O Processamento de Linguagem Natural nos Relatórios de Auditoria Governamental

RYCHARD NUNES GUEDES
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
CATARINA LENICE LOPES PEDROSA
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
JOSEDILTON ALVES DINIZ
Universidade Federal da Paraíba

Resumo

Um dos problemas basilares da contabilidade é comunicar aos seus usuários informações nas quais se verifique compressibilidade. Ter as informações contidas nos documentos contábeis e financeiros de forma estruturada computacionalmente é um fator importante que ensejará o desenvolvimento de novas formas de comunicação com a comunidade, aumentando a compreensibilidade e a efetividade social das práticas contábeis. Entretanto, por se tratar de documentos criados para consumo humano, a sua estruturação automática não é trivial. Assim, o uso de processamentos de linguagem natural (PLN) têm um potencial considerável para estruturar tais dados, possibilitando a criação de painéis gráficos para indicação de desempenho financeiro, orçamentário, patrimonial e operacional, atual e futuro das entidades públicas. No Tribunal de Contas do Estado da Paraíba são criados, quadrimestralmente, relatórios de acompanhamento das gestões municipais apresentados em formato .pdf mas que, apesar de estar disponível digitalmente, possui suas informações desestruturadas computacionalmente. O objetivo dessa pesquisa foi desenvolver um conjunto de ferramentas que aproveita a tecnologia cognitiva para extrair e analisar o texto desses relatórios, criando um veículo de busca capaz de identificar dados significativos em documentos que coletivamente podem ajudar os auditores a identificar áreas de risco, contribuir com planejamento estratégico de auditoria dos tribunais de contas e empoderar a sociedade para o exercício do controle social. Para tal, foram utilizadas ferramentas de conversão de arquivos .pdf para texto puro e, posteriormente, aplicação de técnicas de padronização de texto. Em seguida, utilizando uma amostra da base de relatórios disponível, foram criadas uma série de regras que poderia ser aplicada no restante dos relatórios para identificar as informações que deveriam ser extraídas e estruturadas computacionalmente. Com isso, painéis de visualização foram criados com a utilização dos dados extraídos dos relatórios, auxiliando e trazendo agilidade aos auditores em suas análises.

Palavras chave: Processamento de Linguagem Natural, Auditoria Pública, Automatização.

1. Introdução

As inovações tecnológicas têm proporcionado transformações surpreendentes na forma como as auditorias governamentais são conduzidas. As ferramentas de tecnologia da informação como inteligência artificial, automação de fluxo de trabalho e análise de dados estão eliminando uma série de procedimentos repetitivos, originários de um esforço intensivo no escopo de processos e manuais de auditoria (Raphael, 2017). Destaca-se por importante que a inovação está permitindo a inserção de novos achados de auditoria que antes eram inimagináveis por não existir ferramentas que pudessem operar em um ambiente de big data, efetuando cruzamentos de dados de forma eficiente.

A auditoria governamental tem como escopo final a elaboração de documentos textuais destinados a comunicar uma ampla variedade de mensagens (Fisher, Garnsey, & Hughes, 2016) incluindo, entre outros, o desempenho financeiro, orçamentário, patrimonial e operacional, atual e futuro das entidades públicas, a partir das prestações de contas, da participação da sociedade e dos órgãos de controle, no qual a fiscalização se dá em obediência a padrões de domínio e regulamentos, bem como evidências de conformidade com normas e regulamentos relevantes.

Todo esse compêndio de informação é sintetizado em um relatório de auditoria. Todavia, é de relevo destacar que esse é um componente essencial das auditorias governamentais, merecendo destaque, posto que, por natureza, é desafiador a compreensão desse relatório, por parte da sociedade, principalmente em ambiente de tecnologia de comunicação desenvolvido.

Os relatórios de auditoria governamental geralmente são muito analíticos e detalhados, o que dificulta sobremaneira a compreensão, por parte do cidadão, dos itens sintéticos definidos pela a Constituição Federal, tais como, aplicações mínimas em saúde e educação, gasto com despesas com pessoal, resultado orçamentário, dentre outros. Além da compreensão da sociedade, os órgãos de controle externo, no caso os Tribunais de Contas, precisam definir e exigir melhorias da atividade governamental, o que requer informações sintéticas de todos os seus jurisdicionados para definir, com base dos achados pretéritos, os planejamentos de auditoria futuros.

O Tribunal de Contas tem apresentado seus relatórios de forma não estruturados, ou seja, não possuem um modelo de dados predefinido ou não estão organizados dentro de um padrão comum. Normalmente, ele é um texto e geralmente inclui conteúdo multimídia. Enquanto alguns arquivos podem ter uma estrutura interna, eles ainda são considerados desestruturados porque os dados não se encaixam em um banco de dados (Feldman & Sanger, 2007). Por sua vez os dados estruturados como planilhas são facilmente pesquisáveis por algoritmos básicos. Portanto, é mais difícil analisar documentos textuais não estruturados do que estruturados.

O objetivo dessa pesquisa é desenvolver um conjunto de ferramentas que aproveita a tecnologia cognitiva para extrair e analisar o texto dos relatórios de auditoria, criando um veículo de busca capaz de identificar dados significativos em documentos que, coletivamente, podem ajudar os auditores a identificar áreas de risco, contribuir com planejamento estratégico de auditoria dos tribunais de contas e empoderar a sociedade para o exercício do controle social.

Ao fim dessa pesquisa apresenta-se as informações dos relatórios de auditoria em bases de dados estruturadas, capaz de ser utilizadas em painéis de Business Intelligence (BI), ou em planilhas eletrônicas, capaz de prestar informação para o exercício controle do externo, controle do gerencial e controle do social.

Nessa seara, é de se destacar que as técnicas conhecidas necessitam que os dados estejam estruturados. No contexto do Tribunal de Contas do Estado da Paraíba, locus dessa pesquisa, isso ocorre com os dados declarados pelos órgãos públicos aos quais estão sob fiscalização do tribunal e estruturados em bancos de dados. Entretanto, o mesmo não acontece com os dados auditados, presentes nos documentos de julgamentos, relatórios e pareceres, bem como os dados gerados diariamente na web em mídias sociais, por exemplo.

Devido ao conteúdo dos dados do relatório de auditoria serem, em sua maioria desestruturado, tal fato se contrapõe aos códigos computacionais ou linguagens de maquinas, que não conseguem processar essa larga massa de dados. Esses fatores motivaram o desenvolvimento de um leque de técnicas computacionais para a análise automática e representação dessa linguagem humana, conhecida como Processamento de Linguagem Natural (PLN) (Appelt, Hobbs, Bear, Israel, & Tyson, 1993).

Este estudo complementa uma abordagem inovadora de mineração de texto para sintetizar as informações do relatório de auditoria, mediante técnicas de Processamento de Linguagem Natural (PLN), que são amplamente utilizados nas literaturas de linguística computacional (Grimmer & Stewart, 2013).

Os relatórios de auditoria representam uma fonte de dados não explorados (Nian, Zimmerman, Mccoy, & Mar, 2016). É difícil extrair valor dos dados usando pesquisas manuais demoradas. Por outro lado, os computadores têm capacidade ilimitada para extrair valor de forma eficiente, desde que o texto do documento seja preparado em um formato uniforme que humanos e máquinas possam entender. A aplicação de PLN ao texto pode permitir que os auditores usem cada sentença de cada relatório, gerando um pacote de novas informações.

