Constructing a Text Mining Database for

*Energy Disaster* Business Intelligence

by Tai Nguyen, M.S. Candidate – Data Science

Capstone Project for *Verisk Analytics*

GalvanizeU, University of New Haven

44 Tehama St. San Francisco, CA

email: ngtantai3@gmail.com

**Abstract— From household heating to oilfield disasters, gas explosions are a fatal reality. Insurance underwriters work to mitigate that risk. Verisk Analytics seeks to gain a deeper understanding by compiling an informative database of historical pipeline disaster and gas explosion incidents as well as other energy disaster records reported from online sources. This paper provides an overview for constructing a database infrastructure applied toward analyzing and mitigating unplanned disasters and accidents. A process for constructing Business Impact Analysis (BIA) is presented in this paper to create a feature database for risk analysis and other business intellegence.**

**During the database construction, Beautiful Soup, PDFMiner, Pandas, and Numpy and other python language tools extracted and processed a variety of online resources having inconsistent file formats. Specifically, such construction successfully demonstrates processes for data extraction, architecture for pre-structuring data, and concatenation from various file formats, such as PDF, HTML, XML, and MS Excel. The varied and complex process of online resource extraction further includes such techniques as context-based pattern matching and programmatic website data extraction. This interactive and creative process yields new business insights and value from the data mined from this text mining database for energy disaster business intelligence.**

1. **INTRODUCTION**

Since 2009, the U.S. remains the leading producer of natural gas in the world.1 During the daily course of business activities, energy exploration and production companies are acutely aware of the effects of unplanned gas and pipeline explosions and, therefore, must build safety precautions and budgetary plans for responding to and recovering from such grave events with minimal human, business, and environmental harm. As part of the process of proactively building a text mining database for constructing Business Impact Analyses (BIA) and Disaster Recovery Plans (DRP) for these companies, Verisk seeks to gain an informed, updated understanding of such tragic events to ensure further risks and harms are mitigated for its business customers in the energy exploration and production industry.

DRPs and BIAs are both business intelligence indicators for ensuring that a company provides an optimal plan for addressing an energy exploration and production disaster incident and that any resulting business disruptions are mitigated or even possibly eliminated. To identify critical resources for implementation, a DRP seeks to analytically explore risk and business impact factors of disruptive events. Specifically, as a component of recovery strategies, a DRP provides a general role that addresses the scope of responsive activities, relevant manufacturer documentation, triaging the most serious threats and vulnerabilities, critical asset allocation, disaster recovery plan strategies, emergency response teams, as well as reviewing case history of similar disasters. DRPs can differ considerably with various incident scenarios.

To minimize risk of operational disruption from a pipeline disaster, a BIA provides recommendations for disaster recovery and applies standardized strategies developed from a database of studied approaches that are responsive to vulnerabilities that can lead to potential losses. A BIA requires gathering information that includes metrics providing for a detailed description of disaster incidents. This collection of information for a BIA further includes a mandatory list of date and locations of incidence, technology resources, and repercussions imparted to victims, institutional as well as technology resources.

With a mission devoted toward providing value-added data that assists the insurance industry, Verisk provides critical information used in BIA reporting worldwide. Verisks efforts include building and maintaining an informative database of historical pipeline disaster and gas explosion incidents for this industry sector. The data sources are chosen from a distributed diverse domain of disaster records reported from disparate online sources. Arising from these past disaster events, the recovered raw data is then cleaned, filtered, and structured in the format manner that is comprehensive and ideally suited for data exploration to gain insightful business intelligence and, possibly, predictive modeling in the insurance field.

In a role that provides supporting expertise for the insurance underwriting industry, Verisk creatively implements solutions with proven performance worldwide. Through templates and team effort, Verisk takes proactive action to ensure factors for continued business operations, reducing expense ratios, protecting capital solvency, improving prediction of loss ratios, enhancing growth and profitability as well as financial strength ratings are incorporated with each database created for business intellegence.

