Alocação de Tópicos Latentes — Um Modelo para Segmentação de Dados de Auditoria do Governo de PE

Allocation of Latent Topics - A Model for Segmentation of PE Government Audit Data

João Alberto da Silva Amaral [1] orcid.org/0000-0002-8141-4787
Jairson Barbosa Rodrigues [2] orcid.org/0000-0003-1176-3903

[1] Secretaria da Controladoria do Estado de Pernambuco, Recife, Brasil,
[2] Universidade Federal do Vale do São Francisco, Petrolina, Pernambuco, Brasil.

E-mail do autor principal: João Alberto da Silva Amaral jasa@ecomp.poli.br

A modernização da Gestão Pública trouxe para os órgãos de Governo o desafio de adequar seus processos de auditoria e fiscalização dos recursos públicos para melhor aplicação das informações disponíveis nas bases de dados. A utilização de técnicas que possibilitem a identificação de padrões e permitam descobrir ou prevenir atos de improbidade dos agentes públicos, tornou-se indispensável face ao volume de dados relevantes disponível. O fato de grande parte destes dados se apresentarem em formato textual demanda a utilização de técnicas inteligentes de mineração de texto, para atender a demanda do Controle Interno Estadual. Este artigo apresenta um estudo de caso da técnica de Alocação de Tópicos Latentes (*Latent Dirichlet Allocation - LDA*) aplicada sobre um conjunto de dados composto por mais de 65 mil registros de itens comprados pelo Governo Estadual entre os anos 2008 a 2017. O objetivo foi segmentar os itens adquiridos pelo Governo do Estado, aproximando-os a partir de características extraídas das suas descrições, procurando assim disponibilizar informações úteis às ações de controle. A técnica se mostrou eficaz para detectar tópicos em uma granularidade maior que a classificação humana.

Palavras-Chave: Auditoria Governamental; Controle Interno, Mineração Textual; LDA; Modelagem de Tópicos.

Abstract

The public management modernization brought a challenge to Government entities: to adjust their auditing and inspection processes to better apply the information available in databases. Techniques of patterns identification to discover or prevent misconducting acts of public agents have become indispensable. Due to the significant amount of data is presented in unstructured textual format requires intelligent text mining techniques to meet the demands of State Internal Control. This paper presents a case study of the Latent Dirichlet Allocation (LDA) technique applied to a data set composed of more than 65,000 records of items purchased by the State Government between 2008 and 2017. The objective was to segment the items bringing them closer to the features extracted from their descriptions and provide useful information to control actions. The technique showed effective in detecting topics at a higher granularity than human classification.

Key-words: Government Audit; Internal Control; Text Mining; LDA; Topic Modeling.

DOI: 1025286/repa.v5i1.1179

1 INTRODUÇÃO

A fiscalização das ações de Governo e o combate a ilícitos são atribuições do sistema de controle interno previsto no Art. 70 da Constituição Federal do Brasil **[1]**. Ainda, o surgimento da Lei de Responsabilidade Fiscal (LRF) **[2]** reforça e amplia as competências deste sistema, que em sua essência tem a atribuição de fiscalizar os atos da gestão pública quanto à legalidade, legitimidade, economicidade, aplicação das subvenções e renúncia de receitas.

Em Pernambuco, o órgão responsável pela gestão do sistema de controle interno é a Secretaria da Controladoria Geral do Estado (SCGE). A modernização da Gestão Pública, através da implantação de sistemas informatizados para execução das mais diversas atividades, impôs um novo desafio ao Governo: como utilizar as informações coletadas para gerar conhecimento e melhor atender as necessidades da sociedade?

Entre os anos 2016 e 2017, a despesa estadual executada ultrapassou 30 (trinta) bilhões de reais. Esta despesa é executada de forma descentralizada. Cada uma das 66 unidades gestoras executam seus processos licitatórios e são responsáveis por todas as etapas da execução da despesa pública.

A SCGE utiliza os dados disponíveis nos sistemas corporativos do Governo Estadual para planejar e executar auditorias e fiscalizações dos recursos públicos. A eficiência das equipes envolvidas nestas atividades está diretamente relacionada com a qualidade das informações contidas nestes bancos de dados.

1.1 Definição do Problema

O Sistema de Gestão do Banco de Preços (GBP) é o responsável pelo cadastro e gerenciamento dos itens comprados pela administração pública estadual. No passado, o cadastro de novos itens adquiridos era realizado de forma descentralizada, permitindo que itens semelhantes fossem cadastrados livremente, resultando em inserção inadequada. Tais incoerências prejudicam a comparação de valores em novas compras, bem como, a identificação de outliers em compras já realizadas, sendo esta última, uma das principais ações de controle executada pela equipe de auditoria da SCGE. Atualmente, após mudança no processo, este cadastro ocorre de forma centralizada.

Visando corrigir estas inconsistências, o Estado, através da Secretaria de Administração (SAD), iniciou atividade para tratá-las. No entanto, o trabalho não tem sido auxiliado por nenhuma técnica apurada, consistindo de intervenções manuais.

O grande volume de dados, associado com a multidisciplinaridade dos itens adquiridos, fazem deste um problema de difícil resolução. Este trabalho apresenta uma metodologia baseada em mineração textual para identificar características relevantes em cada item cadastrado (documento), aproximando-os de forma a agrupar itens que atendem a uma mesma necessidade.

