Natural Language Processing of Text in Legal Contracts for Structured Database Construction

Capstone Project for *Orange Silicon Valley*

Rafael V. Baca, Esquire
M.S. Data Science: GalvanizeU, University of New Haven
44 Tehama St. San Francisco, CA
rafabaca.software.esq@gmail.com

Abstract— **While migrating paper document files to cloud-based databases, many businesses seek new methodologies to make their digitized files easily retrievable and searchable. Further complications can arise from specialized documentation that cannot be easily accessed via search due to arcane terminology beyond everyday language, such as legislative documents and business appraisals. Specifically, unlike either numbers or words in every day language that can be discreetly parsed while creating a searchable database, legal natural language associated with business contracts remains infamously problematical, as legal terminology is often quite vague and varied in accordance with each lawyer's individual choice. By compiling an informative database of legacy business contracts scanned in Portable Document Format ("PDF") Orange Silicon Valley seeks to gain a deeper understanding of legal natural language processing ("NLP"). The Orange contract database will include legal phrases from different national legal systems, languages, linguistic character symbols, and with different document formatting (- such that semantic as well as syntactic structuring of legal phrases greatly varies). This paper provides an overview of one process for structuring legal natural language databases for legal and business intelligence.**

During the database structuring, PDFMiner, the Stanford NER Tagger, Beautiful Soup, Pandas, Numpy, and other python natural language tools processed text extracted from contracts with pdf formatting. In particular, the insights gained from structuring legal contracts successfully demonstrate processes for data extraction, architecture for pre-structuring data, and processing legal terminology where meaning is not often exacting. In a greater effort to harmonize information within an Orange legal database, this process of data structuring further includes matching specific legal contracts with standardized international business practices by gathering corporate data profiles from online resources through webscraping for data extraction. To foster new legal and business insights, the process for structuring this database involves gaining great familiarity with the legal and business subject matter of these corporate business contracts to create innovative new database architectures for legal nlp.

I. Introduction

Orange Silicon Valley is the R&D division of the recently rebranded French national telecom having a business presence in over 29 countries serving 265 million customers. Orange Silicon Valley is interested in enhanced harmonization of its business contracts. In particular, a corpus dataset for this capstone project is over 500+ business contracts executed by Orange, each averaging between 50-100 pages beginning with the 1990s to the present.

This project intends to aggregate a database of representative business contracts that are predominantly from paper contracts that were scanned to ".pdf" file format. Processes for a structured database creation, aspects for determining commonly shared legal phrases, requisite elements for creating a legally binding contract, extracting text from various contract formats, extracting a wide variety of text [illegible] legal speech, aspects for pre-structuring a database for future use are addressed in this project.

As a component of this structuring project, hand-selected essential elements of a contract that can be most widely shared across all global business units of Orange will be identified. The legal nlp database is further designed for future experiments during an Orange internship to identify the most common, impactful, and legally binding phrases in the corpus to generate a legal service workflow for efficiently executing internationally robust contracts for company-wide use.

Due to reasons of confidentiality, this project cannot provide detailed information regarding the source data used to generate a database with structured legal natural language. The workflow of this project was thus divided in two components. First, a pilot database is constructed and structured on a local computer using a toy dataset comprised of publically available online business contracts. Given the successful code and workflow outline from the localized first component, the second component consists of repeating the database structuring process by configuring an identical environment while on the firewalled, corporate virtual private network in the time window allotted by the project.

II. Overview

A. Hypothesis

Harvesting key strategic business intelligence from a corpus of internal business contracts for the company's legal and business units will ultimately enhance workflow in future contract construction within the company. The methodologies for structuring digitized legal contracts to make searchable and readily retrievable documents in the legal field are currently being developed - such that further complications with the derived processes may arise due to a lack of standard procedures and nlp tools specific to the legal sector.
The arcane language used in legal documents makes nlp parsing difficult given present limitations of natural language processing tools. This project intends to provide an overview of one process for structuring a legal database infrastructure for harmonized legal and business intelligence.

B. Theory of the solution

In the pursuit of constructing a structured natural language database of business contracts, specific processes for data extraction, architecture for pre-structuring data, and processing legal terminology will be developed in an area of natural language processing where semantic meaning associated with legal documentation is not often exacting by design.