1. **OVERVIEW**
2. Hypothesis

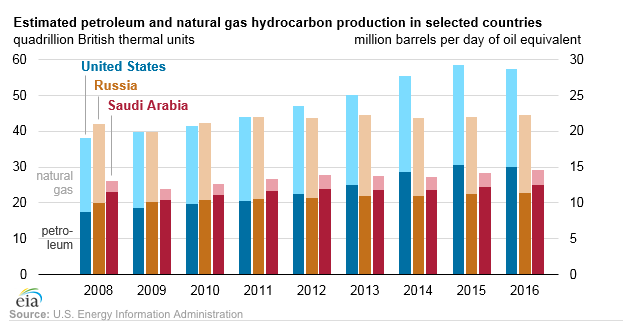
Value from business intelligence is gained by building a database from the right information that is current and accurate for the required industry sector. To ensure the accuracy of information for use with BIA and DRP construction, it is believed that a greater amount of incoming data will be required from a variety of online sources. Furthermore, to ensure uniform compatibility of data from sources with varying formats, it is believed that new, creative data pre-processing and structuring techniques will be required for gaining uniform compatibility while being introduced to a text mining database.

B. Theory of the solution

Before any information is harvested, it is believed that developing an initial workflow process of prescreening online source data websites will be required. This prescreening should be applied early-on in the workflow and should include metrics to evaluate business value and consistency of the data as well as for ease of access with avoidance of online sources having secured data. This process should tag or “weight” useful sites for future access as well as list sites with no apparent data value. Development of a dataset in line with company templates will improve workflow. Overall, this process will need to be implemented on a continuous basis to ensure the timely accuracy and value of the data to Verisk’s energy sector customers.

C. Prior Literature

Due to improved developments in oilfield hydraulic fracturing, the USA since 2011 has surpassed Saudi Arabia and Russia to remain the largest producer of natural gas in the world.8



As a source of energy, natural gas is used domestically for heating and cooking within the home and, when liquefied, is used as an alternative fuel to power many city busses and semi-trucks. In its gaseous state, the mostly methane natural gas readily dissipates outdoors and typically poses little risk for explosion.

The danger comes from occasional leaks of gas that can collect in significant quantities to cause an explosion when accidentally introduced to an ignition source.9 This project builds a database that can serve as a source for business intelligence for underwriters and for other insurance providers servicing the energy exploration and production industry. By its very nature the energy exploration and production industry is a high risk – high reward industry, where large and middle-level operators compete on nearly equal footing. Businesswise, it is advantageous for insurance providers to expand to middle-market operators interested in business intelligence provided by BIAs and DRs that are derived from energy sector databases.

Verisk provides the largest statistical database in the insurance industry, namely ISO.10 The ISO database and related data management service is used for evaluating current markets in many industry sectors to make better underwriting decisions, especially to the energy exploration and production industry in the form of DRs and BIAs.

Illustratively, for BIAs, a variety of questions need to be asked while building a text mining database for the energy exploration and production industry. First, one must be sure to initially ask what needs to be addressed if the disaster impact precludes a business function from no longer being performed.11 Second, ensure that business services continue despite an adverse disaster incident, one must ask what needs to be done before a disaster and what can be deferred until after a disaster occurs.12 Moreover, if the effected business is linked to multiple dependencies, it is also important to consider the impact within the wider network of other business functions while constructing a BIA or other responsive business intelligence plan.13

Further, there is literature for developing insurance databases to serve middle-level business markets with predictive models.14 As there are many middle-level energy exploration and production operators in the industry, it is advantageous to consider building a database with some consideration of future predictive modeling.

Specifically, to provide a solid foundation of future predictive model development for the insurance industry, at least four (4) steps should be observed: 1) generate categories of variables, 2) exploratory data analysis or “EDA”, 3) receive and transform the sourced variables, and 4) pre-partitioning the model set for future model build.15 Further, considerations would include selecting categories of variables that correspond to industry standards such as federal standards for emergency management and business continuity as a result of a hazards disaster.16

To build the database for this project, all of the above considerations from the literature were taken into account.

1. **IMPLEMENTATION**

A. Choice of tools

As this project primarily involves text mining from online sources, the python language offers the flexibility to apply many different applications from webscraping to pre-structuring the data.