1.2 Objetivo Geral

Segmentar itens de compra adquiridos pelo Governo do Estado, aproximando-os a partir de características extraídas das suas descrições.

1.2.1 OBJETIVOS ESPECÍFICOS

- Desenvolver metodologia de exploração de textos para agilizar ações de auditoria;
- Categorizar itens de compra em agrupamentos semânticos mais específicos.

1.3 Trabalhos Relacionados

No contexto de pesquisas relacionadas à utilização de algoritmos de mineração de texto para auxiliar as atividades de auditoria, destacam-se os trabalhos de **[3][4].**

Em **[3]** é apresentado o resultado da aplicação do algoritmo de descoberta não supervisionada de clusters *EM* (*Expectation-Maximization*) **[5]**, utilizando a ferramenta *WEKA* [1] para identificar padrões de comportamento nos dados de processos licitatórios que possam apontar eventuais irregularidades cometidas por empresas e assim subsidiar ações de auditoria governamental.

O estudo **[4]** apresenta análise comparativa dos algoritmos *Naïve Bayes, Naïve Bayes* Multinomial – Frequência Inversa e Similaridade para encontrar o algoritmo mais indicado para classificar sentenças como evidências de irregularidades ou não.

Nos trabalhos **[6] [7]** técnicas de Inteligência Artificial são aplicadas para o cálculo do preço médio de produtos adquiridos pelo poder público. Em **[6]** foram utilizadas técnicas de detecção de anomalia implementadas em R, enquanto que em **[7]** foi

[1] *Weka(Waikato Environment for Knowledge Analysis)* – ferramenta que agrega implementações de algoritmos provenientes de diferentes abordagens/paradigmas na sub-área da inteligência artificial dedicada ao estudo de aprendizagem de máquina.

utilizado o classificador *J48*, baseado em árvores de decisão na ferramenta *WEKA*. A técnica foi aplicada para classificar os itens adquiridos a partir de suas descrições, divergindo deste trabalho pela técnica aplicada e pela limitação do escopo – identificar os empenhos destinados às aquisições dos objetos de gastos: gêneros Alimentícios, Combustível, Material de Expediente, Peças Automotivas e Material para Obras.

Este artigo traz uma contribuição adicional ao apresentar um estudo de caso de segmentação de itens de compra por proximidade semântica utilizando modernas técnicas de mineração textual sobre dados de auditoria governamental.

2 BASE CONCEITUAL

Nesta seção, serão apresentados conceitos necessários para o entendimento do problema, tais como: auditoria governamental, mineração textual e algoritmos de descoberta de tópicos, bem como as técnicas de estimativa de partições e avaliação da qualidade da separação.

2.1 Auditoria Governamental

Conjunto de técnicas aplicadas sobre determinadas atividades, objetivando constatar sua conformidade com as normas, regras, orçamentos e objetivos. Sua execução está diretamente relacionada ao acompanhamento das ações efetuadas pelos órgãos e entidades que compõem a administração direta e indireta das três esferas de Governo. Constitui-se em um importante instrumento de controle para garantir melhor alocação de recursos públicos, transparência, prevenção e combate à corrupção **[8]**.

2.2 Clusterização

Segundo **[9]**, clusterização é a tarefa na qual se procura identificar um conjunto finito de categorias, ou *clusters,* para descrever uma informação, sem a existência de um professor ou crítico que indique o que cada padrão representa.

2.3 Mineração Textual

Do inglês, *Knowledge Discovery in Text (KDT)*, é o processo de descoberta de conhecimento potencialmente útil e não trivial, previamente desconhecido, em bases de dados desestruturadas **[10]**.

Trata-se de um campo multidisciplinar, envolvendo recuperação de informação, análise textual, extração de informação, agrupamento, categorização, visualização e tecnologias de bases de dados. Abrange as seguintes fases **[9]**:

- **Coleta de Documentos:** definição dos documentos relevantes para o negócio. É a etapa responsável pela escolha da fonte mais confiável e das colunas mais relevantes em um conjunto de dados;
- **Pré-processamento**: os documentos serão processados, para que seja definida uma estrutura, a qual será utilizada na próxima etapa;
- **Extração de Conhecimento:** serão utilizadas técnicas para detectar os padrões não visíveis nos dados;
- **Avaliação e Interpretação dos Resultados**: os usuários utilizarão o conhecimento gerado para apoiar as suas decisões.

2.3.1 PRÉ-PROCESSAMENTO

Etapa realizada imediatamente após a fase de *coleta* para definir uma estrutura para a massa textual. Pré-processar textos é, por muitas vezes, o processo mais oneroso do *KDT*, uma vez que não existe uma única técnica que possa ser aplicada para a obtenção de uma representação satisfatória em todos os domínios **[11]**.

