is now “the largest and most trusted free crowdfunding platform.” This [TechCrunch article](#) breaks down the crowdfunding acquisition details and possible motivations.

The following is a list at some of the primary reasons that have been reported as to why people create a GoFundMe campaign, though this list is not exclusive nor exhaustive.²⁶ The goal of this project to gain new insight into these needs by visualizing the campaign data:

- lack of transparency on pricing (pricing varies with health insurance provider, hospital, and state)

²⁶ (Bluth, Rachel. 2019)

- lack of coverage period
- gaps in coverage (i.e., time off, new job, etc)
- unexpected or recurring costs like travel for loved one
- research or clinical trials not covered by health insurance
- drugs that a drug company can not make money off of so there is little to no access
- costs beyond what healthcare covers
- home care or special accommodations
- being out of work for a while
- loss of health insurance

Data Collection


The data collection was a multi-step process that was complicated by the fact that GoFundMe does not have an api for dynamic data consumption. After researching a few other projects that had accomplished the feat, I took a similar approach. For starters I visited the [GoFundMe robots.txt](#) to ensure that the individual campaign pages were not explicitly declared in the robots.txt file and that I could access the data. Also mindfulness was paid to the speed and timing with which the data collection process took place, which ended up significantly limiting the amount of campaign data I had access to by the end of the semester. I have documented the alternative methods I would use in the future more in detail in the conclusion and future steps section of this thesis. Essentially, a python tool like [Scrapy Doc](#) or [selenium-python](#) would have increased the collecting process significantly, but out of concerns of time management and needing at a minimum a sample of data, I resorted to a javascript data collection framework, [puppeteer](#), and the parsing framework [cheerio](#).

A significant reference that first introduced me to selenium as a tool, was the work of Phillip Hopen published in the NYC Data Science Academy blog, [“Scraping GoFundMe”](#), which also helped inform other limitations of the GoFundMe platform interface.


The campaigns are sorted by categories which users can then access through a collections page. My project aimed to use campaigns explicitly tagged “Medical” and “Emergency” particularly because often time health issues become an emergency and because this was a human-centered design approach, I wanted to be inclusive rather than exclusive.

Some of the initial limitations I encountered were the individual campaign category collection pages which only load the individual campaign cards twelve at a time and were javascript


enabled, which placed limitations on my proposed method of collecting individual campaign urls. While most of the campaigns loading on these specific category pages were active campaigns and recently created, another limitation to this process was that there was no specific way to ensure I was collecting a specific date range of campaigns.



DALY CITY, CA
Help Andy walk again
At 8:30pm on March 23, 2019, Fernando (Andy) Teodoro parked h...
Last donation 1w ago
\$148,280 raised of \$100,000



SANTA MONICA, CA
Help Sam Lloyd Beat Cancer
By all accounts, 2019 was off to a roaring start for our dear friend, S...
Last donation 2d ago
\$146,521 raised of \$100,000



NEW YORK, NY
Pat Cleveland fund
Many of you have asked about Pat... While in Paris this week, Pat was ...
Last donation 1d ago
\$139,854 raised of \$150,000

[Show More](#)

As an alternative method I resorted to downloading the [sitemap.xml](#) which listed all the campaigns on the website, their url, and the date they were last modified. This modification parameter did not indicate anything significant in terms of creation, as many of the campaigns that were recently modified were created years ago. My goal was to collect the campaign data of all campaigns since the last election cycle, from November 2018 until the current day, which was April 2019. In order to filter the campaign categorized as “medical” or “emergency” that were created between “November 2018” and “May 2019” i had to first extract the campaign urls from the twenty-five sitemap.xml files, each of which had 50,000 campaigns, and filter by last date modified is greater than October 31, 2018, this allowed me to at least filter out any campaigns that haven't even been updated.

Once those urls were collected I wrote a fetch function with puppeteer that would individually grab each url and extract the data from the browser and download it as a .txt file. This process of using puppeteer was not very efficient. Of the more than 150,000 campaigns that had been modified or created since November 1, 2018, I was able to extract the data for only 15,000 campaigns. This sample size alone was somewhat sufficient, although further filtering for the explicit “created-by” date and once outliers and null data was removed, that sample became just

under 2000, at 1995 entities. According to GoFundMe's website, 1 in 3 campaigns²⁷ are medical, but my data showed that this was likely an overestimation and the more accurate assessment based on the sample size collected is closer to 1 in 5.

Data Analysis

After completing the parsing and cleaning of the data, I used jupyter notebooks and pythto run the tf-idf and k-means cluster algorithms. The jupyter notebook analysis can be accessed on the project's github repository under the process subdirectory. The TF-IDF method is one of the most widely used methods for document word extraction, and this was the method applied to the "story" parameter, this is the section of the GoFundMe campaign content that is the least structured and seemed to have the greatest opportunity for machine learning insight enhancements.

"TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF-IDF numbers imply a strong relationship with the document they appear in. "²⁸ K-means is the most important flat clustering algorithm. Its objective is to minimize the average squared Euclidean distance of documents from their cluster centers where a cluster center is defined as the mean or centroid of the documents in a cluster. "²⁹

Since I had no former experience with either of these methods, I referenced a [published example](#)³⁰ that guided the process in a general sense and then included secondary analysis elements that further research revealed as relevant, including applying the elbow criterion method and removing outliers from the data. This was an imperfect process and future considerations would be more effective in collecting a larger data set by allowing more room for error and utilizing some of the tools and methods I outlined above, and further elaborate on in the future directions subsection of the conclusion.

First steps were to remove any stopwords applying NLTK method to the model, so any common words such as "the" and "or" would not be considered for the corpus. The corpus is then tokenized and stemmed so it can be enumerated and applied to the tf-idf vector. After importing sklearn feature extraction the tf-idf matrix was created and the cosine similarity was then applied to the metric for the k-means clustering.

²⁷ "GoFundMe Medical Fundraising."

²⁸ (Ramos, Juan. 2003)

²⁹ ("The Stanford Natural Language Processing Group")

³⁰ (Rose, Brandon.)

“K-means initializes with a pre-determined number of clusters Each observation is assigned to a cluster (cluster assignment) so as to minimize the within cluster sum of squares. Next, the mean of the clustered observations is calculated and used as the new cluster centroid. Then, observations are reassigned to clusters and centroids recalculated in an iterative process until the algorithm reaches convergence.”³¹

“To get a Tf-idf matrix, first count word occurrences by document. This is transformed into a document-term matrix (dtm). This is also just called a term frequency matrix. An example of a dtm is here at right. Then apply the term frequency-inverse document frequency weighting: words that occur frequently within a document but not frequently within the corpus receive a higher weighting as these words are assumed to contain more meaning in relation to the document.”

	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

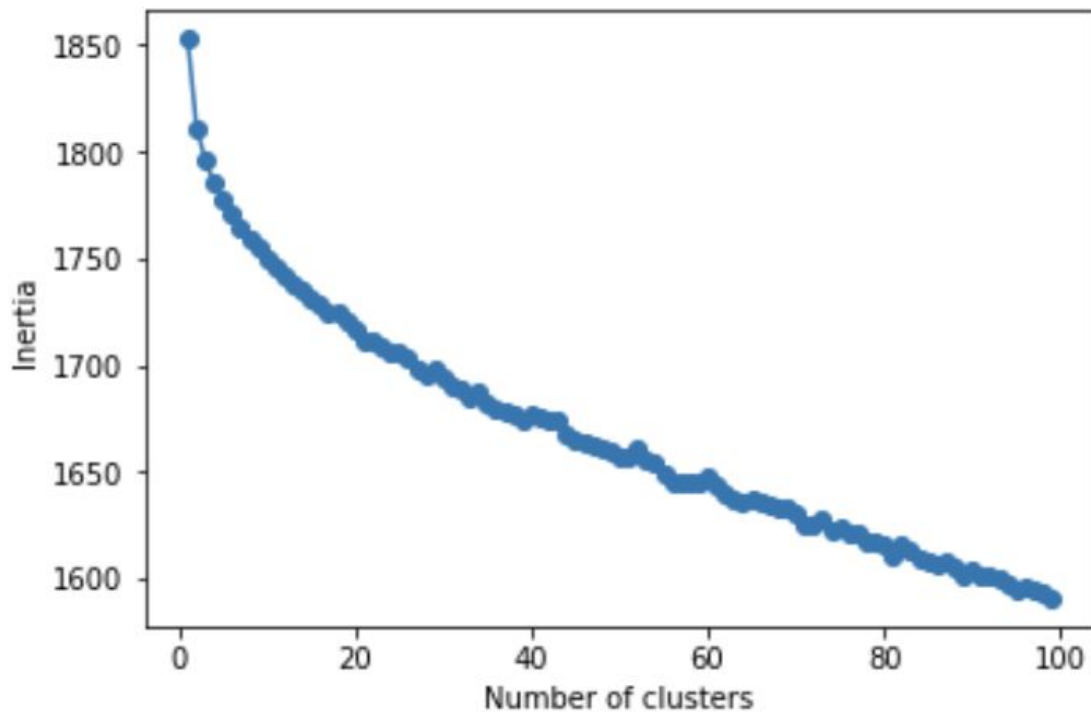
After creating the matrix and running the k-means cluster analysis on the matrix, varying the cluster number in an effort to reach local optima, it yielded the following results for ten clusters and the top term ranked terms in each. Here we can start to see patterns emerge, and interestingly, the emergency campaigns and campaigns in a language other than English, were also grouped into clusters together:

