Inferência de tópicos para identificação de subáreas temáticas de projetos culturais

Nádia Felix Felipe da Silva [I]
http://orcid.org/0000-0002-3875-2211

Núbia Rosa da Silva [II]
http://orcid.org/0000-0003-1982-5144

Kátia Kelvis Cassiano [III]
http://orcid.org/0000-0002-9389-9187

Douglas Farias Cordeiro [IV]
http://orcid.org/0000-0002-5187-0036

[I] *Universidade Federal de Goiás, GO, Brasil.*
Doutorado em Ciência da Computação e Matemática Computacional pela Universidade de São Paulo.

[II] *Universidade Federal de Catalão, PR, Brasil.*
Doutorado em Ciência da Computação e Matemática Computacional pela Universidade de São Paulo.

[III] *Universidade Federal de Goiás, GO, Brasil.*
Doutorado em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro.

[IV] *Universidade Federal de Goiás, GO, Brasil.*
Doutorado em Ciência da Computação e Matemática Computacional pela Universidade de São Paulo.

http://dx.doi.org/10.1590/1981-5344/3500

Os dados abertos governamentais podem ser vistos como uma importante iniciativa de órgãos e instituições da

sociedade civil, voltados à promoção da transparência e permitindo, além disso, sua reutilização como insumo no desenvolvimento de projetos de inovação. Entretanto, é comum que determinados conjuntos de dados demandem a aplicação de tratamentos específicos, para que os mesmos possam ser utilizados de forma mais eficaz, como é o caso da necessidade de classificação destes dados através de Mineração de Dados. Neste cenário, este trabalho apresenta uma proposta de inferência de tópicos automática utilizando o método *Latent Dirichlet Allocation* para a classificação de projetos culturais em áreas temáticas, por meio da identificação da similaridade entre seus dados. Os resultados apresentados demonstram a viabilidade da abordagem no contexto de dados abertos governamentais.

Palavras-chave: *dados abertos governamentais. Inferência de tópicos. Mineração de dados. Projetos culturais.*

Inference of topics with Latent Dirichlet Allocation for Open Government Data

Open government data can be considered as an important initiative of institutions of civil society, promoting transparency and allowing its reuse as an input in the development of innovation projects. However, it is common for certain databases to require the application of specific treatments, so that the data can be used more efficiently, such as the case of classification using Data Mining. In this scenario, this paper presents an automatic topic inference proposal using the Latent Dirichlet Allocation method to classify cultural projects in their thematic areas, by identifying the similarity in their data. The results demonstrate the feasibility of the approach in the context of open government data.

Keywords: *open government data. Topic inference. Data mining. Cultural projects.*

Recebido em 02.04.2018 Aceito em 15.03.2021

1 Introdução

O significado de "aberto", de acordo com *Open Knowledge Foundation* (2010), em termos como "dados abertos" ou "conteúdo aberto", refere-se à facilidade de acesso, uso, modificação e compartilhamento para qualquer propósito. Os dados abertos tem motivado positivamente várias iniciativas e tem sido estimulado por movimentos globais sobre a transparência nos registros dos gastos, culminando no conceito de Dados Abertos Públicos ou Governamentais.

Nessa perspectiva, o governo brasileiro criou o dados.gov.br que disponibiliza dados governamentais segundo os princípios da abertura de dados. A abertura de dados ocorre em vários estágios: coleção, armazenamento, distribuição e reuso. Neste contexto, a distribuição é a fase que requer especial atenção pois implementa o provisionamento dos dados de interesse aos usuários. Portanto, é necessário usar ferramentas de publicação dos dados, denominadas APIs, as quais possibilitam a realização de consultas específicas e direcionadas.

A transparência na Administração Pública, a melhoria da qualidade dos dados governamentais e a criação de novos negócios são bons exemplos de usabilidade deste recurso, mas ainda é um desafio maximizar a compreensão sobre o conteúdo textual disponível.

Neste trabalho é apresentada uma proposta de utilização de ferramentas de mineração de dados (*data mining*) para realizar análise de dados. Empiricamente, este trabalho explora a inferência de tópicos por meio da aplicação do algoritmo *Latent Dirichlet Allocation – LDA* (BLEI, 2003; BLEI, 2012) - em projetos culturais submetidos ao Ministério da Cultura do Brasil através do Salic (Sistema de Apoio às Leis de Incentivo à Cultura).

No escopo da literatura acadêmica, existem vários estudos sobre LDA porém não foram encontradas abordagens que tratem a inferência de tópicos sobre dados abertos governamentais. O objetivo deste trabalho é, então, desenvolver uma solução para análise de dados abertos governamentais, em português, realizando todos os passos de preparação de dados, extração de características, representação e visualização de resultados. Dessa forma, com este estudo, se espera enfatizar os benefícios que as ferramentas automatizadas de mineração de textos podem trazer aos tomadores de decisão.

2. Trabalhos Relacionados

2.1 Dados Abertos Governamentais

Os dados podem ser considerados qualquer elemento que representa algo, e que embora originalmente seja apresentado em um formato aparentemente sem valor, intervenções específicas podem transformá-los em informação, agregando valor e gerando conhecimento (LAUDON, 2011).

Este conceito está diretamente relacionado à tecnologia, pois avanços em ciência da computação e áreas afins desencadearam um aumento exponencial na capacidade de manipulação, recuperação e utilização dos dados sob diversas perspectivas.

Embora existam dados que necessitam ser protegidos contra divulgação e acesso não autorizado, há os dados abertos que podem ser livremente publicados, utilizados, reutilizados e distribuídos para diferentes propósitos, estando sujeitos - no máximo - à exigência de atribuição à fonte original e ao compartilhamento pelas mesmas licenças em que foram apresentados (OPEN DEFINITION, 2017).

A partir do momento que os dados são abertos, diferentes organizações e sistemas podem trabalhar de forma colaborativa, ampliando a comunicação e o desenvolvimento eficiente de sistemas complexos. Os dados abertos tem impulsionado, portanto, movimentos globais como a definição do conceito de Dados Abertos Públicos ou Governamentais, ligado ao conceito de democracia.

Neste contexto, dados públicos são disponibilizados eletronicamente, incluindo, mas não limitado, a documentos, bancos de dados e gravações audiovisuais. Por serem públicos, não estão sujeitos a limitações de privacidade, segurança ou controle de acesso, são primários (publicados como coletados na fonte, sem transformações e com a melhor granularidade possível), atuais (disponibilizados o mais rápido possível como forma de preservar o seu valor), acessíveis, processáveis por máquina (razoavelmente estruturados para permitir o processamento automatizado), não tem formato proprietário nem estão sujeitos a licenças, direitos autorais, marcas comerciais ou patentes.

O OGP (*Open Government Partnership*) argumenta que um "governo aberto" deve atender os seguintes requisitos: disponibilizar informações sobre as atividades governamentais, promover a participação social, implementar alto padrão de integridade no campo da Administração Pública e promover o acesso à tecnologias que viabilizem a transparência e prestação de contas.

De acordo com Dietrich *et al*. (2012), Dados Abertos Governamentais são dados produzidos durante o curso das atividades usuais e não identificam indivíduos, sendo, portanto, um subconjunto de

informações do setor público. Para Ribeiro e Almeida (2011, p.2571), o conceito Dados Abertos Governamentais pode ser entendido como "o esforço para publicar e disseminar informação do setor público na *web*, permitindo reutilização e integração de dados".

