Springboard Data Science
CAPSTONE PROJECT #2

by Jason Green-Lowe

Executive Summary

What types of citations against healthcare facilities are most likely to lead to large fines? Can the amount of a fine be predicted based on the narrative summary of the accident or injury that led to the fine? Based on a publicly available dataset from the California Department of Public Health, the answer is yes: incidents involving cardiovascular emergencies or elder abuse were associated with an average fine of about $12,000, compared with an average of only $1,800 for incidents involving diabetes or missing paperwork.

Using the natural language processing tools in the Spacy and Gensim libraries, I converted the text of 2,883 incident reports from California nursing homes into lists of lemmatized trigrams, i.e., short phrases made up of meaningful root words. I then used Scikit-learn's vectorizer tools to convert these trigrams into sparse arrays. With the arrays, I was able to make both continuous and categorical predictions about the size of the fine associated with each narrative report using three different machine learning techniques. The most successful categorical model was a Gradient Boosting Machine, which achieved an F1 score of 0.929 when distinguishing high-fine narratives ($5,000 or more) from narratives with medium or low fines. The most successful continuous model was the Random Forest Generator, which achieved a correlation of r = 0.746 between its predictions and actual test data that was held in reserve for validation.

In addition to predicting the amount of these fines, I also generated Latent Dirichlet Allocation (LDA) models to automatically identify the primary topic of each narrative. The topics had a coherence score between 0.4 and 0.49 for all eighty of the sets of hyperparameters that I tested, suggesting that the topics identified were persuasive and human-recognizable. In fact, most of the topics do appear fairly clear. For example, one of the topics had "rash," "scabies," "infection-control," "outbreak," and "cream" as its top five most characteristic words, all of which work together to paint a coherent narrative. Another topic had "ad," "trust-account," "fund," "money," and "$" as its top five most characteristic words, suggesting that this topic deals with fines that were imposed for misuse of patient funds. The LDA models generated these topics using only a general language database, with no medical knowledge or training.

Data Sources and Data Cleaning

All data for this project was obtained from the California Department of Public Health through California's Open Data Portal. A general database on healthcare facility enforcement actions lists the address, date, fine amount, and facility name of each incident where a facility was issued a fine based on problems with patient safety between 2007 and 2017. Another specialized database contains the full text of narrative reports from the more serious safety violations that occurred between 2012 and 2017. These narrative reports are linked by incident report numbers. Although the name of the hospital sometimes varies slightly from one database to the next (e.g. Castro Valley Kaiser vs. Kaiser of Castro Valley), the incident report number appears to be sufficient to uniquely identify each incident. There were 2,885 incident reports in the full-text database, of which 2,883 had valid numerical data in their "fine amount" column.

The data as presented by California's Open Data Portal was already very clean. The only changes that I needed to make to get a usable Pandas DataFrame were to convert some of the columns from text into date/time or from currency into numerals. Other than the 2 rows that were missing their fine amounts (which I simply dropped from the database), there was only one column that had any null values: "Class_Assessed_Final." The reason for the null values was that sometimes the fine category (e.g. AAA for a very severe fine, C for a very minor fine) did not change or was not appealed, in which case the column was intentionally left blank. This did not present a problem for my research.

Natural Language Processing (NLP)