Top 10 Terms for each Cluster

Cluster 0 words:	Cluster 1 words:	Cluster 2 words:	Cluster 3 words:	Cluster 4 words:	Cluster 5 words:	Cluster 6 words:	Cluster 8 words:	Cluster 7 words:
get,1674	cancer,1557	fire,163	de,1033	kidney,286	car,358	surgery,1638	help,4386	years,1852
help,4386	treatment,1393	home,1055	que,445	transplant,267	accident,265	help,4386	family,2524	help,4386
go,1621	help,4386	house,372	la,423	dialysis,69	help,4386	years,1852	friend,1221	treatment,1393
time,2143	breast,247	lost,358	en,356	kidney,553	broken,143	work,2021	please,1119	life,1331
work,2021	diagnosed,748	family,2524	el,308	kidney,361	injury,225	tumor,464	know,1501	disease,396
day,1235	breast,1804	help,4386	para,241	donor,69	recovery,492	eyes,185	love,1166	dogs,189
know,1501	fighting,611	flooded,80	un,238	help,4386	work,2021	get,1674	support,1202	cost,805
back,1112	family,2524	everything,410	su,189	failure,75	car,623	insurance,737	time,2143	service,179
us,1040	stage,322	rebuild,59	por,184	family,2524	family,2524	cover,576	expenses,954	support,1202
weeks,1097	chemotherapy,345	clothes,71	unas,152	work,2021	road,268	remove,382	work,2021	time,2143

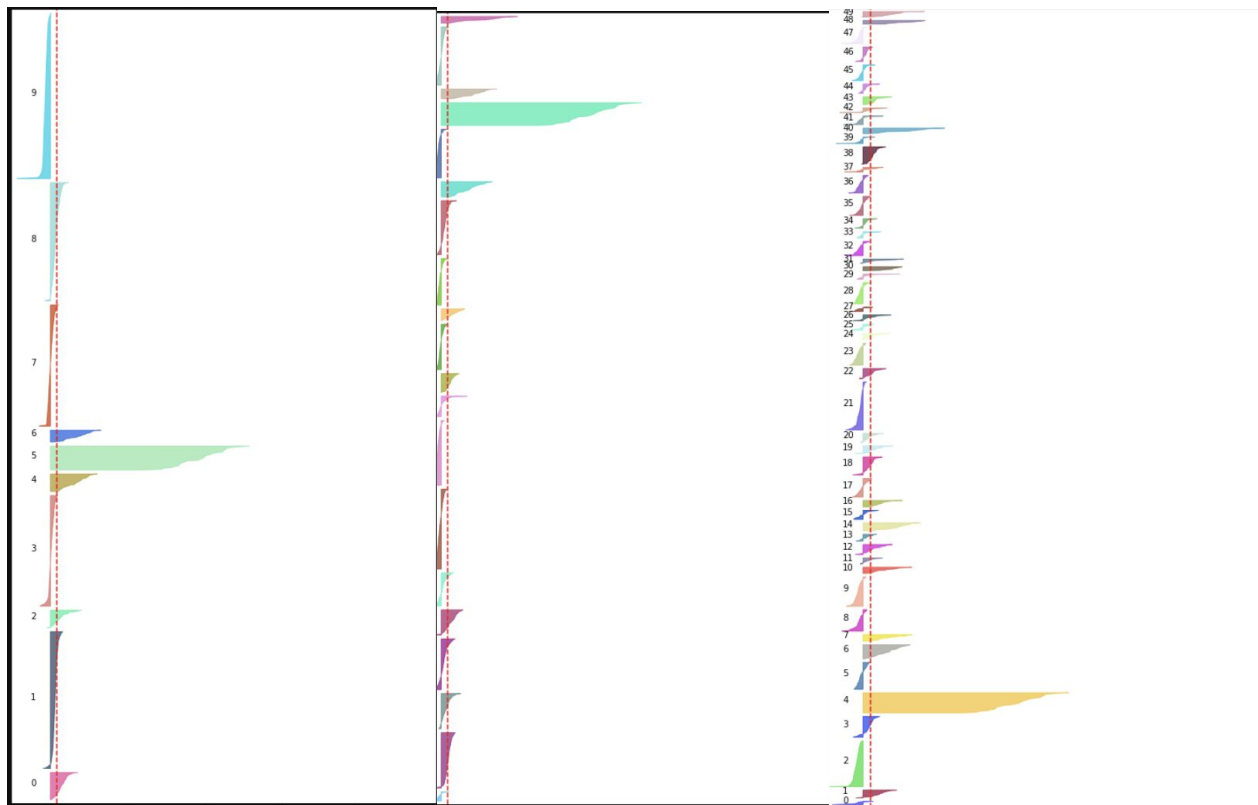
³¹ (Rose, Brandon.)