Neste sentido, uma consulta em uma página *web* ou um aplicativo (por exemplo, previsões do tempo em tempo real, informações sobre transporte público ou localização) provavelmente tem como fonte dados abertos e, porque não, dados abertos governamentais pois estes dados são, em sua maior parte, disponibilizados por recursos cujas origens são instituições do governo. Tal fato, portanto, sugere que os cidadãos comuns já utilizam dados abertos governamentais cotidianamente.

Os Dados Abertos Governamentais representam um cenário de revolução dos dados, tendo como base a integração das estatísticas sociais na tomada de decisão pública e privada e a promoção de uma relação de confiança entre a sociedade e o governo por meio da prestação de contas e transparência.

Davies (2014) considera que estes aspectos da revolução dos dados provocam um ponto de tensão na evolução dos Dados Abertos Governamentais. O processo de abertura dos dados não é uma tarefa fácil para os governos democráticos em todo o mundo.

O termo Dados Abertos Governamentais, como uma categoria no universo de dados disponíveis e de interesse público, tornou-se evidente pouco a pouco sendo fortalecido em 2008 após uma publicação do então presidente dos Estados Unidos, Barack Obama, sobre Transparência e Dados Governamentais e a criação de um portal no qual os dados do governo norte-americano poderiam ser acessados na Internet por qualquer cidadão.

No Brasil, o acesso à informação é garantido constitucionalmente - a Constituição Federal de 1988, em seus artigos XIV e XXXIII, garantiu o direito de acesso à informação e o direito de receber informações de orgãos públicos, com exceção daqueles cuja privacidade é essencial para a segurança da sociedade e do Estado.

Embora já estivesse previsto na Constituição Federal de 1988, outros arcabouços legais e ações específicas foram necessários para a abertura de dados no Governo Brasileiro. Existe a necessidade de implementar padrões, adequar legislações e criar políticas que garantem a continuidade das ações, suprindo as necessidades daqueles que consomem os dados, mapeando as informações para torná-las disponíveis, adotando padrões para segurança, qualidade, confiabilidade e utilidade. Toda esta infraestrutura tecnológica e de comunicação leva tempo e consome recursos, portanto, não é trivial.

Os Dados Abertos Governamentais Brasileiros fazem parte da política de acesso à informação do governo federal (INDA, 2012) e, neste

cenário, o governo brasileiro criou o portal dados.gov.br que reúne, mantém e disponibiliza dados governamentais segundo os princípios dos dados abertos, sem mecanismos de controle e restrições de forma que tanto pessoas físicas quanto jurídicas podem explorá-los de forma livre.

Tendo em vista o potencial para criar novos negócios, os Dados Abertos Governamentais Brasileiros têm despertado o interesse de empresários e pesquisadores, promovendo o desenvolvimento de soluções que produzem conhecimento a partir do cruzamento de dados, provendo alto nível de transparência e possibilitando aos cidadãos maior controle sobre as ações do governo, bem como aumentando a geração de empregos no país. Neste contexto, a Figura 1 apresenta aspectos relevantes da abertura de dados governamentais.

Figura 1. Potencial dos Dados Abertos Governamentais.

Uma das mais eficientes formas de garantir a abertura de dados é o provisionamento de serviços de consulta por meio de APIs RESTful (CHOLIA *et al*., 2010), que possibilita análises sobre os mais diversos requisitos.

No entanto, se for tomado apenas os dados disponibilizados por estas APIs, seguindo o esquema padrão apresentado, pode se haver uma insuficiência para uma análise eficiente e precisa. Portanto, a utilização de mecanismos auxiliares e de maior capacidade analítica, tais como ferramentas de mineração de dados, torna-se uma alternativa interessante, além de fornecer resultados satisfatórios.

2.2 Inferência de Tópicos

A Mineração de Texto, também conhecida como KDT (*Knowledge Discovery from Textual Databases*) refere-se ao processo de extrair padrões interessantes e não triviais com objetivo de descoberta de conhecimento em documentos ou dados textuais não estruturados (FELDMAN; DAGAN, 1995).

Tal abordagem requer a aplicação de técnicas e ferramentas específicas, capazes de processar grandes volumes de textos e identificar características implícitas, possibilitando assim a análise qualitativa e a melhor compreensão do conteúdo disponível em documentos.

Estes textos podem estar representados das mais diversas formas (*e-mails,* arquivos em formato *pdf*, conteúdo de páginas *web*, textos eletrônicos digitalizados) e o conhecimento pode ser extraído por meio de análise semântica, a qual é baseada na funcionalidade e relevância dos termos, ou estatística, baseada na frequência dos termos do corpus textuais (EBECKEN *et al*., 2003 *apud* MORAIS; AMBRÓSIO, 2007).

Enquanto a análise semântica tem como base identificar a função de um termo na estrutura do documento, a análise estatística ignora a ordem em que os termos aparecem no texto, bem como qualquer informação estrutural. No entanto, possibilita a extração de informações sobre associações existentes entre os termos e assim pode ser realizado o agrupamento por padrões (*clustering*).

A Modelagem de Tópicos pode ser entendida como uma das diversas técnicas para a Mineração de Texto. Resulta-se em um conjunto de algoritmos que utilizam métodos estatísticos para analisar as palavras ou termos que formam uma coleção de documentos textuais e identificar os temas ou tópicos existentes, bem como as relações entre as três unidades canônicas – documentos, tópicos e termos.

Desta análise resulta a representação de cada documento em uma coleção, por meio de descrições que mantém relações estatísticas para classificação, sumarização e recuperação de informações, sem a necessidade de rotular os documentos *a priori*.

O algoritmo LDA (*Latent Dirichlet Allocation*) é um desses algoritmos para Modelagem de Tópicos, que parte do princípio de que documentos são representados como um misto de tópicos latentes e cada tópico é caracterizado pela distribuição dos termos ou palavras (BLEI, 2003).

Para o LDA, um tópico é uma distribuição multinomial sobre os termos em um vocabulário do corpus textual. Para interpretar um tópico, uma análise preliminar poderia se basear na lista dos mais prováveis termos naquele tópico. No entanto, os tópicos inferidos por meio de LDA não são facilmente interpretados por humanos e, por isso, uma abordagem tem se baseado na definição de medidas quantitativas para

mensurar a interpretabilidade e, dessa forma, estabelecer um ranking dos termos de um tópico.

Por meio da identificação dos tópicos, documentos que tratam de um mesmo assunto ou assuntos equivalentes são agrupados pela quantidade de palavras similares e frequentes. No entanto, não é trivial a identificação de similaridade de significado entre os termos tendo em vista as variações morfológicas, sintáticas e semânticas, de uma língua ou idioma.

Para reduzir este problema, faz-se necessária a aplicação de técnicas de normalização dos documentos textuais, a qual pode ser realizada a partir do uso de um vocabulário controlado, específico para as áreas temáticas do conjunto de dados, remoção de palavras não relevantes ou *stop-words* (preposições, pronomes, artigos, advérbios) - palavras comuns a todos os documentos que apesar da alta frequência não permitem discriminá-los.

O *Thesaurus* (DIAS-DA-SILVA, 2003) é um dicionário eletrônico que pode ser acoplado a ferramentas computacionais e a WordNet.BR é uma base de dados *on line*, desenvolvida pelo Núcleo de Linguística Computacional da USP, com léxicos significativos para o português brasileiro (ARANHA, 2007). Um tópico, então, pode ser definido formalmente como uma distribuição de termos sobre um vocabulário controlado, o qual é caracterizado pelo conjunto de todos os termos relevantes em uma coleção de documentos textuais.