Nesse contexto, extrai informações importantes que estão presentes dos Relatórios de Acompanhamento da Gestão Municipal, definidos na Resolução Normativa RN-TC N 01/2017 do Tribunal de Contas do Estado da Paraíba. Com essas informações dispostas de forma estruturada, determinadas auditorias serão facilitadas, assim como será possível criar painéis de visualização buscando a democratização da informação pública.

2. Referencial Teórico

2.1 A transparência na auditoria governamental

Como o advento do Estado Democrático de Direito foram implementados alguns aspectos acerca do controle social na administração pública. Conforme argumenta Matias-Pereira (2010, p.97) “A democratização do Estado tinha como um dos pressupostos o controle do seu aparelho pela sociedade civil. Assim, a transparência do Estado, expressa na possibilidade de acesso do cidadão à informação governamental constituía um requisito essencial”.

De uma forma ampla, a transparência deve caracterizar os atos dos administradores públicos, na forma que os cidadãos tenham acesso e compreensão daquilo que os governamentais têm realizado, após o poder de representação que lhe foi confiado (Cruz, de Souza Ferreira, da Silva, & da Silva Macedo, 2009).

No Brasil, a transparência no setor público se inicia com a Constituição Federal de 1988, mas ela só ganha mais força nos anos 2000 com o surgimento da Lei de Responsabilidade Fiscal (LRF) que estabelece a transparência como um princípio de gestão fiscal. E se torna mais concreta com o surgimento da Lei 131/2009 (Lei da Transparência) e a Lei 12/527 (Lei de Acesso à Informação) que tentam aprimorar o controle social na administração pública brasileira.

Para se tornar mais efetiva à transparência é necessário que as informações apresentadas não sejam limitadas aos dispositivos legais. Conforme argumente Cruz et al (2010) “a divulgação das informações acerca dos atos de gestão pública não deve se limitar aos relatórios previstos em dispositivos legais (em geral, relatórios fiscais e financeiros), mas também de informações qualitativas que reportem desempenho, projetos e atingimento de metas em áreas relevantes para a sociedade, tais como saúde, educação, cultura, transporte, saneamento e outras.”

A transparência não consiste apenas em divulgar a informação com fim nela mesma. Faz-se necessário que o cidadão tenha uma compreensão nítida dos eventos que ocorrem na gestão pública. Muito embora, há de se pensar que a clareza da informação é de responsabilidade do gestor apenas, o relatório de auditoria é o documento em que a sociedade espera que possa aferir as informações prestadas pela gestão.

Assim, o relatório de auditoria, anteriormente denominado parecer de auditoria, é a nova expressão atribuída pelas normas internacionais de auditoria ao meio pelo qual o auditor apresenta a conclusão de sua auditoria, na forma de uma opinião (Longo, 2011). Já as normas de auditoria expõem que o posicionamento da auditoria deve expressar claramente sua opinião por meio de um relatório de auditoria escrito. Nesse sentido a forma como a opinião é expressa vai fazer toda a diferença para quem vai fazer uso dele.

Na prática o que se observa é que o relatório foi escrito para entender a formalidade processual dentro do processo de prestação de contas anuais, e muitas das vezes o seu conteúdo não é plenamente entendido, seja por problema de compreensão textual ou devido a sua extensão ou por falta de objetividade.

Nesse interim, a tecnologia da informação pode ser bastante útil, na medida que é capaz de ler e apreender os aspectos mais relevantes que podem ser utilizados, pelos julgadores do Tribunal de Contas, pelo gestor e pela a sociedade que vai executar o controle social dos recursos públicos aplicados.

2.2Uso de tecnologia nos relatórios de auditoria

O avanço da tecnologia vem permitindo inovações cada vez mais criativas. Nos últimos anos, um crescente comum em diversas áreas tem sido o uso de tecnologias cognitivas, que permitem máquinas realizar atividades antes restritas a seres humanos (Schatsky, Muraskin, & Gurumurthy; 2015). Dentre estas tarefas, as mais comuns estão relacionadas a visão computacional e PLN.

No âmbito desta última, esforços vem sendo direcionados para diversas subáreas, como análise de sentimento do autor de um texto ou criação de resumos. Isso ocorre especialmente ao avanço do poder computacional, tornando factível a aplicação de vários métodos complexos que fazem análises na semântica do texto. Entretanto, o foco inicial dos pesquisadores da área foi em desenvolver técnicas voltadas para sintaxe, já que esta teria aplicabilidade mais direta em algoritmos que seriam capazes de fazer máquinas aprenderem sobre determinados tópicos (Cambria & White; 2014).

Dessa forma, o interesse na extração de informação surge no princípio das pesquisas em PLN. Essa área consiste na identificação e extração de informações textuais, sendo possível o armazenamento posterior na forma estruturada, o que torna possível a exploração mais eficiente da enorme quantidade de dados ali presente.

As conferências nomeadas Message Understanding Conferences (MUCs), patrocinadas pelo Defense Advanced Research Projects Agency (DARPA) e organizadas pela

US Naval Ocean System Center, com sua primeira edição em 1987, foram os principais eventos que tornaram a atenção para este tópico (Du, Pivovarova, & Yangarber, 2016).

Na MUC-4 foi apresentado o FASTUS (do inglês Finite State Automata-based Text Understanding System), um sistema para extração de informação pré-especificada de textos, com alta velocidade e acurácia, sendo classificado entre os melhores sistemas desenvolvidos no contexto da conferência. Seus destaques eram a capacidade de processar longos textos; e o curto espaço de tempo para a adaptação do sistema a um conteúdo textual específico (no caso da MUC-4, que buscava extrair informações relacionadas a terrorismo, cerca de 3 semanas) (Appelt et al, 1993).

De acordo com seus criadores, sistemas baseados em semântica foram testados, obtendo resultados razoáveis. Entretanto, devido ao curto espaço de tempo para desenvolvimento e processamento dos dados, esses sistemas foram descartados por desprenderem muito esforço computacional em parte irrelevantes do texto, as quais não continham nenhuma ou quase nenhuma informação significativa. Assim, em contraste com aos sistemas baseados em semântica, que buscam o entendimento completo do texto, o FASTUS é baseado em três pilares:

- Apenas uma pequena porcentagem do texto é relevante;
- A informação deve ser mapeada em representações simples, rígidas e pré-definidas;
- As nuances de significado e objetivos do escritor com seu texto pouco ou nada importam.

No FASTUS, o texto é inicialmente pré-processado para um formato padrão. Dessa forma, várias operações podem ser realizadas, como a capitalização de todas as letras, correção ortográfica e remoção de acentos e pontuação. Posteriormente, o texto é dividido em trechos, os quais são inseridos no sistema propriamente dito, que é constituído de 4 fases:

- Acionamento: por meio da presença de determinadas palavras chaves, um trecho será definido como relevante;
- Reconhecimento de frases: identificação de verbos, substantivos e uma série de classes críticas de palavras, como preposições, conjunções e pronomes relativos, dentro de um trecho acionado/relevante;
- Reconhecimento de padrões: são definidos padrões, codificados como máquinas de estado finito que tem seu estado atual variado de acordo com uma sequência de entradas, para reconhecimento de pontos de interesse;
- Mescla de incidentes: a medida que cada trecho é processado, novas estruturas vão sendo criadas, as quais geralmente estarão incompletas pela baixa probabilidade de um trecho conter todas as informações buscadas. Ao final, todas as estruturas são mescladas para a criação de uma estrutura final, a qual será a mais completa possível.