To test the validity of the model, the elbow method was applied to the clusters, showing a subtle tapering at 10 clusters, but essentially not distinguishable enough to confidently claim that this is an ideal model. Either a larger data set would need to be analyzed or a custom stopwords corpus could be created to minimize similarity based on initial findings and improve the effectiveness of the model.



Elbow criterion results

Using this [silhouette sample example](#), I analyzed the k-means cluster results for model efficiency and the results were also relatively inconclusive in terms of validating the model. None of the silhouette test results tested fully indicated an optimum cluster grouping, as we can see from the silhouette test results figures from the analysis below, many of the campaigns are wrongly grouped from clusters ranging from 5-50. This indicates that k-means might not be the most effective way to cluster this data. Or that the sample size may be too small, and that the groups do not distinguish far enough from a cluster centroid and simultaneously as far from the nearest neighbor centroid.



Silhouette results for max number of clusters, left to right: 10, 20, 50

This is not entirely surprising as these are all campaigns were already categorized as medical and aim to solicit money, we can assume some consistency in language many of the terms will be similar in nature. However after analyzing the term frequency results per cluster there was a perceivable distinction between the clusters. The variations that occurred may be distinct but less obvious in a frequency count methodology. An improved model might include a predetermined corpus based off an initial tf-idf analysis of the campaign story data. These findings are what led to a hybrid approach of structured and unstructured learning to identify and estimate 10 as the optimal number of clusters.

The data was then reformatted into a new dataframe to include the cluster grouping for each campaign and exported as a csv for use in the front-end interface.

Design

Crowdfunding is successful due to user participation, and this project aims to leverage that engagement into civil engagement and systemic change by providing new insights, creating transparency, and providing resources to new data sources.

The aggregated data from the collected and parsed data, merged with the new cluster count provided the primary data source for my visualization. The application was built using d3.js and hosted on github pages. The primary data source, after normalizing value formats and eliminating null values and outliers, is represented by the following summary:

title	url	cluster	city	state	story	raised	goal	engagement	success	donors	shares	length
Deb and Tom Carie Heart Transplant	https://www.gofundme.com/	9	JASPER	IN	Ded Carie's husband	1575	10000	11	16	27	188	5
Byron’s brain tumor trial air f	https://www.gofundme.com/	7	MELBOURNE	FL	Please help me fund r	8310	20000	15	42	101	474	8
A new smile for Conor	https://www.gofundme.com/	9	WOODSTOCK	GA	My son Conor has rec	350	2100	18	17	7	26	6
Jordan family	https://www.gofundme.com/	2	PARADISE	CA	My cousin Dawn and	2725	10000	13	27	24	136	5
Healing to Cherie Giambalvo	https://www.gofundme.com/	9	CHICO	CA	Cherie was working o	7685	10000	14	77	80	417	14
Loving Father in Need	https://www.gofundme.com/	8	TACOMA	WA	My name is Eduard G	2610	50000	9	5	27	249	5
Lewallen’s Medical Despair	https://www.gofundme.com/	0	SANCLEMENTE	CA	I would say my family	2010	20000	33	10	32	30	10
Help Kara through her recovery	https://www.gofundme.com/	5	GARDNER	MA	On Sunday afternoon	2185	5000	5	44	52	1000	5
Rita Cargill Ellerd Cancer Fund	https://www.gofundme.com/	2	BRAZORIA	TX	Hi everyone my name	480	10000	8	5	7	72	5
Help Me Fight Abuse, Cancer, & amp;	https://www.gofundme.com/	7	SAINTLOUIS	MO	My name is Chris O'L	3712	11000	40	34	51	27	13
One of our own needs a hand !	https://www.gofundme.com/	8	SEBRING	FL	Her grand babies is w	857	5000	8	17	26	276	4
Help Ludia Get Rehabilitation/Dialysis	https://www.gofundme.com/	4	JERSEYCITY	NJ	Greetings everyone M	786	10000	9	8	6	54	5
A trailer to live	https://www.gofundme.com/	5	PARADISE	CA	My name is JD I have	1465	3000	13	49	23	130	5

Additional data from the [Health Care Cost and Utilization Report](#), and [United States Census bureau](#) were also incorporated into the project, but extracted and processed manually. All data can be found in the [project repository data subdirectory](#) for reference.

The project design consists primarily of three parts, the first representing the tf-idf findings mapped to the cluster associated with those terms. By introducing the terms initially and as selected groupings, it sets the user with an expectation that these grouping will be used repeatedly throughout the project. The color scheme was determined combining three different [colorbrewer](#) sequential color schemes combined to create a custom gradient palette. Warm calming colors of pink, purple, and orange were selected as a way to avoid the project feeling too cold or unapproachable and instead inviting the user to explore and interact with the data visualizations.