A Figura 2 apresenta uma representação de um documento a partir do conceito de inferência de tópicos, sendo os termos relevantes característicos destacados em três cores distintas (azul, verde e púrpura).

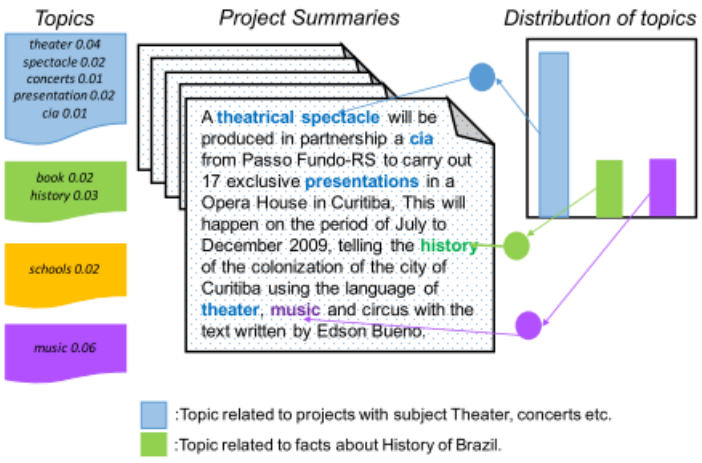

Figura 2. Representação de um documento em tópicos.

3. Nossa proposta

Nesta seção, é apresentada uma proposta de aplicação de Mineração de Textos sobre dados abertos governamentais, por meio do uso do algoritmo LDA, proporcionando a extração de padrões sobre documentos textuais referentes a projetos culturais brasileiros, para possibilitar a inferência de tópicos correspondentes a diferentes temas, especialidades ou assuntos abordados naqueles projetos.

Em geral, os documentos em formato texto possuem tamanho expressivo, da ordem de milhares de palavras ou termos, muitas das palavras são repetidas, expressam o mesmo significado ou tem significado irrelevante, nem todas as palavras ou termos de um texto tem a mesma importância. Para viabilizar a manipulação de bases de dados textuais e garantir a eficiência do processo de mineração de textos, foi realizada a etapa de preparação dos dados.

Inicialmente, foi definida a base de dados de interesse, por meio da seleção do campo Resumo de cada documento, entendendo que as informações constantes no mesmo são suficientes para expressar o tema ou assunto abordado, ou seja, constituem amostras representativas dos documentos. Assim, foi reduzida a dimensionalidade do corpus textual.

Posteriormente, um pré-processamento sobre esta nova base de dados foi realizado, preparando-a na língua português brasileiro, por meio da remoção de *stop-words* – palavras sem conteúdo semântico significante no contexto e, portanto, irrelevantes para a análise.

Foram, então, removidas *tags html,* palavras auxiliares e conectivas (preposições, pronomes, artigos) sem informação discriminativa no conteúdo. Geralmente estas palavras apresentam uma incidência muito alta na coleção de dados, mas são insignificantes do ponto de vista semântico.

As *stop-words* foram removidas com base em uma lista provida pela ferramenta NLTK (LOPER, 2002), adaptada para o domínio cultural.

Um modelo espaço vetorial foi definido para a representação do conjunto de dados. Neste modelo, cada documento foi representado por um vetor onde seus elementos correspondem aos termos[1] da coleção de documentos sendo atribuído um peso para cada termo, que representa o seu grau de importância no documento. Isso resulta em uma distribuição de probabilidades sobre um vocabulário específico, provendo um conceito semântico para a análise (AGGARWAL, 2012; FELDMAN, 2006).

Basicamente, a conversão de um texto em um vetor ou qualquer outra representação que evidencie características relevantes é uma parte importante do processo de mineração de textos, pois os resultados dos algoritmos de aprendizado de máquina são diretamente proporcionais à representação da coleção de documentos textuais.

Desta maneira, cada documento foi modelado em um vetor de *n* elementos (onde *n* é a quantidade de palavras que compõem a coleção de documentos, e não somente aquelas presentes no documento em questão) na forma (*termo 1, peso 1*), (*termo 2, peso 2)...(termo n, peso n*). Dessa forma, palavras que o documento não contém recebem grau de importância zero e palavras constantes no documento tem o peso calculado segundo uma métrica de importância que, para este trabalho, é baseada na frequência do termo (*tf*) naquele documento.

Os projetos culturais do Ministério da Cultura do Brasil (MinC) são classificados em sete áreas temáticas, a saber: Audiovisual, Humanidades, Música, Artes Cênicas, Artes Integradas, Patrimônio Cultural e Artes Visuais. Uma determinada área temática (por exemplo, Artes Cênicas) possui subáreas (Dança, Circo, Ópera, Teatro) que representam peculiaridades daquela área no contexto cultural.

A Figura 3 apresenta um grafo da estrutura atual dos projetos da base de dados SALIC. Existe uma problemática que reside na falta de um consenso na identificação das subáreas temáticas, uma vez que tal processo é realizado pelo próprio usuário quando do cadastramento do projeto no MinC.

Logo, a estrutura está sujeita à interpretação semântica dada pelo usuário ao projeto, resultando em polissemia - multiplicidade de sentidos na definição de áreas e subáreas temáticas. É o caso, por exemplo, da existência de uma área temática "Áreas Integradas" e subárea também denominada "Áreas Integradas", conforme pode ser observado na Figura 3.

Tendo em vista essa problemática, a proposta deste trabalho é obter uma melhor identificação das subáreas temáticas a partir da inferência de tópicos. Para o uso do algoritmo LDA foi definida a quantidade de tópicos (*k=15*).

Por meio da análise da relevância dos termos nos tópicos inferidos e do agrupamento dos mesmos, espera-se que seja possível melhor entendimento semântico e redesenho da estrutura dos projetos culturais, no que tange à identificação de subáreas para as áreas temáticas existentes.