Além do FASTUS, existem outros sistemas mais modernos que buscam extrair informações de textos incialmente criados para consumo humano. Du, Pivovarova e Yangarber (2016) propuseram uma arquitetura de sistema, denominada PULS, que provê suporte a tomada de decisão com base em informações extraídas de fontes textuais online, como portais de notícias e redes sociais, analisando-as, ao contrário do FASTUS, semanticamente com uso de algoritmos de aprendizagem de máquina, criando um compilado conciso e de uso simplificado.

Técnicas de PLN também são aplicadas por Yan, Yu, Liu e Wu (2018) para analisar as associações que podem existir entre o risco identificado do cliente e os textos corporativos divulgados ou relatório de auditoria. De acordo com os autores, dados não estruturados, como texto, raramente são utilizados nessas associações, especialmente pelo fato das técnicas necessárias para tal serem relativamente novas ao mundo de contabilidade e auditoria. Dessa forma, Yang et al. introduzem um sistema de mineração de texto que mede o risco do cliente em quatro aspectos: financeira; estratégica; operacional; e graves riscos provenientes de relatórios anuais.

Fisher et al. (2016) realizaram uma revisão de literatura no âmbito de contabilidade e auditoria, classificando as pesquisas relacionadas em quatro tipos distintos: análise manual do texto; mineração de texto básica; PLN + Aprendizagem de Máquina; e revisões de literatura. É fácil imaginar que as análises manuais tiveram grande destaque no início das pesquisas, mas que foram perdendo força ao passo que técnicas mais robustas e completas foram surgindo, sendo isso evidenciado na revisão dos autores. Entretanto, um fator que deve ser destacado é que a mineração de texto básica vem crescendo ao mesmo passo que técnicas mais complexas, como uso de PLN + Aprendizagem de Máquina, sem perder destaque e se contrapondo às análises manuais. Com essa análise, é possível inferir que, mesmo com técnicas modernas e robustas, algoritmos simples ainda são bastante eficientes de acordo com o objetivo final.

3. METODOLOGIA

Para alcançar os objetivos dessa pesquisa de desenvolver ferramentas de tecnologia cognitiva para extrair e analisar o texto dos relatórios de auditoria, foram utilizados os Relatórios de Acompanhamento da Gestão Municipal dos quadrimestres do exercício de 2017. Tais relatórios foram instituídos pela Resolução Normativa RN-TC N 01/2017 do Tribunal de Contas do Estado da Paraíba que disciplina a instauração, no primeiro dia útil de cada exercício financeiro, processos de acompanhamento relativos à gestão dos Prefeitos Municipais, entre outros.

Tais processos devem ser atualizados a cada quadrimestre, sendo criado então um novo relatório que será sequência do anterior (permanecendo assim a mesma identificação). Esses relatórios contêm informação previamente auditadas acerca da gestão financeira de cada município, possuindo informações relativas a gastos e despesas em geral, assim como gastos específicos com Educação, Saúde, Pessoal e Previdência.

A Resolução em comento, está alinhada com os normativos constitucionais e infra constitucionais. destacando As informações principais constantes nos relatórios são as seguintes:

- Execução Orçamentária
  - Receita arrecadada
  - Despesa executada
- Receita de Impostos e Transferências;
  - Previsão
  - Executado
- Receita Corrente Líquida
- Educação
  - Total das receitas do FUNDEB
  - Total das aplicações em magistério
  - Percentual das aplicações em magistério

  - Total das aplicações em outras despesas
  - Total das aplicações em MDE
  - Total das receitas de impostos e transferências
  - Percentual das aplicações em MDE
- Saúde
  - Base de cálculo para as ações e serviços públicos de saúde
  - Total das aplicações em saúde
  - Percentual das aplicações em saúde
- Lei de Responsabilidade Fiscal
  - Total das despesas com pessoal (Executivo, Legislativo e Município)
  - Total das despesas com pessoal do Ente
  - Percentual das despesas com pessoal
- Contribuições ao Regime Geral de Previdência
  - Obrigações patronais estimadas
  - Obrigações patronais do exercício pagas
  - Estimativa do valor não recolhido

Além desses dados, é imprescindível identificar a qual município e a qual período um determinado relatório faz parte. Dessa forma, informações como Nome do Município, Número do Processo e Período de Análise também são extraídos. Na Figura 1 é possível verificar a disposição de alguns dos dados de interesse, os quais estão destacados em amarelo. Todos os dados, com exceção da identificação da Prefeitura e do período, são numéricos e dispostos em tabelas.

Para efeitos de teste e validação, foram levados em consideração apenas Relatórios de Acompanhamento da Gestão Municipal dos dois primeiros quadrimestres de 2017, o que deveria resultar em 446 relatórios dado que a Paraíba tem 223 munincípios, no entanto, 3 municípios apresentaram problemas particulares e não tiveram relatório emitido no segundo quadrimestre, restando então um total de 443 relatórios. As regras foram criadas com base em 35 relatórios do primeiro quadrimestre, escolhidos de forma aleatória. Posteriormente, as regras foram validadas com todos os relatórios.

Todos os relatórios são gerados, inicialmente, de forma automática por um sistema do Tribunal de Contas do Estado da Paraíba, fazendo uso dos dados que foram declarados pela própria prefeitura. Esse relatório é disponibilizado no formato .doc para os auditores, os quais realizam sua auditoria baseando-se nos empenhos e suas descrições, incluindo-os ou excluindo-os. Dessa forma, é realizado um ajuste no valor sugerido inicialmente pelo sistema (linhas azuis na Figura 1). Esse relatório é então disponibilizado na web no formato .pdf, o qual é um arquivo totalmente otimizado para impressão, encapsulando uma completa descrição do conteúdo, como imagens, textos e tabelas.

Para extração dessas informações, o arquivo .pdf, precisa ser transformado em um arquivo textual, sendo utilizando a ferramenta Poppler em sua versão 0.72.0 para realização desta etapa. Todas as etapas posteriores foram realizadas utilizando a linguagem de programação Python, na versão 3.6.1.

Posteriormente, deve ser realizada a padronização do texto, com remoção de acentos e transformação de todas as letras em maiúsculas. Além disso, foi observado que a maioria dos dados que estão em tabela são separados horizontalmente por uma quebra de linha (computacionalmente falando, a sequência de caracteres “\n”) e verticalmente por mais de um espaço em branco. Então, sequências que contenham dois ou mais espaços em branco são substituídas pelo caractere especial pipe (|) por meio de expressões regulares, discutidas mais a frente, com objetivo de facilitar a identificação dos dados em passos posteriores.