A workflow process for successfully extracting text data from each legal contract will be constructed and applied. The workflow will serve as a pipeline for future extraction of legal text to create a structured database for internal legal and business intelligence.

C. Prior Literature

There are only a handful of references that specifically address the development of practical tools for drafting business contracts and of the infamous ambiguity associated with legal

the legal sector, such as E-law, are not publicly available but are currently being developed to possibly include information retrieval with legal context awareness in the future.[2]

Illustratively, in the pursuit of developing a corpus of 30 Australian legal contracts, the tools of python and nltk hand-coded tagging were used to create features in a machine learning binary statistical classification experiment determining whether sentences can be predictably recognized as part of legal clauses of a contract.[3] To limited success, a support vector machine ("SVM") model was developed using *Weka* data mining software to carry out machine learning.[4] The ultimate goal of the Australian experiment is to recognize as many ambiguities as possible associated with the language of legal contracts and to form an understanding for future development of practical tools to assist in contract drafting.[5]

Other studies applied similar goals of studying the structure of a law sentence in terms of other implied equivalences as well as specifically applied word clustering algorithms based on uniform information from bigrams.[6] Interestingly, with a corpus of 500,000 legislative bills for enacting into law (- as distinguished from a corpus of legal contracts), similar research on legal legislation sought to extract a uniquely canonical version of shared text between clustered documents to synthesize a baseline legislative bill despite heavy use of obtuse legal language throughout.[7] Specifically, as similar to business contracts, model (legal) legislation is often unintelligible to the public in both source and content such that a large corpus in this academic literature study identifies clusters of state legislation from which prototype sentences that summarize the similarity and variation of text represents each corresponding type of clustered legislative bill.[8]

From the academic literature, a general intuition is gained as to the inherently difficult nature of performing meaningful NLP processes on legalese - although the prior literature does positively chart a course for machine learning classification and clustering of legal clauses to some successes. In particular, each of these studies encountered difficulty with the meaning of legal language but used some available open source tools to limited success, such as nltk tagging and standard machine learning models for clustering and classification.

III. Implementation

A CRISP-DM Synthesization: The approach for this project follows the CRISP-DM methodology, as an industry standard process for data mining with a focus on directly meeting the business needs of the stakeholder. The following sections are the steps provided by the CRISP-DM process and form the framework for the entire capstone project.

A. Business Understanding

Having a multinational footprint, Orange is continuously interested in harmonizing its business activities in the most efficient way. In its commitment to providing quality products and services, Orange legal department is interested in automating its workflow processes using the latest Data Science processes for analyzing its corpus of contracts to derive an experimental baseline with this project. Specifically, comprehensive list of corporate participants from each Orange business contract in the corpus and of their respective fields of industry. A second phase, out of scope of this particular project, a baseline of commonly shared and legally required legal clauses will be identified from the corpus of business contracts. Further techniques in engineering for ambiguous and inconsistent legal context in the current literature will be reviewed in depth at the second phase to construct a successful and reproducible model for legal nlp of business contracts. Orange will use these derived baselines as tools for future activities across the globe. Moreover, the harmonized baselines will serve to increase internal worker productivity and reduce cost.

B. Data Understanding - Choice of data

In a first phase, an initial set of business contracts were hand-selected by a committee of Orange stakeholders and the author of this project for general review to identify commonly shared legal meaning and preferred structure to serve as a (minimum) baseline of preferred clauses that define a legally enforceable contract. Each selected contract from the initial set ranges between 15 to nearly 100 pages in length and applied in industry to various international business scenarios by Orange in the past.

In a second phase, in the final weeks of this current project, the bulk of experimental data was provided having a corpus of 500+ Orange business contracts with each predominantly in English but including many non-English business characters symbols arising from activities with international business partners. Because of the highly confidential nature and trade secrets of this second-phase, substantive data for the current project could only be accessed at while at Orange during limited weekly business hours. Therefore, to continue project development, a toy dataset of publicly available online "master service agreement" business contracts from similar telecom corporations was created to continue work on the project off-site.