A second phase of the project involves a barcode chart rendering horizontally across the screen and creating a vertical interaction for the thousands of campaigns to be plotted individually. The allows for vertical comparisons against other grouped cluster and also internal value comparisons within groups. The hover interactive property makes its nearly impossible for the user to avoid discovering a story and a click function allows them display the individual campaign metadata

and further click on an external link directing them to the campaign's web page on GoFundMe where they can see the latest updates or explore other interactions. Throughout the design process, user feedback continually referred to success rates as the most intriguing and so this became the default filter feature by which the barcode chart is spread across the x axis, the success rate was calculated from the ratio of "amount_raised" to "goal". Additionally the qualitative data parameter of each campaign's title is included in the hover tooltip to help guide the user to campaigns that most interest them and allows them to sort through multiple campaigns on an individual level without need to click or scroll indefinitely.

The color threshold for the bar charts indicates where on the spectrum the campaign lies, either above or below the group's average, or exceeding the goal. This was an important indicator to include because if a user only visit GoFundMe.com the most successful campaigns are often trending first, even if not explicitly entering the site through the trending pages. The data analysis I conducted revealed that the majority of campaigns do not meet, let alone exceed their goal. This finding is elaborated on within the "Findings" subsection. Additionally, the default opacity was set to a decimal to allow users to see where the campaign distribution was most consolidated. Sparse bars or light bars imply few campaigns, whereas less spaced bars or combined bars with a darker opacity show heavy distribution of campaigns at that success rate threshold.

If the barcode chart focused more on the qualitative, the last visualization implementation was designed to be more quantitative and also conclusive. It is both explanatory and informative so as to give the user concrete data, averaged across all groups to gain a better sense of the corpus as a whole. The project was designed to converge, diverge and converge again, allowing the user to walk through the data set from the big picture down into the metadata of an individual campaign.

Findings

Some of the primary issues for someone to launch a GoFundMe Campaign are lack of transparency on pricing which varies with health insurance hospital and state, lack of coverage period (though those with coverage struggle as well), gaps in coverage - ie time off, new job, etc, and costs beyond what healthcare covers extra costs like travel for loved one, research or clinical trials not covered³², which can be seen in the chart where "family", "time" and "work" are used

³² ("GoFundMe Medical Fundraising.")

even more frequently than “cancer” and “surgery”, showing the recurring need for support in these areas of people’s lives when confronting a medical issue.

The average success rate, amount of goal raised, across all groups is 40.9% and the average engagement rate, the percent of people who donate to a campaign out of all the users who interact in an explicit way, via sharing or liking or donating, is 14.5%.

Using k-means clustering to group the campaigns and tf-idf to find the most relevant unique terms within the corpus and each group, we can see not surprisingly that the term “help” is number one. Interestingly the non-medical terms “work” and “family” had similar rankings to the more medical terms of “surgery” and “cancer”.

Referring back to one of the inspirations for this project, where Andrew Thompson, President and CEO, Proteus Digital Health said,

"We don't have a healthcare system. We have a sickcare system. And it's important to note that it was built in the last century to do a very important job, which was to deal with acute disease and trauma....Today, we have very different challenges, 75% or 85% of what we need to deal with is chronic disease that's dealt with in community settings, not in hospitals. So, we need to supplement, and in many ways magnify the power of this magnificent sickcare system with a healthcare system."

When the campaigns groups were then regrouped into two categories, one being what Thompson would call the “sickcare”- cancer, treatment, hospital, surgery, and the other what he would call “healthcare” - work, family, service, expenses - to see if there was a significant difference in campaign success for those categories, the results showed that the groups had similar success rates, with the “healthcare”- family, work group having slightly more campaigns that met or exceeded their goal and an overall success rate of 43% slightly higher when compared to the more medical, or “sickcare” group that had an average rate of 40%- This tells us that both categories have evident needs as well as contributors to both types of campaigns. We can also see that most campaigns concentrate in distribution around the average on which we can see from the darker opacity. Additionally, the raw data showed that clusters 8 and 1 represented the most campaigns at 372 and 307 respectively. This shows a clear separation of the groups but also an equal need. Where group 8 represented more of the new “Healthcare” i., e., community based needs associated with terms “work”, “family”, “expenses”, “family”, “friends”, “love”, “support”, and “time”; as compared to the referenced “Sick Care” needs that are hospital or medical professional related like “hospital”, “surgery”, “cancer”, “treatment”, “help”, “diagnosed”, “stage”, “fighting”, “family”, and “chemotherapy”.

A third category that emerged from the groups were the campaigns associated with accidents, featuring terms “fire” and “car”- and the campaigns in non-english languages- as they did not explicitly fit into either of the two groupings based on the structured algorithm term extraction, but there is an opportunity as a next step to re-analyze the data within these groups, by applying translation mechanism to the model as a next step to be able to incorporate all campaigns into the model I apply the categories to all data groups.

