Revista de Contabilidade e Organizações
www.rco.usp.br

DOI: http://dx.doi.org/10.11606/issn.1982-6486.rco.2023.197181

Journal of Accounting and Organizations
www.rco.usp.br

Clusterização de precedentes de IRPJ no CARF

Cluster analysis of IRPJ precedents in CARF

Fabiano de Castro Liberato Costa[a], Antonio Lopo Martinez[b], Roberto Carlos Klann[a]

[a] *Universidade Regional de Blumenau - Brasil*
[b] *Universidade de Coimbra - Portugal*

Palavras-chave
Tributação de rendimentos corporativos.
Clusterização.
Jurisprudência tributária.
Conselho Administrativo de Recursos Fiscais.

Resumo

O objetivo deste estudo foi agrupar acórdãos do Conselho Administrativo de Recursos Fiscais (CARF) relacionados ao Imposto de Renda Pessoa Jurídica (IRPJ), prolatados entre 2016 e 2020, empregando técnicas de aprendizado de máquina (ML) para a clusterização de documentos textuais. A análise resultou em 13 *clusters* exclusivos, um achado inédito na literatura contábil tributária no Brasil. Essa identificação é relevante para o CARF, contribuintes, administração tributária e profissionais contábeis e tributaristas envolvidos em questões contábeis e tributárias relacionadas ao IRPJ. Os algoritmos de ML utilizados mostraram-se eficientes na resolução de problemas complexos de processamento de linguagem natural (PLN), como criar representações vetoriais de termos e identificar temáticas em dados não estruturados, fornecendo contribuições valiosas para o entendimento de matérias controversas no IRPJ à luz da jurisprudência administrativa. A clusterização de precedentes se traduz em maior acessibilidade e análise de padrões nos julgamentos, facilitando a tomada de decisões na contabilidade tributária.

Keywords
Corporate income taxation.
Clustering.
Taxation jurisprudence.
Administrative Council of Tax Appeals.

Abstract

The objective of this study was to cluster judgments of the Administrative Council of Tax Appeals (CARF) related to corporate income tax (IRPJ) rendered between 2016 and 2020, employing machine learning (ML) techniques for the clustering of textual documents. The analysis resulted in 13 unique clusters, an unprecedented finding in the tax accounting literature in Brazil. This identification is relevant for the CARF, taxpayers, tax administration, and accounting and tax professionals involved in accounting and tax issues related to the IRPJ. The ML algorithms used proved efficient in solving complex natural language processing (NLP) problems, such as creating vector representations of terms and identifying themes in unstructured data, providing valuable contributions to understanding controversial IRPJ issues in light of administrative case law. The clustering of precedents translates into greater accessibility and analysis of patterns in judgments, facilitating decision-making in tax accounting.

Informações do artigo
Recebido: 01 de maio de 2022
Aprovado: 27 de abril de 2023
Publicado: 02 de junho de 2023
Editor responsável: Prof. Dr. Sílvio Hiroshi Nakao

Implicações práticas

A clusterização beneficia: i) Órgãos julgadores, identificando matérias litigiosas em contabilidade e tributação, otimizando recursos; ii) Contribuintes, esclarecendo controvérsias contábeis do IRPJ e facilitando conformidade tributária; iii) Autoridades fiscais, direcionando capacitação, aprimorando fiscalização e orientação tributária e contábil; iv) Profissionais tributários e contábeis, impulsionando qualificação e especialização em temas relevantes.

Autor correspondente: Tel. +351 239 859 802/01.
E-mail: fclcosta@furb.br (F. C. L. Costa); antoniolopo@terra.com.br (A. L. Martinez); rklann@furb.br (R. C. Klann).
Faculdade de Direito da Universidade de Coimbra. Pátio da Universidade, 3004-528, Portugal.

1 INTRODUÇÃO E PROBLEMA

Os acórdãos do CARF relativos ao IRPJ são decisões administrativas que estabelecem critérios para a apuração e tributação do IRPJ, sendo importantes para a área contábil. Esses acórdãos fornecem informações valiosas sobre as principais causas que levam às controvérsias e sobre os critérios que as autoridades fiscais usam para verificar a escrituração contábil de uma empresa. Além disso, os acórdãos do CARF também ajudam a esclarecer outras questões, como a aferição do lucro real da empresa e a dedução de despesas.