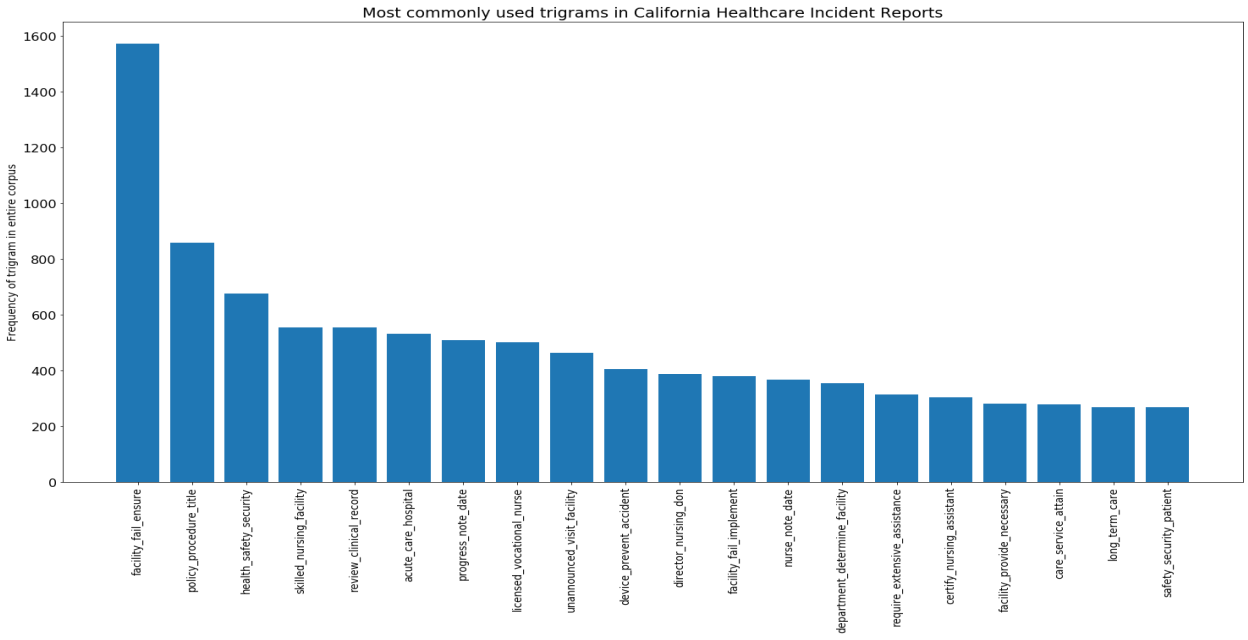
After cleaning the data, I processed all of the narratives using Spacy's medium web corpus. The Spacy library contains tools that allowed me to automatically remove stopwords (commonly occuring English words like "the," "and", "for," etc.), punctuation, whitespace, capitalization, and conjugations. I also removed very rare words and words that were very common in this particular set of narratives (e.g. "patient," "resident," "chart," etc.). This left me with a denser vocabulary, i.e., each word that appeared in the vocabulary at all was more likely to occur multiple times per document, to make up a large share of a document, and to occur in multiple documents across the corpus.

My primary approach to the natural language processing was to use the "bag of words" technique, i.e., to ignore the order in which words appear in any given document. The downside to this approach is that it is unable to account for syntactic meaning; for

example, the sentence "The patient received no benefit from a large dose of antibiotics" would be treated the same as the sentence "The patient received a large benefit from no dose of antibiotics," even though those two sentences have opposite meanings. The advantage to this approach is that it radically reduces the memory needed to hold the corpus, allowing machine learning models to be rapidly trained and optimized using many iterations. Moreover, when removing 'stopwords' like "no," "not," and "however," most of the syntactic meaning would be lost even if the word order were preserved.

To help recapture some of the meaning lost by ignoring word order, I used Gensim's "Phrases" tool to identify frequently-occuring bigrams and trigrams in the corpus. If two words (or three words) appeared adjacent to each other for a large fraction of the total appearances of all of those words, then the two- or three-word phrase was treated as an additional token that could be added to the vocabulary.

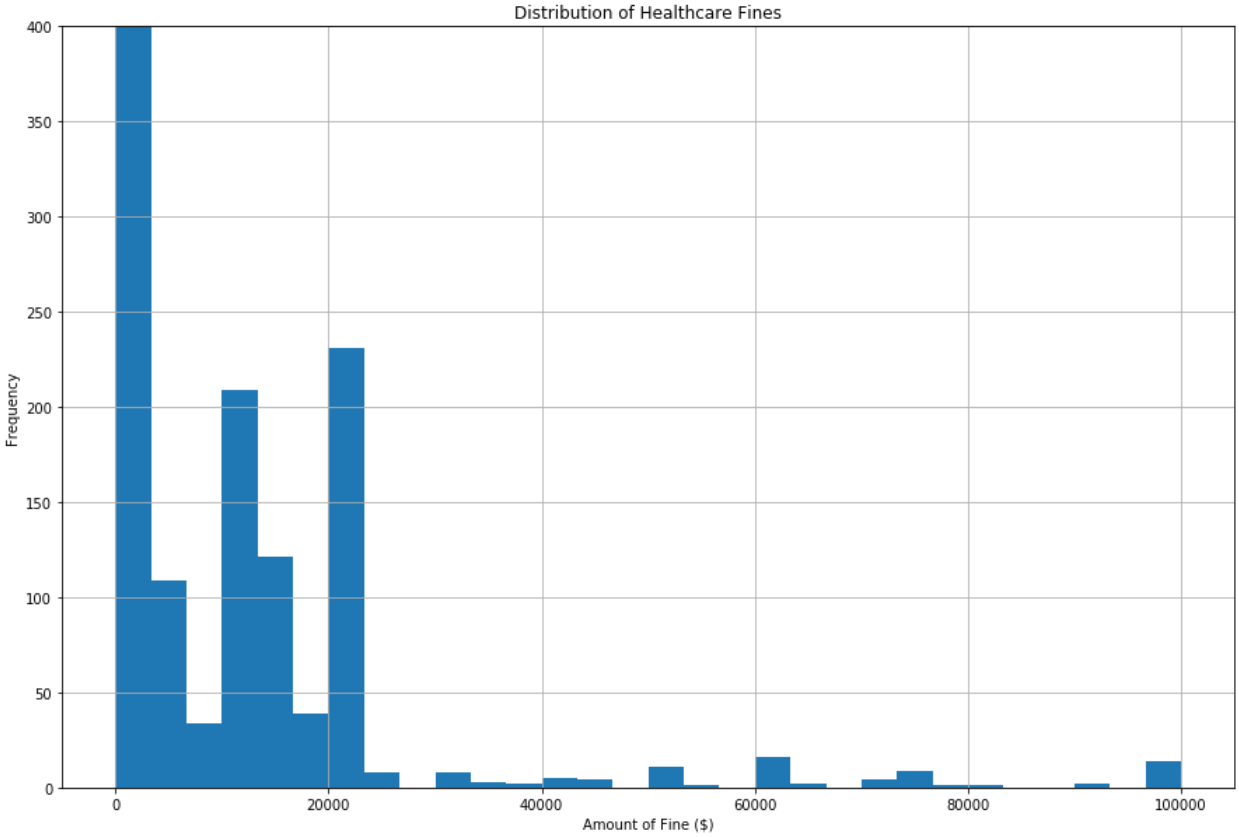
For example, the words "skilled," "nursing" and "facility" appear together so often in the California Public Health corpus that the Phrases tool hyphenated them as "skilled-nursing-facility." This is appropriate because those three words together do carry a unique meaning that is not fully captured by adding each of the three words on its own. Many of the most frequently-occuring trigrams in the corpus were surprisingly natural-sounding phrases that occur in everyday conversation about healthcare, e.g., "licensed-vocational-nurse," or "acute-care-hospital," or "progress-note-date." Even the phrases that were somewhat more awkward were still immediately intelligible, e.g., "facility-fail-prevent" is almost certainly a contraction for "The facility failed to prevent."