PROCESSO N.º: 00110/17
JURISDICIONADO: Prefeitura Municipal de João Pessoa
PRODUTO: Relatório de Acompanhamento da Gestão Municipal
PERÍODO: Janeiro a abril de 2017
RELATOR: Conselheiro Antônio Nominando Diniz Filho

Aplicações em FUNDEB	Valor
1. Receita do FUNDEB (cota-parte + complementação)	65.002.270,64
2. Receita de Rendimentos de Aplicação	0,00
3. Ajustes da Auditoria	567.520,83
4. Total das Receitas (Base de Cálculo) (1+2+3)	65.569.791,47
5. Despesa com Remuneração dos Profissionais do Magistério	63.043.203,63
6. Restos a Pagar inscritos no Exercício sem Disponibilidade Financeira de Recursos do FUNDEB (60%)	0,00
7. Ajustes da Auditoria	0,00
8. Total das Aplicações em Magistério (5+6+7)	63.043.203,63
9. Percentual de Aplicação em Magistério (8/4*100)	96,15%
10. Outras Despesas	10.549.281,98
11. Restos a Pagar inscritos no Exercício sem Disponibilidade Financeira de Recursos do FUNDEB (40%)	0,00
12.Ajustes da Auditoria	0,00
13. Total das Aplicações em Outras Despesas (10+11+12)	10.549.281,98

Fonte: Balancete – Despesa Liquidada até o período de análise

Gastos com Pessoal (R$)					
Elemento de Despesa	**Adm. Direta do Executivo**	**Adm. Indireta**	**Poder Executivo**	**Poder Legislativo**	**Município**
Contratação por Tempo Determinado (1)	185.256.547,17	120.116.341,04	305.372.888,21	0,00	305.372.888,21
Vencimentos e Vantagens fixas (2)	390.814.241,75	82.528.161,46	473.342.403,21	41.320.961,65	514.663.364,86
Outras Despesas Variáveis Pessoal Civil (4)	6.966.065,49	9.281.869,00	16.247.934,49	0,00	16.247.934,49
Outras de Pessoal Contratos de Terceirização (5)	0,00	0,00	0,00	0,00	0,00
Ajustes da Auditoria (6)	0,00	0,00	0,00	0,00	0,00
Total das Despesa com Pessoal (7) (1+2+4+5+6)	583.036.854,41	211.926.371,50	794.963.225,91	41.320.961,65	836.284.187,56
Obrigações Patronais (3)	115.639.135,93	29.968.311,29	145.607.447,22	7.979.612,96	153.587.060,18
Diferença positiva com inativos e as receitas de contribuições (8)					60.346.000,00
Total das despesas de Pessoal do Ente					1.050.217.247,74
Receita Corrente Líquida					1.866.480.791,54
% da despesa com Pessoal			50,39%	2,64%	56,27%
Limite Legal			54%	6%	60%

Fonte: Balancetes do mês de referência (abril/2017) e os onze anteriores

Figura 1 Exemplos dos dados de interesse de uma amostra aleatória

O Tribunal de Contas do Estado da Paraíba, ao disponibilizar a base dos relatórios, entregou uma parcela de outros relatórios ou processos que não eram objeto de estudo. Dessa forma, antes de processar o texto inteiro, é necessário identificar se o arquivo corresponde a um Relatório de Acompanhamento da Gestão Municipal. Isso foi realizado por meio da criação de uma expressão regular que contém regras relacionadas ao título que o relatório deve ter, similar à etapa de acionamento do FASTUS.

Caso realmente seja um relatório de interesse, o texto é completamente processado, buscando as tabelas que contém as informações definidas anteriormente. Essa busca é realizada por outros processos de acionamento encadeados, buscando definir inicialmente

sequências de textos maiores que contém a tabela, como um capítulo, e posteriormente refinar dentro dessa seleção até encontrar apenas o texto correspondente à tabela.

Todos esses acionamentos são feitos por meio de expressões regulares, as quais não possuem alto nível de complexidade já que o texto original é proveniente de um modelo pré-definido. Os auditores, dispostos desse modelo, possuem liberdade para alterá-lo por completo, já que sua edição é realizada em arquivos .doc. Porém, por definição em um Procedimento Operacional Padrão interno, os dados que são buscados devem estar disponíveis no formato de tabela. Por ventura, pode ocorrer de alguns auditores alterarem ou incluírem certos termos. Com a observação desses casos, regras mais amplas devem ser criadas para que essas situações fiquem previstas no escopo das regras.

Com a identificação das tabelas, mais uma vez expressões regulares são aplicadas, linha a linha, para identificar se o texto em questão possui informações que são de interesse da extração. Caso positivo, esses dados são inseridos em uma estrutura pré-definida.

Como forma de metrificar a taxa de acerto da metodologia em questão, foram utilizados dois métodos:

- Existem redundâncias em alguns dados. Como exemplo, pode-se citar o trecho de Aplicações em FUNDEB da Figura 1, no qual a linha 9 é resultado de uma operação matemática entre as linhas 4 e 8. Dessa forma, a fim de validação, todas essas informações são extraídas e, tomando proveito desse fato, é realizada uma comparação entre os dados diretamente correlatos, sendo considerada incorretas as informações extraídas que não estejam em concordância, considerando possíveis erros de arredondamento de casas decimais;
- Comparação, por meio de um processo manual, do resultado da extração e do relatório em si, feito com uma pequena amostragem de 25 arquivos que não tenham apresentado o problema relatado no item anterior.

Expressões Regulares

Uma expressão regular pode ser definida como uma fórmula que descreve um conjunto específico de palavras sobre algum alfabeto, consistindo de símbolos individuais deste, além de operadores como:

- Pipe (|): operador OU;
- Asterisco (*): operador kleene estrela, que permite 0 ou mais repetições de um determinado padrão;
- Mais (+): operador kleene mais, que permite 1 ou mais repetições de um determinado padrão;
- Ponto (.): coringa, que pode representar qualquer símbolo do alfabeto.

Além de ser uma importante noção em teoria de linguagem, expressões regulares são amplamente usadas em diversas áreas do conhecimento para definir padrões de busca. Formalmente, dada uma expressão regular p e um texto t, o objetivo em uma busca é verificar se existe uma subsequência do texto t que corresponde ao padrão definidor por p, podendo ainda definir a quantidade de subsequências que existem e sua posição no texto t (Backurs & Indyk; 2016). Como exemplo, dada a expressão regular abc|d e o texto abababdbbababc, serão encontrados os grupos ----abd----abc, já que a expressão regular define que serão encontradas as subexpressões abc ou abd.

Por ser de ampla utilidade, expressões regulares estão definidas em diversas linguagens de programação, como Python, Java e JavaScript. Podem ser usadas nas mais variadas áreas, como mineração de dados, redes de computadores e biologia computacional.

Para disponibilização das informações de forma clara e mais compreensível foi utilizada a ferramenta de Business Intelligence – Tableau. A partir desse font end a aplicação que é capaz de interagir diretamente com os usuário das informações geradas

4. RESULTADOS

Como primeira etapa do processo, a conversão do formato .**pdf** para .**txt** de 1743 arquivos levou cerca de 9,21 minutos. Em seguida, a etapa de pré-processamento para esta mesma quantidade durou cerca de 8 segundos. Já o passo de identificação dos 443 Relatórios de Acompanhamento da Gestão Municipal dentre os 1743 arquivos, com período referente ao ano de 2017, durou menos de 1 segundo. Em seguida, a etapa de extração das informações desses relatórios durou 4 segundos. Ao todo, o processo durou menos de 10 minutos. Diversas regras foram criadas no sistema, deixando inviável a apresentação de todas neste artigo. Entretanto, com fins de exemplificação, é apresentado o caso de extração das informações referentes aos gastos com Saúde.