Preliminary Understanding of Data:

This paper's author, a "subject matter expert" providing professional experience as a U.S. lawyer, first compared each contract from the initial set of hand-selected set of business contracts against a Master Service Agreement template further provided by Orange (hereinafter "template"). From a thorough comparison of legal affect and identification of nlp aspects for optimal data science statistical modeling, a preliminary understanding of a business contract baseline is then gained. Some of the comparative factors applied during the review included: identifying legal meaning having robust legal impact (from a global perspective); manually determining what legal elements were shared by all contracts; manually determining what elements in some contracts could possibly improve all contracts by eventual insertion back in to the template, and determining how or if the elements should play a meaningful role in statistical modeling.

In effect, an underlying baseline for the entire project was manually developed from this comparison against the template provided by the stakeholders to gain a preliminary understanding for construction of this project's workflow. In particular, each provision or contractual "clause" of the

entire project so as to comparatively measure all other contracts within the corpus to thus eventually form a structured database for binary classification of legal clauses from other contracts as either corresponding to the hand-derived legal nlp baseline or not.

Accordingly, the hand-derived baseline labels for the comparison with legal clauses from other contracts within the corpus include the following (potentially) supervised labels: (1) *Parties;* (2) *Recitals;* (3) *Performance;* (4) *Compensation*; (5) *Intellectual Property;* (6) *Confidential Information;* (7) *Terms (contract duration) and Termination;* (8) *Warranties;* (9) *Superseding Contract, Authority & Jurisdiction/ Governance;* (10) *Signature Blocks; and* (11) *Effective Date.* These clause labels derived from the manual comparison have a dual effect in that each label collectively acts to satisfy the minimum requirements that define a legally binding contract and further include the minimum standard clauses commonly used by Orange while conducting its business operations internationally.

Currently, lawyers as a matter of real-world business practice often apply a time-consuming *ad hoc* approach to each contract that can differs from the Master Service Agreement template which serves as the underlying legal nlp metric for this project. The cumulative effect of many *ad hoc* responses in contractual construction is inefficient and leads to greater costs for the legal department and business units alike while gaining a harmonized understanding of these business documents. The proposed experimental results allows a practicing lawyer to know what basic provisions of a contract should be and what industries are most at risk to deviate from the master service agreement template to addressed them accordingly with greater efficiency and at reduced costs. In practice, a lawyer would be identifying the risk of certain provisions and particular industries that consistently deviate from the norm early-on and adjust their subsequent workflow accordingly.

Notably, there are further observations gained from this manual exercise of preliminary understanding: (1) comparison of other contract clauses against the template (master service agreement) was made using overall semantic meaning and not strictly with a specific legal term or legalese terminology (- for the following reason): one contract may establish a legal element in a few words or legal term whereas another contract may take a few pages and address other scenarios under that particular notion of subject matter and, moreover, a strict legal term or phrase can have different meanings or actions within their surrounding local context (for example, the word: "terms"); and (2) the number of contracts in the finally received and confidential corpus of contracts for this project will remain fixed (although a small toy dataset of varying size of pdf contract documents was developed of similar online contracts to initially test the open source tools) (- for the following reason): a corpus of 500+ should be a satisfactory number without expending further company resources.

Process Workflow Understanding:

Process:

The revised, present goal for this project is to enable contracts for this project. In the second phase (outside the scope of this project) the totality of all other hand-derived baseline labels for comparison with legal contract clauses within the corpus will be studied for business intelligence insights.

The process for structuring the legal text extracted with natural language processing is achieved in two primary phases: 1) using the phase one data and 2) subsequently using phase two data (i.e. the bulk corpus of 500+ contracts discussed above). For each phase of the workflow, data will be mined in that text data will be extracted from the corpus of pdf contracts to create a single structured column of data for each of the eleven (11) derived baseline labels.

For the scope of this project, a minimum viable product consisting of a single baseline label – "parties" – will be structured in to a column. As the stakeholders have expressed a preliminary interest to identify its partner companies to the Orange contracts with their respective industries, text data for the "parties" label will be a first column of the eleven (11) clause labels to be created for this current phase one of this structuring process. In a future project, data from the corpus will be extracted and inserted into structured data columns for the remaining ten (10) of eleven (11) clause labels derived from the preliminary understanding in phase one.