Exploring the campaigns by group shows us patterns of needs as well as who is contributing to what, and maybe more importantly, which campaigns are struggling the most. The average amount raised per campaign is \$5,449. The campaigns that raised the most had top terms: “cancer”, “treatment”, “surgery”, “work”, and “expenses”. The average campaign fundraising goal ranged from 10k-20k, with campaigns fundraising for top terms “heart”, “liver”, “doctor”, “bills”, “hospital”, “chemotherapy”, “cancer”, and “family” being the highest. The average amount of donors per campaign across all groups was 56, with an average donation of \$97. Surprisingly the campaigns that raised the least were those associated with “kidney”, “donor”, “dialysis”, “transplant” and “failure”, also with mentions of “family” and “work”. These sorts of findings offer an opportunity to explore the data with medical professionals to identify patterns within the data that take professional medical knowledge to identify. One examples as a case study, that could be considered for further exploration, is particularly on display in the kidney focused group. This is just another benefit to visualizing this corpus of data and using the hybrid methodology of both structured and unstructured learning. By revealing the cluster top terms, and applying some research methods I was able to discover that at home kidney dialysis is something more and more healthcare providers are encouraging and patients are requesting.

“Once you get someone home, they feel better at home than they do with in-center dialysis,” said Dr. Leslie Spry, a spokesman for the National Kidney Foundation and medical director of the Nebraska-based Dialysis Center of Lincoln. Over the past five years, Spry said he has seen a rise in the use of home dialysis. “Most people who are in in-center dialysis will tell you that they are usually very much incapacitated after going in for their treatment.” Home dialysis users tend to recover in about 30 minutes, he said.

But Dr. Rajnish Mehrotra, a professor of medicine at the University of Washington, said providers face financial disincentives that prevent many from offering home-based services. “Most patients do not have access to home dialysis because either their insurance, mostly Medicare, is not going to pay enough to cover the costs, or (the patients) lack sufficient socio-economic resources to be on it,” he said.

Some experts say more end-stage renal disease patients would be served by home dialysis if more clinicians recommended it. Though the number of patients using peritoneal dialysis has increased by about 40% over the past five years, a 2011 study published in the American Journal of Kidney Diseases found that up to half of dialysis patients choose home treatment.

“Most programs that train nephrologists don't have enough people who are on home dialysis,” Mehrotra said. “As a result, the nephrologists are not comfortable offering home hemodialysis, which in turn it feeds into the cycle of it not being offered enough.”³³

Since kidney failure is a chronic disease and requires almost daily treatment intervention for survival, it can be taxing on patients to travel to treatments facilities and medical professionals have seen improved results and also just a better overall experience when patients, when medically appropriate to do so, can receive their treatment at home with loved ones, in a familiar environment and surrounded by their community. It also requires less travel and time away from work or doing what they love. However, few health insurance providers are willing to cover the costs of at home treatments. This is just one example of a case study in the medical field where better treatments plans can be implemented with a more integrative health care system, so that insurance providers don't become a barrier to better health and visualizing the data allowed for this type of case study to emerge.

Conclusion

Digital literacy should not be a prerequisite to accessing affordable healthcare, but with more than 1 in 3 campaigns on the GoFundMe platform being for Medical expenses of all kind, it is slowly becoming one. This is not a sustainable solution for a healthcare substitute, and many are being left out of the conversation particularly due the the stigma of crowdfunding, people with diseases perceived to be “choices” or “blameworthy” diseases like mental health and addiction or “getting my life together as a single mom with a sick kid”, as compared to “faultless” diseases such as cancer or a genetic disorder.

The social element also cannot be ignored, we know from the findings presented above that the campaigns with higher engagement yield higher success rate, so those without a built-in network

³³ (Johnson, Ross. 2014)

or community to help promote their campaign are more likely to suffer. “If there’s already a hierarchy of affliction and need in society, then crowdfunding often works to exacerbate it.”³⁴

As we evolve and technology makes things like home dialysis treatments an option, our healthcare system needs to evolve and adapt to both those capabilities and also people's needs. There are predictions that the percent of the workforce working from home in 2020 could be as high as 50%.³⁵ The data reveals that people’s needs are not financially astronomical and not always about the latest treatment, but on average, more likely about access to treatment and managing the balance of their lives- work, family, friends - when health issues arise. As we approach the 2020 elections, healthcare will be a topic that every politician will be addressing if not forced to address. Whether the solution is revival of the Affordable Healthcare Act or a single payer healthcare system, we are in a period of redesign and considerations towards new possibilities, and a human-centered solution are possible. By looking to the stories and needs of people first, this data can provide healthcare policymakers and civic-minded innovators new meaningful ways to consider a more integrative healthcare system.