A definição da estrutura aplicada aos documentos impactará diretamente os resultados do trabalho em andamento, ou seja, as próximas etapas serão diretamente afetadas pelas técnicas utilizadas aqui. Foram selecionadas as seguintes técnicas de pré-processamento:

- ***Tokenização*** **(Atomização):** consiste na divisão de um texto em um conjunto de termos **[12]**. Nesse passo, são removidos caracteres especiais e pontuações, dado que não contribuem para classificação. Além disso, os caracteres maiúsculos são convertidos para minúsculos **[4]**;
- **Radicalização (*Stemming*)**: palavras variantes morfologicamente serão combinadas em uma única representação, o radical **[12]**. Por exemplo, os termos *livro* e *livrinho* são reduzidas para o radical *livr*;

DOI: 1025286/repa.v5i1.1179

- **Remoção de *Stop Words***: nesta etapa os termos com pouca ou nenhuma relevância serão removidos. Estas palavras são, em sua maioria, pronomes, preposições, artigos, numerais e conjunções **[13]**;

palavras quando analisados, dão indícios da existência de um tema ou assunto que ocorre em um subconjunto de documentos. A expressão tópico é usada levando-se em conta que o assunto tratado em uma coleção de documentos é extraído

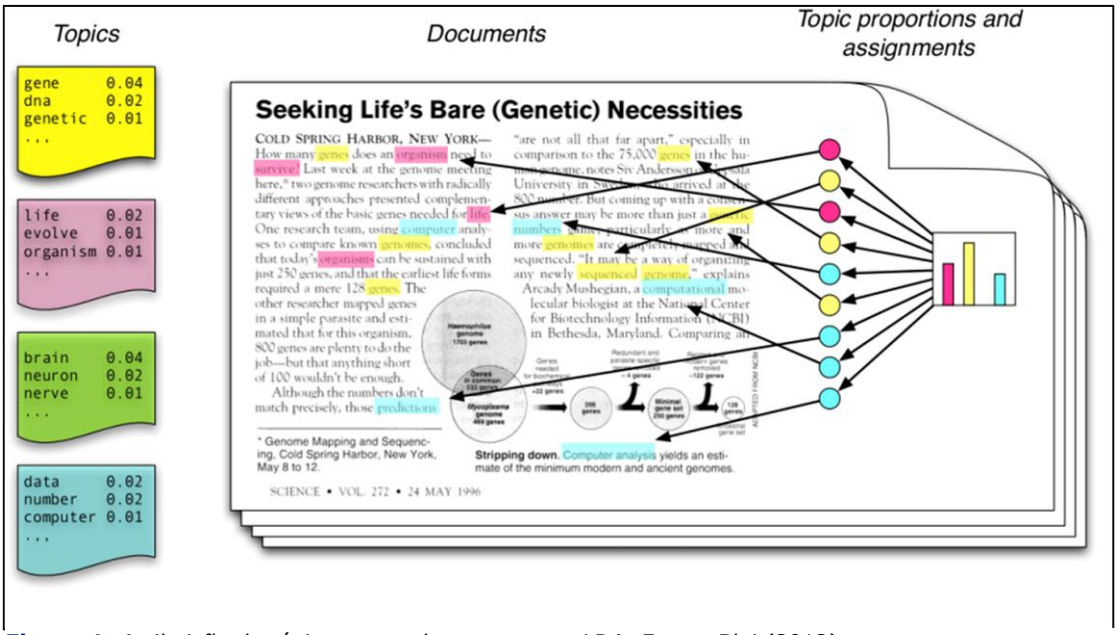

Figura 1: Atribuição de tópicos a um documento em LDA. Fonte: Blei (2012)

- **Correção Ortográfica:** erros ortográficos são comuns quando se trabalha com grandes massas de dados, especialmente se geradas a partir de digitação manual por seres humanos **[11]**. Para auxiliar o trabalho de identificação destes erros foram utilizadas a matriz de frequência dos termos (*TF*), bem como a seleção dos termos para cada tópico da LDA (Seção 2.4).

- **Indexação:** atribui-se uma pontuação para cada termo, garantindo uma única instância do termo no documento **[10]**.

2.4 Latent Dirichlet Allocation (LDA)

Os modelos probabilísticos de tópicos (*Probabilistic Topic Models*) **[14]** surgiram como nova área de pesquisa, a partir de 2003, motivados pela necessidade de maior eficiência na extração de informações textuais.

A exploração de grandes volumes de dados é simplificada pelos modelos probabilísticos na descoberta dos tópicos. Os tópicos são estruturas com valor semântico que formam grupos de palavras que frequentemente ocorrem juntas. Esses grupos de automaticamente, ou seja, um tópico é definido como um conjunto de palavras que frequentemente ocorre em documentos semanticamente relacionados **[15]**.

O *Latent Dirichlet Allocation (LDA)* é um algoritmo para modelagem de tópicos que utiliza a abordagem bayesiana para aprender a estrutura latente de tópicos que compõem um documento e segmentar uma coleção de documentos a partir de tópicos dominantes em cada documento submetido. Baseia-se em um modelo generativo assumindo que cada documento em uma coleção é uma mistura de tópicos.