Data Exploration

After converting the narrative texts into lemmatized trigrams and the fine amounts into integers, I explored the data to see which sizes of fines were most common, how frequently each trigram appeared, and whether there was any obvious connection between the size of a fine and the most-commonly-appearing words in a narrative.

The distribution of fines was highly irregular, as shown in the histogram below. The maximum fine was $100,000, but the vast majority of fines were less than $25,000, and a handful of standard fine amounts like $1,000 and $2,000 appeared very frequently. In fact, $2,000 was actually the median fine in the dataset.



The most frequently occurring words in the corpus were "resident" (93,195 instances), "facility" (42,661 instances), and "state" (37,519 instances). About 90% of the most commonly occurring words were neutral with respect to the likely size of the fine and were not obviously connected with any particular topic. This makes sense because a word that is strongly connected to a particular topic is only likely to appear in narratives about

that topic, so it is not likely to be one of the most commonly appearing words across the entire database. The fact that the most commonly-occuring words were not very informative about a narrative's topic or likely fine size helps justify the decision to remove very-commonly-occurring words from the predictive analysis in the next section.

Although there were many bigrams and trigrams identified by the Gensim Phrases tool, these n-grams were not very common compared to unigrams. The 50th-most-common unigram ("hour") had 5,192 instances, compared to only 4,508 instances for the 1st-most-common bigram ("care-plan"). The 10th-most-common bigram ("family-member") was roughly as common as the 1st-most-common trigram ("facility-fail-ensure"), with both phrases having about 1,550 instances each.

Unigrams

word	frequency
resident	93195
facility	42661
state	37519
patient	31776
indicate	24292
care	18806
staff	18395
date	17792
review	17593
client	16772
interview	15119

Bigrams

word	frequency
care_plan	4508
policy_procedure	3640
clinical_record	2535
note_date	2076
fail_ensure	2059
diagnosis_include	1999
progress_note	1757
30_p	1736
facility_fail_ensure	1573
family_member	1538
trust_account	1533

Trigrams

word	frequency
facility_fail_ensure	1573
policy_procedure_title	859
health_safety_security	677
skilled_nursing_facility	555
review_clinical_record	555
acute_care_hospital	532
progress_note_date	509
licensed_vocational_nurse	501
unannounced_visit_facility	464
device_prevent_accident	405
director_nursing_don	387