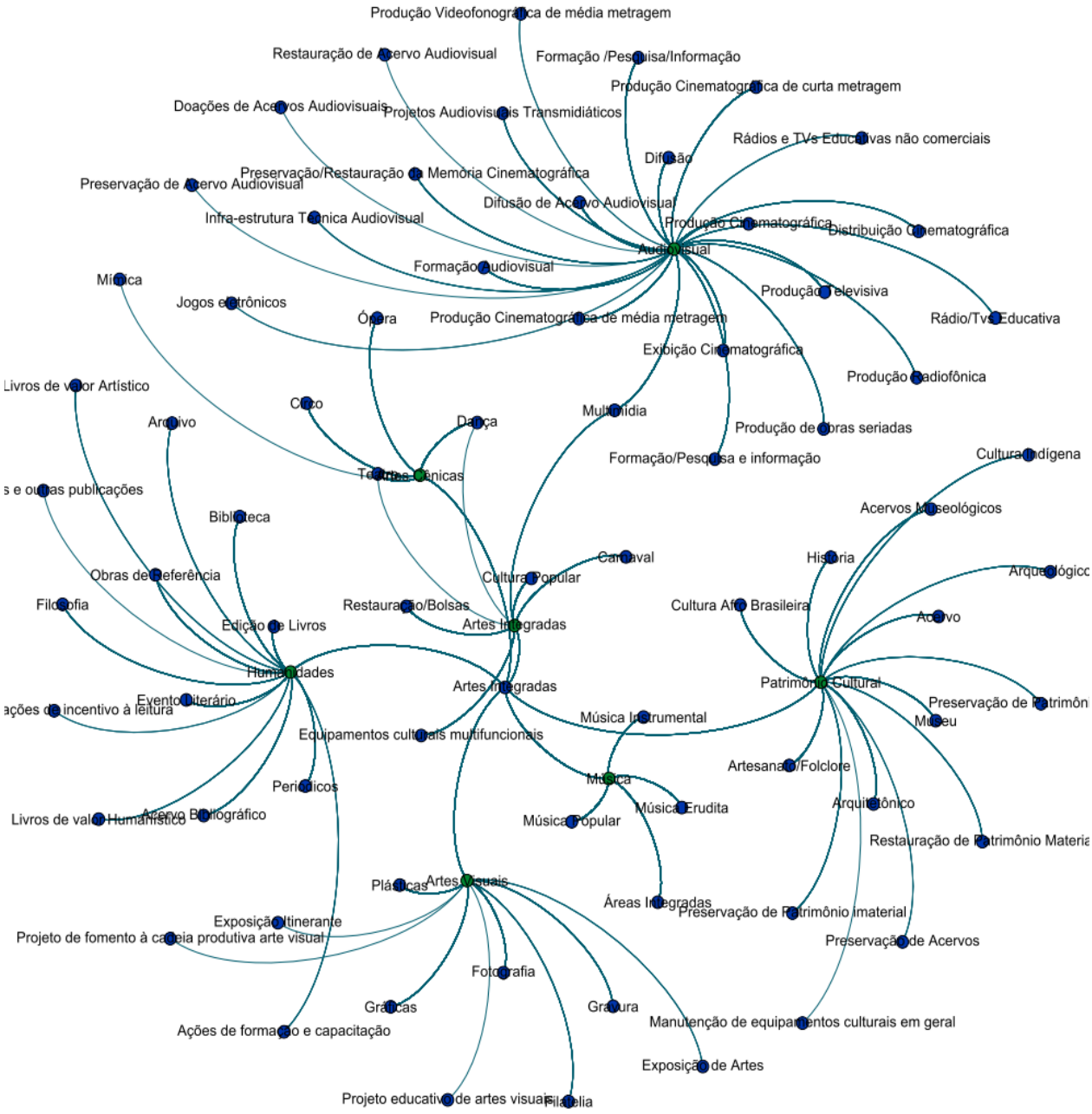

Figura 3. Áreas e segmentos temáticos dos projetos culturais.

4. Desenvolvimento Experimental

4.1 Base de Dados, Ferramentas e Técnicas

A base de dados experimental foi obtida através de uma API Restful desenvolvida pelo Núcleo de Pesquisas em Gestão, Políticas e Tecnologias da Informação da Universidade Federal de Goiás, através do Laboratório de Dados Abertos[2], para o Ministério da Cultura do Brasil (MinC), com o propósito de abertura dos dados de projetos culturais gerenciados pelo sistema denominado SALIC Web (SALIC, 2016). A base de dados

associada à API do SALIC possui cinco entidades principais: *Proponente*, *Proposta*, *Projeto*, *Incentivador* e *Fornecedor*. A cada uma destas entidades, estão associados um conjunto de atributos. Para o presente trabalho, foi considerada, especificamente, a entidade Projeto, a qual contém, entre outros, os seguintes atributos: PRONAC, nome, UF, município, valor do projeto, objetivos etc. A Figura 4 apresenta o esquema geral associado a esta entidade.

Figura 4. Esquema geral da base dados.

A partir disso, foi construída uma base de dados experimental a partir do atributo *Resumo* da entidade *Projeto*, sendo construída uma amostra contendo os resumos de 5000 projetos culturais.

Foi utilizada a linguagem de programação Python, módulo NLTK e *scikit-learn*, para desenvolvimento do algoritmo para processamento de dados textuais – LDA – em duas abordagens: primeiramente, foi realizado processamento de toda a base de dados para inferência de 100 tópicos (BLEI, 2013) e depois foi realizado o processamento considerando os projetos pertencentes a cada uma das sete áreas temáticas e considerando a quantidade de tópicos (k) igual a 15, ou seja, cada tópico inferido representa uma subárea a qual se enquadra um determinado projeto cultural.

Para apresentação dos resultados foi utilizada a ferramenta de visualização LDAvis (SIEVERT; ILLVI, 2014). Tal ferramenta possibilita interatividade entre os tópicos inferidos e os termos, segundo a relevância dos mesmos. A relevância é uma métrica definida por Sievert e Illvi (2014) para a interpretabilidade dos tópicos. Tal métrica permite estabelecer um ranking dos termos em ordem de sua utilidade no processo de interpretação dos tópicos. A relevância é definida com base na frequência do termo:

$$(w,k|\lambda)=\lambda\log(\phi_{kw})+(1-\lambda)\log\left(\frac{\phi_{kw}}{p_w}\right)$$

onde λ define um peso para a significância do termo w no tópico k, pw é a probabilidade marginal do termo w no corpus textual e ϕ_{kw} é um

estimador da frequência do termo w no tópico k, utilizado pelo LDA para a inferência do referido tópico.

O valor de lambda está diretamente relacionado à interpretabilidade dos resultados. Para $\lambda=1$ os termos são apresentados em ordem decrescente de suas probabilidades estimadas pelo algoritmo no processo de aprendizado para um tópico específico. Se $\lambda=0$ então os termos são apresentados unicamente pela sua frequência relativa. Sievert e Illvi (2014) demonstraram empiricamente que o valor ótimo para interpretabilidade dos tópicos considerando a relevância dos termos é $\lambda=0,6$.

4.2 Resultados e Discussão

Inicialmente, na Figura 5 é apresentada uma nuvem de palavras para um dos 100 tópicos inferidos a partir do processamento de toda a base de dados de interesse. A nuvem de palavras possibilita uma noção de relevância dos termos de um determinado tópico do ponto de vista da frequência dos mesmos no *corpus* textual – o tamanho das palavras é proporcional à sua relevância.

Para o tópico 98, observa-se que os termos mais frequentes são "temporada" e "crianças", seguido por "espetáculo", "teatral", "social".

Por meio da identificação dos termos relevantes e tomando como base as áreas temáticas dos projetos culturais brasileiros, ao tópico 98 poderia ser atribuído um significado e a partir do mesmo ser identificado, por exemplo, como "Teatro Infantil".

Figura 5. Nuvem de palavras para o tópico 98.

Trata-se, portanto, de um tipo de análise exploratória da base de dados, a qual permite uma compreensão das categorias existentes. Assim, um resultado poderia ser a definição da granularidade da estrutura documental – quantas áreas temáticas existem, em quantas subáreas se dividem, por exemplo.

O processamento da base de dados por área temática resultou no agrupamento (*clustering*) dos projetos culturais em 15 (quinze) tópicos inferidos ou subáreas, conforme proposta apresentada na seção anterior.

A Figura 6 apresenta os *clusters* obtidos para a área temática 176 ("Artes Cênicas"). Esta figura provê uma perspectiva global dos tópicos inferidos sobre a base de dados. Cada círculo representa um tópico e, a partir de então, uma subárea temática. As áreas dos círculos são proporcionais à frequência relativa dos tópicos no corpus textual e a distribuição dos círculos no espaço representa o grau de diferenciação dos tópicos, segundo a distância Jensen-Shannon e considerando as 2 (duas) primeiras componentes principais.

A análise de componentes principais é uma técnica utilizada em análise exploratória e preditiva que promove uma transformação linear nos dados de forma que os mesmos sejam representados em um novo sistema de coordenadas. Neste sistema, as componentes mais relevantes correspondem às duas primeiras dimensões, em eixos denominados principais (JOLLIFFE, 2002). Nesta representação, portanto, a maior variância dos dados é definida ao longo da primeira coordenada e denominada PC1 (primeira componente principal).