Ou seja, aplicando uma técnica como a LDA, é possível organizar e agrupar um subconjunto de documentos pelos seus respectivos temas. O modelo é baseado em duas premissas:

a) **Todo tópico é uma mistura de palavras**: cada tópico é constituído por palavras relevantes identificadas nos documentos de uma coleção. Por exemplo, o tópico A pode ser constituído pelos termos palavra I, palavra II e palavra III. Enquanto o tópico B, de uma mesma coleção, pode ser composto pelos termos palavra III, palavra IV e palavra V. É importante ressaltar que os termos podem ser compartilhados entre os tópicos (ver Figura 1**)**.

b) **Todo documento é uma mistura de tópicos:** cada documento pode conter palavras de vários tópicos em proporções específicas. Por exemplo, um determinado documento em uma coleção de produtos poderia ser composto por uma mistura de 50% sobre o tópico A, 30% sobre B e 20% sobre C; enquanto outro documento poderá ser composto por 30% sobre o tópico A, 20% sobre B, 30% sobre C, e 20% sobre D (ver Figura 1).
Um modelo *LDA* é definido por dois parâmetros:

- **α:** a frequência média da ocorrência de cada tópico em um determinado documento. Um alto valor de α significa que cada documento provavelmente conterá uma maior mistura de tópicos. Um valor baixo para α indica maior probabilidade de os documentos conterem mistura de poucos tópicos.
- ***β:*** a distribuição de palavras por tópico. Da mesma forma, um valor alto para β significa que cada tópico terá alta probabilidade de conter misturas de várias palavras. Um valor baixo sugere que o tópico será formado por poucas palavras.

Os tópicos são $\beta_{1:k}$, onde cada β_k é uma distribuição sobre o vocabulário fixo. As proporções dos tópicos para o d-ésimo documento são θ_d, onde $\theta_{d,k}$ é a proporção do tópico *K* no documento *d*. As atribuições de tópicos para o d-ésimo documento são z_d, onde $z_{d,n}$ é a atribuição do tópico para a n-ésima palavra no documento d. Finalmente, as palavras observadas para o documento d são w_d, onde $w_{d,n}$ é a n-ésima palavra no documento d, a qual é um elemento do vocabulário fixo. Com essa notação, o processo generativo em LDA corresponde à distribuição conjunta das variáveis observadas e ocultas representada na formulação do processo generativo da LDA, sob a forma de uma distribuição **[17]**.

No treinamento do modelo, o objetivo é encontrar parâmetros α e β, que maximizem a probabilidade do *Corpus*[2] ser gerado pelo modelo. A LDA pode ser formalmente descrita através da Equação 1.

$$p(\beta_{1:k}, \theta_{1:k}, z_{1:k}, w_{1:k}) = \prod_{i=1}^{k} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(Zd, n, |\theta d)\, p(wd, n \mid \beta 1{:}k, Zd, n) \right) \quad (1)$$

2.4.1 INFERÊNCIA DA LDA VIA MÉTODO GIBBS

O problema computacional definido pela LDA reside na inferência de estruturas temáticas que correspondem às distribuições de probabilidades relacionadas com as relações documento-tópico e tópico-palavra **[15]**.

Dentre os métodos de inferência, o método *Gibbs* é o mais popular, principalmente pela facilidade de implementação e sua aplicação em diversos problemas **[16]**. O método é um caso especial da simulação de Monte Carlo em Cadeia de *Markov*, capaz de emular distribuições de probabilidades com alta dimensionalidade **[15]**.

2.4.2 SELEÇÃO DO NÚMERO DE TÓPICOS

O modelo LDA pode capturar correlações de palavras em uma coleção de documentos textuais com um conjunto de distribuição multinominal de baixa dimensão, chamado tópicos. No entanto, é importante selecionar o número apropriado de tópicos para um conjunto de dados específico **[18]**.

Dentre as estratégias para obter tal estimativa destacam-se as técnicas que tentam maximizar ou minimizar P (**w** | φ, θ) para encontrar estimativas de máxima ou mínima verossimilhança de ϕ e θ. Onde W é o conjunto de palavras do vocabulário; θ a distribuição de tópicos por documentos; e φ a distribuição dos tópicos sobre as palavras de todo o vocabulário. Estas técnicas podem ser conferidas em **[18-21]**.

Foram selecionadas quatro técnicas para estimativa do número de tópicos nos documentos analisados: *Griffiths2004* **[20]** e *Deveaud2014* **[21]** (métricas de maximização; *CaoJuan2009* **[18]** e *Arun2010* **[19]** (métricas de minimização).

2.4.3 MÉTRICAS DE AVALIAÇÃO

a) Análise Humana

Nesta técnica, é utilizado o conhecimento do especialista para verificar se os documentos de um

[2] Corpus - é um conjunto de documentos, onde cada documento é uma mistura de tópicos e cada tópico é uma mistura de palavras.

DOI: 1025286/repa.v5i1.1179

mesmo tópico foram segmentados de forma coerente, ou seja, se valor semântico entre documentos de um mesmo agrupamento. Para auxiliar a análise humana aplicamos uma técnica de identificação dos tópicos, que consiste em utilizar as duas palavras mais relevantes de cada tópico para intitulá-lo **[22]**.

b) Tamanho dos Tópicos:

Uma forma de avaliação utiliza o tamanho dos tópicos como métrica, onde os agrupamentos que possuem mais documentos associados são melhores **[22]**.