I then split the corpus into four sections: narratives associated with no fine ($0), narratives associated with a small fine ($1 - $1,999), narratives associated with a medium fine ($2,000 - $9,999), and narratives associated with a large fine ($10,000 - $100,000). I counted the frequencies of words in each of these smaller corpuses and subtracted the relative frequency of those same words in the entire corpus (adjusting for the relative size of each corpus) to see if the most frequently appearing words in each sub-corpus would offer a clue as to what types of incidents led to the highest fines. However, at this stage of the analysis, the word frequencies were not very informative. It is possible to make guesses as to why words are appearing in one column but not another; for example,

perhaps "p.m." appears frequently in the "high fine" column because emergencies where delivering care by a particular hour of the day is critical were cases that led to more deaths and therefore higher fines. Similarly, words like "write," "clarify," and "clinical-record" in the no-fine category suggest that facilities who successfully appealed their fine were let off with a warning because the only real problem with those facilities was a deficiency in their paperwork. Ultimately, though, these are only guesses. There are too many words in these lists that could appear in any type of narrative to make any firm statements about what kinds of narratives are leading to high fines from the raw word frequencies alone.

no_fine	small_fine	med_fine	large_fine
discharge	client	facility	resident
notice	account	abuse	1
write	patient	transfer	indicate
transfer_discharge	shall	discharge	's
record	dcs	administrator	care
facility	ad	allegation	fall
reason	o	transfer_discharge	date
clinical	trust_account	notice	review
clinical_record	trust	ssd	state
resident	abuse	adm	physician

Latent Dirichlet Allocation (LDA) Modelling

To pursue the question about what kinds of narratives lead to high fines, I excluded the most common words in this corpus, filtering out any words that appear in 15% or more of the incident reports. I then trained a series of Latent Dirichlet Allocation models to classify the filtered corpus based on notional "topics" that are automatically detected within the data. Although an LDA model does not have any semantic understanding of what a document is about, it is able to develop and assign topics to documents based on the words that are most *characteristic* of a particular group of documents. For example, if the words "pigeon," "dove," and "sparrow" are all found at 5 times their ordinary frequency in documents 3, 6, 7, 12, and 34 in a corpus, then an LDA model might generate a topic labelled Topic 4 that is defined as ["pigeon", "dove", "sparrow"] and then describe documents 3, 6, 7, 12, and 34 as belonging with high probability to Topic 4.

A human observing the output of the LDA model can look at the words in the topic definition and determine that Topic 4 is, e.g., a topic about birds, and that documents 3, 6, 7, 12, and 34 are therefore also likely to contain an above-average amount of information about birds.

Three of the most important parameters for LDA models are:

- The number of distinct topics used in the model,
- *Alpha*, the strength of the prior distribution of topic-assignments for each document, and
- *Eta,* the strength of the prior distribution of topic-assignments for each word.