Once the corporate parties for each contract from Orange's corpus of contracts is identified through nlp text mining extraction, determining the corresponding industry associated for that identified corporate partner triggers a separate process in this current project that requires further structural design. In particular, as it is highly likely that each extracted corporate name is a publicly traded company, then the industry characterizing each selected publicly traded company that is partner to an Orange contract can be obtained through a standardized industry classification used for business, such as the Standard Industrial Classification (SIC) system. An online source that pairs a named company with an SIC classification could be extracted through a Dunn & Bradstreet (D&B) pay API or UK Government's Companies House free API among other resources. These APIs should provide a SIC classification number and corresponding industry description for each identified Orange contract partner.

As a matter of standard business practice, each corporate entity registered with D&B or the UK Companies House is assigned an SIC number as a means for classifying the nature of goods and services that each company provides within a particular category of industry. As the original international system for classification, most other industry classifications are backward compatible to the SIC, such as the North American Industry Classification System (NAICS) currently used by the United States and the United Nation's International Standard Industrial Classification (ISIC) system. Moreover, for this current project, the process of webscraping the SIC values from the UK Companies House is advantageous as it is publicly available (i.e. no internal confidentiality concerns) and free of charge.

Assumptions:

(1). For a written business contract to be in legal effect for at least common law countries (and civil law (European) countries) assume the following elements are necessary: offer, acceptance, mutual assent (with authority), consideration and (commonly mutual -) performance. Assume these legal ingredients are embedded within the various written clause labels, topics, "clauses" or "provisions" of a contract. Assume in business practice there never is really a standard boilerplate template for what, where, and to the extent provisions are placed within a written contract. (– giving rise to the adage "always read your contract carefully".) Therefore, assume each of the eleven (11) clause labels derived from the preliminary understanding in phase one above satisfies the legal requirements for a valid business contract.

(2). Assume that the master service agreement provided by Orange is the template contract and baseline to judge other business contracts in the corpus. In other words, the master service agreement is the norm by which to baseline from. In terms of metrics, deviation from the master agreement would be some level of quantifiable risk from a legal and business standpoint.

Specifically, assume that only some hand-selected provisions of the Orange Master Service Agreement to be the actual baseline for this analysis and not the entirety of language within the agreement itself. The key provisions comprising the eleven (11) clause labels derived in phase one of the project represent foundational or "base-line" contractual elements that should ideally be shared by all other contracts in the corpus.

(3). Assume each of the eleven (11) clause labels derived from the preliminary understanding in phase one arises from rigorous semantic legal, business, and natural language processing analyses. From a natural language text processing standpoint, this assumption is key to this project in that there are obvious disparities in language quantity, content, and syntax where a fuzzy analysis is preferred – such as topic modeling for example, given the meaning of each provision is desired over a literal comparison against the Master Service Agreement template.

(4.) The scope of contracts comprising the corpus of contracts are business-to-business and business-to-institutional contracts, as opposed to direct-to-consumer contracts.

C. Data Preparation - Choice of Tools

As this project primarily involves programmatically extracting text from a large corpus of contracts, the python language offers the flexibility to apply many different applications from webscraping to pre-structuring the data.

(1) Data Extraction:

a.) The initial corpus of contracts was provided primarily in Adobe Portable Document Format (PDF) with some in Microsoft (MS) Word ".doc" format. Firstly, all MS Word contracts were converted to ensure the entire corpus is set to PDF format. Furthermore, some non-English language characters that were converted to facilitate full UTF-8 character compatibility.

b.) Numpy is a library for mathematical computation with programmatic instructions provided in python and also serves for the creation of array data structures.[9] Another open source ideal for exploratory data analysis as well as furnishes a variety of data structures for python.[10]

c.) As portable document format, "PDF", is actually not a document file format but rather a graphics file format (i.e. image file), PDFMiner[11] is an open source Optical Character Recognition (OCR) package that helps to selectively extract the text content from a PDF image file and output a character string. The PDFMiner OCR module techniques are programmatically applied to convert at least some elements embedded within the image to an alphanumeric text string. Thereafter, while in a text format (i.e. ".txt"), the derived character strings for each legal contract document in the corpus will be parsed by each word. Although open source software, PDFMiner conversions are not necessarily true to the text shown on the original pdf document image, as the resulting initially converted text string is often a confused combination of alphanumeric text and unintended encoding errors - also known as "*Mojibake*".