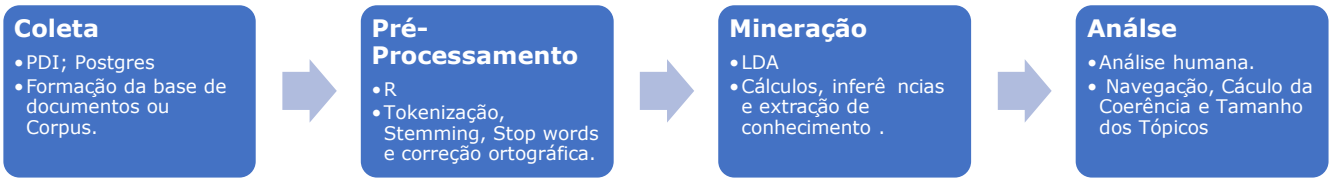

Figura 2: Diagrama ilustrativo dos métodos e materiais aplicados. *Fonte: o Autor,2018.*

No entanto, esta conclusão entra em conflito com a possibilidade de utilização do modelo para encontrar agrupamentos especializados e, portanto, com poucos documentos relacionados.

c) Coerência dos Tópicos:

Os tópicos obtidos como saída da LDA são formados por palavras estatisticamente relacionadas. No entanto, isto não significa que elas possuam relacionamento semântico claro entre si, ou seja, que façam sentido para um avaliador humano que as leia. As medidas de coerência atribuem um valor ao grau de similaridade semântica, de forma a tornar possível a automatização do processo de validação daqueles tópicos **[23]**. Ou seja, a partir do cálculo da coerência é possível verificar, sem a atuação humana, a existência de valor semântico nos tópicos. Calculamos a coerência através da função *FitLdaModel* disponível na biblioteca *textmineR* **[24]**, na linguagem R.

3 MATERIAIS E MÉTODOS

Esta seção descreve as etapas executadas neste trabalho, conforme Figura 2.

3.1 Coleta de Documentos

Foi utilizada a ferramenta *Pentaho Data Integration (PDI)* para auxiliar no processo de *Extract, Transform and Load* (ETL[3]). Os dados foram extraídos do Portal da Transparência de Pernambuco e carregados no SGBD *Postgres*. Nesta etapa, foram selecionados 65.461 (sessenta e cinco mil quatrocentos e sessenta e um) documentos, que correspondem aos diferentes itens adquiridos pelo Governo do Estado de Pernambuco entre os anos de 2008 a 2017.

A base de dados selecionada possui 63 atributos, em sua maioria com informações relacionadas ao processo de compra. Por exemplo: dados sobre a entidade compradora, o fornecedor, valor praticado, número do processo licitatório, entre outros. As informações relacionadas ao item adquirido se restringem a quatro atributos: descrição, nome, classe e grupo.

A partir da análise exploratória dos dados, foi selecionada apenas a descrição dos itens comprados como documento submetido às próximas etapas do presente trabalho. A decisão foi tomada para reduzir impacto dos erros humanos no processo de cadastro de cada processo de compra e para permitir a comparabilidade de itens adquiridos por outras entidades, uma vez que os demais atributos não teriam sua correlação assegurada em outras bases de dados.

Esta escolha trouxe também um desafio ao modelo, o tamanho dos documentos analisados varia entre textos curtos com apenas cinquenta caracteres e medianos com até dez mil caracteres.

3.2 Pré-processamento

Considerando que os documentos coletados nesta pesquisa foram redigidos por diferentes atores, sem supervisão ou controle de qualidade, foram aplicadas

[3] ETL é o processo que trata da sistematização da Extração de dados de fontes externas, Transformação para atender às necessidades de negócios e Carga dos dados dentro de um SGBD para posterior utilização.

as técnicas de pré-processamento: *tokenização*, *stemming*, indexação, correção ortográfica e remoção de *stop words*. Todas estas técnicas foram implementadas na linguagem R.

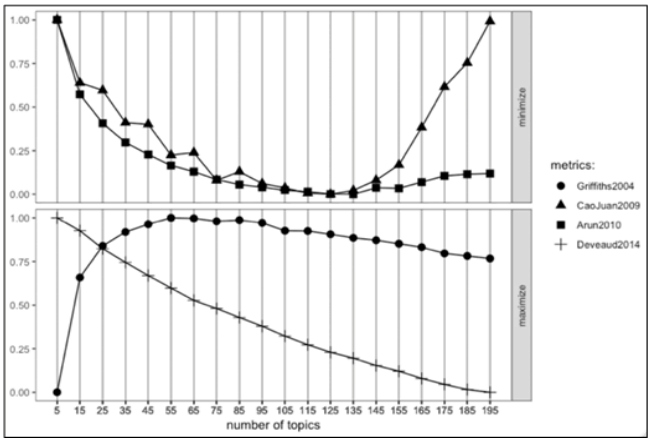

Figura 3: Resultado das métricas para escolha do K (número de tópicos) ideal. Fonte: o Autor, 2018.

Após a *tokenização* foi aplicada a técnica de indexação, onde, por meio do *TF-IFD* **[13]**, as palavras de cada documento receberam um valor indicando sua relevância. Esta informação, associada ao conhecimento sobre o domínio do negócio, auxiliou na definição das *stop words.* Na sequência, com o auxílio da frequência das palavras, cruzada com a informação do *TF-IDF*, foram corrigidas palavras relevantes que se encontravam com grafia errada.

3.3 Extração de Conhecimento

Nesta etapa foi utilizado a LDA (Seção 2.4), implementado na Linguagem R, disponível na biblioteca *topicmodels* **[25]**.

3.4 Avaliação e Interpretação dos Resultados

Para a avaliação dos resultados, a análise humana foi aplicada para avaliar os desempenhos das técnicas descritas em b) e c). Foram selecionados vinte tópicos para cada técnica, os dez de melhor pontuação e os dez de pior, nas respectivas métricas.