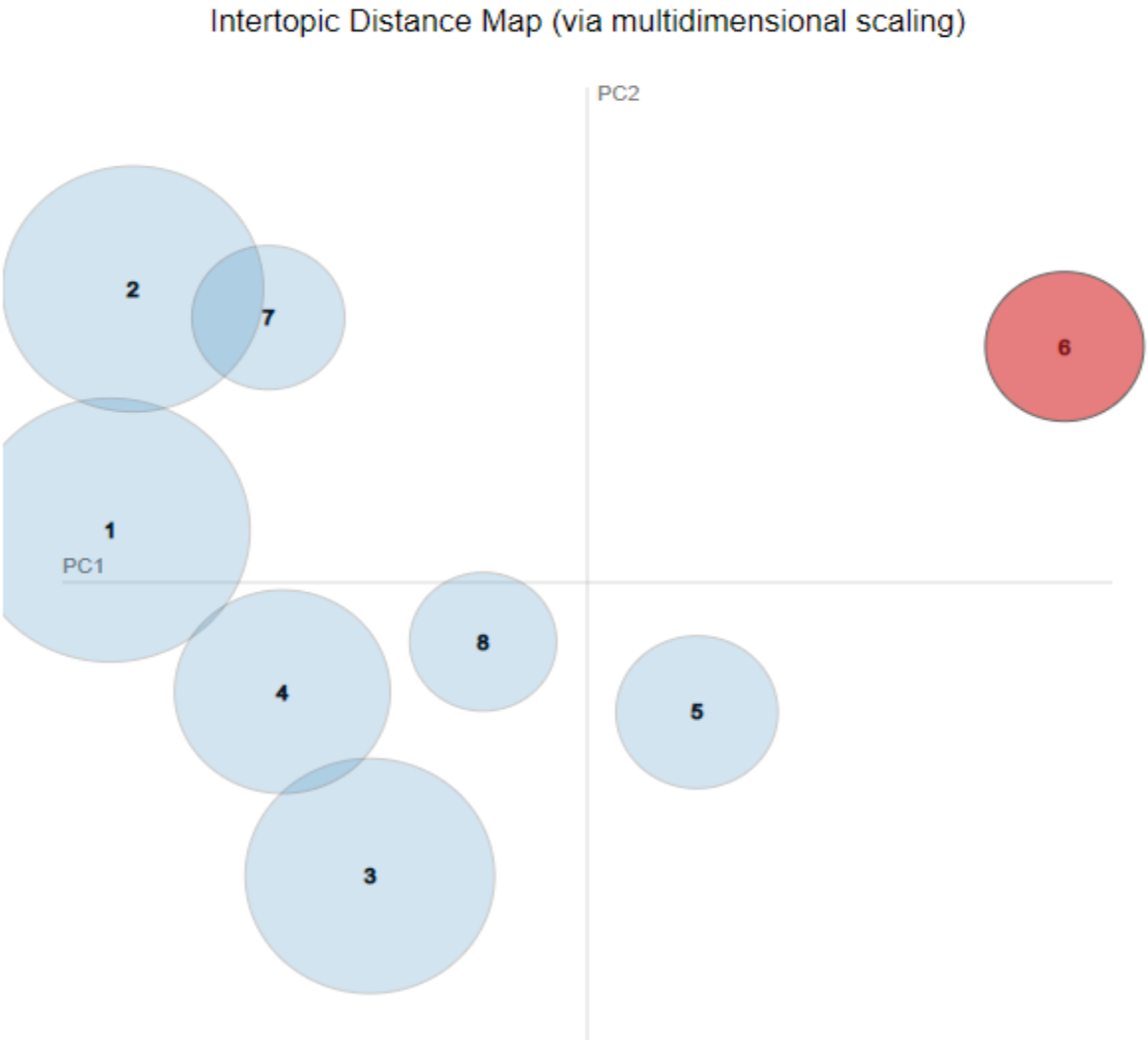
If *alpha* is set too high, then the model will rigidly assume that each document discusses many different topics and will be too reluctant to assign a document to a particular topic. If *alpha* is set too low, then the model will rapidly conclude that a document is focused on a particular topic, and will guess at each document's primary topic based on too little evidence, inaccurately excluding other possibilities. Similarly, if *eta* is set too high, then the model will rigidly assume that any given word could belong to any given topic, and if *eta* is set too low, then the model will rapidly assign words to topics even when the words do not really belong as part of those topics.

I tested a grid of plausible hyperparameters to try to identify which model would be the best fit for this particular corpus. I tried models with 3, 5, 8, 10, and 20 distinct topics, *alphas* of 0.01, 0.1, 0.3, and 0.9, and *etas* of 0.01, 0.1, 0.3, and 0.9, generating a total of 80 sets of hyperparameters. I then tested each model's coherence score using the C_v metric, which assesses the semantic similarity of the most common words in each topic to ensure that the algorithm has actually identified a pattern of words with a common semantic meaning, rather than a mere statistical artifact.

