Uso de técnicas de inteligência artificial para subsidiar ações de controle

Luís André Dutra e Silva é servidor do Tribunal de Contas da União, Bacharel em Ciência da Computação pelo UniCeub, com certificação em Engenharia de Software pelo IEEE e em Gestão de Projetos pela Universidade de Stanford.

RESUMO

Os serviços cognitivos são uma alternativa baseada em Inteligência artificial para a obtenção de soluções que são capazes de detectar padrões de qualquer tipo em textos, imagens ou qualquer outra fonte de dados. A seguir, é feita a descrição dos experimentos de utilização de técnicas de inteligência artificial em bases não estruturadas, realizados pelo TCU. As etapas do projeto são detalhadas e relacionadas às aplicações e possibilidades futuras. Os resultados desses experimentos foram muito promissores e pretende-se adotar as técnicas utilizadas nesse período em novas iniciativas tecnológicas para o TCU e para a Administração Pública em geral.

Palavras-chave: Serviços cognitivos. NLP. Inteligência artificial. *Machine Learning*. Mineração de textos.

1. INTRODUÇÃO

Grande parte das informações consultadas e produzidas pelo TCU são recebidas de órgãos jurisdicionados, registradas em relatórios e instruções processuais, votos, acórdãos, despachos e outros documentos. Esses registros são textuais e complexos e exigem recursos sofisticados de interpretação para se obter o conhecimento linguisticamente representado, especialmente em razão de os dados não estarem estruturados. Essa característica requer inúmeras análises e combinações

para exploração e agregação de valor às informações e ao processo de tomada de decisão.

Esse fato implica significativo esforço por parte dos servidores do Tribunal para estruturação desses dados e sistematização de conhecimentos para uso e tomada de decisão. Exemplo disso é o esforço requerido para a realização de determinadas atividades de gestão e controle, como o monitoramento de deliberações do TCU e as iniciativas de classificação e triagem de processos de Tomada de Contas Especial (TCE).

Nesse contexto, o uso de ferramentas e algoritmos amparados em modelos de *machine learning* para automatização da interpretação de documentos revela-se essencial e estratégico para classificação e extração automática de informações contidas em fontes de dados não estruturados.

As técnicas de inteligência artificial exploradas e sistematizadas nesse contexto permitem o aprendizado de máquina das características mais complexas de conceitos presentes em diferentes documentos. Com isso, pretende-se estruturar e tornar disponíveis e úteis informações inicialmente dispersas em diferentes documentos e formatos.

A título de informação, o desenvolvimento de experimentos iniciais com o uso de técnicas de *Deep Learning* revelou-se bastante promissor. O protótipo foi realizado no período de 1º de julho a 31 de dezembro de 2015, e a prova de conceito realizada utilizou como base de treinamento cerca de 257 mil acórdãos de 1993 a 2013 e classificou deliberações contidas no texto de 5.300 acórdãos proferidos entre 2014 e 2015. O resultado obtido revelou precisão média de mais de 96%.

2. DESENVOLVIMENTO DOS TRABALHOS

Dado o caráter exploratório e o pioneirismo de uso de técnicas de inteligência artificial no âmbito do TCU, o desenvolvimento dos trabalhos foi dividido em etapas de experimentação com foco inicial para processos de tomada de contas especial. Dado o caráter experimental, muitas das análises realizadas foram descartadas para que a melhor solução surgisse. Em que pese essa orientação, o trabalho sempre teve presente a ideia de replicação para outros objetos de controle e desafios de negócio.

