Applied Topic Modelling and Text Mining to Improve Gas Related Building Safety

Darius Mehri
NYC Department of Buildings
New York City, NY, USA
dmehri@buildings.nyc.gov

ABSTRACT
Since the Second Avenue explosion, the NYC Department of Buildings (DOB) has collaborated with NYC utility companies to prevent future incidents. This paper will show how topic modelling, an advanced natural language processing technique, can be used to understand and prevent future gas related incidents. The topic modelling algorithm automatically generates topics (or "themes") from notification notes. The results show that topics account for a particular pattern of word use that correlate with patterns of occurrence and observation of risk. Text mining is then used to develop a risk tool to classify incidents as high or low risk. Since the implementation of the tool, the DOB inspected 1562 high risk buildings and 25% of the inspections resulted in enforcement actions.

1.INTRODUCTION
On March 26, 2015 a gas explosion occurred at 121 Second Avenue in the East Village. The fire that resulted from the explosion caused two deaths, nineteen injuries and destroyed four buildings. The explosion was caused by an illegal tap into the gas main and the bypass of the gas using a garden hose.

To help prevent future catastrophic gas incidents, in 2016 City Hall enacted local law 154 requiring utility companies to notify the DOB about gas incidents that are a danger to the public. Since the evaluation of risk by the utility company can be subjective, the agency decided to use natural language processing to analyze notification notes to "triage" the incidents to ensure that risk is properly classified. An objective of the project therefore is to reclassify a critical number of misclassified incidents according to DOB criteria.

Topic modeling is a statistical method that automatically analyzes the words in a body of texts to discover the themes that run through them, how themes in the texts are connected to each other, and how they can change over time (Blei, 2012). Topic modeling is particularly useful for the analysis of large amounts of unstructured documents, and in finding patterns in the texts without prior knowledge about the content of the documents. An advantage for using the method is that it provides an automated (rather than a labor intensive and manual) procedure for coding texts.

In recent years, topic modelling has become a popular method in the social sciences and the humanities. It has been used by a number of academics to uncover themes in a large body of documents. These studies include the analysis of newspaper articles to identify the linguistic contexts that surround policy domains (DiMaggio, 2013), journal abstracts to understand the changing contours of academic fields (McFarland, 2013), and measuring political engagement in massive open online courses (Reich, 2016). This research applies the topic modelling methods developed in these academic papers but does so to a case where the results are used to develop a risk tool to inform inspection strategy that targets unsafe buildings.

The development of the risk tool presented in this paper has three steps. First, topic modelling was used to automatically generate themes from a collection of notes in the utility notification reports. Second, the topics were analyzed by the plumbing unit to determine what topics and their key words determined high risk according to agency criteria. Third, by combining the topic keywords and previous DOB building violations, a tool was developed using text mining to determine if a building is high risk and rank the level of risk.

2. DATA AND METHODS
The collaboration between the DOB and the utility companies is a multi-step process. Utility companies are notified about gas incidents where they are required to shut off the gas at buildings over a range of issues, such as gas leaks, fires, and observation of illegal activity. If the utility company observes illegal activity (i.e. diversion or theft of service) they are required to correct the problem and then notify the DOB within 24 hours to inspect the building. If the utility company does not observe illegal activity and restores service after repairing the gas apparatus (i.e. non-illegal gas leaks), the utility companies are required to notify the DOB about these incidents at the end of the month.

All of the notifications, whether sent to the DOB daily or at the end of the month, include notes about the incident. The corpus included 1135 observations of original notes from the notifications. The topic modelling analysis was conducted on the first few months of notes from utility company notifications. The reason a small sample was used is that at the time of the analysis, the cooperation with the utility companies had just begun. The topic modelling was therefore useful as an exploratory tool to uncover and understand the themes in the utility notes.

Bloomberg Data for Good Exchange Conference.
16-Sep-2018, New York City, NY, USA.

The topic modelling algorithm generates topics based on the co-occurrence of key words in a corpus (a collection of documents). A topic is defined as a probability distribution over a fixed vocabulary. For example, if "illegal" occurs alongside "service", "valve" and "shut" with high probability, it is considered a topic. In this case, the topic relates to a gas shut-off due to illegal work. All of the documents in a corpus share the same set of topics, however, each document reveals the topics in different proportions (Blei, 2012). Since topics emerge from the original texts automatically, they do not require prior annotation or labelling.