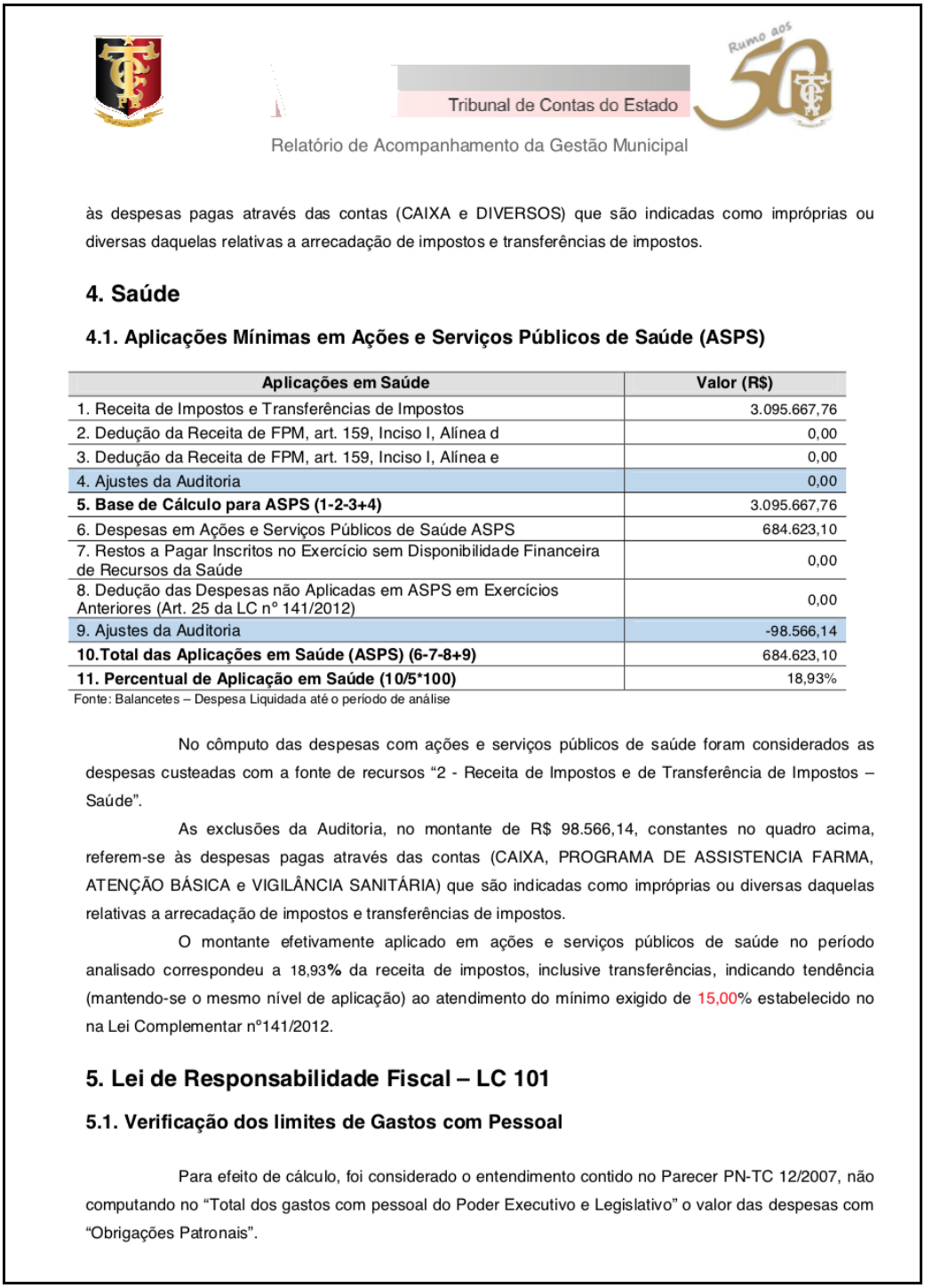
Conforme citado na seção anterior, o primeiro passo é encontrar o trecho do texto que contém a tabela com as informações que são buscadas. Na Figura 2 é mostrada disposição da informação no arquivo .pdf. Já na Figura 3 é mostrada a disposição da informação no arquivo textual já pré-processado.

Para encontrar o trecho do código que contém a tabela em questão, foi utilizado, inicialmente, a expressão regular "\d\. SAUDE". Tal expressão encontra grupos que contenham um dígito qualquer (0-9) seguido de um ponto, espaço em branco e a palavra SAUDE, em maiúscula. Para o caso descrito nas Figuras 2 e 3, o caractere "d" poderia ser substituído pelo caractere "4" na expressão regular. Entretanto, foi observado que o auditor poderia inserir ou remover algum capítulo no relatório, alterando a numeração. Assim, optou-se uma regra mais generalista.

A segunda etapa é identificar onde a tabela acaba. Isso pode ser alcançado com a expressão regular "FONTE". Vale salientar que essa segunda expressão é utilizada na parte do texto que possui início após a primeira expressão. Com o trecho do texto que contém a tabela selecionado, a próxima etapa é identificar as linhas que contém a informação que deve ser extraída. Em Saúde, são buscadas informações que estão em três linhas diferentes:

- Base de cálculo para as ações e serviços públicos de saúde;
- Total das aplicações em saúde;
- Percentual das aplicações em saúde.

Como pode ser visto na Figura 3, todas as colunas de uma mesma linha ficam antes da quebra de linha. Dessa forma, ao identificar o texto "BASE DE CALCULO PARA(AS)* ASPS", por exemplo, basta separar a linha em duas, usando como separador o caractere pipe, extraindo a segunda parte. Vale salientar ainda que a sub-expressão "AS)*" é usada por ter sido observado que alguns auditores optam por inserir o artigo "as" no texto, enquanto outros omitem isso.



Tribunal de Contas do Estado

Relatório de Acompanhamento da Gestão Municipal

às despesas pagas através das contas (CAIXA e DIVERSOS) que são indicadas como impróprias ou diversas daquelas relativas a arrecadação de impostos e transferências de impostos.

4. Saúde

4.1. Aplicações Mínimas em Ações e Serviços Públicos de Saúde (ASPS)

Aplicações em Saúde	Valor (R$)
1. Receita de Impostos e Transferências de Impostos	3.095.667,76
2. Dedução da Receita de FPM, art. 159, Inciso I, Alínea d	0,00
3. Dedução da Receita de FPM, art. 159, Inciso I, Alínea e	0,00
4. Ajustes da Auditoria	0,00
5. Base de Cálculo para ASPS (1-2-3+4)	3.095.667,76
6. Despesas em Ações e Serviços Públicos de Saúde ASPS	684.623,10
7. Restos a Pagar Inscritos no Exercício sem Disponibilidade Financeira de Recursos da Saúde	0,00
8. Dedução das Despesas não Aplicadas em ASPS em Exercícios Anteriores (Art. 25 da LC nº 141/2012)	0,00
9. Ajustes da Auditoria	-98.566,14
10.Total das Aplicações em Saúde (ASPS) (6-7-8+9)	684.623,10
11. Percentual de Aplicação em Saúde (10/5*100)	18,93%

Fonte: Balancetes – Despesa Liquidada até o período de análise

No cômputo das despesas com ações e serviços públicos de saúde foram considerados as despesas custeadas com a fonte de recursos "2 - Receita de Impostos e de Transferência de Impostos – Saúde".