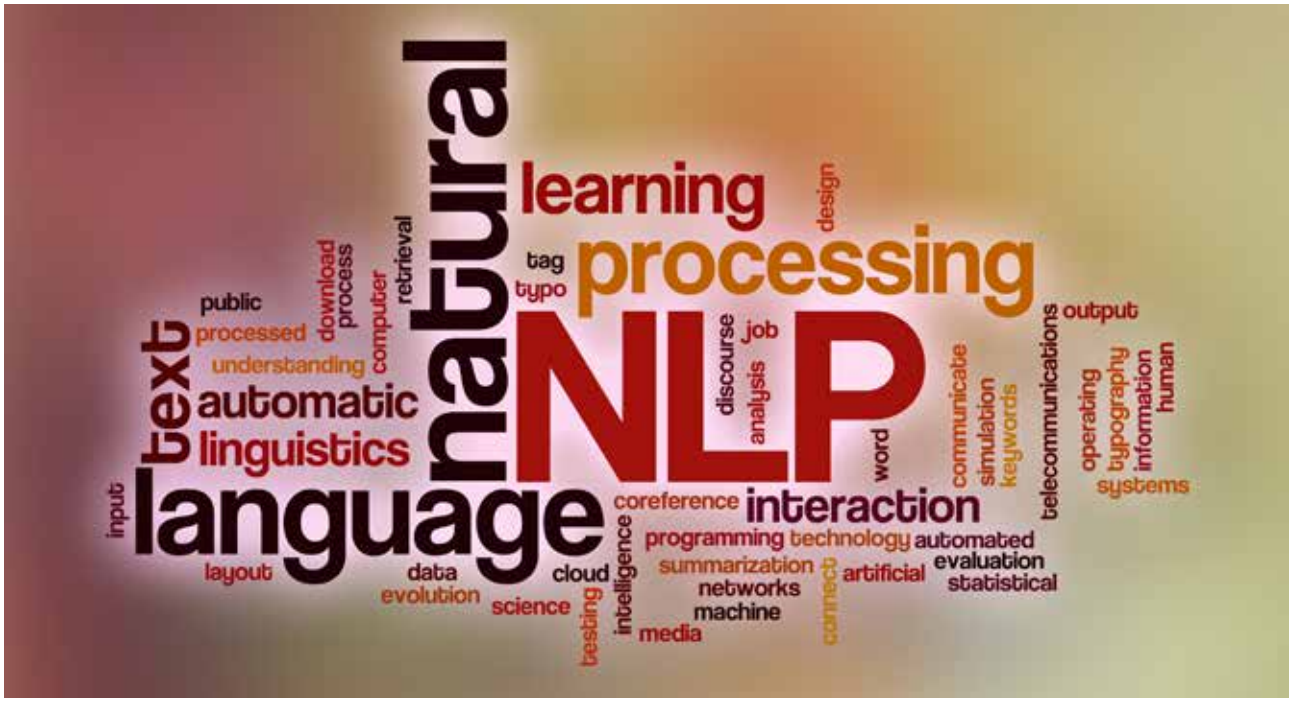
A primeira etapa consistiu na obtenção dos documentos de processos de Tomada de Contas Especial diretamente do repositório do GED (Gerenciamento Eletrônico de Documentos), com o propósito de estabelecer uma base local para futura mineração das peças de TCE. O volume de documentos utilizados nessa etapa foi de cerca de 680 mil peças, e, para a extração deles foi utilizado um algoritmo em linguagem Python para acesso ao banco de dados de desenvolvimento e ao GED em produção com a finalidade de realizar download de todas as peças em pastas relativas a cada processo de TCE. Essa etapa foi realizada no período de 20 a 24 de fevereiro de 2016, e o principal desafio enfrentado foi o fato de que o download das peças foi frequentemente interrompido devido a problemas de conexão, e não foi possível, no

início, baixar todas as peças. No entanto, ao final, todas as peças de todos os processos de TCE foram baixadas.

Na segunda etapa, foi realizada a obtenção da base de dados do Cadastro de TCE em Apex e da base de Cadastro de TCE da CGU (Controladoria-Geral da União), bem como os documentos de relatórios da CGU em formato Word, com o propósito de compor a base de dados correlatos aos processos de TCE para realização de mineração de texto nos documentos obtidos. O volume de documentos utilizado nessa etapa foi de cerca de 2 mil peças e foi utilizado um script em SQL (*Structured Query Language*) para realização dos *dumps* dos bancos de dados e acesso aos dados compartilhados da CGU por compartilhamento da rede local. As ações dessa etapa foram realizadas entre 2 e 3 de março de 2016 e teve como principal desafio o fato de que nem todos os documentos em formato Word da CGU possuíam metadados correspondentes no sistema de cadastro deles. Os dados de cadastro do sistema em Apex correspondiam a uma pequena fração do passivo existente de processos de TCE. Como resultado dessas ações, os dados baixados ficaram disponíveis para análise posterior.