Specifically, the PDFMiner package initially proved difficult to run on a python 3.+ environment, so a local shell environment based on python 2.7 was created for the project. Procedurally, all other packages were subsequently migrated to the local 2.7 shell environment and the process was further duplicated on the Orange Corporation's firewalled virtual private network. Thereby, from a toy set of PDF contracts set on a local 2.7 environment, PDF Miner was used to extract text from the image files of publicly available master service agreement contracts found online. The toy set was programmatically loaded through TextBlob for PDFMiner text extraction. Encoding or *Mojibake* errors were encountered and cleaned by programmatically inserting regular expression syntax to the text on the local toy set and detrimentally to the entire corpus of contracts as discussed below.

d.) 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.[12] Regular Expressions apply pattern matching techniques to identify and eliminate the patterns of *Mojibake* encoding errors intertwined with the generated natural language text sequence.

e.) Due to a redirection in procedural preferences by the stakeholders, initially structuring the parties from the entire corpus of contracts is of immediate great business value for initial extraction with the PDFMiner tool. Thereafter, to obtain the particular industry sector of each party, it was assumed that obtaining a Standard Industry Code (SIC) associated with each company name would successfully draw the pairing of company name to industry sector. As a free online source, the UK government offers an SIC code and related industry description for every company that is a legally registered to do business within that country.[13] In the future, Orange Business may optionally consider using its D&B corporate subscription to gain access to a fee-based comprehensive international company registry with richer industry description from the D&B API.

f.) For webscraping, the Beautiful Soup python library[14] is an open source tool for pulling data out of online HTML files. Along with the exploratory data analysis 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

government website as this library ensures minimal encoding or metadata errors during the transfer.

D. Modeling - Choice of Models

Although an aspect of the CRISP-DM approach, the choice of machine learning models is out of scope of this phase of the project -- as this project involves the natural language processing of text in legal contracts for construction of a structured database. With optimism gained from the above discussion of prior literature, there is some prescient for machine learning classification and clustering models with legal text – although to limited, arduous prior successes.

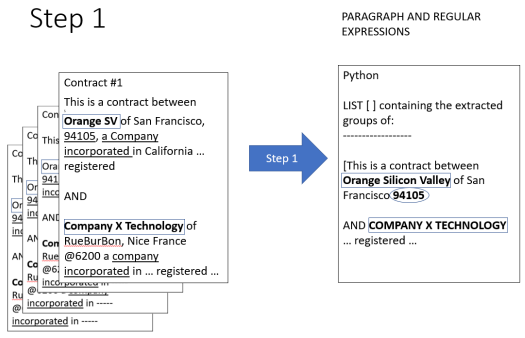
E. Evaluation and Deployment

[To fit the actual factual circumstances and subsequent responsive adaptions implemented for this project, the CRISP-DM elements of Model Results and Choice of Metrics will be substituted with Results from Implemented Procedure and Procedural Assumptions made.]

A major discovery is that various users of the open source tool PDFMiner have reported that PDFMiner alone does not necessarily extract meaningful text from every PDF.[15] From direct observation, about 60% of meaningful text was successfully extracted from the local toy dataset of online contracts. However, with same code exported to the internal, firewalled corporate network only about 10-15% of the contracts exhibited meaningful text for extraction from the 479 successfully converted by PDFMiner from the original corpus of 500+ contracts.

1.) Results from Implemented Procedure

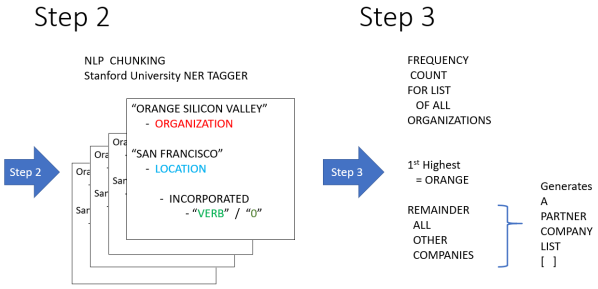
The revised, present goal for this project is extract Orange's business partners for each contract from the corpus of contracts set for this project and determine that partner's corresponding industry description. By successfully executing this procedure, a "proof-of-concept" will be established for programmatically extracting text from Orange's corpus of PDF contracts for the column labeled as "parties". This preliminary understanding acts to confirm that the remaining ten (10) clauses can be extracted using the same procedure derived from the established workflow pipeline for this present project (phase one). Thereafter, with data extracted for all eleven clauses, a structured database will be formed for the corpus of business contracts that will permit future natural language text modeling of legal text. Accordingly, the procedural steps for extraction (applied to both the local (non-confidential) and (confidential) virtual private network environments) are as follows:



STEP 1:

containing relevant, targeted text from each contract that has been successfully processed by PDFMiner. After PDFMiner successfully processes all PDF documents, one might need to initially, manually verify whether the desired text is available for extraction as some converted contracts may be entirely comprised of encoding errors or many contain no received text strings.