As exclusões da Auditoria, no montante de R$ 98.566,14, constantes no quadro acima, referem-se às despesas pagas através das contas (CAIXA, PROGRAMA DE ASSISTENCIA FARMA, ATENÇÃO BÁSICA e VIGILÂNCIA SANITÁRIA) que são indicadas como impróprias ou diversas daquelas relativas a arrecadação de impostos e transferências de impostos.

O montante efetivamente aplicado em ações e serviços públicos de saúde no período analisado correspondeu a 18,93% da receita de impostos, inclusive transferências, indicando tendência (mantendo-se o mesmo nível de aplicação) ao atendimento do mínimo exigido de 15,00% estabelecido no na Lei Complementar nº141/2012.

5. Lei de Responsabilidade Fiscal – LC 101

5.1. Verificação dos limites de Gastos com Pessoal

Para efeito de cálculo, foi considerado o entendimento contido no Parecer PN-TC 12/2007, não computando no "Total dos gastos com pessoal do Poder Executivo e Legislativo" o valor das despesas com "Obrigações Patronais".

Figura 2 Informações referentes à Saúde no relatório em formato .pdf

As informações foram removidas em formato textual. Porém, para que esteja bem estruturada, é necessário que valores numéricos sejam armazenados como tal. Então, mais regras são aplicadas para que os valores sejam convertidos, tais como remoção de separadores de milhar, substituição do separador decimal vírgula por ponto e remoção do símbolo de porcentagem. Tudo isso se faz necessário por existir uma padronização computacional para que uma máquina entenda esse conteúdo como valores numéricos.

```
RELATORIO DE ACOMPANHAMENTO DA GESTAO MUNICIPAL
 AS DESPESAS PAGAS ATRAVES DAS CONTAS (CAIXA E DIVERSOS) QUE SAO INDICADAS COMO IMPROPRIAS OU
 DIVERSAS DAQUELAS RELATIVAS A ARRECADACAO DE IMPOSTOS E TRANSFERENCIAS DE IMPOSTOS.
 4. SAUDE
 4.1. APLICACOES MINIMAS EM ACOES E SERVICOS PUBLICOS DE SAUDE (ASPS)
APLICACOES EM SAUDE|VALOR (R$)
1. RECEITA DE IMPOSTOS E TRANSFERENCIAS DE IMPOSTOS|3.095.667,76
2. DEDUCAO DA RECEITA DE FPM, ART. 159, INCISO I, ALINEA D|0,00
3. DEDUCAO DA RECEITA DE FPM, ART. 159, INCISO I, ALINEA E|0,00
4. AJUSTES DA AUDITORIA|0,00
5. BASE DE CALCULO PARA ASPS (1-2-3+4)|3.095.667,76
6. DESPESAS EM ACOES E SERVICOS PUBLICOS DE SAUDE ASPS|684.623,10
7. RESTOS A PAGAR INSCRITOS NO EXERCICIO SEM DISPONIBILIDADE FINANCEIRA
0,00
DE RECURSOS DA SAUDE
8. DEDUCAO DAS DESPESAS NAO APLICADAS EM ASPS EM EXERCICIOS
0,00
ANTERIORES (ART. 25 DA LC NO 141/2012)
9. AJUSTES DA AUDITORIA|-98.566,14
10.TOTAL DAS APLICACOES EM SAUDE (ASPS) (6-7-8+9)|684.623,10
11. PERCENTUAL DE APLICACAO EM SAUDE (10/5*100)|18,93%
FONTE: BALANCETES|DESPESA LIQUIDADA ATE O PERIODO DE ANALISE
NO COMPUTO DAS DESPESAS COM ACOES E SERVICOS PUBLICOS DE SAUDE FORAM CONSIDERADOS AS
 DESPESAS CUSTEADAS COM A FONTE DE RECURSOS 2 - RECEITA DE IMPOSTOS E DE TRANSFERENCIA DE IMPOSTOS
 SAUDE.
AS EXCLUSOES DA AUDITORIA, NO MONTANTE DE R$ 98.566,14, CONSTANTES NO QUADRO ACIMA,
 REFEREM-SE AS DESPESAS PAGAS ATRAVES DAS CONTAS (CAIXA, PROGRAMA DE ASSISTENCIA FARMA,
 ATENCAO BASICA E VIGILANCIA SANITARIA) QUE SAO INDICADAS COMO IMPROPRIAS OU DIVERSAS DAQUELAS
 RELATIVAS A ARRECADACAO DE IMPOSTOS E TRANSFERENCIAS DE IMPOSTOS.
O MONTANTE EFETIVAMENTE APLICADO EM ACOES E SERVICOS PUBLICOS DE SAUDE NO PERIODO
 ANALISADO CORRESPONDEU A 18,93% DA RECEITA DE IMPOSTOS, INCLUSIVE TRANSFERENCIAS, INDICANDO TENDENCIA
 (MANTENDO-SE O MESMO NIVEL DE APLICACAO) AO ATENDIMENTO DO MINIMO EXIGIDO DE 15,00% ESTABELECIDO NO
 NA LEI COMPLEMENTAR NO141/2012.
 5. LEI DE RESPONSABILIDADE FISCAL|LC 101
 5.1. VERIFICACAO DOS LIMITES DE GASTOS COM PESSOAL
PARA EFEITO DE CALCULO, FOI CONSIDERADO O ENTENDIMENTO CONTIDO NO PARECER PN-TC 12/2007, NAO
 COMPUTANDO NO TOTAL DOS GASTOS COM PESSOAL DO PODER EXECUTIVO E LEGISLATIVO O VALOR DAS DESPESAS COM
 OBRIGACOES PATRONAIS.
```

Figura 3 Informações referentes à Saúde no relatório em formato .txt pré-processado

Dos 443 arquivos analisados, foi possível verificar que em 27 deles não foi possível extrair ao menos uma das informações. Além disso, foi observado, de acordo com a primeira métrica de validação descrita na seção anterior, que 23 relatórios apresentaram algum problema. Entretanto, ao analisar esses 23 relatório, é possível observar que o preenchimento por parte do auditor em 12 deles foi feito de forma errônea, como inserção de vírgula em lugar de ponto em números ou cálculo errado de porcentagem, não sendo assim classificado como um erro de extração.