Ao estudar os acórdãos do CARF relativos ao IRPJ, os profissionais da área contábil podem obter uma compreensão mais profunda das leis e regulamentos fiscais e como eles se aplicam aos negócios. Isso os ajuda a fornecer conselhos precisos e eficazes aos seus clientes sobre questões fiscais e a minimizar o risco de sanções fiscais. Além disso, os acórdãos do CARF também oferecem *insights* sobre o planejamento tributário e ajudam os profissionais da área contábil a identificar oportunidades para maximizar as deduções fiscais de seus clientes.

Este artigo objetiva analisar acórdãos emitidos no CARF referentes ao IRPJ, utilizando aprendizagem de máquina (*Machine Learning* - ML) não supervisionada e técnica de agrupamento (*clustering*). A clusterização de textos busca identificar grupos de documentos similares em um conjunto maior, dividindo-o em partes semanticamente heterogêneas (Serras, 2021).

O objetivo principal desta pesquisa foi empregar algoritmos de aprendizado de máquina para agrupar processos similares julgados no CARF, facilitando a criação de decisões uniformes com base em precedentes semânticos e colaborando para a otimização da gestão tributária. Os objetivos específicos englobam a análise da clusterização, visando identificar o assunto principal de cada grupo e a quantidade de acórdãos proferidos. Além disso, o artigo busca ressaltar que a análise dos resultados obtidos pelo modelo de aprendizado de máquina requer conhecimento em contabilidade e tributação, e que a aplicação desta tecnologia pode gerar impactos significativos na área contábil, especialmente no planejamento tributário, controles internos e processos internos das empresas.

Este estudo visa melhorar a compreensão das temáticas relacionadas à contabilidade que são objeto de disputas administrativas entre contribuintes e a Receita Federal no CARF. Embora haja esforços acadêmicos para agrupar documentos jurídicos, persiste uma lacuna a ser explorada quanto à aplicação dessas técnicas na triagem processual de acórdãos do CARF, especialmente em questões relacionadas ao IRPJ. Assim, esta pesquisa busca preencher essa lacuna investigativa, contribuindo para o conhecimento contábil, assegurando segurança jurídica e auxiliando o CARF em sua missão.

O CARF, órgão colegiado vinculado ao Ministério da Economia, julga em grau de recurso autuações impostas pela Secretaria Especial da Receita Federal do Brasil (RFB) a contribuintes pessoas físicas e jurídicas. Sucessor do antigo Conselho de Contribuintes, o CARF é organizado em câmaras e turmas (Rêgo, 2020). É imprescindível aprimorar seus processos, reduzindo tempos médios de julgamento e melhorando a prestação administrativa à sociedade. Novos sistemas informatizados devem automatizar a atividade de julgamento e proporcionar meios ágeis para apresentação de recursos pelos contribuintes (Serpa, 2021).

Esta pesquisa apresenta relevância para a literatura contábil em diversas dimensões. Inicialmente, sob uma perspectiva prática, auxilia profissionais e estudiosos da tributação na identificação, classificação e agrupamento sistemático das principais questões objeto de litígio tributário referente ao IRPJ. A contribuição científica reside na proposta de uma abordagem metodológica inovadora, fundamentada em modelo de aprendizado de máquina (ML) para agrupamento de documentos, proporcionando eficiência e confiabilidade na pesquisa dos precedentes administrativos tributários. Essa abordagem oferece um enorme potencial como ferramenta para análises práticas e futuras investigações nas áreas contábil-tributária e financeira.

Neste artigo, discutimos o conceito de clusterização e as principais contribuições práticas deste procedimento aplicado aos acórdãos do CARF. Em seguida, apresentamos de maneira sucinta a metodologia e o modelo ML adotado. Posteriormente, exibimos os resultados obtidos com a aplicação do modelo desenvolvido, identificando os principais '*clusters*' de acórdãos relacionados ao IRPJ. Por último, analisamos as implicações da pesquisa e apresentamos algumas conclusões relevantes para o campo da contabilidade.

2 CLUSTERIZAÇÃO DE ACÓRDÃOS DO CARF DE IRPJ E CONTABILIDADE

O objetivo de qualquer processo de agrupamento, independentemente do algoritmo utilizado, é estruturar ou dividir um conjunto de objetos não estruturados em grupos (*clusters*). A finalidade principal é minimizar as distâncias entre objetos dentro de um agrupamento e maximizar as distâncias dos elementos fora dele, ou seja, agrupar elementos similares (Calambás et al., 2015). Dito de outra forma, os documentos dentro de um *cluster* devem ser o mais semelhantes possível, enquanto documentos em diferentes *clusters* devem ser bastante distintos

(Martins, 2018).