Considerations and Future Directions

The data collection process was limited both by my lack of experience with data scraping tools and the consideration for acknowledging terms and conditions of the crowdfunding platforms I was accessing. Future development would likely include the use of scrapy or selenium-python, both of which are python browser-based data retrieval tools. My limited experience with Python tools has revealed its power and future goals as an extension of my work in this program is to better learn python and both its data collection and analysis capabilities. I often found myself resorting to the technical workaround of using more complicated or time consuming javascript functions to accomplish a task that would be significantly faster, more efficient and more accurate in Python. But often I needed to abandon the Python option if I could not figure it out quickly or risk not completing the task at all. This happened throughout my curriculum, and was mostly evident when working with large datasets that needed to be processed quickly, but the limitations of this was none more glaring than during the data collection phase of my thesis project. Before embarking on the continuation of this project or any like it, my Python skills will need to be developed.

³⁴ "People are Raising \$650 Million on GoFundMe each Year to Attack Rising Healthcare Costs."

³⁵ (Markets Business Insider. 2018)

After analyzing the data I realized I probably could have left out the “emergency” category campaigns, as the structured learning k-means algorithm grouped them together, because they were in essence the farthest neighbor from the other medical campaigns and the nearest neighbor to each other. More extensive machine learning algorithms would need to be applied to the data to test this model more in depth, and I imagine a larger data set would also have an effect on these groupings. Regardless, the tf-idf grouped the campaign with the top terms “fire”, “car”, “accident”, “family” together and so even without the knowledge that these campaigns had already been pre-grouped by the “emergency” category, the k-means cluster process was successful in recognizing their relation.

Collecting the entire corpus of campaigns listed on the sitemap.xml would also improve the analysis process, as more patterns would begin to reveal themselves over time, and the analysis would not be limited to technical limitations but explicit random samplings if a smaller dataset is still preferred.

There are arguments against using k-means as the most effective algorithm for clustering data as it is based on a Euclidean distance, when the suggested best practice is tf-idf and cosine similarity.

“K-means is designed for Euclidean distance.

The key problem is the mean function. The mean will reduce variance for Euclidean distance, but it might not do so for a different distance function. So in the worst case, k-means will *no longer converge, but run in an infinite loop* (although most implementations support stopping at a maximum number of iterations).

Furthermore, the mean is not very sensible for *sparse* data, and text vectors tend to be *very sparse*. Roughly speaking the problem is that the *mean* of a large number of documents will no longer look like a real document, and this way become dissimilar to any real document, and more similar to other mean vectors. So the results to some extent degenerate.

For text vectors, you probably will want to use a different distance function such as cosine similarity.

And of course you first need to compute number vectors. For example by using relative term frequencies, normalizing them via **TF-IDF**.

There is a variation of the k-means idea known as **k-medoids**. It can work with arbitrary distance functions, and it avoids the whole "mean" thing by using the *real* document that

is most central to the cluster (the "medoid"). But the known algorithms for this are much slower than k-means."³⁶

The model I created calculated the cosine distance as measure of similarity, and was accurately supported for the purpose and scope of this project. However, this discovery deserves consideration and more advanced research and testing to then possibly apply a different cluster method in future iterations and compare results to find the most optimum model.

A future direction would be to collect the data for multiple years and do a comparison over time, and also to collect secondary data set on state costs for top procedures and examine relationships between the GoFundMe corpus groups and cross reference with state data.

³⁶ (Stack Overflow. 2012)

Bibliography

- "Bankruptcy Statistics.", <https://www.debt.org/bankruptcy/statistics/>.
- "GoFundMe Medical Fundraising.", accessed May, 2019, <https://www.GoFundMe.com/start/medical-fundraising>.
- "People are Raising \$650 Million on GoFundMe each Year to Attack Rising Healthcare Costs." c. <https://www.forbes.com/sites/carolynmcclanahan/2018/08/13/using-GoFundMe-to-attack-health-care-costs/#692e432a2859>.
- "The Stanford Natural Language Processing Group", <https://nlp.stanford.edu/>.
- Alkon, Cheryl. 2018. "How to Crowdfund Your Kid's Medical Expenses." *Today's Parent*, "Mar 26, ". <https://www.todayparent.com/kids/kids-health/how-to-crowdfund-your-kids-medical-expenses/>.
- Anony-Mousse. 2012. *How can I Cluster Document using K-Means (Flann with Python)?*.
- Bassani, Gaia, Nicoletta Marinelli, and Silvio Vismara. 2018. "Crowdfunding in Healthcare." *Journal of Technology Transfer*: 1-21. doi:10.1007/s10961-018-9663-7.
- Berliner, Lauren S. and Nora J. Kenworthy. 2017. "Producing a Worthy Illness: Personal Crowdfunding Amidst Financial Crisis." *Social Science & Medicine; Social Science & Medicine* 187: 233-242. doi:10.1016/j.socscimed.2017.02.008.
- Bluth, Rachel. 2019. "GoFundMe CEO: 'Gigantic Gaps' in Health System Showing Up in Crowdfunding." *Kaiser Health News*, "Jan 16, ". <https://khn.org/news/gofundme-ceo-gigantic-gaps-in-health-system-showing-up-in-crowdfunding/>.
- Board of Governors of the Federal Reserve System. 2018. *Federal Reserve Board Issues Report on the Economic Well-being of U.S. Households*.
- Bureau of Labor Statistics. 2018. *Employee Benefits in the United States*.
- Burch, Gordon and Jason Chan. 2019. "Investigating the Relationship between Medical Crowdfunding and Personal Bankruptcy in the United States: Evidence of a Digital Divide." *MIS Quarterly* 43 (1): 237. doi:10.25300/MISQ/2019/14201.
- Centers for Medicare and Medicaid Services. 2017. *National Health Expenditures 2017 Highlights*.
- Collins, Sara R. and David C. Radley. 2018. *The Cost of Employer Insurance is a Growing Burden for Middle-Income Families*.