Na terceira etapa, foi realizado o agrupamento das peças de processos de TCE de acordo com função de governo e tipos de irregularidades constantes no sistema de cadastro em Apex, com o propósito de tentar classificar automaticamente as peças não anotadas, que compõem a maioria dos processos, com a finalidade de acelerar o processo de triagem e distribuição dos processos. O volume de documentos utilizados foram cerca de 680 mil peças, analisadas por meio de redes neurais do tipo *Self-Organizing Maps*, ou *Kohonen Maps*, mediante o software livre Somoclu em ambiente Linux. Essas ações foram realizadas entre 4 e 10 de março de 2016 e o principal desafio enfrentado foi o fato de que a grande quantidade de documentos analisados demandou vários dias consecutivos de processamento e a proporção entre número de documentos que possuíam classificação prévia para o número de documentos sem classificação seria muito discrepante para a obtenção de um resultado útil. Não obstante as dificuldades encontradas, foi possível classificar grande parte dos documentos por função de governo usando por base o cadastro obtido do sistema Apex utilizado até então para as TCEs.

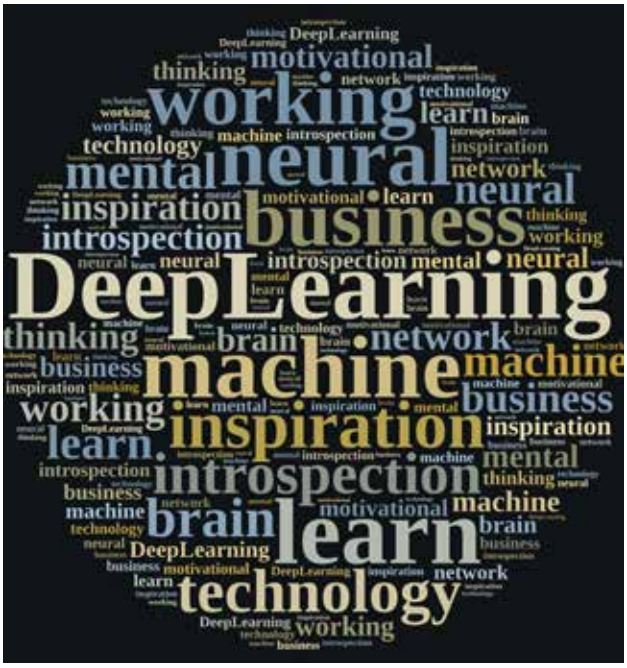
A quarta etapa foi a realização da análise das irregularidades dos processos da função de governo Saúde, com o propósito de classificar os processos enviados em forma de planilha Excel que estavam pendentes de distribuição. O volume de documentos analisados foi de cerca de 3.146 peças por meio da técnica de clustering utilizando *Self-Organizing Maps*. Essa etapa foi realizada entre 11 e 21 de março de 2016 e teve como principal desafio o fato de que devido ao pequeno número de processos classificados previamente em relação ao total existente e à conformação difusa das classificações de irregularidades, provavelmente o resultado não seria



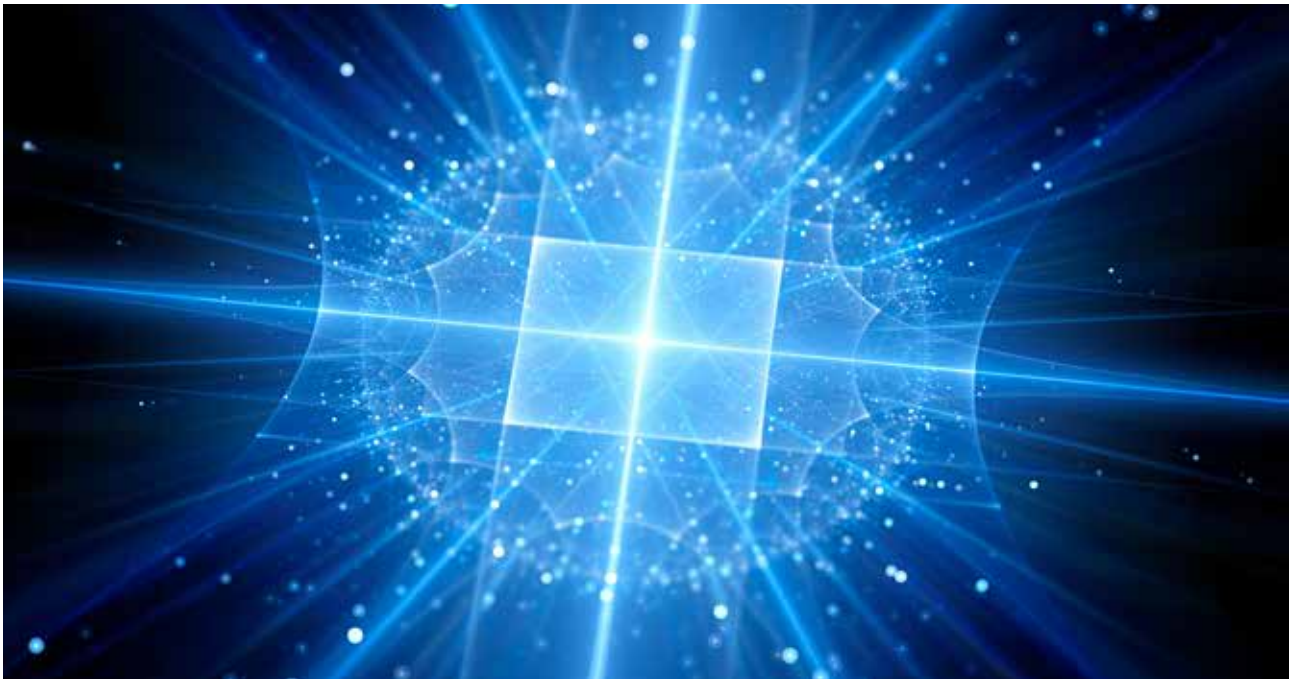
muito preciso, com isso, dos 3.146 processos identificados na função de governo Saúde, todos estariam com os tipos de irregularidade II, III e IV, e a maioria também com o tipo V. Nenhum desses 3.146 teria o tipo I, o que segundo os especialistas deveria ser detectado para cerca de pelo menos 10% dos processos.