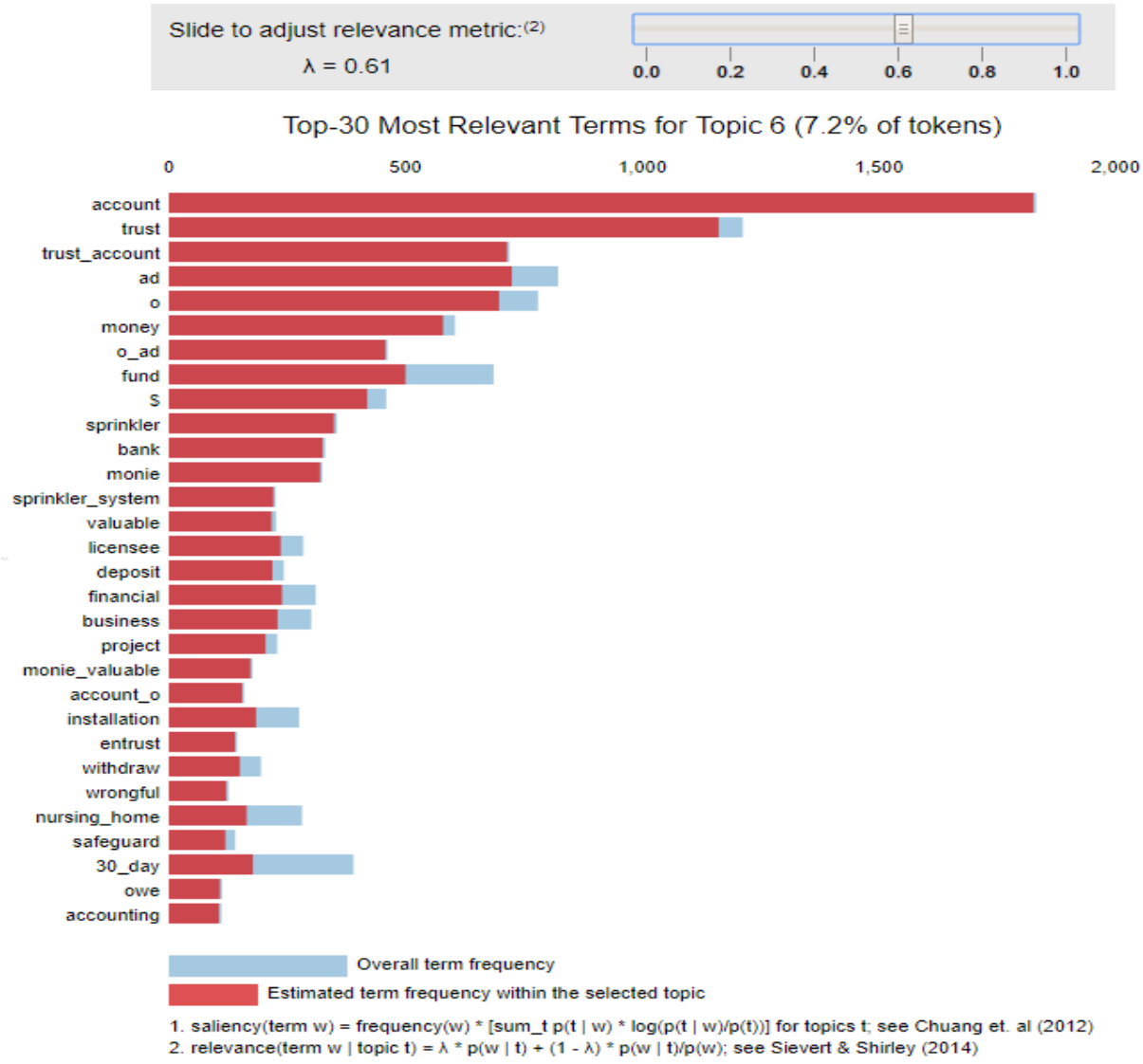
The coherence score was relatively robust to variance in the hyperparameters, with all 80 specifications returning a coherence score between 0.40 and 0.49. Hyperparameter sets at the high end of this range were available for all different numbers of topics, so I chose to work with a model that had 8 different topics. Based on exposure to healthcare narratives from my work as a medical malpractice attorney, I subjectively felt that 8 topics would be enough to cover the likely range of narratives without creating topics that overlap each other or that have been artificially merged together. The highest coherence score for 8-topic models was at *alpha* = 0.9 and *eta* = 0.9, so I used those hyperparameters to build a 'clean' model that could be exploited for visualization and prediction. I re-trained this clean model with ten times the number of iterations and passes to help make the model more precise.

The end result was a satisfying distribution of documents into relatively crisp, meaningful topics, as shown in the visualization below:

The two dimensions of the graph represent a principal component analysis of the topics, which simplifies the variation between the topics into only two dimensions for ease of visualization. Surprisingly, even these simplified dimensions appear to have an obvious meaning: topics that are further to the right-hand side of the graph relate to finances and money. Topic 6 (highlighted in red) is all about fraud, embezzlement, and trust accounts. Topic 5, near the center of the graph, includes a mix of words about payment and treatment. All other topics, on the left-hand side of the graph, deal with medical issues rather than financial issues. The top-left corner of the graph is more focused on emergencies and cardiovascular issues, and the bottom-left corner of the graph is more focused on psychological issues and sexual abuse.



The LDAvis library provides a visualization tool that allows users to see at a glance which words appear most frequently in each topic to get a sense of what those topics are about. By hovering the mouse over different topics, the user can generate different lists of commonly appearing words. A slider bar allows users to determine what weight to assign to global probability versus topic-specificity: moving the slider to the left displays more words that appear disproportionately often in one particular topic (which can sometimes show statistical artifacts that are unique but meaningless, e.g., a particular date), and moving the slider to the right displays more words that are common both in this topic and across the entire corpus (which can sometimes fail to adequately differentiate the topic). In the chart below, a moderate setting of the slider bar allows us to see terms that are clearly specific to the topic of financial fraud, although a few words about sprinklers seem to have snuck in by chance.



Topic-Based Fine Prediction

Using the topics developed above in the Latent Dirichlet Allocation models, I assigned each document in the corpus to exactly one topic -- whichever topic was most closely associated with that document. I then calculated the average fine assessed for each document that was assigned to that topic. This generated strikingly different average fines: as shown in the chart below, narratives about cardiovascular emergencies generally had fines that were five times larger than narratives about missing paperwork or dietary problems.