Outros dez tópicos foram selecionados aleatoriamente para uma avaliação independente dos critérios descritos nas técnicas já avaliadas. No total cinquenta tópicos avaliados, o que representa 40% do total de tópicos identificados. Esta atividade foi executada pelo autor deste artigo, que atua na SCGE, como Gestor Governamental de Controle Interno, desde 2010.

4 RESULTADOS

Por se tratarem de dados brutos nunca tratados, a etapa de pré-processamento apresentou impacto significativo nos resultados obtidos durante os experimentos realizados. Após esta etapa, a escolha do número de tópicos que seria utilizado na modelagem da LDA foi escolhido com base na análise da Figura 3. Assim foram selecionados 85 e 125 para valores de *K*.

Quadro 1 - *Resumo das coerências dos tópicos*

k-Tópicos	Mínimo	Máximo	Média
85	3,2%	44,8%	14,7%
125	1,5%	58,4%	14,0%

Fonte: o Autor, 2018.

Com base no Quadro 1 observamos que há uma equivalência na média das coerências dos tópicos, no entanto o experimento com maior quantidade de tópicos apresenta também o pior e o melhor tópico. Apesar do experimento com 85 tópicos se aproximar da classificação humana pré-existente na base de dados, que é de 78 classes, optamos por analisar os resultados obtidos no experimento com 125 tópicos, pois este representa um ganho significativo no objetivo deste trabalho.

Quanto ao tamanho dos tópicos (ver Figura 4), foi observado equilíbrio na distribuição dos documentos, onde 60% dos agrupamentos possuem tamanho próximo a 500 documentos.

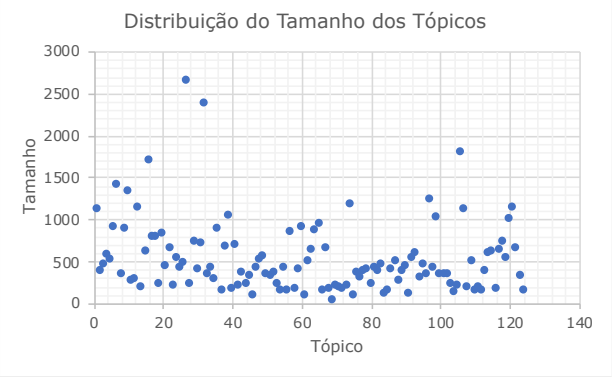

Figura 3: Distribuição do tamanho dos tópicos.
Fonte: o Autor, 2018.

DOI: 1025286/repa.v5i1.1179

Quadro 2 – Seleção dos 5 melhores tópicos pelo critério da coerência

Nome Sugerido	agulh_descartavel	cab_conexa	manutenca_servic	liqu_ativ	automot_veicul	impressa_resoluca
Palavra 1	agulh	cab	manutenca	liqu	automot	impressa
Palavra 2	descartável	conexa	servic	ativ	veicul	resoluca
Palavra 3	atox	conector	corret	composica	pec	veloc
Palavra 4	protetor	par	equip	álcool	original	folh
Palavra 5	sering	red	pec	basic	ano	minim
Palavra 6	luer	mach	prevent	frasc	genuin	impressor

Fonte: o Autor, 2018

Quanto a distribuição das coerências (ver ***Figura 4***) observamos uma maior concentração dos tópicos para uma coerência inferior a 0,2 (20%), a figura demonstra o resultado onde 80% dos tópicos apresentaram coerência inferior a 20%. Neste experimento obtivemos o melhor tópico, *agulha_descartavel (*ver Quadro 2) com coerência igual a 0,584 (58,4%).

Apesar de termos observado a relação direta entre a qualidade do valor semântico dos tópicos e sua coerência, em nossa análise exploratória, identificamos tópicos com melhor valor semântico que obtiveram valores inferiores nos cálculos de suas coerências. Como exemplo, o tópico "Extintor_Incendio", apresentou coerência igual a 4,31% e tamanho igual a 239.

Quadro 3: Exemplos de inconsistências

Item	Documento	Tópico
1	COCO SECO - GRANDE	Extintor_Incendio
2	COLA PARA FORMICA – RESINA FENOLICA, PARA SER UTILIZADA EM FORMICA, BORRACHA, NA COR AMBAR, APRESENTADO COMO LIQUIDA, APLICACAO POR PINCEL OU ESPATULA, EMBALADO EM 740 GRAMAS.	Agulha_descartavel

Fonte: o Autor, 2018

Na análise humana, identificou-se apenas vinte documentos agrupados inadequadamente. O resultado equivale a 8% do total de documentos pertencentes a este tópico. A mesma análise foi realizada para o tópico "Agulha_descartavel" e o resultado foram trinta e três documentos agrupados inadequadamente de um total de 311. O resultado, 10%, sugere uma leve vantagem para o tópico "Extintor_Incendio" segundo a análise humana, mesmo havendo diferença nos valores das respectivas coerências.