A quinta etapa consistiu em criar o serviço de extração de deliberações e teve como propósito disponibilizar um serviço web de extração de deliberações de textos de acórdãos com a finalidade de realizar a inserção dos tipos de deliberação Determinação a Órgão/Entidade de forma semiautomática, no sistema Radar, por meio de interface Apex. Cerca de 200 acórdãos são proferidos por semana, para analisá-los foi utilizado *Deep Learning* em Python com o framework Theano/Keras, POS tagging usando a biblioteca livre Aelius, expressões regulares e o framework Flask em Python com a infraestrutura Werkzeug. Essa etapa foi realizada no período de 22 a 28 de março de 2016 e, durante sua execução, foi encontrado o desafio de que existem muitas ambiguidades linguísticas a serem tratadas por meio de regras específicas na extração de atributos de tipos de deliberações de acórdãos. Algumas dessas ambiguidades necessitam de algum tipo de compreensão do texto em nível próximo ao da inteligência humana. Não obstante, a maior parte dos itens de deliberação do tipo Determinação a Órgão/Entidade é identificada com precisão maior do que 96% e seus atributos, segundo os especialistas, também são extraídos com alta precisão. Sendo assim, o tipo de deliberação escolhido para demonstrar a utilidade desse serviço é considerado o mais adequado para que haja um elevado grau de sucesso na iniciativa de automatizar o processo de extração.

A sexta etapa foi dedicada ao propósito de estabelecer correlação entre os textos das instruções de TCE realizados pelo TCU e os 37 parágrafos-padrão, correspondentes a cada tema, disponíveis como auxílio no portal. O volume de documentos utilizado foi de cerca de 5.997 instruções processuais de TCE, e as técnicas utilizadas nessa abordagem foram similaridade léxica e semântica utilizando o algoritmo de distância de edição Levenshtein e cosseno entre os vetores de *Bag of Words*. Essas ações ocorreram entre 1º e 7 de abril de 2016 e tiveram como principal desafio o fato de que a similaridade entre cada tema e as instruções é elevada qualquer que seja o tema escolhido, sendo assim, é muito difícil distinguir com exatidão a presença dos parágrafos-padrão em cada instrução analisada. Portanto, o resultado foi o fato de que o percentual de similaridade apurado para cada instrução/parágrafo, conforme indicados pelos especialistas na área de TCE, foi insatisfatório, pois as similaridades ficaram muito próximas umas das outras. Sendo assim, o experimento foi considerado encerrado e, portanto, essa abordagem foi descartada.