Regular expressions implemented with the Re module are proven as quite helpful in extracting only desired text. For example, target words such as "registered", "company incorporated", a US or European zip code of 4-5 successive numerical characters will provide a focal point for the subsequent extraction. Thereafter, in practice, all text is extracted before the focal point to the beginning of the sentence.



STEP 2:

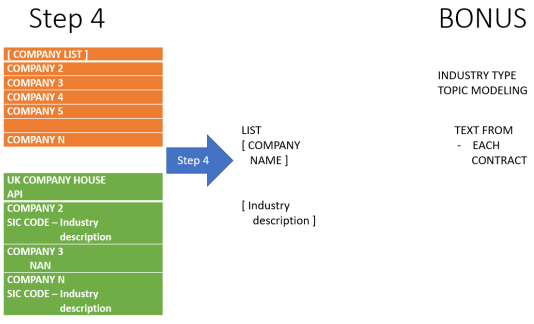
Chunking, also called shallow parsing, is a natural language process for identifying or tagging parts of speech in short phrases. One specific form of chunking, developed by Stanford University, is called the Named Entity Recognition (NER) tagger which beneficially provides named entities tags from noun phrases.

In operation, the extracted blurb paragraphs derived from Step 1 are programmatically subjected to chunking using the optional NER tagger, via the nltk module, to identify "Organizations" for each named company this is a legal party that participates in each contract from the corpus.

In step 2, a resulting python list is generated for the company names that are NER tagged as "organization" for a blurb extracted each contract of step 1.

STEP 3:

A frequency count is programmatically applied to the list of company names from step 2. The highest frequency company in the list, Orange is removed so as to create a list of unique names of partner companies to each of the contracts comprising the Orange corpus.



STEP 4:

Each company name from the list of partner companies in step 3 is matched to their corresponding description provided by the UK Companies House online government database description, if actually listed by UK Companies House. This is programmatically accomplished by inserting each partner company's name into the corresponding html path for the html profile of that company provided by the UK Companies House website.

Thereafter, the SIC number and accompanying industry description is extracted by Beautiful Soup from the html page accessed for each partner company in the list generated in step 3. Notably, as many large multinational companies may have a variety of subsidiary companies, the UK Companies House Profile of a company having the closest word match with the derived partner company from step 4 is deemed a proper match for this particular procedure.

STEP 5 (Optionally):

Each contract with successfully extracted text can further be subjected to topic modeling to gain further insights regarding the corpus. As discussed below, assume that a statistical "topic model" will conduct fuzzy discovery for each provision in each contract in the corpus to match the meaning of a corresponding provision in the Master Agreement.

Using the text feature extraction tf-idf sklearn machine learning tool for topic modeling, a trial run was conducted for contracts in the toy data set. The extracted topics from the output generally conveying a topic for each entire contract as: a master service agreement for technology services.

2) Procedural Assumptions Made

Discrepancies in extracting text from a variety of PDF sources will be explored on a deeper level. There is an initial hypothesis that various encoding errors may arise depending on the manner by which the original document was converted to PDF file format such as by either scanner or by PDF options provided by word processing software, such as MS Word.

Moreover, the documents themselves can be inherently attributable to encoding errors. For example, it was discovered that contract documents with inked hand signatures and redactions made with Sharpie markers were far less likely to have meaningful text extracted from the corresponding PDF file with the PDFMiner tool. Non-Adobe software conversions to PDF file format, such as MS Word, were shown to have consistent encoding errors with the open source PDFMiner tool while on the Orange virtual private network as well.