Numpy is a library for numerical computation with given scripted python instructions and provides for the creation of array data structures.2 Another open source library, the Pandas package is built on the numpy library and is ideal for EDA as well as provides a variety of data structures and tools for exploring data.3

For webscraping, the Beautiful Soup python library4 is an open source tool for pulling data out of HTML and XML files. Along with the EDA tools and data frames from the pandas library, Beautiful Soup is an ideal and flexible tool for retrieving and subsequently structuring online data. Beautiful soup is especially valuable when retrieving HTML tables from online sources as this library ensures minimal encoding or metadata errors during the transfer, and thus leaving the table as true as possible to the originally presented online table.

PDFMiner is widely popular online tool for extracting alphanumeric text from documents having a portable document format (PDF) file format. PDF documents are actually image files where optical character recognition (OCR) techniques are programmatically applied to convert at least some elements embedded within the image to an alphanumeric text string. Although open source software, PDFMiner conversions are not necessarily true to the text shown on the image original document, as the resulting initially converted text string is often a confused combination of alphanumeric text and unintended encoding errors also known as “Mojibake”.5

The python Re package provides for the use of regular expressions that are used to eliminate the Mojibake so as to clean the text to a desirable and recognizable alphanumeric form true to natural language.6 Regular Expressions apply pattern matching techniques to identify and eliminate the patterns of Mojibake intertwined with the generated natural language text sequence.

Capturing data from a variety of online sources further leads to complications that require harmonization of file types provided by these various online sources into a single text string for pre-processing purposes. A variety of online file types are retrieved from a variety of online sources including PDF, HTML, XML, MS Excel and others. The python Glob library assists with file concatenation to return a python list as well as provides an iterator for processing big batches of incoming files.7

B. Choice of data

For pipeline and natural gas disasters in the energy sector, Verisk is interested in collecting information that can form the foundation of standardized incident briefings for the insurance industry. Information for constructing incident reports can be from various online sources and may include for each incident: date of the incident, country, state, name of owner, name of affected business installation, factual details of the incident, cause of loss as well as details of the loss.

Furthermore, Verisk is interested in constructing historical timelines for each disaster incident for the energy exploration and production industry. Each timeline includes but is not limited to the most recent data retrieved online as it is important to quickly gain a comprehensive and accurate accounting of the event as possible.

For data retrieval from a variety of websites, an initial workflow process of screening online data for the business value and consistency of the relevant data is created. This process requires assigning a tag or “weight” to useful websites early-on in the workflow for future access as well as lists those sites with no apparent business value assigned to the data. Each of the weighted sites contains about three to ten thousand (3,000 – 10,000) data points with a variety of encoded file formats that will be pre-processed to form business intelligence products ranging from incident briefings to BIAs.

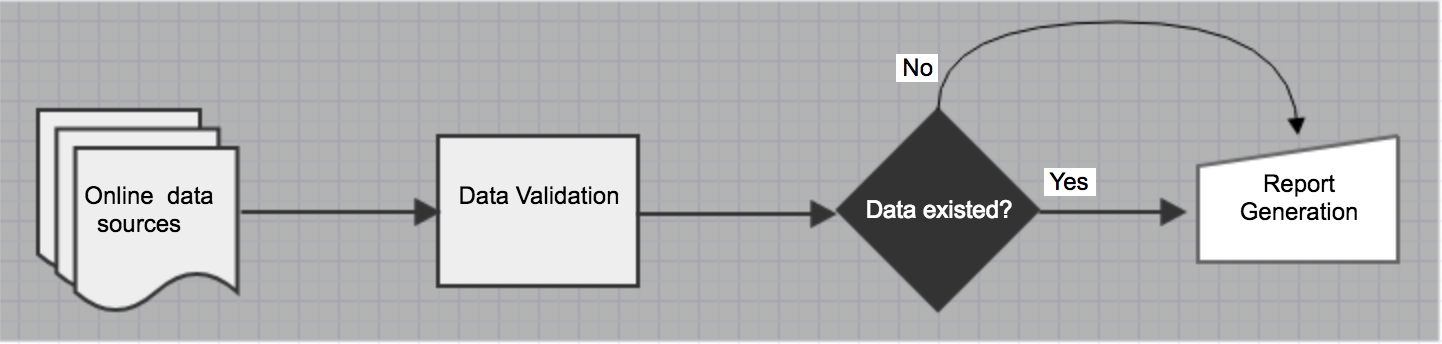
Due to confidentiality, this project cannot provide detailed information regarding this pre-screening process of data sources. Overall, the resulting process developed will need to be implemented on a continuous basis to ensure the timely accuracy and value of the data to Verisk’s energy sector customers.