A sétima etapa foi um experimento com os propósitos de: analisar a estrutura dos relatórios da CGU e definir quais informações podem ser extraídas e estruturadas; aplicar algoritmos de correção de textos com elevado nível de ruído proveniente de OCR (*Optical Character Recognition*); elaborar algoritmo de extração de informações semelhante ao utilizado para o Extrator de Deliberações; disponibilizar interface de validação para os especialistas em TCE; definir critérios de classificação e encaminhamento com base nas informações extraídas; classificar e sugerir encaminhamento por meio da interface de validação; e validar a classificação e encaminhamento pelos especialistas. O volume de documentos envolvido foi de cerca de 3.800 relatórios, e a técnica utilizada foi *Information extraction* usando expressões regulares e cruzamento de dados estruturados com base nas entidades extraídas dos documentos. Essa etapa foi realizada no período de 8 a 26 de abril de 2016 e teve como principal desafio o fato de que a qualidade dos documentos digitalizados não é suficiente para garantir extrações que correspondam ao mesmo grau de confiabilidade alcançados pela leitura realizada pelos auditores que instruem TCE. Ao final, essa etapa teve os seguintes resultados: com a ajuda de especialistas em TCE, foram definidas as informações a serem estruturadas e extraídas dos relatórios originários da CGU em formato



Word; foram realizadas extrações de informações nesses relatórios que permitiram preencher com boa precisão uma minuta de instrução para o tipo de irregularidade "Omissão no dever de prestar contas"; foram realizados cruzamentos das informações extraídas com sistemas estruturados, tais como o Sisobi, Siconv, Siafi; as mesmas extrações realizadas em documentos em formato Word passou a ser realizada também para os relatórios digitalizados em PDF, que são de menor qualidade devido ao processo de OCR; foi disponibilizada uma interface de validação do extrator para os especialistas em TCE. As extrações realizadas por esse experimento foram consideradas úteis na gestão do estoque de TCE. Segundo os especialistas, o trabalho que é realizado de forma manual para este fim pode ser parcialmente automatizado, reduzindo, assim, o tempo necessário para a distribuição/triagem de TCE. No entanto, a geração automática de minutas foi considerada problemática pelos especialistas devido ao fato de apresentarem dados que não preenchem totalmente as necessidades dos auditores que instruem TCE.

Na oitava etapa, realizada entre 27 de abril e 25 de maio de 2016, foram feitas as primeiras entregas com o propósito de disponibilizar as primeiras versões dos serviços previstos no cronograma do projeto. Conforme cronograma do projeto de criação de serviços cognitivos, estão disponíveis para validação e acesso os seguintes produtos em sua primeira versão: serviço REST de extração de deliberações de acórdãos; interface de validação do Extrator de Deliberações; serviço REST de extração de entidades de relatórios de TCE; interface de validação do Extrator de Conteúdo de TCE.

A nona etapa, realizada entre 8 e 15 de julho de 2016, teve como propósito a melhoria de documentos digitalizados por meio da utilização de serviços terceirizados. Foi utilizada uma amostra de dez relatórios com grandes problemas de digitalização, e a API Google Vision foi utilizada para essa tentativa. O desafio principal era melhorar a qualidade do texto extraído dos documentos digitalizados de TCE sem recorrer a novo processo de digitalização. No entanto, o resultado alcançado não foi melhor, em termos de quantidade de erros, do que o software de OCR adotado no âmbito do TCU (Adobe).

A décima etapa, entre 16 e 25 de julho de 2016, teve como propósito o desenvolvimento da primeira versão do serviço de reconhecimento de entidades mencionadas, NER (*Named Entity Recognition*), capaz de extrair de qualquer texto os nomes de pessoas físicas, pessoas jurídicas, CPF, CNPJ e referências normativas. O volume de documentos utilizados para treinar os modelos neurais de reconhecimento de entidades foi de cerca de 58 mil acórdãos e a base Amazônia da Universidade de Lisboa, contendo as anotações manuais de cada tipo de entidade mencionada. Para esse fim, foi utilizado o framework de *machine learning* Apache OpenNLP em Python e Java, com anotações manuais e automáticas em textos de acórdãos. Com a utilização de algoritmos de *machine learning* no lugar de regras pré-definidas, a precisão e as demais

métricas associadas podem não ser as ideais enquanto não houver uma grande quantidade de textos anotados manualmente. A primeira versão do extrator de entidades genérico foi exposta como serviço web na infraestrutura de produção. Como foi separada uma base de validação não treinada de 10 mil sentenças aproximadamente para avaliar a qualidade do serviço web NER, foi obtido um score F1 de cerca de 81% para pessoas físicas, o que corresponde ao estado da arte em termos de extração de entidades mencionadas. A métrica F1 é a mais adequada para demonstrar o equilíbrio entre precisão e recall, que deve ser o objetivo a ser buscado.