As a possible solution, while moving move forward in the Orange internship phase of this project, other modules for extracting properly encoded text from PDF image files will be explored before entirely constructing the structured database for natural language processing of legal text. Otherwise, 10-15% of the confidential contracts with successfully extracted text will be used to form the structured database for future predictive modeling and search as a "minimum viable" proof-of-concept.

As there is limited advanced knowledge regarding the natural language processing of legal text, it should be added that initial priorities for this project changed as new aspects were learned during the course of the capstone period and as the final data for the corpus was received later-on in the capstone project timeline. However, it is highly likely that a structured database will be developed and used as described above to at least a limited dataset while on the company's virtual private network.

IV. CONCLUSION AND FUTURE RESEARCH

A. Conclusion

Despite the inherent difficulties with natural language processing of legal text, the project exercise for constructing a structured natural language database of business contracts successfully demonstrates processes for data extraction, architecture for pre-structuring data, and processing legal terminology where meaning is not often exacting. Despite many complex moving factors involved with this project, a workflow process for successfully extracting text data for a single column for the "parties" clause label was successfully constructed. Moreover, a successful means to identify the industry sector from each company name has been demonstrated using the corresponding SIC code description for each company name, a standardized classification method typically used in business and accounting.

In a greater effort to harmonize information within the resulting database, this process of construction further includes matching legal contracts with internationally standardized online resources requiring webscraping for data extraction. To foster new legal and business insights, the process for building this database involved imagining new architectures with great familiarity with the legal subject matter of business contracts.

B. Future Research

Given the above results and some prescient for machine learning classification and clustering models with legal text from current literature, with eleven (11) columns of labeled clauses a structured database will be formed as a baseline for determining whether each contract sufficiently conforms the baseline contract or poses a either a legal or business risk in need of addressing. Although now out of scope for this project, phase two of the project (the Orange internship) will concatenate the eleven (11) derived columns for each of the clause labels to create a structured database for predictive machine learning techniques. In this second phase, the process goal will be to determine how much of a quantifiable business/legal risk does each individual contract from the corpus deviate from the norm (Master Agreement template).

Specifically, it is understood that classification modeling will be applied to the corpus of contracts with respect to the template where each of the eleven (11) clause labels will be set to a binary classifier for determining if each row (i.e. contract) contains the semantic natural language characterized by the corresponding clause label derived from the template (i.e. master service agreement). In totality, binary classification for each of the eleven (11) clause labels will assess whether each contract overall corresponds to the semantic legal meaning defined by the provided template (master service agreement).

II. REFERENCES

[1]. Corpus Based Classification of Text in Australian Contracts by Michael Curtotti and Eric McCreath. In Proceedings of Australasian Language Technology Association Workshop 2010, pages 18−26, at 19 and 25.

[2]. E-law Module Supporting Lawyers in the Process of Knowledge Discovery from Legal Documents, by Kozkowsi *et al*. In Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing (Hissar, Bulgaria, 10–11 September 2015), pages 46–48 at 46. https://aclanthology.info/pdf/W/W15/W15-5308.pdf , *see also* http://opi-lil.github.io/nlp/2015/08/25/E-law-module.html.

[3]. Corpus Based Classification of Text in Australian Contracts Id. at p. 20.

[4]. Id. at p. 21, *see also* https://github.com/fracpete/python-weka-wrapper .

[5]. Id. at 25.

[6]. Supervised and Semi-Supervised Sequence Learning for Recognition of Requisite Part and Effectuation Part in Law Sentences by L Nguyen et al. Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing (Blois (France), July 12-15, 2011) pages 21–29, at 23, 25. https://aclanthology.info/pdf/W/W11/W11-4404.pdf .

[7]. Prototype Synthesis for Model Laws, by M. Burgess et al. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Berlin, Germany, August 7-12, 2016). pages 1579–1588 at 1579 – 81, https://aclanthology.info/pdf/P/P16/P16-1149.pdf .

[8]. Id. at p1579, 1587.

[9]. http://www.numpy.org/

[10]. http://pandas.pydata.org/about.html

[11]. https://pypi.python.org/pypi/pdfminer/

[12]. https://pymotw.com/2/re/

[13]. https://beta.companieshouse.gov.uk/

[14]. https://www.crummy.com/software/BeautifulSoup/

[15]. https://stackoverflow.com/questions/33271509/pdf-to-word-doc-in-python