Diante da impossibilidade de analisar prontamente grandes conjuntos de dados, especialmente sem correlações aparentes, métodos estatísticos, computacionais e de análise auxiliam no reconhecimento de padrões. Com o aumento exponencial de dados textuais, analisar padrões em documentos legais tornou-se desafiador. Um desafio na área jurídico-tributária é responder rapidamente à crescente demanda por questões tributárias. Utilizando mecanismos de agrupamento, é possível distribuir trabalho entre assessores, considerando a semelhança entre documentos (Oliveira & Nascimento, 2021).

O uso de palavras-chave é essencial para analisar grandes conjuntos de dados não estruturados, como os apresentados no CARF. Em geral, o especialista faz a triagem dos documentos e distribui os processos a serem julgados entre os membros da equipe, desviando-se da atividade principal do especialista (Oliveira & Nascimento, 2022).

Por meio do uso de algoritmos de aprendizagem de máquina, é possível aprimorar os processos de busca de casos de IRPJ no CARF, com o apoio de processos de agrupamento e categorização. Identificaram-se termos e referências legais, analisando expressões regulares para descobrir referências legais, que juntas correspondem às principais características utilizadas como parâmetros de agrupamento. A formação do *cluster* é realizada por um processo de divisão, no qual é inicialmente definido um *cluster* geral, e os documentos nele contidos são iterativamente analisados. Se a semelhança do documento for baixa em relação ao centróide do *cluster*, o documento é separado para formar um novo *cluster* (Silva et al., 2021). A mineração distribuída de dados é contemplada no campo da pesquisa que envolve a aplicação do processo de extração de conhecimento sobre grandes volumes de informações armazenadas em bancos de dados publicamente disponíveis do CARF (Liu & Chen, 2017). As novas ferramentas de análise de dados proporcionam oportunidades para realizar previsões, classificações e outras tarefas online em suas bases de dados localizadas em diferentes nós interligados através da internet (Rodríguez, 2015).

A clusterização dos assuntos abordados nos acórdãos do CARF relacionados ao IRPJ é interessante por diversos motivos, incluindo a conexão com a contabilidade:

i) Do ponto de vista do órgão julgador - CARF: identifica-se quais matérias demandam mais horas de trabalho dos conselheiros, permitindo uma melhor alocação e distribuição de processos entre as turmas, facilitando a especialização e, consequentemente, a eficiência na análise de questões contábeis e tributárias.

ii) Do ponto de vista dos contribuintes: é possível conhecer as matérias relativas ao IRPJ mais julgadas pelo CARF, o que permitiria que dedicassem maior atenção àquelas mais recorrentes, impactando diretamente na gestão contábil e fiscal das empresas. Tal conhecimento, se possível segregado por porte ou ramo de atividade, poderia nortear eventuais planejamentos tributários (lícitos, evidentemente) executados pelos contribuintes.

iii) Do ponto de vista da autoridade tributária (RFB): com a clusterização, é possível direcionar esforços para a) capacitação do corpo funcional em relação às matérias com maior litigiosidade, abordando também aspectos contábeis relacionados; b) criação de grupos especializados de auditores em determinadas matérias, incluindo as questões contábeis envolvidas.

iv) Do ponto de vista dos profissionais contábeis e jurídicos: a partir da identificação dos temas mais sujeitos a disputas, os profissionais podem se preparar e se qualificar para lidar com eles, aprofundando sua especialização em determinados temas contábeis e tributários, nos quais desenvolveriam vantagens comparativas em relação a outros profissionais.

Considerando a importância do IRPJ não apenas no contexto do contencioso administrativo, mas também na arrecadação federal e na contabilidade das empresas, optou-se por delimitar o escopo deste projeto à análise de decisões em processos relativos a esse tributo. Em geral, esses casos são julgados pela 1ª seção do CARF.

Outra consequência da limitação ao IRPJ é que a análise estará, naturalmente, restrita às autuações realizadas junto a pessoas jurídicas. Isso implica em focar em questões contábeis e fiscais que afetam diretamente a gestão e o planejamento tributário das empresas, contribuindo para uma melhor compreensão das implicações desses tributos na contabilidade corporativa.