Neste contexto, no escopo deste trabalho, a utilização dos dois primeiros componentes principais (PC1 e PC2), permite uma representação significativa da coleção de dados e viabiliza a análise da distribuição dos tópicos inferidos no espaço bidimensional e das relações existentes, por meio das características mais relevantes.

Na Figura 6 a área dos círculos é proporcional à relevância dos tópicos na área em questão. Para a identificação dos tópicos, é necessária uma análise das características das unidades canônicas associadas ao contexto e, portanto, uma análise da distribuição dos termos nos tópicos é factível.

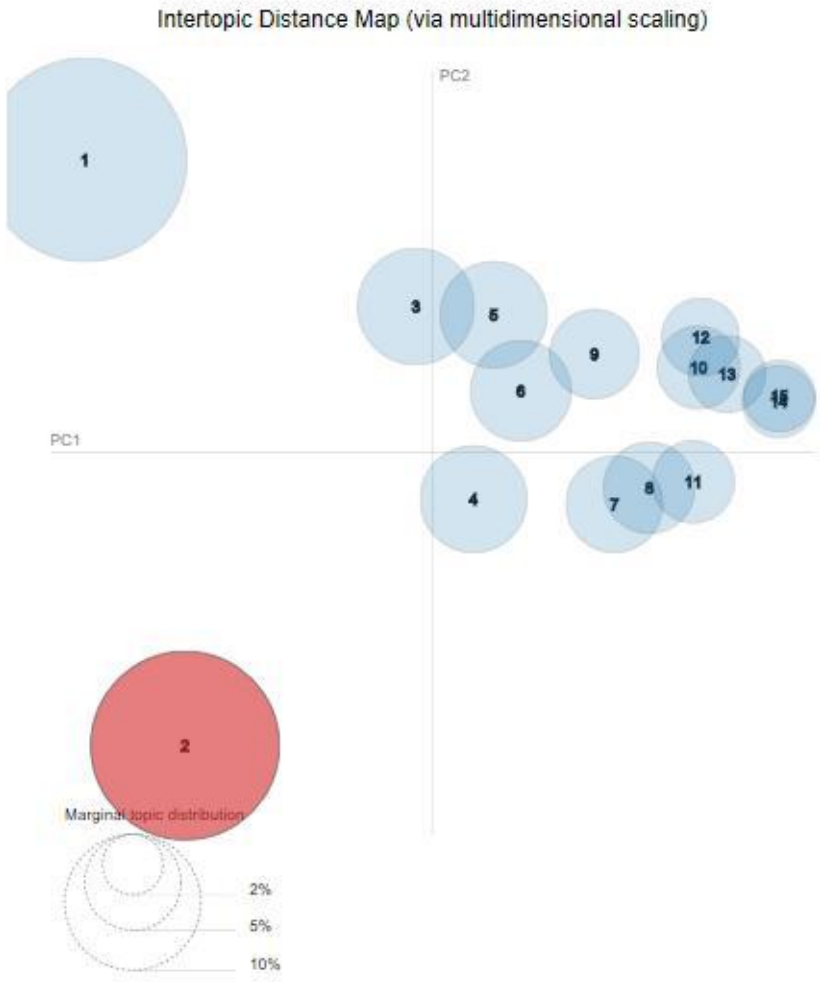

Figura 6. Distribuição dos tópicos inferidos para os projetos da área temática 1 - "Artes Cênicas".

Conforme descrito na seção anterior, a ferramenta de visualização utilizada (LDAvis) trata a relevância em função da frequência do termo, mas em uma abordagem mais coerente e robusta, pelo fato de ter um peso associado à importância do termo na definição do tópico, ou seja, a sua significância.

A ferramenta possibilita explorar os tópicos de forma interativa. Desta maneira, o tópico 2 da Figura 6 é selecionado, e informações dos termos relacionados ao mesmo são apresentados ao usuário (Figura 7).

A análise da Figura 7 deve ser realizada por meio da comparação das larguras das barras vermelha e azul para um determinado termo. É possível entender quão relevante é um termo para o tópico selecionado

(uma proporção alta de vermelho sobre o azul) ou ainda, pode-se verificar sua frequência no tópico (largura absoluta do vermelho).

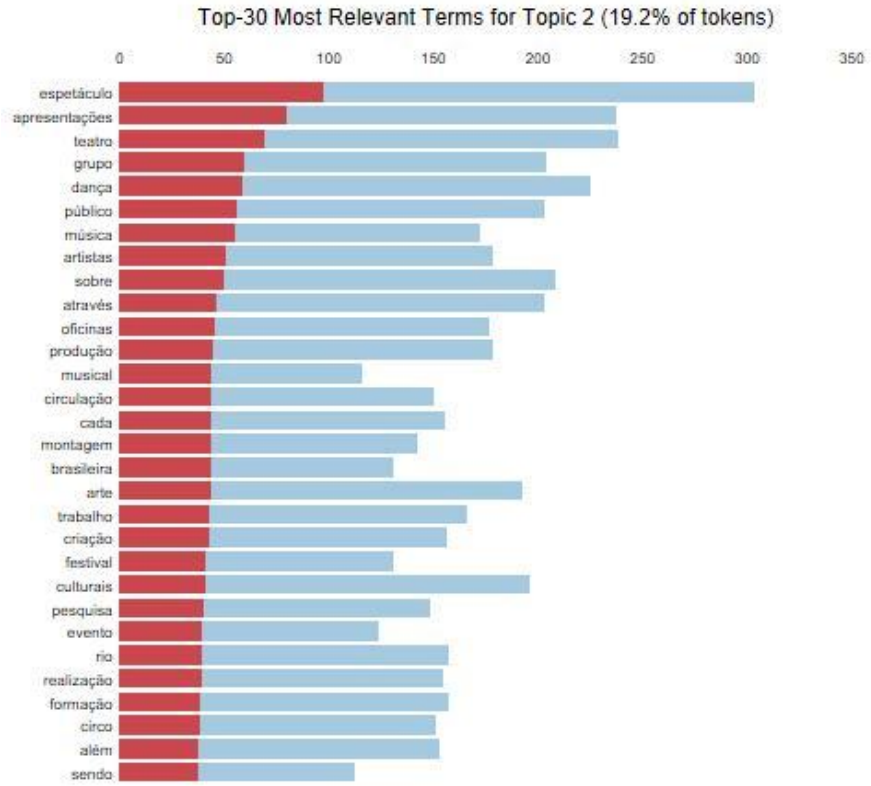

Figura 7. Frequência e Relevância dos termos para o tópico 2.

Dessa forma, na Figura 7 observa-se que para o tópico 2 da área temática 1 (“Artes Cênicas”), os cinco termos mais relevantes são “espetáculos”, “apresentações”, “teatro”, “música”, “musical”.

Uma vez que a área temática é conhecida, a análise específica dos termos relevantes proporciona maior clareza na identificação das subáreas ou nomeação dos tópicos inferidos. Por exemplo, sendo a área 1 denominada “Artes Cênicas”, a subárea referente ao tópico 2 poderia ser nomeada “Teatro Musical”, uma especificidade do teatro que combina música, dança e atuação artística.

