**I AM OPEN Inc.**

**Technical Report**

**in response to**

**Request for Proposal (RFP24062-19-096)**

**FOR THE PROVISION OF**

**Artificial Intelligence Insights into Regulations**

**FOR**

**The Canada School of Public Service**

**Closing Date for the RFP:**

**September 14th 2018**

**15:00 pm EDT.**

**Executive Summary**

In response to the Canada School of Public Service’s requirement for acquiring artificial intelligence insights into regulations, I AM OPEN Inc. proposes an end-to-end integrated solution framework with a set of text analytic tools. We call this solution framework “Enterprise Text Analysis” (ETA). ETA is designed to enable regulation writers to enhance existing regulations and develop new regulations with insights gained from advanced AI tools leading to greater human machine interactions. We believe that this is the most trustworthy approach to improve regulations that better achieve the government’s objective of supporting innovation, growth and competitiveness while protecting the health, safety, and well-being of Canadians, and the environment.

The text analytic tools are built with our company’s expertise in advanced data analytics (ADA) and machine learning (ML). We also leverage system support and cloud infrastructure of the Southern Ontario Smart Computing Innovation Platform (SOSCIP) to deliver our proposal demo. To facilitate human machine interaction, we solicit help from the NAViGaTOR visualization team of the Igor Jurisica Lab at Krembil Research Institute with their expertise in gaining insights from graph network. It is the intention of I AM OPEN Inc., SOSCIP and the NAViGaTOR team to build an industry grade ETA production environment should the government contract us to develop such an AI based solution.

For human staff to work effectively with extensive and complex textual documents that are highly formal in nature represents a key set of challenges. The terminology often carries specialized vocabulary that require great consistency while maximizing the desired effect of the language. The existence of over 2600 federal regulations can create a challenge for human experts to quickly find relevant regulations, concepts, and terminology for applying regulations, modifying regulations, or drafting new regulations. To help address these challenges, we propose as part of our ETA solution a set of human assisting regulatory text tools that focus on search and comparisons between documents, terms, or sentences in a manner that allows expert staff to approach their search in the most meaningful way possible.

The first tool is the ‘Summaries’ tool which enables users that are less familiar with regulations to quickly find appropriate documents from automatically generated summaries based on keyword search.

The second tool is the ‘Documents Graph’ tool which enables users to obtain a graph visual representation of how regulatory documents relate to each-other based on semantic significance.

The third tool is the ‘Context Search’ tool which enables users to read through entire regulatory documents and select a significant term that has a context of interest. This allows users to quickly find other regulations that have common phrasing to help with editing or exploring regulatory documents.

The fourth tool is the ‘Word Cloud’ tool which enables users to quickly scan through the most commonly used terms in all regulatory documents and quickly identify where a term may be inappropriately used without knowing what term to search for to begin with.

Overall, these tools aim to demonstrate an initial set of key regulatory text tools that make use of natural language processing, machine learning, and advanced analytics to assist human experts. By employing a modular solutions approach, additional tools based on similar or new methods can be added to our ETA solution overtime.

**Analysis Report**

(Suggestion: itemize results from various tools or combination of tools)

The first tool is the ‘Summaries’ tool which enables users that are less familiar with regulations to quickly find appropriate documents based on keyword search. Upon entering keywords of interest, the most relevant search documents are retrieved and their automatically generated summaries are made available for rapid human interpretation and document identification.

The second tool is the ‘Documents Graph’ tool which enables users to obtain a visual representation of how regulatory documents relate to each-other based on semantic significance. Each document is represented by a node and an edge connection links it to other documents that share meaningful relationships. This further allows users to easily filter the degree of relatedness they are interested in and quickly view the relevant content of selected regulations.

The third tool is the ‘Context Search’ tool which enables users to read through entire regulatory documents and select a significant term that has a context of interest. Upon selecting the term, the tool will search and find all other documents that have a similar context to the initial selected term. This enables expert users to quickly find other regulations that may need updating or simply to identify common sentence phrasing.

The fourth tool is the ‘Word Cloud’ tool. This enables users to quickly scan through the most commonly used terms in all regulatory documents. By selecting a term of interest, the user then obtains a report on all the documents that make use of the term. This tool thus helps highlight overused or underused terms to quickly identify where a term may be inappropriately used without knowing what term to search for to begin with.

**Methodology of Analytics**

(Human machine interactions via graph network visualization is just one of the methods, right?)