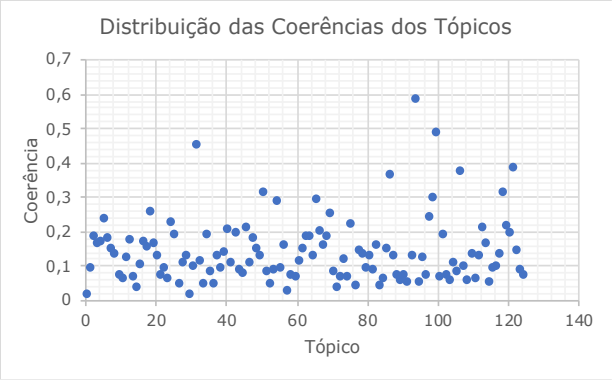

Figura 4: Distribuição das coerências dos tópicos
Fonte: o Autor, 2018.

O Quadro 3 apresenta alguns exemplos das incoerências verificadas nos agrupamentos, no item 1 um documento com tamanho inferior a 20 caracteres foi alocado no tópico Extintor_Incendio. Enquanto o item 2, cola para fórmica, está agrupado em Agulha_descartavel. Os exemplos apresentados refletem a fragilidade do modelo em trabalhar com documentos muito pequenos.

A LDA apresentou bom desempenho quando submetido ao agrupamento de documentos com formação consistente e tamanho adequado, no Quadro 4 é fácil perceber a existência de um valor semântico entre os documentos de um mesmo grupo.

5 CONCLUSÕES

Este artigo descreveu a aplicação de um modelo de Alocação de Tópicos Latentes aplicado sobre a base de dados de itens adquiridos pelo Governo do Estado de Pernambuco, com o objetivo de agrupá-los a partir de características extraídas de suas descrições.

Quadro 4: Exemplos de tópicos

Item	Documento	Tópico
1	RECARGA PARA EXTINTORES – GAS CARBONICO (CO2).	Extintor_Incendio

2	EXTINTOR DE INCENDIO A PO QUIMICO SECO – DO TIPO VEICULAR E MARITIMO, CLASSE ABC, 1ª-5B, (GARANTIA DE 05 ANOS), 1KG, PRESSURIZADO.	Extintor_Incendio
3	DISCO DE SINALIZACAO PARA EXTINTOR – EM ALUMINIO, 25CM DE DIAMETRO, CIRUCULO INTERNO NA COR AMARELA, COM INDICACAO FONE 193 E CIRCUNSCRITO POR OUTRO NA COR VERMELHA, USO EM EXTINTOR TIPO CO2, INCLUSIVE PARAFUSOS E BUCHAS.	Extintor_Incendio
4	SERINGA DESCARTAVEL – EM PLASTICO, ATOXICO, APIROGENICO INTEGRO E TRANSPARENTE, APRESENTANDO RIGIDEZ E RESISTENCIA MECANICA NA SUA UTILIZACAO, CORPO COM GRADUACAL MILIMETRADA, EMBOLO BORRACHA ATOXICA NA PONTA, BICO LATERAL LUER, OXIDO DE ETILENOSILICONIZADA, COM CAPACIDADE DE 20 ML, SEM AGULHA, CONF. NBR-09752, ART.31 L.8078/90 E PORT.N.A/96-M.S.	Agulha_descartavel
5	CATETER INTRAVENOSO RADIOPACO ESTERIL – EM POLIURETANO, RADIOPACO, ESTERIL, ACESSO PERIFERICO, DESCARTAVEL, COM FILTRO HIDROFOBO, COM CONECTOR LUER LOCK, E CONECTOR LUER LOCK COM BISEL TRIFACETADO C/CAMARA DE REFLUXO SANGUINEO QUE FACILITE A EMPUNHADURA, COM SISTEMA TRAVA DE SEGURANCA (PROTECAO DA AGULHA), TAMANHO VARIANDO DE ACORDO COM O USO TAMANHO 18G, PADRONIZACAO DE CORES DE ACORDO COM A NORMA ABNT 10555-2, EMBALADO EM TRANSPARENTE, INDIVIDUAL, ATOXICO, ROTULAGEM RESPEITANDO O DECRETO LEI 79094/77 ROTULAGEM RESPEITANDO ACORDO COM LEGISLACAO VIGENTE.	Agulha_descartavel

Fonte: o Autor, 2018

Quanto aos resultados obtidos, as técnicas de pré-processamento se mostraram muito valiosas na manipulação e qualificação dos documentos submetidos. No entanto, o tamanho curto da descrição dos itens analisados prejudicou o desempenho do algoritmo, tornando-se um desafio a ser superado em trabalhos futuros.

A base utilizada possuía uma classificação prévia que sugeria a existência de 78 tópicos. A identificação dos 125 tópicos extraídos pela LDA disponibiliza para gestores e auditores agrupamentos mais especializados. Foram verificadas. No entanto, inconsistências causadas, principalmente, pelo curto tamanho dos documentos submetidos ao modelo.

Ainda, por se tratar de técnica que utilizou apenas dados das descrições dos itens adquiridos, a comparabilidade destes tópicos com outras bases pode trazer grandes benefícios à Administração Pública, pois tornaria viável a comparação de itens independente da estrutura de dados.

Por fim, o Governo do Estado de Pernambuco lançou, em 2018, ferramenta que utiliza as informações contidas nas notas fiscais eletrônicas, emitidas pelo sistema fiscal da Secretaria da Fazenda do Estado, para auxiliar o cidadão a identificar o melhor preço de compra de produtos. Esta funcionalidade, útil para a sociedade, teria grande impacto se aplicada para as compras estaduais. A solução proposta neste trabalho, possibilita o desenvolvimento de uma solução equivalente à disponibilizada ao cidadão e com possibilidade de atualização e verificação do valor de mercado em tempo real.