**IV. CONSTRUCTING A TEXT MINING DATABASE FOR BUSINESS INTELLIGENCE**

The creation of a database for information pertaining to the energy exploration and production industry is summarized in five (5) steps: 1) The prescreening of data sources, 2) data extraction from online data source, 3) context-based pattern identification, 4) pre-structured text mining and 5) polishing. As Verisk demands the most current, relevant data as possible, it is further understood that that these five steps are repeatable in a continuous loop.

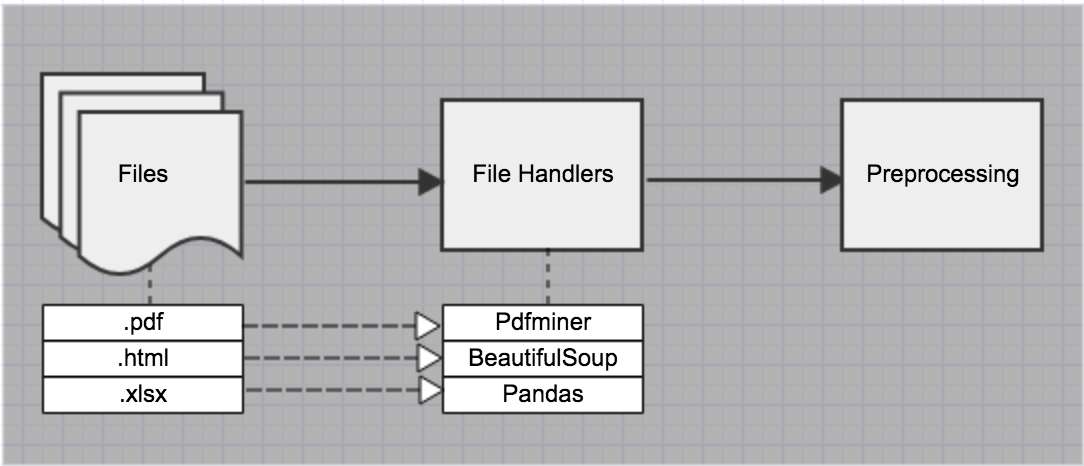
STEP 1 – DATA SOURCE PRESCREENING:

As opposed to a programmatic process, this step is an iterative manual process involving subject matter experts relating to the field of the E&P industry. As a general approach for examining each potential source of data consider the following aspects. When reviewing the potential source, determine if the subject matter or “topic” is relevant and desired for adding to the newly generating database. It is important that not only one understands the subject matter but whether such information is of potential business value to the industry sector, such as the E&P industry. Next aspect is to determine whether the source provides sufficient enough data for programmatic extraction. One further aspect to be considered is whether data can be accessed with little or no restrictions. Further, a tag or “weight” is assigned to the most useful sites for future access as well as list sites with no apparent data value.



STEP 2 – DATA EXTRACTION:

Extracting data from any given source must be done by first choosing the correct tool. If the information is formatted in PDF file format, the open source PDFMiner tool will be applied. If the source information is in XLSX format, then the Pandas package can be used to extract. Lastly, if the information is formatted in HTML, then Beautiful Soup is the ideal tool for this task. With these tools, all retrieved data is converted from its native format and appended to a corresponding python text string and encoded in UTF-8.



STEP 3 – CONTEXT BASED PATTERN SEARCHING:

This step is quite interesting as it involves creatively providing a solution as a regular expression or a series of regular expressions to the proposed problem. The problem will always vary depending on the specific information that must be extracted from the data retrieved from the source. As regular expressions principally apply specific instructions for matching exact character or word patterns with text data, it often becomes a puzzle game to see what combination of regular expressions will be needed to ultimately retrieve the desired source of information. Given this perspective, the context of the data is critical as one will need to find a single word, number or symbol that will serve as an “anchor” by which to extract nearby words that include the desired information. It is possible that many creative iterations of this puzzle game – like process may need to be applied so as to eventually get the desired information.