The researcher must specify the number of topics, and tune two parameters: exclusivity and semantic coherence (Roberts, 2016). Exclusivity is defined as how unique a topic is compared to the others. The topics do not need to be mutually exclusive, but key words that are ranked high in one topic typically should not be ranked high in another topic. Semantic coherence is defined as how meaningful a topic is to the researcher. To achieve high semantic coherence, the researcher must change the number of topics and evaluate them along with the original text to ensure they are meaningful to the research question. The reason this process is important is that in topic modelling the documents are the observations and the topic structure - the words assigned to the topics - are the hidden structure. The goal of the researcher is to interpret the hidden structure (Blei, 2012).

The data provided by the utility companies includes notes associated with the incident. The notes are generated freeform by the utility company technician and therefore contain many misspellings. This problem adversely impacts accuracy since the topic modelling algorithm considers misspelled words as distinct. To resolve this issue, an automatic spell corrector developed by Peter Norvig was implemented. When the auto spell correction is complete, stop words are removed from the text and the words are lemmatized and reduced to their root form through stemming..

3.RESULTS

Optimal exclusivity and semantic coherence was achieved at twelve topics. The topics can be broadly categorized as incidents related to boilers, customer complaints of service shut off, gas leaks, gas meter issues, illegal activity, failure of an integrity test and unsafe piping. Below is a complete list of the twelve topics, their key words, topic proportions and their correlations.

Topic 1: Boiler and gas related issues (includes physical harm to residents)
Key Words: boiler, acv, found, plug, spillage, unit

Topic 2: Customer water, heat and gas complaints
Key Words: water, customer, hot, heat, turn, call

Topic 3: Leak in apartment related to cook and/or riser
Key Words: apartment, cook, riser, valve, plumber, notification

Topic 4: FDNY called to location, structural damage
Key Words: FDNY, fire, locate, shut, due

Topic 5: Found one or multiple gas leaks, service shut off.
Key Words: leak, found, shut, odor, locate

Topic 6: Checked for gas leaks due to odor, found/ not found
Key Words: gas, check, vent, flue, made, odor, read

Topic 7: Meter problems, locked gas due to safety issues
Key Words: meter, lock, found, case, turn, safety

Topic 8: Found illegal activity and service shut off
Key Words: service, valve, shut, locate, head, illegal, hose

Topic 9: Failed integrity test
Key Words: test, fail, integrity, left, meter

Topic 10: Piping unsafe due to shoddy and/ or unsupported work
Key Words: pipe, tag, house, left, issue, lock, safe

Topic 11: Gas leaks due to unknown reasons
Key Words: meter, time, amp, correct, remark

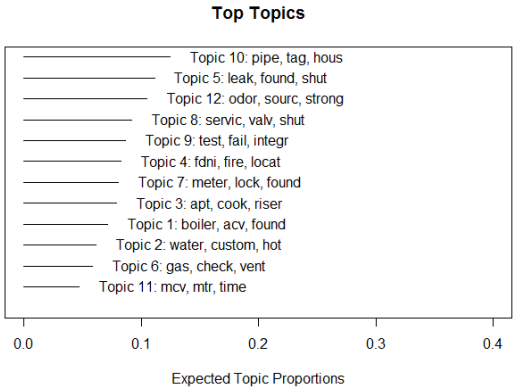
Topic 12: Odor strong in apartment, source known or unknown
Key Words: odor, source, strong, faint, apartment, found

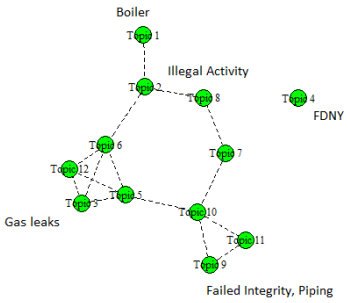
3.1. RANK, CORELATION AND CLUSTERING OF TOPICS

Topics are ranked according to the probability that they occur in the corpus, Figure 1 shows the ranking of the topics. These results show that illegal activity is a high ranking topic but other topics that do not necessarily involve illegal activity, dominate the proportion of topics in the corpus. As can be seen in the figure, the top three topics include Topic 10 (unsafe pipping) at 12%, Topic 5 (multiple gas leaks) at 11%, and Topic 12 (strong odor in apartment) at 10%. The topic with illegal activity, Topic 8, is ranked fourth and with an 8% probability of it occurring in the corpus.

A network map was generated to show topic ties and the ways in which they are clustered. Topics that share ties are closely correlated (they share similar words). The figure below shows Topic 8 (illegal activity) is mostly correlated with Topic 2 (customer complaints) and Topic 7 (meter problems). This high correlation shows that illegal activity is associated with meter issues and that it may be uncovered and reported to the utility companies via of customer complaints. Interestingly, topics clustered together are those associated with "gas leaks" and "failed integrity, piping" incidents.