	Train	Test
Life Support	$13,831	$13,195
Diabetes	$10,403	$11,097
Sores	$9,852	$8,346
Escape	$9,240	$12,406
Abuse	$2,361	$2,243
Rashes	$2,040	$2,347
Administration	$1,988	$2,414
Theft	$1,709	$1,786

As a further validation of the choice of topics, I held out 20% of the documents as test data when training the LDA model, and then averaged the fines for these 20% of the corpus separately, so as to calculate an average fine for the training data in each topic, and an average fine for the test data in each topic. As shown in the chart above, the average fine per topic was nearly the same in both datasets, suggesting that (a) these topics reflect real patterns in the data, and (b) these topics are associated with real differences in the average fine per topic. The training data average fine amounts and test data average fine amounts per topic have a Pearson correlation of r = 0.963, which is extremely strong and indicates that the averages are very closely associated within each topic.

Categorical Fine Prediction

To extract further detail from the data, I then set aside the idea of LDA topics and modelled each document as a vectorized bag of words, without trying to specify which words were important or relevant. I then used several different unsupervised learning techniques to try to predict (a) which documents would have a fine of more than $1,000,

and (b) which documents would have a fine of more than $5,000. Each result was treated as a binary variable, i.e., the document either is or is not over the threshold, and the machine learning algorithm tried to predict whether the document's fine was over the threshold by using only the vectorized bag of words from the document's narrative.

The specific thresholds of $1,000 and $5,000 were chosen based on the distribution of the data: approximately three-quarters of the fines were above $1,000, and approximately one-quarter of the fines were above $5,000, so these thresholds gave the models a chance to make predictions in both 'easy' and 'hard' conditions, i.e., in cases where the naive model that always predicts that the fine is over the threshold will usually be correct or usually be incorrect.

As shown in the chart below, the Gradient Boosting Machine ("GBM") classifier was more successful than either the Random Forest Generator or the Multinomial Naive Bayes classifier. The GBM generated the highest F1 scores in both the $1,000 (easy) and $5,000 (hard) test conditions and was able to correctly identify an impressive 96% of the cases where the fine was over $5,000, even though only 27% of the total cases were over $5,000.

	MultinomialNB	RandomForest	GradientBoost
Train >$1,000	0.866	0.738	0.919
Test >$1,000	0.761	0.732	0.843
F1 >$1,000	0.828	0.844	0.895
Naive >$1,000	0.727	0.727	0.727
Train >$5,000	0.869	0.806	0.974
Test >$5,000	0.785	0.794	0.962
F1 >$5,000	0.679	0.403	0.929
Naive >$5,000	0.270	0.270	0.270