3 METODOLOGIA

3.1 Coleta de dados

Os dados das decisões do CARF foram recuperados diretamente da página de pesquisa do órgão[1]. A pesquisa dos acórdãos pode ser realizada pelo número do processo ou do acórdão, pelo nome do relator, pelo nome ou CPF/CNPJ do contribuinte, ou por palavras-chave presentes na ementa ou no texto completo da decisão. A coleta de dados foi realizada usando a biblioteca *Selenium* do Python e a implementação *ChromeDrive*[2]. A forma de acessar o conteúdo das páginas foi o XPATH. A página do CARF possui a seguinte configuração:

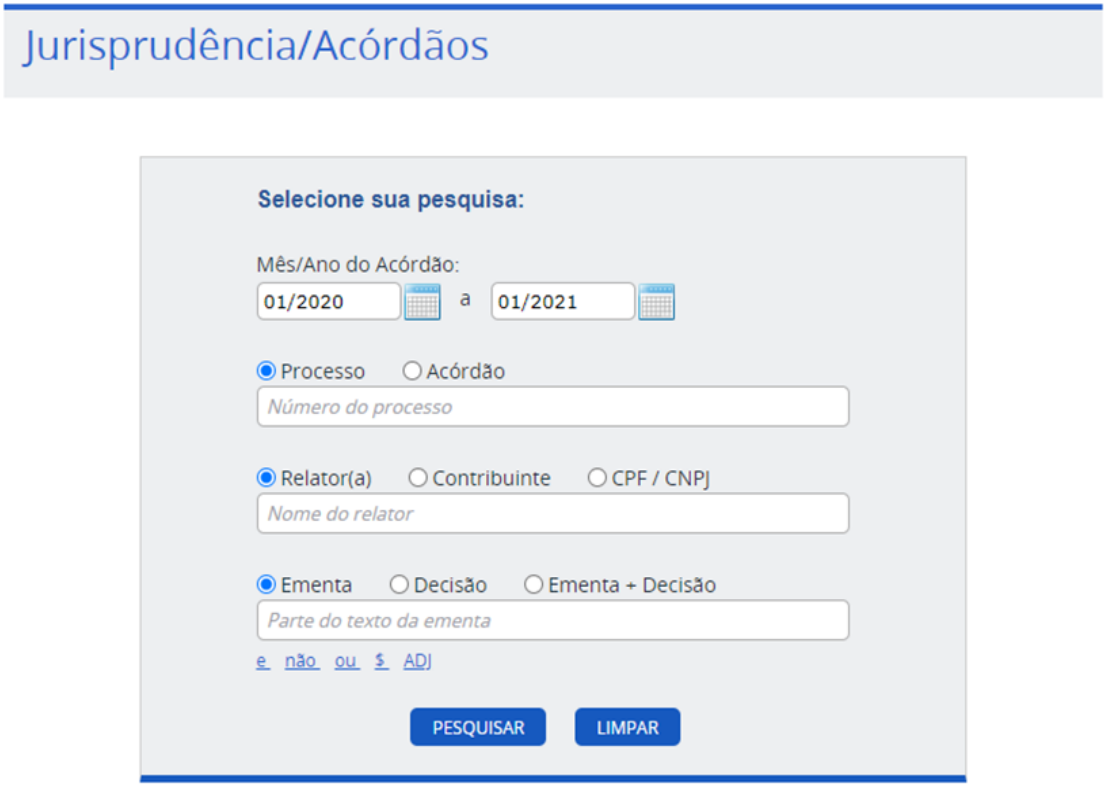

Figura 1. Coleta de dados do CARF

Considerando que o objetivo deste trabalho é analisar as decisões relacionadas ao IRPJ, optou-se por buscar acórdãos que apresentassem esse termo em sua ementa. De maneira geral, nota-se que as ementas podem abordar diversos assuntos distintos. Dentro de cada tópico, são informados os exercícios ou anos-calendário referentes aos fatos geradores ou às infrações, seguidos da apresentação da decisão no formato "PREÂMBULO <quebra de linha> Dispositivo". A teoria e a prática jurídicas estabelecem que a ementa é composta por duas partes: o preâmbulo, também conhecido como verbetação, e o dispositivo. O preâmbulo geralmente é escrito em letras maiúsculas (caixa alta) e consiste em uma sequência de palavras-chave ou expressões que indicam o assunto debatido no julgamento. O dispositivo é a regra resultante do julgamento do caso específico, devendo ser objetivo, conciso, afirmativo, preciso, unívoco, coerente e correto (Freitas, 2011).