A análise da Figura 7 possibilita, ainda, entender o significado da relevância do termo neste contexto. Observa-se, por exemplo, a proporção das barras vermelhas sobre as azuis para os termos “dança” e “musical”: aproximadamente 25% (55 de 220 *tokens*) para o termo dança e 42% (50 de 120 *tokens*) para o tema musical, de forma que apesar de não ser mais frequente que o termo dança, o termo musical é mais relevante, oferecendo, portanto, maior contribuição na definição semântica do tópico selecionado.

Tal observação corrobora com o fato de que ser mais frequente não significa, necessariamente, ser mais relevante. Mesmo com a remoção das *stop-words* e o pré-processamento realizado, existem termos muito comuns nos projetos que não tem significado semântico suficiente para a diferenciação e/ou identificação dos tópicos.

Ainda sobre a Figura 7 é possível observar que os termos "música" e "musical" apresentam praticamente a mesma relevância para o tópico 2. Semanticamente, no contexto da área temática, apresentam também o mesmo significado, ou seja, pertencem à mesma categoria. Trata-se de um problema de vocabulário – sinonímia – não tratado nesta implementação.

A ferramenta de visualização permite, ainda, selecionar um termo e verificar a distribuição condicional do mesmo sobre os tópicos. As Figuras 8 a 11 apresentam seleções na área temática 1 ("Artes Integradas"), cujo termo mais frequente é "carnaval".

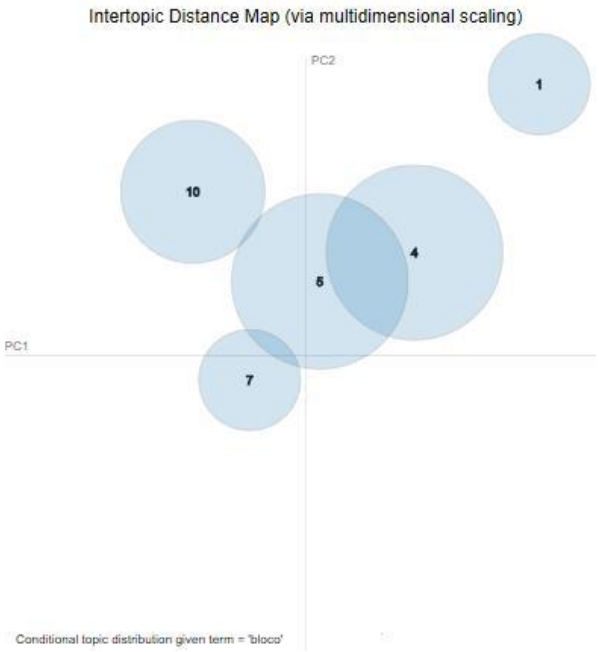

Figura 8. Seleção do termo "bloco".

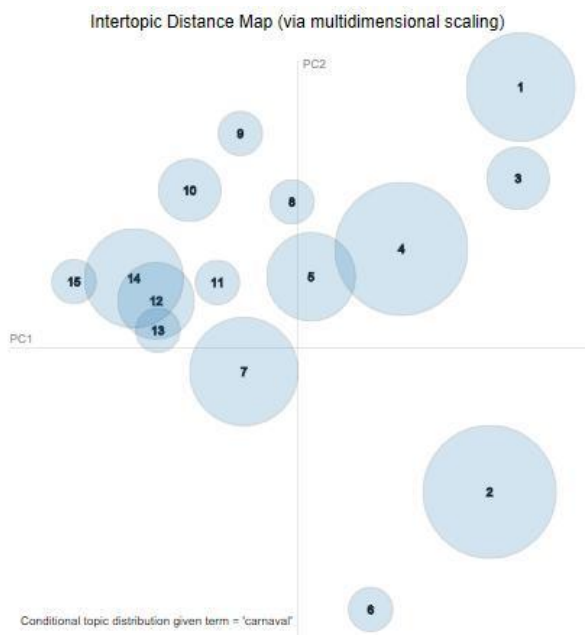

Figura 9. Seleção do termo "carnaval".

Analisando as figuras acima, observa-se que ao selecionar o termo "bloco", os círculos referentes aos tópicos 4, 5 e 10 apresentam maior área na Figura 9, em comparação à Figura 8. Neste tipo de visualização, a área dos círculos é alterada proporcionalmente à frequência do termo no tópico. Assim, é possível analisar o resultado apresentado pelo LDA no que tange fidelidade no agrupamento de tópicos similares.

Tendo em vista a área temática abordada e as características apresentadas quando da seleção do termo "bloco" seria coerente, por exemplo, afirmar que os tópicos 4 e 5 estão relacionados a uma subárea que pode ser categorizada como "carnaval de rua".

De forma análoga, os tópicos 8 e 5 podem ser associados a uma subárea "festa junina", por meio da interpretação das características apresentadas quando da seleção dos termos "quadrilha" (Figura 10) e "forró" (Figura 11).

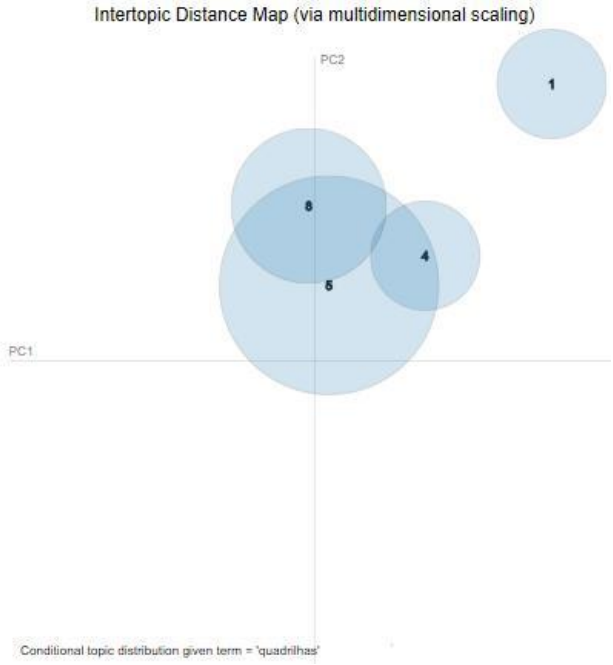

Figura 11 Seleção do termo "quadrilhas".

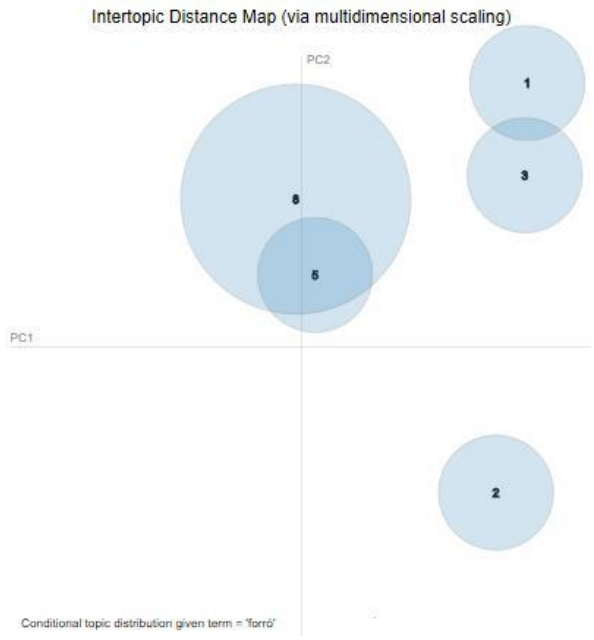

Figura 10 Seleção do termo "forró".

Estas observações podem, ainda, auxiliar na definição da granularidade da estrutura da base de dados documental, no que tange à quantidade de subáreas a serem consideradas, com base nas características específicas dos tópicos e no significado semântico atribuído.