5.1 Trabalhos Futuros

- Submeter os resultados a um grupo maior de especialistas para expansão da validação subjetiva;
- Aprofundar processo de análise com inclusão de técnicas de visualização de tópicos;
- Aplicar outras técnicas de Mineração Textual para comparar com a LDA;
- Evoluir o trabalho para identificar o preço médio de cada item adquirido pelo Estado e adicionar dados do mercado privado e de outras entidades.

REFERÊNCIAS

[1] BRASIL (País). **Constituição da República Federativa do Brasil**. Senado, Brasília. Disponível em:www.planalto.gov.br/ccivil_03/014.constituicao/constituicao.htm, Acesso em 10/10/2018.

[2] BRASIL (País). **Lei Complementar nº 101, de 4 de maio de 2000**. Estabelece normas de finanças públicas voltadas para a responsabilidade fiscal e dá outras providências. Diário Oficial da União, 05 de maio de 2000.

[3] SILVA, C. V. S.; RALHA, C. G**. Detecção de cartéis em licitações públicas com agentes de mineração de dados**. Revista Eletrônica de Sistemas de Informação, v. 10, n. 1, p. 1-19, 2011.

[4] SANTOS, BRENO SANTANA; *et al.* **Comparing Text Mining Algorithms for Predicting Irregularities in Public Accounts**. Proceedings of

DOI: 1025286/repa.v5i1.1179

the XI Brazilian Symposium on Information Systems (SBSI 2015).

[5] NEAL, R.; HINTON, G. E. **A view of the em algorithm that justifies incremental, sparse, and other variants**. In: Learning in Graphical Models. [S.l.]: Kluwer Academic Publishers, 1998. p. 355–368.

[6] CARVALHO, R. N. **Banco de preços: metodologia para cálculo de preços médios nas compras do governo brasileiro, 2014.**

[7] SOARES, A. M. 2010. **A mineração de texto na análise de contas públicas municipais**. Master's thesis. State University of Ceará, Fortaleza, Brazil.

[8] ARAÚJO, I. DA P. S. **Introdução à auditoria: breves apontamentos de aula aplicáveis à área governamental**. Egba, Salvador, BA. 1998.

[9] REZENDE, S. O. J. B; PUGLIESI, E. A; MELANDA, M. F. P. **Mineração de dados**. In: Sistemas Inteligentes: Fundamentos e aplicações. São Paulo: Manole, 2003.

[10] FELDMAN, R. AND DAGAN, I. 1995. **Knowledge Discovery in Textual Databases (KDT)**. (1995). Retrieved February 1, 2015.

[11] CARRILHO JUNIOR, JOÃO RIBEIRO; PASSOS, EMMANUEL PISECES LOPES (ADVISOR). **Development of a Methodology for text Mining**. Rio de Janeiro, 2007. 96p. MSc Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

[12] WEISS, S. M., INDURKHY, N. and ZHANG, T. 2010. **Fundamentals of predictive text mining**. Springer London, New York, NY.

[13] SÁ, H. R. de. 2008. **Seleção de características para classificação de texto**. Federal University of Pernambuco, Recife, PE.

[14] BLEI, D. M.; NG, A. Y.; JORDAN, M. I. **Latent dirichlet allocation**. J. Mach. Learn. Res., JMLR.org, v. 3, p. 993–1022, mar. 2003. ISSN 1532-4435.

[15] FALEIROS, T. P.. **Modelos probabilísticos de tópicos: desvendando o Latent Dirichlet Allocation**. 2016. – Instituto de Ciencias Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

[16] GEMAN, S.; GEMAN, D. **Stochastic relaxation, Gibbs distributions and the Bayesian resto- ration of images**. IEEE Transactions on Pattern Analysis and Machine Intelligence, Tay- lor & Francis, v. 6, n. 6, p. 721–741, nov. 1984.

[17] BLEI, D. M. **Probabilistic topic models**. Communications of the ACM, 55 (2012).

[18] CAO JUAN, XIA TIAN, LI JINTAO, ZHANG YONGDONG, AND TANG SHENG. 2009. **A density-based method for adaptive LDA model selection**. Neurocomputing — 16th European Symposium on Artificial Neural Networks 2008 72, 7–9: 1775–1781.

[19] ARUN, R.; *et al*. **On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations**. In Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, 391–402, 2010.

[20] DEVEAUD R.; SANJUAN, E.; BELLOT, P. **Accurate and effective latent concept modeling for ad hoc information retrieval**. Document numérique 17, 1: 61–84, 2014.

[21] GRIFFITHS, T. L.; STEYVERS, M. **Finding scientific topics**. Proceedings of the National Academy of Sciences 101, suppl 1: 5228–5235, 2004.

[22] MIMNO, D. *et al.* **Optimizing semantic coherence in topic models**. In EMNLP, 2011.

[23] NEWMAN, D. *et al.* **Automatic evaluation of topic coherence**. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

[24] LDA models parameters tuning, Disponível em: https://github.com/nikita-moor/ldatuning.

[25] Tools and Classes for Statistical Models. Disponível em https://CRAN.R-project.org/package=modeltools .