Os "assuntos" mencionados explicitamente na ementa referem-se apenas ao tributo ou às normas gerais aplicáveis de forma genérica ao caso concreto, não sendo adequados para caracterizar, com precisão, a matéria objeto de análise. Dessa forma, para os propósitos deste trabalho, adotar-se-á a expressão "matéria" para se referir aos diversos temas tributários analisados, evitando confusão com os "assuntos" expressamente indicados nas ementa.

O resultado da coleta compreendeu 10.162 acórdãos proferidos no CARF no período de 2016 a 2020.

[1] https://carf.fazenda.gov.br/sincon/public/pages/ConsultarJurisprudencia/consultarJurisprudenciaCarf.jsf
[2] https://chromedriver.chromium.org/home

Considerando que o projeto visa a clusterização com base na *feature* "ementa", a análise exploratória dos dados focou-se nesse aspecto. O período analisado não incluiu os anos de 2021 e 2022 devido a mudanças no regimento do CARF, que afetaram a dinâmica das decisões. Além disso, uma greve dos conselheiros representantes da Fazenda afetou significativamente o andamento dos processos relacionados ao IRPJ, especialmente aqueles que envolviam grandes créditos tributários.

3.2 Tratamento de dados

Este projeto foi realizado integralmente na linguagem Python, versão 3.7.3, utilizando o ambiente Jupyter Notebook da plataforma Anaconda. A técnica escolhida para efetuar a clusterização foi o método de particionamento denominado *k-means*, pertencente à biblioteca *scikit-learn* do Python, ajustado sobre a matriz esparsa gerada pela aplicação do método tf-idf (*term frequency - inverse document frequency*) no texto da ementa, com redução de dimensionalidade.

Foram utilizadas as ferramentas *scikit-learn*[3] e NLTK[4] (*Natural Language Toolkit*), duas das principais bibliotecas para *machine learning* e NLP (*Natural Language Processing*) disponíveis para a linguagem Python. Para a parte gráfica, foram utilizadas as bibliotecas Matplotlib e Seaborn. Após a importação das bibliotecas, foram importados e concatenados em um único *DataFrame* os arquivos gerados nos dois *notebooks* anteriormente mencionados.

Inicialmente, analisou-se o "tamanho" da *feature* "ementa" em cada observação, em relação à quantidade de palavras. Utilizando a função "*describe*", foi possível verificar que a menor quantidade de palavras encontrada em uma ementa foi de 9, enquanto a maior foi de 1.784. A média de palavras por ementa é de 166,6, com desvio-padrão de 152,6. A mediana da distribuição é de 114 palavras. A plotagem do histograma foi realizada com o uso das bibliotecas Matplotlib e Seaborn.

Da análise dos termos extraídos nas ementas, percebe-se claramente os radicais das palavras “compensação”, “multa”, “receita”, “provisão”, “despesa”, “estimativa” e “omissão”, entre outras. Tais palavras já fornecem um primeiro *insight* sobre as matérias mais tratadas nos acórdãos. Algumas palavras (ou melhor, seus radicais) chamam a atenção: "ágio", "creditório", "decadência", "homologação", "DCOMP", "origem", entre outros. Esses radicais também são indicativos das matérias mais analisadas pelo CARF, relacionadas ao IRPJ.

3.3 Criação do modelo de *machine learning*

O aprendizado de máquina tem como objetivo criar algoritmos capazes de identificar padrões em grandes conjuntos de dados, com mínima intervenção humana. Dessa forma, os algoritmos "aprendem" a partir dos dados. O propósito deste estudo foi utilizar um modelo de aprendizado de máquina para agrupar ementas nas decisões do CARF, identificando as matérias julgadas e a relevância dessas matérias de acordo com o tamanho do grupo.

Existem três tipos de aprendizado de máquina: supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, os dados são rotulados e o modelo aprende a classificar ou prever valores para novos registros (Thangaraj & Sivakami, 2018). Esse tipo de modelo é adequado para situações em que há um conjunto de dados rotulados disponíveis para treinamento. A classificação ocorre no âmbito do aprendizado supervisionado, onde o sistema passa por treinamento e teste antes da classificação em si. Já no aprendizado não supervisionado, os dados rotulados não estão disponíveis. Apesar de ser um processo mais complexo e que pode apresentar problemas de desempenho, ele é adequado para lidar com grandes volumes de dados.