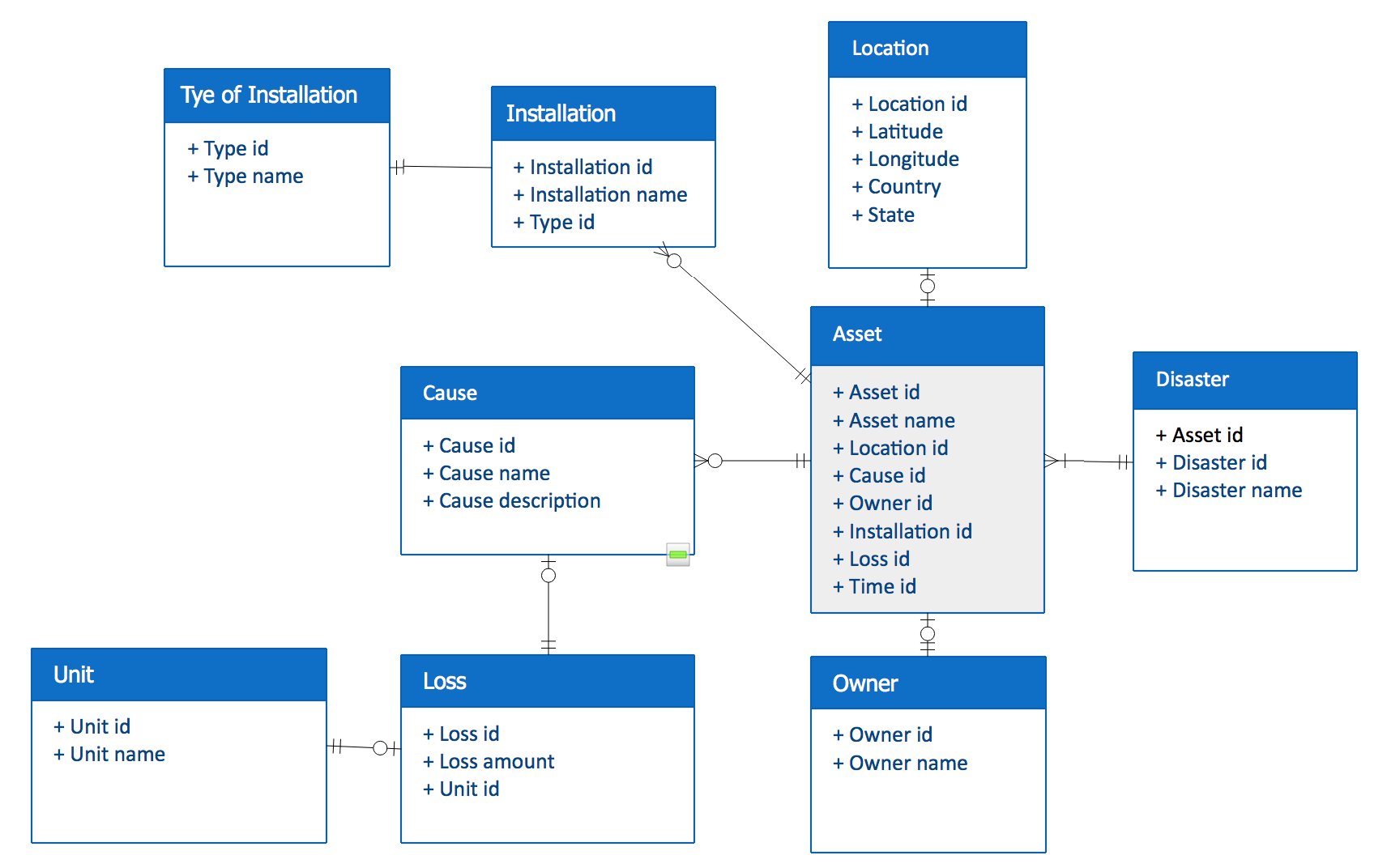
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Applied Regex Spectrum** | | | | | |
| Characters I | **\d** | **\w** |  |  |  |
| Characters II | **.** | **\.** | **\** |  |  |
| Quantifiers I | **+** | **{}** | **\*** | **?** |  |
| Character Classes | **[ … ]** | **[^x]** | **[^x-y]** | **-** | **[\d\D]** |
| Logic | **|** | **( … )** | **(?: …)** |  |  |
| Lookarounds | **(?= ...)** | **(?<=…)** | **(?: …)** |  |  |
| Anchors | **^** | **$** |  |  |  |
| Inline Modifiers | **(?i)** | **(?s)** | **(?d)** |  |  |

STEP 4 – PRESTRUCTUED TEXT MINING

This step considers structuring the data in a variety of categories or metrics. The manner by which the data is finally structured is out of scope of this project, however some consideration of structure at this time is necessary to prepare the data for practical use by Verisk, hence this step is called “Prestructured” Structured Text Mining.