Figure 1: Top Topics and Correlations





4. THE RISK TOOL AND ITS IMPACT

4.1 RISK TOOL

After the topic modelling analysis was complete, a meeting was held with the DOB Plumbing Unit to develop the risk analytics criteria. The unit was most interested in topics that could help prevent future gas explosions, therefore, the focus was on topics related to the illegal tapping and bypass of gas. Topics most of interest to the group therefore were Topic 8 (illegal activity), Topic 7 (meter problems), and Topic 9 (failed integrity test) and Topic 11 (gas leaks unknown reasons). Table 2 below shows the final list of key words from the topics.

A program in Python was written to mine the notes for these high risk key words. An incident is determined to be high risk if it has one of the key words. A "high risk ranking" was created based on a count of the key words and the incorporation of DOB building data. The DOB data included previous complaints and stop work orders. An incident is ranked high if it has more than one key word, and/or a complaint, and very high if it has a key word and a stop work order. The list of high risk sites are sent to the plumbing unit on a daily basis and at the beginning of the month, and they are added to the building inspection route sheet.

Table 2: Risk Analytics Criteria

Illegal or illegal bypass
Illegal piping
Illegal piping bypass
Flex or flex hose
Broken meter
Meter bypass
Defective meter
Meter disconnected

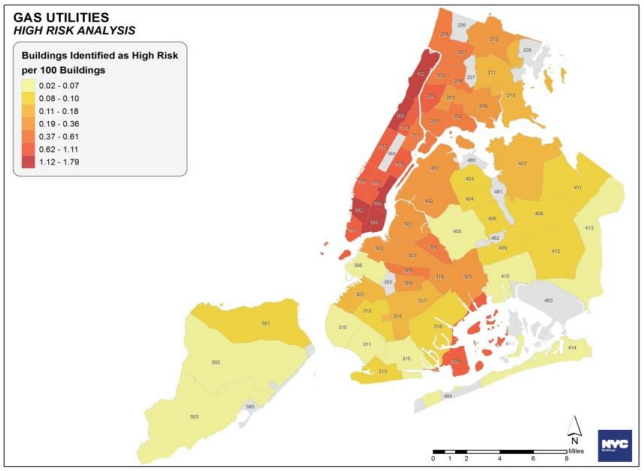
4.2 IMPACT

From August 2016 through June 2017, the DOB received 13,217 notifications of gas incidents from the utility companies. From this sample, 1562 resulted in high risk notifications to the DOB. The text mining added and additional 391 to the high risk classification (an additional 25%). Table 3 shows the breakdown of enforcement actions comparing the high risk call-in notifications and the notifications deemed high risk using text mining. The text mining shows similar violation results to the call-ins, validating the use of text mining as an effective method to classify risk. Figure 4 shows a map of all of the high risk incidents (call-ins and text mining) per 100 buildings by community board. The highest density of high risk buildings are located Manhattan.

Table 3: Comparison of Violation Rates

Enforcement Action	Call-in Notification	Text Mining Classification
DOB Violation	24.7%	22.5%
No Violation Warranted	39.3%	52.9%
Unable to Gain Access	35.9%	24.6%

Figure 4: Map of High Risk Incidents



5.CONCLUSION

This case study shows that text analysis can be used as an effective tool to triage the evaluation of risk. Topic modelling was used in the exploratory phase of the analysis, and then coupled with a text mining procedure to develop a risk tool. This process shows that the analysis of text adds value to a working relationship with a goal to improve city services.

6. REFERENCES

Blei, D. M. (2012). Probabilistic Topic Models. Communications of the ACM, 77-84.

DiMaggio, P. M. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of the U.S. government arts funding. Poetics, 41, 570-606.

Editorial. (2013). Introduction - Topiic Models: What they are and why they matter. Poetics, 41, 545-569.

McFarland, D. D. (2013). Differentiating language usage through topic models. Poetics, 41, 607-625.

Reich, J. B. (2016). The Civic Mission of MOOCs: Measuring Engagement across Political Differences in Forums. Proceedings of the Third (2016) ACM Conference on Learning at Scale, (pp. 1-10).

Roberts, M. B. (2016). A topic model for experimentation in the social sciences. Journal of the American Satistical Association, 111, 988-1003.

Tangherlini, T. P. (2013). Trawling in the Sea of the Great Unread: Sub-corpus topic modelling and Humanities Research. Poetics, 41, 725-749.