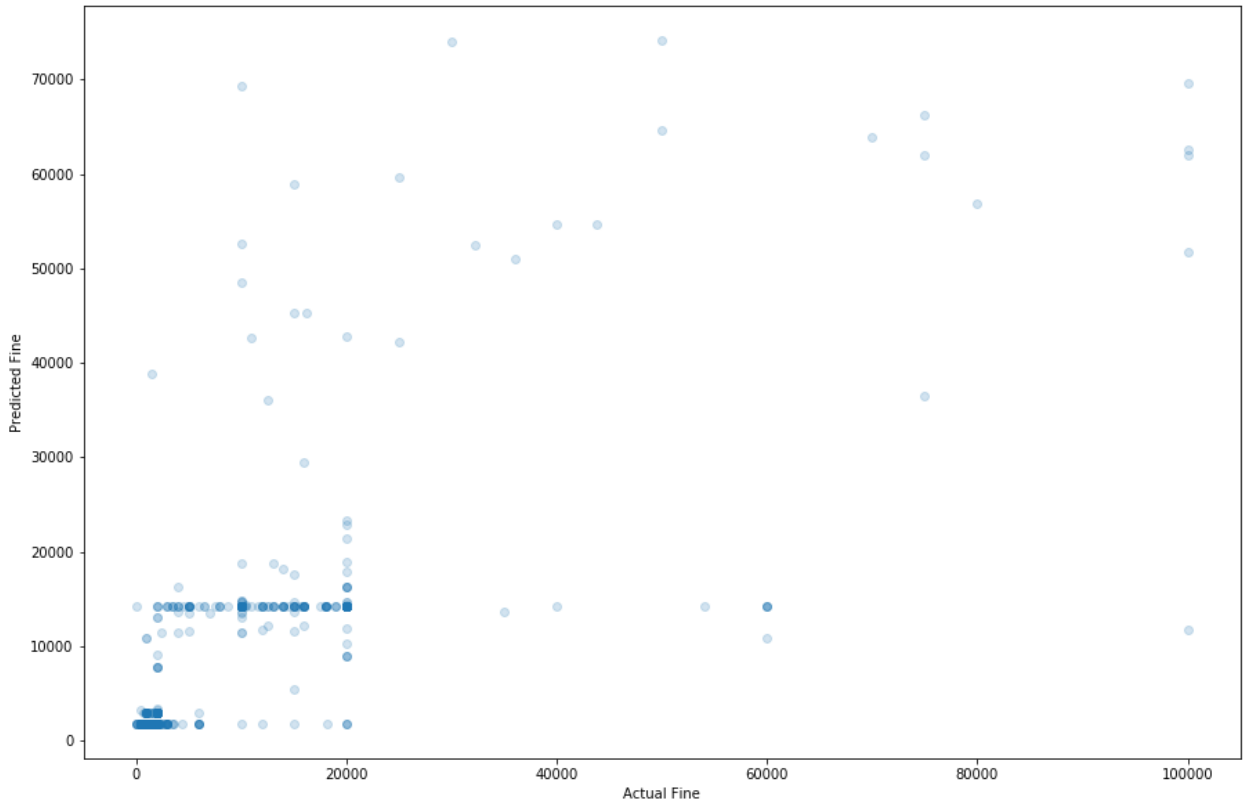
Continuous Fine Prediction

Some clients might want the ability to estimate the exact size of the likely fine on a narrative report, rather than just to classify the expected fine as small, medium, or large. To continuously estimate the size of each fine, I used regressors rather than classifiers, and coded the y-variable as a linear estimate of each fine, rather than as a binary variable stating whether the fine was over a given threshold.

As shown in the chart below, the results were somewhat less compelling than the results for the categorical fine prediction, and this time the Random Forest Regressor was the most accurate predictor.

	MultinomialNB	RandomForest	GradientBoost
Train	0.600	0.818	0.704
Test	0.454	0.542	0.311

The scatterplot below graphs the actual fine (x-axis) against the predicted fine (y-axis) using the best available random forest regressor model. Each point is rendered at 20% transparency to allow for easier visualization of dense and sparse areas of the graph. As you can see, the model was able to accurately predict very high fines for some of the fines that were in fact very large, and accurately predict very low or zero fines for some of the fines that actually were at or close to zero, but the model struggled with intermediate data: the model predicted $15,000 fines for a very large fraction of the data, even though very few fines were exactly $15,000. The model also struggled to accurately predict $20,000 fines, with several fines that were actually $20,000 being estimated as $0 or $25,000.



Applications and Ethics

The kinds of predictions made by the models discussed in this paper could be useful to a wide variety of clients. The ability to predict the likely amount of a fine for a given health incident would help inform medical directors who are trying to reduce the rate of serious health errors, litigation directors who are trying to reduce the liability for their hospital or nursing home, insurance adjusters who are trying to set an appropriate fee for malpractice insurance, defense attorneys who are trying to figure out which cases to settle instead of defending in court, and government prosecutors who are trying to figure out which cases to prioritize in order to have a maximum deterrent impact on health care companies.

As with many kinds of data analysis, the ability to rapidly gain insight into thousands of cases can be a clarifying and useful tool, but it is also prone to abuse: people who want to know how the government assigns liability in order to maintain their behavior while minimizing their fines can get just as much use out of these tools as people who want to try to change their behavior to minimize the harm they are causing.

The economist Charles Goodhart is famous for claiming that "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." As applied to this context, Goodhart's Law means that the more hospitals know about how their performance is being evaluated, the more they will be able to "cheat" at those evaluations by creating a superficial appearance of safety and compliance that does not match the real character of their operations. Agencies who have access to insights from machine learning are like teachers who can see the exact questions that will appear on a standardized test -- there is a strong temptation to narrowly focus the curriculum (or the quality control procedures) around only those topics that are likely to be evaluated.

One way to cope with this problem is to try to instill a code of professional ethics among data scientists. Although most statistical patterns **can** be exploited for narrow commercial gain, not every kind of data science project **should** go forward, and it might be possible to set up a norm against certain kinds of unethical research, especially in the public sector. A key question for practitioners to ask is whether a given research project seeks to aid a client in complying with the law, or whether the project is merely a tool to help a client reduce the consequences of breaking the law.

Another way to cope with this problem is to try to ensure that all stakeholders in the system have access to machine learning techniques: if regulators can see the patterns and

regularities in their own fines and inspection procedures, then they may be able to randomize or add more variety to those procedures so as to more broadly measure the full spectrum of hospital behavior. As the scope of a regulatory evaluation becomes broader and broader, it becomes more and more pointless to try to game the system: eventually, the conduct being evaluated is almost coterminous with the conduct that the government wants to promote, so a hospital who tried to reduce its fines would necessarily wind up reducing patient harm as well.