- Cunha, Darlena. 2015. "Americans are Crowdfunding Health Care. they Shouldn'T have to." *Time*, "Apr 30, ". <http://time.com/3831505/crowdfunding-health-care/>.
- Gonzales, Amy L., Elizabeth Y. Kwon, Teresa Lynch, and Nicole Fritz. 2018. "'Better Everyone should Know our Business than we Lose our House': Costs and Benefits of Medical Crowdfunding for Support, Privacy, and Identity." *New Media & Society* 20 (2): 641-658. doi:10.1177/1461444816667723.
- Healthcare Cost Institute. 2019. *2017 Health Care Cost and Utilization Report*.
- Henry J. Kaiser Family Foundation. 2018. *Key Facts about the Uninsured Population*.
- Johnson, Ross. 2014. "Home Dialysis Grows Despite Cost and Logistical Hurdles." *Modern Healthcare*, "Oct 11, ". <https://www.modernhealthcare.com/article/20141011/MAGAZINE/310119932/home-dialysis-grows-despite-cost-and-logistical-hurdles>.
- Lagazio, Corrado and Francesca Querci. 2018. "Exploring the Multi-Sided Nature of Crowdfunding Campaign Success." *Journal of Business Research; Journal of Business Research* 90: 318-324. doi:10.1016/j.jbusres.2018.05.031.
- Macdonald, Claire. 2018. "Healthcare: Pouring a Little Cold Water on Crowdfunding." *OECD Obs*.
- Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval* Cambridge University Press.
- Markets Business Insider. 2018. "Fifty Percent of U.S. Workforce Will be Remote by 2020; ULTATEL's Cloud-Based Technology Paves the Way." , "Feb 14, ".
- Mathur, Aparna. 2018. "Exposing the Myth of Widespread Medical Bankruptcies." *Forbes*, "Apr 9, ". <https://www.forbes.com/sites/aparnamathur/2018/04/09/exposing-the-myth-of-widespread-medical-bankruptcies/#3ae7916dc2a1>.
- National Conference of State Legislatures. 2018. *Health Insurance: Premiums and Increases*.
- Payne, Emily and Chris Nicholls. 2019. "The Dismantling of the ACA: An (Updated) Timeline." . <https://www.benefitspro.com/2019/03/22/the-dismantling-of-the-aca-a-timeline/?slretur n=20190416014145>.
- Qaiser, Shahzad and Ramsha Ali. 2018. "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents." *International Journal of Computer Applications*. https://www.researchgate.net/publication/326425709_Text_Mining_Use_of_TF-IDF_to_Examine_the_Relevance_of_Words_to_Documents.

- Ramos, Juan. 2003. "Using TF-IDF to Determine Word Relevance in Document Queries" .
[https://www.semanticscholar.org/paper/Using-TF-IDF-to-Determine-Word-Relevance-i
n-Queries-Ramos/b3bf6373ff41a115197cb5b30e57830c16130c2c](https://www.semanticscholar.org/paper/Using-TF-IDF-to-Determine-Word-Relevance-in-Queries-Ramos/b3bf6373ff41a115197cb5b30e57830c16130c2c).
- Rose, Brandon. "Document Clustering with Python." , <http://brandonrose.org/clustering>.
- Techonomy Media Inc. 2017. "Transcript from Techonomy Health in NYC." .
[https://techonomy.com/wp-content/uploads/2017/06/Data-Driven-Healthcare_transcript
-2.pdf](https://techonomy.com/wp-content/uploads/2017/06/Data-Driven-Healthcare_transcript-2.pdf).
- Tozzi, John and Zachary Tracer. 2018. "Sky-High Deductibles Broke the U.S. Health Insurance System." *Bloomberg*, "June 26, ".
[https://www.bloomberg.com/news/features/2018-06-26/sky-high-deductibles-broke-the
-u-s-health-insurance-system](https://www.bloomberg.com/news/features/2018-06-26/sky-high-deductibles-broke-the-u-s-health-insurance-system).
- World Health Organization Global Health Expenditure database. 2017. *Current Health Expenditure (% of GDP)*.