The applications which we have built identify relationships between federal regulations using natural language methodologies. The benefit of knowing these relationships is that a person knows exactly which regulations relate to a regulation of interest and how strongly they relate. The advantage of using natural language methodologies is that we can identify contextual similarities between regulations that may not be obvious to the human eye when reading regulations. Generally, the only references humans have which relate one regulation to another are citations that are placed within the text. In addition, a machine can use the same standard and method of comparing regulations across all regulations, as opposed to a human whose method of comparison may change from regulation to regulation. Additionally, machines can compare and link regulations much faster than humans for whom it would be challenging to compare each regulation with every other regulation to discover any relationships.

Below we describe the methods used to prepare data for the 4 applications discussed above.

**Processing:**

In order to compare regulations in a meaningful way we first need to parse text within each regulation to remove all words and characters which do not add substantial meaning in differentiating two regulations. These include most punctuations and stop words (e.g is, the, a, at). We also stem words which is a method to reduce inflected words to their root (e.g running and run).

**Vectorization:**

In order for machines to compare words or sentences, we must first convert words into numerical forms. There are multiple ways to vectorize text, but the method we used is called term frequency – inverse document frequency (tf-idf) vectorization. Tf-idf are a way to convert the textual representation of information into vector space models, or sparse features. Each word in a document is given a score that reflects how important the word is to the document in the collection of documents (corpus of federal regulations). The tf-idf value of a word in a document increases proportionally to the number of times the word appears in that document and is offset by the number of documents in the corpus that contain the word.

**Top words and Summaries:**

The top word in each document is the word with the highest tf-idf score. This word also represents the main topic of the document.

To get the summary for a document we calculate the average tfidf score for each sentence in the document. The n sentences with the highest average tf-idf scores are chosen as the main summary to represent what the regulation is about.

**Word2Vec:**

Word2Vec is another method for converting textual representation of information into vector space models. The method extracts features to create word embeddings. The Word2Vec model is a neural network take in inputs of target and context words where the target word represents the word of focus and the context words are the words found around (in context) of the target word in the text. We define a context window which determines the number of context words around a target word that the model will be trained on. Each unique word in the corpus is assigned a vector in space. Word vectors are placed in the vector space so that words that share common contexts in the corpus are located close to one another in the vector space.

Word2Vec word vector representations are useful because they enable us to compare words or sentences by the context they are used in. For example, a human may conclude that two regulations are related if they see the same recurring word such as “immigration” in both. But in fact the regulations may not be related at all if the word “immigration” is used in a totally different context. Another example, a simpler comparison method which checks if two regulations have the same topic word may not find any links between regulations where the topic words are not the same but are synonyms.

**Document Network Graph:**

In order to build the network graph for regulations we compared the word embeddings (word2vec) for the summaries of each document that we created before. To compare word embeddings, we use a method called cosine similarity which basically measures the cosine of the angle between the vectors of both summaries in the vector space. If two vectors have an angle of 0 degrees than the cosine similarity will be 1, in which case the two vectors are maximally similar.

We calculate the cosine similarity of the word embedding for the summary of each document with the word embedding of the summary of every other document and rank the similarity of the other documents to the document between 1 and 0. Then we place a cut-off for the minimum cosine similarity required for two documents to be considered related.

**Word Cloud:**

The word cloud presents all the top words within the corpus and all the documents related to each top word. At this point we have already narrowed down the top words for each document using tf-idf methodology. In order to find the documents which relate to a particular top word, we first get all the documents which have this top word in their top word list. Next, we get all the documents which have a contextually similar top word above a certain threshold in their top words list.

**Context Search:**

Context search is useful to find other regulations where a particular word is used in the same context as in the regulation of interest. When a context search query is run we first narrow down the list of documents that we will query. To do this, we first check if the word of interest is a word in the word cloud, if yes, we take the list of documents related to that word in the word cloud. Next, we get the list of documents connected to the document of interest in the document network graph.

We go through each document and the sentences within them which contain the target word. If the contextual similarity (cosine) of the word embeddings for a sentence is above a threshold then the sentence is considered a relevant sentence and it is returned as a reference. If a sentence in a document does not contain the target word, we first check if the sentence and the target sentence are contextually similar above a certain threshold. Then, we do a comparison of the pattern of words within the target sentence with the sentence in question.

For pattern comparison we use a method of comparing n-grams for each sentence. An n-gram is a contiguous sequence of n items from a given sample of text. The items can be phonemes, syllables, letters, words or base pairs. In our case, we are creating n-grams of words. If there is a match in at least one n-gram in the sentence in question with the target sentence, we will consider the sentence to be similar to the target sentence.

**Improvements in the future:**