Ainda, a análise da distribuição dos termos provida por esta visualização pode ser útil na diferenciação dos tópicos em casos de polissemia. A Figura 12 apresenta os termos quando o tópico 1 da área "Artes Cênicas" é selecionado.

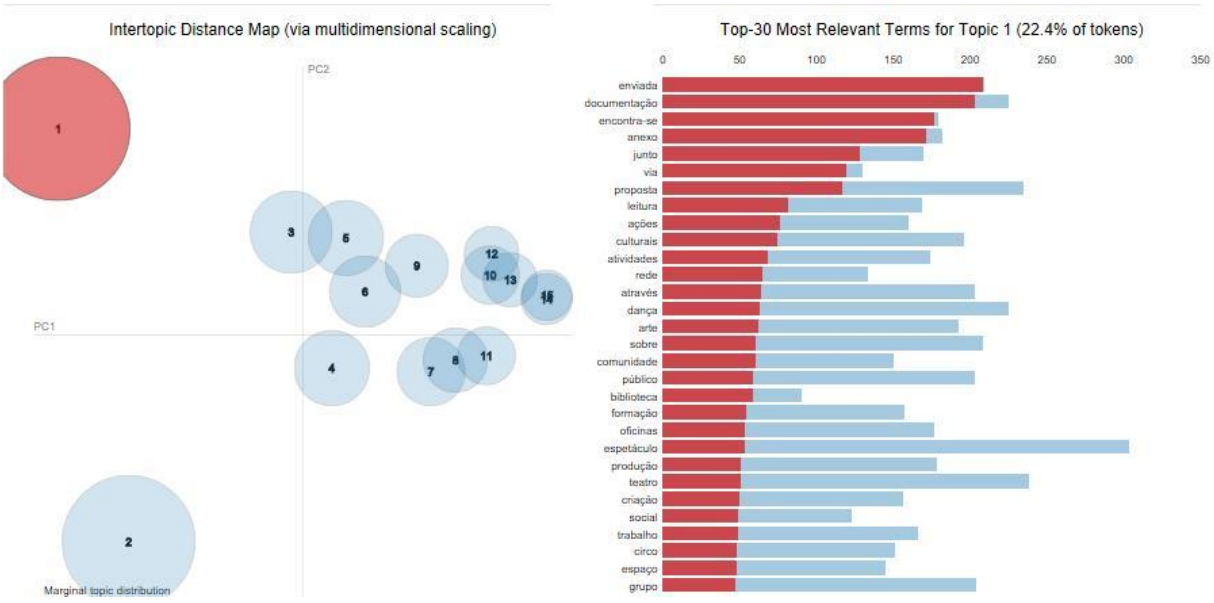

Figura 12. Termos relevantes para o tópico 1 da área "Artes Cênicas".

Algumas características merecem destaque quando a Figura 11 é analisada. Primeiramente, os termos "enviada", "documentação", "encontra-se", "anexo", "junto", "via" são apresentados com alta relevância, porém no contexto não são significantes do ponto de vista semântico. Verifica-se, portanto, uma falha no pré-processamento do corpus textual, a qual pode ser atribuída à definição das *stop-words*.

Uma vez que estes termos são eliminados do escopo da análise, é possível observar o termo "teatro" está presente nos tópicos 1 e 2, com praticamente a mesma relevância. Em um primeiro momento, sugere-se um caso de polissemia, pois graficamente é nítida a diferenciação dos tópicos – estão distantes, ou seja, o mesmo termo foi utilizado em sentidos diferentes.

Para a correta identificação do tópico, portanto, outros termos relevantes devem ser considerados e então pode-se observar que o tópico 1 tem grande influência dos termos "social" e "circo" (Figuras 13 e 14) – aumento significante da área do círculo correspondente ao tópico 1, em relação à área do mesmo na Figura 12, quando estes termos são selecionados.

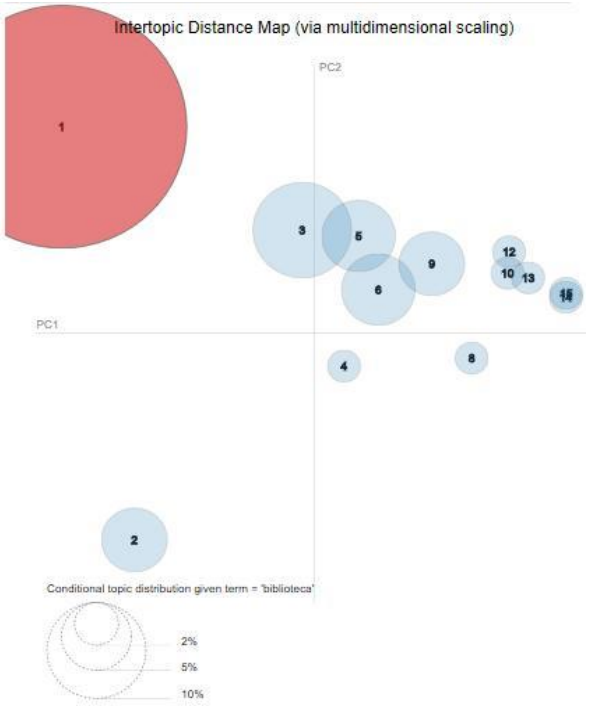

Figura 14. Seleção do termo "social".

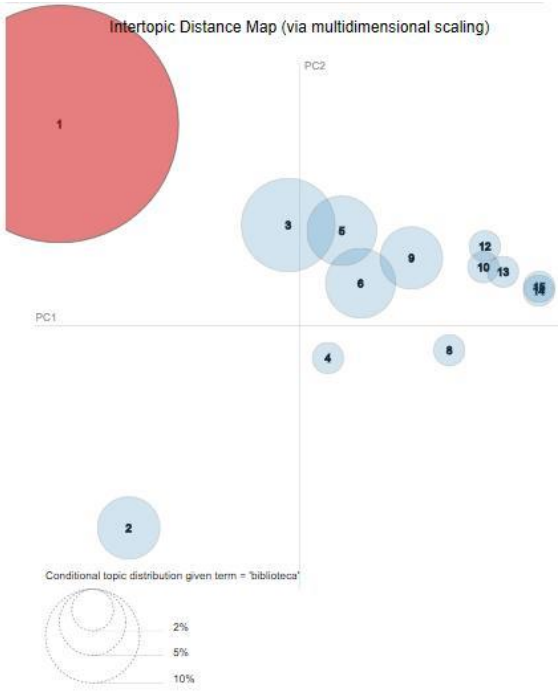

Figura 13. Seleção do termo "circo".

A análise exploratória realizada sobre os termos relevantes do tópico 1 permite sugerir uma identificação para a subárea temática relacionada e, assim, poderia ser denominada “Espetáculo Circense”.

Dessa forma, a análise das características dos tópicos 1 e 2 inferidos possibilitou a identificação de duas subáreas (“Espetáculo Circense” e “Teatro Infantil”, respectivamente), da área temática “Artes Cênicas”. De forma análoga, outras subáreas podem ser identificadas.

5. Considerações Finais

O trabalho apresentou uma implementação do algoritmo LDA como parte de um processo automatizado de mineração de textos, com o objetivo de inferir tópicos e por meio destes identificar subáreas temáticas dos projetos culturais gerenciados pelo Ministério da Cultura do Brasil.