Na décima primeira etapa, entre 1º de agosto e 14 de setembro de 2016, ocorreu o desenvolvimento da segunda versão do extrator de entidades genérico em Java puro, utilizando nativamente a biblioteca OpenNLP e os modelos treinados anteriormente para a primeira versão do NER. O principal desafio foi o fato de que a portabilidade de código Python para Java nem sempre é possível em muitas customizações. Portanto, essa iniciativa poderia não ser bem-sucedida caso um ou mais frameworks utilizados em Python não existissem de forma similar em Java. Não obstante, o serviço NER, de acordo com a arquitetura padrão de serviços do TCU, Arquitetura de Referência 8, foi desenvolvido e encontra-se em produção no ambiente JBoss 6 EAP.

A décima segunda etapa, ocorrida entre 16 de setembro de 2016 e 11 de outubro de 2016, teve como objetivo a construção de um serviço capaz de detectar possíveis erros materiais em acórdãos proferidos pelo TCU antes de serem oficializados. Foram utilizados cerca de 600 documentos para o teste da ferramenta dentre cerca de 1 mil acórdãos apontados por uma inspeção realizada pela Corregedoria do TCU como possuidores de erros materiais em sua elaboração. O resultado alcançado pelo serviço cognitivo foi uma captura de cerca de 40% dos erros materiais presentes nos acórdãos, após a verificação da grafia correta dos nomes dos responsáveis, CPF e CNPJ inexistentes ou pertencentes a pessoas físicas já falecidas ou a pessoas jurídicas inativas.

A décima terceira etapa, em andamento desde 18 de outubro de 2016, tem como objetivo a disponibilização de um serviço de extração automática de ontologias em formato OWL, de documentos de texto, utilizando algoritmos de *machine learning*. São utilizados cerca de 280 mil documentos de jurisprudência do TCU. É utilizada a técnica *Latent Semantic Indexing* realizando a decomposição da matriz W de documentos x conceitos, obtida pela técnica *Bag of Words* e normalização TF-IDF, usando *Singular Value Decomposition*. *Latent Semantic Indexing* é um método estatístico que liga termos em uma estrutura semântica útil, sem análise sintática ou semântica e sem intervenção manual. Usando esse método, cada documento é representado não por termos, mas por conceitos que são verdadeira e estatisticamente independentes de uma forma que os termos não são.

3. CONCLUSÃO

Os serviços cognitivos desenvolvidos podem ser utilizados de forma bem-sucedida por outros sistemas, com a finalidade de aprimorar os processos de trabalho do TCU e da Administração Pública, pois é inerente ao trabalho do TCU e outros órgãos a necessidade de estruturar textos produzidos continuamente e que dependem da classificação e extração das informações contidas nessas bases não estruturadas.

Portanto, a disponibilização desses serviços pode servir para conferir maior exatidão ao trabalho de instrução de processos, bem como para colocar à disposição, para os profissionais de Controle Externo, contextos pertinentes a todo trabalho desenvolvido de forma textual.

Ao elaborar determinado relatório, por exemplo, toda a jurisprudência existente sobre o tema desenvolvido poderia ser automaticamente relacionada ao texto nascente de forma a simplificar o processo de busca por informações relevantes ao conteúdo a ser elaborado. Outro exemplo de uso seria a elaboração de resumos de textos recebidos de órgãos jurisdicionados de forma automática, para agilizar o processo de análise desses documentos externos.

Além disso, podem-se elaborar, por meio de serviços cognitivos, clippings personalizados a respeito de cada assunto tratado em determinado trabalho. Assim, os auditores terão sempre as informações mais atualizadas que subsidiem a elaboração dos relatórios de fiscalização e todos os demais documentos necessários.

REFERÊNCIAS

NOVELLI, Andreia, OLIVEIRA, José. Simple Method for Ontology Automatic Extraction from Documents. International Journal of Advanced Computer Science and Applications. Vol 5. No. 12. 2012

MADDI, Govind, VELVADAPU, Chakravarthi. Ontology Extraction from text documents by Singular Value Decomposition. ADMI. 2001

BERRY, Michael, DUMAIS, Susan, O'BRIEN, Gavin. Using Linear Algebra for Intelligent Information Retrieval. SIAM Review. Vol 37. No.4, pp-573-595, December 1995.