Os relatórios analisados como parte da segunda métrica descrita na seção anterior apresentaram acerto de 100% das informações extraídas. Dessa forma, em 443 relatórios analisados, apenas 38, correspondente a 8,6% do total, apresentaram algum problema que de fato é relacionado a extração de informação

Com a estruturação dos dados, foi possibilitada a construção de diversos painéis de acompanhamento das contas públicas dos municípios paraibanos. Nessa pesquisa foi utilizada a ferramenta Business Intelligence – Tableau, para tabulação e apresentação das informações obtidas a partir da Programação de Linguagem Natural. Na Figura 4, é possível observar, por exemplo, que o município paraibano menos eficiente no que tange a execução orçamentária é São Sebastião do Umbuzeiro, com despesas que alcançam mais que o dobro do valor das receitas

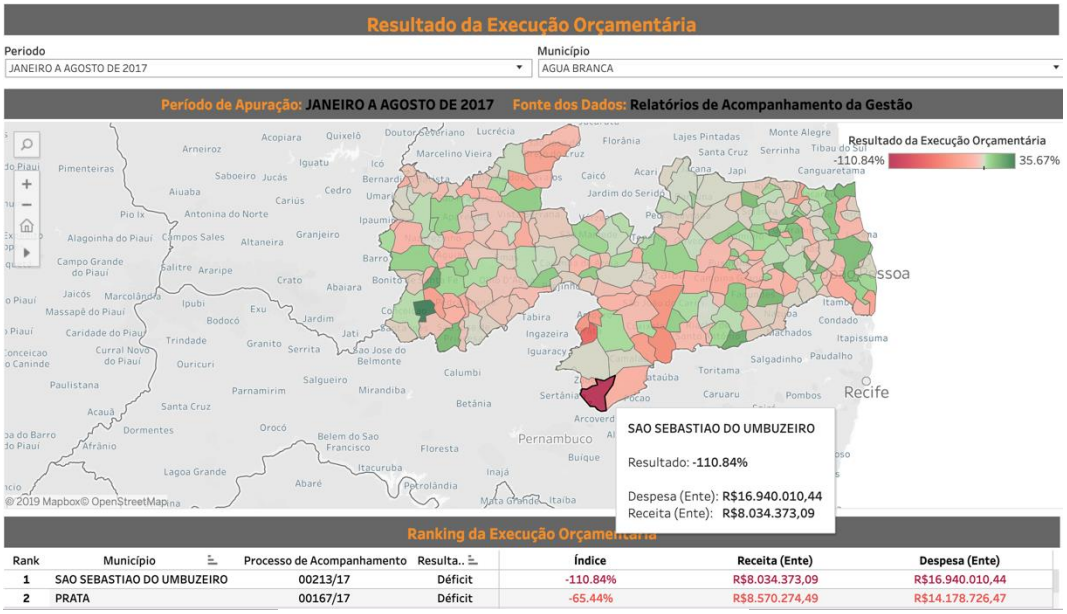


Figura 4 Execução Orçamentária na Paraíba

Em contrapartida, no que tange as aplicações em Saúde e Educação, destacadas nas Figuras 5, 6 e 7, grande parte dos municípios paraibanos seguem as determinações definidas pela lei. Apesar disso, algumas prefeituras merecem atenção especial, como Pocinhos, Cacimba de Areia e São Miguel de Taipu, no quesito Aplicações do FUNDEB; Cubati, Conde e Campim, no quesito Aplicações do MDE; e Boa Vista, Bom Sucesso e Marizópolis, no quesito Aplicações em Saúde.

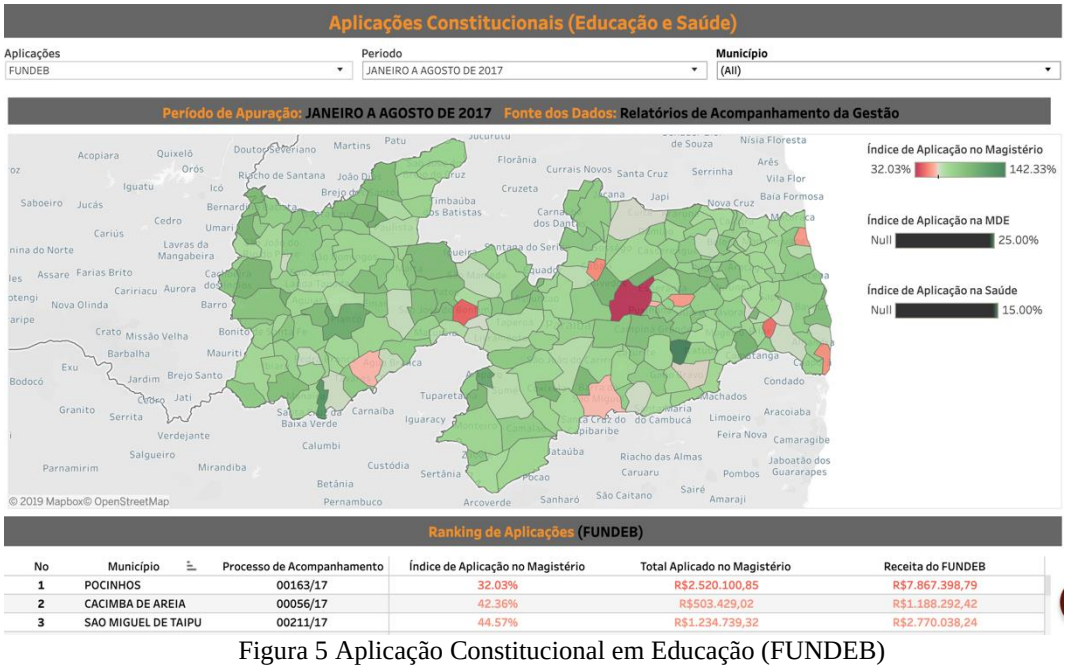


Figura 5 Aplicação Constitucional em Educação (FUNDEB)

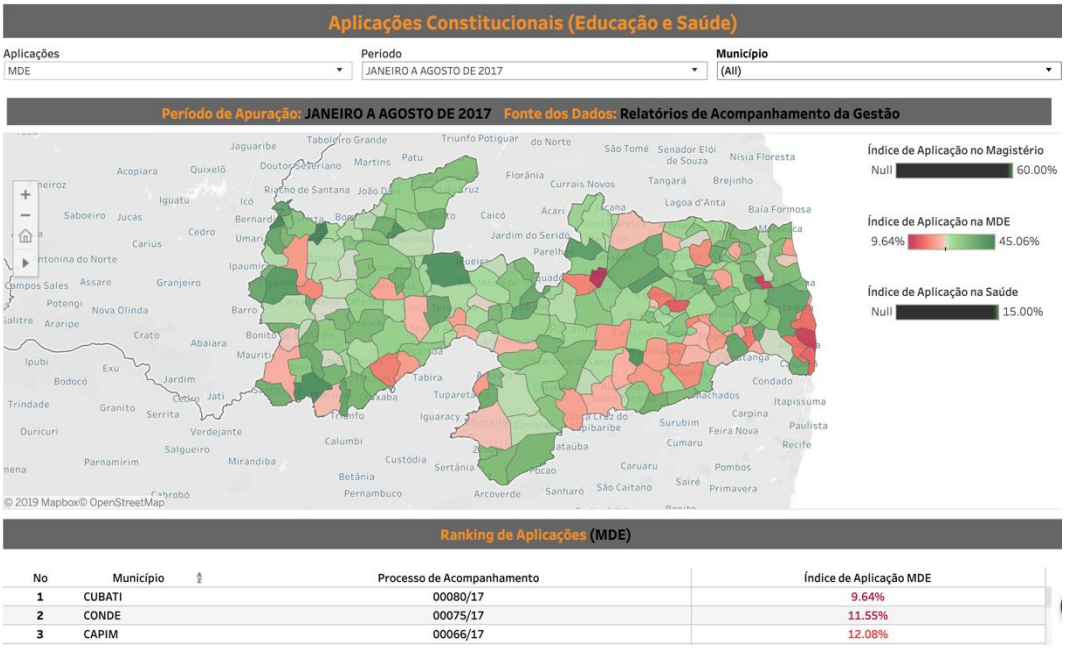


Figura 6 – Aplicação Constitucional em Educação (MDE)