Como resultado principal, obteve-se o agrupamento por padrões ou *clustering* dos projetos em 15 tópicos e o usuário pode explorar, interativamente, cada um destes tópicos e proceder análise da frequência e relevância dos mesmos e de seus termos .

A aplicação de técnicas de inferência de tópicos sobre Dados Abertos Governamentais é uma área nova, especialmente para a língua Portuguesa, não tendo sido encontrados trabalhos relacionados a este cenário. Portanto, os resultados obtidos com o algoritmo LDA, embora

preliminares, confirmam que a técnica é promissora. É, portanto, importante concentrar esforços no sentido de reforçar o uso da mesma.

O processo de identificação dos tópicos neste trabalho foi realizado de forma empírica. Os nomes das áreas temáticas as quais se enquadram os projetos culturais são conhecidos e a identificação das subáreas teve como pressuposto o conhecimento do domínio e julgamento do próprio usuário.

Sabe-se que atribuir um significado aos termos em um corpus textual é um desafio pois muitas vezes ele só existe no momento do uso, além de ser fortemente dependente do contexto. Em termos computacionais, a análise semântica para identificação das subáreas não é trivial.

Como evolução deste trabalho, propõe-se desenvolver um modelo de classificação que contemple análise semântica dos termos. Nesse modelo para cada termo deve existir uma distribuição de probabilidades, variável conforme a usabilidade do mesmo, e o algoritmo LDA deve estimar, na fase de treinamento, a probabilidade de um documento pertencer a uma categoria de termos.

Problemas de vocabulário, tais como sinonímia apresentado nos resultados da Figura 6 por exemplo, podem ser tratados por técnicas de pré-processamento dos dados, para melhor desempenho da classificação.

Uma possibilidade é a utilização de um dicionário, tendo como base a WordNet.BR ou o aplicativo *Thesaurus*, de forma que seja definido um vocabulário controlado com representação dos sinônimos, hierarquias e relacionamentos associativos entre termos.

Assim, o dicionário poderia ser utilizado com a finalidade de substituir todos os termos com mesmo significado por um termo só ou selecionar termos segundo uma hierarquia pré-definida.

Outra possibilidade é a utilização de *stemming*, uma técnica de análise linguística na qual cada termo é considerado isoladamente e reduzido à sua provável raiz, eliminando sufixos (musical/ música), indicando formas verbais e/ou plurais.

Um aspecto importante que pode, ainda, ser tratado como evolução deste trabalho está relacionado à inferência de tópicos de forma não supervisionada, ou seja, independente de conhecimento especializado, *a priori*, do domínio.

Entende-se que automatizar a definição de significado de um termo em um contexto específico ainda é um desafio pois está atrelado às particularidades específicas da língua.

Os resultados apresentados neste trabalho evidenciam também a importância do pré-processamento no âmbito da mineração de textos para aumentar a interpretabilidade em modelos discriminantes, como o proposto.

Neste sentido, o trabalho apresenta uma contribuição significativa tendo em vista que a análise exploratória dos projetos culturais brasileiros possibilitou a identificação de subáreas temáticas de acordo com as características dos tópicos inferidos pelo algoritmo LDA e, ainda, sugeriu implementações no que tange à melhor utilização deste algoritmo na mineração de textos no âmbito dos dados abertos governamentais.

Referências

AGGARWAL, C.; ZHAI, C. *Mining Text Data*. London: Springer Publishing Company, Incorporated, 2012.

ARANHA, C. N. *Uma abordagem de pré-processamento automático para mineração de textos em português*: sob o enfoque da inteligência computacional**.** 144p. Tese de Doutorado em Engenharia Elétrica. Pontifícia Universidade Católica do Rio de Janeiro, 2007.

BLEI, D. M., NG A.Y., JORDAN, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993-1022, 2003.

BLEI, D. M. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.

CHOLIA, S.; SKINNER, D.; BOVERHOF, J. NEWT: A RESTful service for building High Performance Computing web applications. *In*: GATEWAY COMPUTING ENVIRONMENTS WORKSHOP (GCE**)**, 1-11, 2010.

DIAS-DA-SILVA, B. C. e Moraes, H. R. A construção de *thesaurus* eletrônico para o português do Brasil. *Alfa*, v.47, n.2, p.101 - 115, 2003.

DAVIES, T. *Open Data in Developing Countries – Emerging insights from Phase I*. Web Foundation, 2014. Disponível em: http://www.opendataresearch.org/content/2014/704/open-data-developing-countries-emerging-insights-phase-i.html. Acesso em: 18 mar. 2021.

DIETRICH, D., GRAY, J., MCNAMARA, T., POIKOLA, A., TAIT J., POLLOCK, R., ZIJLSTRA, T. *Open Data Handbook Documentation Release 1.0.0*. London: Open Knowledge Foundation, 2012.

EBECKEN, N; LOPES, M; COSTA, M. *Mineração de Textos***,** São Paulo: Manole, 2003.

FELDMAN, R., DAGAN, I., Knowledge discovery in textual databases (KDT). *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD-95), 1, 1995. *Proceedings […]*, Montreal, Canada, August 20-21, AAAI Press, 112-117. Disponível em: https://www.aaai.org/Conferences/KDD/kdd95.php. Acesso em: 18 mar. 2021.

INDA, Infraestrutura Nacional de Dados Abertos. Instrução Normativa nº 4 de 12 de abril de 2012. Disponível em:http://dados.gov.br/pagina/instrucao-normativa-da-inda. Acesso em: 22 jan. 2018.

JOLLIFFE, I. T. *Principal Component Analysis.* 2. ed. New York: Springer, 2002.

Laudon, K.; Laudon, J. *Management Information Systems: Managing the Digital Firm*. New Jersey: Pearson, 2011.

LOPER, E.**;** BIRD, S. Nltk: The natural language toolkit. *In*: ACL-02 WORKSHOP ON EFFECTIVE TOOLS AND METHODOLOGIES FOR TEACHING NATURAL LANGUAGE PROCESSING AND COMPUTATIONAL LINGUISTICS – V.1, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. *Proceedings […]* Association for Computational Linguistics.

MORAIS, E. A. M., AMBROSIO, A.P.L. *Mineração de Textos*. Relatório Técnico INF_005/07. Instituto de Informática. Universidade Federal de Goiás, 2007.

OPEN DEFINITION. *Open Definition 2.1***.** 2017. Disponível em: http://opendefinition.org/od/2.1/en/. Acesso em: 19 de jan. 2018.

OPEN KNOWLEDGE FOUNDATION. *Open Data Handbook*. 2010. Disponível em http://opendatahandbook.org/guide/en/. Acessado em 19 de janeiro de 2018.

RIBEIRO, C. J. S.; ALMEIDA, R. F. . Dados Abertos Governamentais (Open Government Data): Instrumento para Exercício de Cidadania pela Sociedade. *In*: ENANCIB - POLÍTICAS DE INFORMAÇÃO PARA A SOCIEDADE, 12, 2011. *Anais[...]* Brasília: Thesaurus, 2011, p. 2568-2580.

SALIC, 2016. Disponível em http://novosalic.cultura.gov.br.

SIEVERT, C., SHIRLEY, K. E. LDAvis: a method for visualizing and interpreting topics. *In:* WORKSHOP ON INTERACTIVE LANGUAGE LEARNING, VISUALIZATION, AND INTERFACES. *Proceedings […], 2014.* p. 63-70, 2014.

1Termos, atributos e palavras são utilizadas como sinônimos neste artigo e representam unidades canônicas de um documento textual
2Laboratório de Dados Abertos - UFG: https://www.gi.fic.ufg.br/lda/