Initially the required information is exactly and succinctly defined as a category label, also called a metric or variable. At times Verisk provides a specific metric name but other times a name must be chosen *ad hoc* as an implicit description of the desired data. Next, given the defined column name, the desired information is extracted from the appended text string generated in Step 2. This text mining extraction is accomplished by the regular expressions derived in Step 3. It should be added, that to improve workflow, templates were externally provided on occasion and either included a defined column name or implicitly characterized a column name.

Importantly, for each category label, a list string is created. The list includes information values provided from every extracted data source. This information is typically in either a string or a numeric data type. Thereafter, a list comprehension is created that comprises a 2D vector of a list of category names and a list of corresponding information values, i.e. desired natural language text. Overall, this process will need to be implemented on a continuous basis to ensure the timely accuracy and value of the data to Verisk’s energy sector customers.



STEP 5 – CLEANING THE NATURAL LANGUAGE TEXT :

Cleaning refers to cleaning the text and removing extraneous information from the desired natural language text. It also includes reformatting the information to add uniformity for each column. Cleaning further includes formatting the list comprehension to a desired file type for external export.

* 1. **CONCLUSION AND FUTURE RESEARCH**

1. CONCLUSION

This project shows a successful workflow process for screening online source data to build a database for the energy exploration and production industry. Each source is pre-evaluated for relevant business value, quantity, and consistency. This pre-evaluation incorporated an iterative manual process for determining whether the subject matter or “topic” is relevant and desired for adding to the newly generated database. A weight is also assigned to the most useful sites for future access as well as lists websites with no apparent data value.

Context based pattern searching and prestructured text mining techniques are applied to generate a dataset that is of relevant business value to the target industry and inline with company templates to improve workflow. To ensure the timely accuracy and value of the data to Verisk’s energy sector customers, this process will need to be repeated on a continuous basis.

B. FUTURE RESEARCH

As stated, the literature definitely points out that such databases are the ideal resource for predictive modeling relating to insurance underwriting and other related business intelligence. Future experimentation to determine what machine learning model would be best suited for the insurance industry would clearly be the next step.

Specifically, subject matter topics would be predicted for a dataset created from the above constructed database based on Verisk’s needs for the energy exploration and production industry. Next, a machine learning clustering algorithm can be applied to the dataset to predict the most important subject matter topics.

REFERENCES

1 . <https://www.eia.gov/todayinenergy/detail.php?id=31532>

2. <http://www.numpy.org/>

3. <http://pandas.pydata.org/about.html>

4. <https://www.crummy.com/software/BeautifulSoup/>

5. <https://en.wikipedia.org/wiki/Mojibake>

6. <https://pymotw.com/2/re/>

7. <https://docs.python.org/2/library/glob.html>

8. <https://www.eia.gov/todayinenergy/detail.php?id=31532>

9. <https://en.wikipedia.org/wiki/Natural_gas>

10. <https://www.verisk.com/insurance/capabilities/actuarial/data-and-statistical-services/>

11. <http://searchdisasterrecovery.techtarget.com/feature/A-business-impact-analysis-checklist-10-common-BIA-mistakes>

12. Id.

13. [www.preparednessllc.com/assets/preparedness\_bulletin\_business\_impact\_analysis.pdf](http://www.preparednessllc.com/assets/preparedness_bulletin_business_impact_analysis.pdf)

14. <https://www.soa.org/Files/Research/Projects/research-life-auto-underwriting.pdf> FIG 7

15. <https://www.soa.org/files/pdf/research-pred-mod-life-batty.pdf>

16. <http://www.nfpa.org/codes-and-standards/all-codes-and-standards/list-of-codes-and-standards/detail?code=1600>