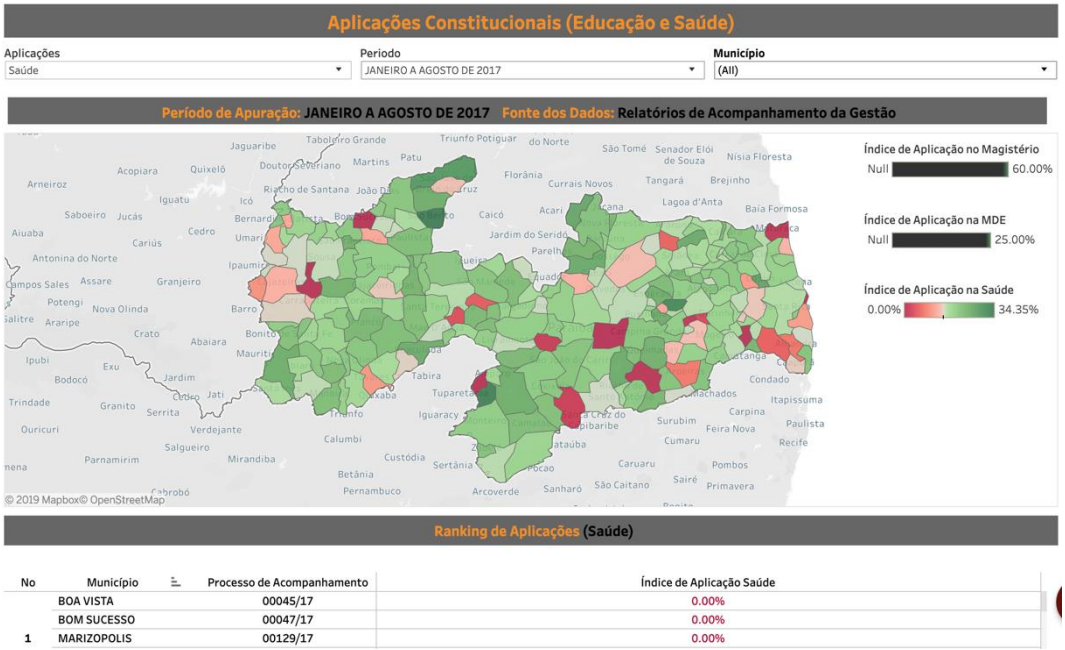


Figura 7 – Aplicação Constitucional em Saúde

5. CONCLUSÕES

As informações geradas pelo modelo PLN mostrou que é essencial, a partir de um texto que apresentam muitas páginas, sumarizar as informações em quadros e gráficos que facilitam a compreensão dos principais aspectos das contas governamentais apreciadas.

Em outro momento da história seria inimaginável que uma máquina pudesse ler um texto longo, apreender como o ser humano (auditor) dispões as informações das PCAs no

relatório de auditoria e em seguida catalogar essas informações em gráficos e tabelas, facilitando muito compreensão da contas da gestão pública.

Assim, como foi desenvolvido no processo metodológico neste trabalho, foi possível verificar que uma pequena amostra de 7,9% da base de relatórios resultou em regras, aplicadas por meio de expressões regulares, capazes de extrair mais de 91% das informações presentes nos Relatórios de Acompanhamento da Gestão Municipal.

Além desses aspectos deve-se ainda levar em consideração que essas regras podem ser ampliadas com a observação e identificação das causas dos erros, podendo, em uma estimativa otimista, extrair até 100% das informações. Ainda, tal sistema obteve um tempo de processamento ínfimo, sendo a maior parte (cerca de 97,7%) desprendida na conversão de relatórios no formato .pdf para .txt, dos quais apenas 25,4% faziam parte do grupo de relatórios de interesse.

Acrescente-se que a metodologia, apesar de não usar técnicas recentes de Aprendizagem de Máquina, se mostrou um instrumento eficiente que pode ser utilizado para verificar as informações inseridas nos relatórios, garantindo a qualidade destes. Pode-se afirma que o sistema apresenta boa acurácia e eficiência, sendo uma ferramenta importante na democratização da informação acerca dos gastos públicos, potencializando sua relevância quando aliada a ferramentas de visualização dados.

Dessa forma, auditores de contas públicas podem se munir de ferramentas que podem trazer agilidade em suas análises, podendo definir como estratégia operacional atacar os casos mais críticos, por exemplo. Da mesma forma, a sociedade no geral também terá facilidades ao fiscalizar se o dinheiro público está sendo investido de forma correta e condizente com a realidade do município.

Referências

Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., & Tyson, M. (1993, August). FASTUS: A finite-state processor for information extraction from real-world text. In *IJCAI* (Vol. 93, pp. 1172-1178).

Backurs, A., & Indyk, P. (2016, October). Which regular expression patterns are hard to match?. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)* (pp. 457-466). IEEE.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, *9*(2), 48-57.

Cruz, C. F., Ferreira, A. C. D. S., Silva, L. M. D., & Macedo, M. Á. D. S. (2012). Transparência da gestão pública municipal: um estudo a partir dos portais eletrônicos dos maiores municípios brasileiros. *Revista de Administração Pública*, *46*(1), 153-176.

Du, M., Pivovarova, L., & Yangarber, R. (2016, July). PULS: natural language processing for business intelligence. In *Proceedings of the 2016 workshop on human language technology* (pp. 1-8). Go to Print Publisher.

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 157-214.

Francis, J., LaFond, R., Olsson, P., & Schipper, K. (2005). The market pricing of accruals quality. *Journal of accounting and economics*, *39*(2), 295-327.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267-297.

Longo, C. G., Jiménez, A., & Marcos, S. (2000). *Manual de Auditoria E Revisão de Demonstrações Financeiras: Novas Normas Brasileiras E Internacionais de Auditoria* . Editora Atlas SA.

Mendel, T. (2009). *Liberdade de informação: um estudo de direito comparado*. Unesco.

Nian, X., Zimmerman, D. X., Mccoy, M., & Mar, S. (2017). Intelligent assessments: government auditors are using cognitive technology to help identify high-risk areas. *Internal Auditor*, *74*(1), 16-18.

Pereira, J. M. (2010). *Governança no setor público*. Editora Atlas.

Raphael, J. (2017). Rethinking the Audit: Innovation Is Transforming How Audits Are Conducted-and Even What It Means to Be an Auditor. *Journal of Accountancy*, *223*(4), 28.

Schatsky, D., Muraskin, C., & Gurumurthy, R. (2015). Cognitive technologies: The real opportunities for business. *Deloitte review*, *16*, 115-129.