Neste estudo, a escolha do modelo de aprendizado não supervisionado foi a melhor opção, uma vez que não havia dados rotulados disponíveis. Além disso, esse modelo é adequado para situações em que é necessário modelar a distribuição dos dados para aprender mais sobre eles ou identificar padrões e agrupamentos.

Em algoritmos de agrupamento, a análise busca criar grupos de objetos similares, diferentes dos outros grupos. Neste estudo, técnicas de clusterização de documentos em formato de texto foram usadas para identificar matérias julgadas pelo CARF a partir das ementas dos acórdãos. Os passos a seguir são (Panagopoulos, 2020): i. Pré-processar texto: Inclui tokenização, *stemming* e remoção de *stopwords*. *Scikit-learn* faz a tokenização e a remoção de *stopwords* no *k-means*. ii. Representar documentos como vetores: Usar TF-IDF para transformar texto em vetor numérico, considerando a frequência e relevância das palavras. iii. Executar clusterização: Aplicar o algoritmo *k-means*, que se baseia na distância euclidiana e minimiza a soma dos erros ao quadrado (SSE) dos centróides de cada *cluster*. Os passos do *k-means* são: especificar o número de *clusters* (k), escolher k centroides

[3] https://scikit-learn.org/stable/
[4] https://www.nltk.org/

aleatórios, atribuir pontos ao centroide mais próximo, calcular o novo centroide de cada *cluster* e repetir até que a posição do centroide não mude, e iv. Avaliar resultado: Analisar as características dos *clusters* e verificar quais documentos estão em cada grupo. Ferramentas como "nuvem de palavras" (*word cloud*) são úteis.

Ao final do artigo se disponibilizam para livre acesso os dados extraídos do site do CARF e *notebooks* utilizadas para extração e tratamento de dados.

4 APRESENTANDO E DISCUTINDO OS RESULTADOS

Os *clusters* obtidos da análise da ementa dos acórdãos foram descritos com base nas matérias abordadas em cada um deles. A partir dessa análise, foi possível associar cada *cluster* à principal matéria nele contida, conforme descrito abaixo. As "palavras caracterizadoras do *cluster*" são aquelas mais relevantes no contexto e não necessariamente as mais frequentes. Antes da avaliação e visualização dos resultados em si, foi realizada a contagem dos acórdãos pertencentes a cada *cluster*.

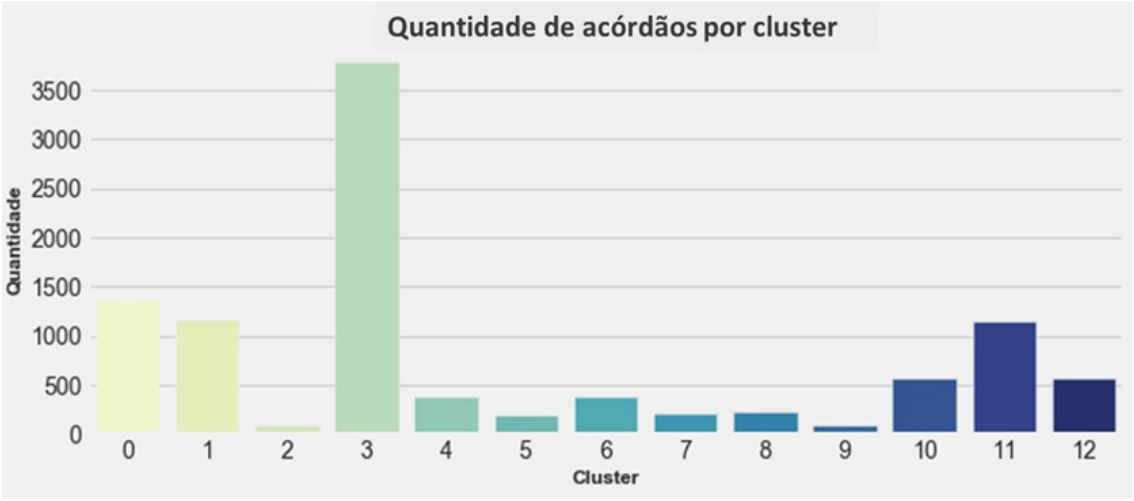

Figura 2. Distribuição da quantidade de acórdão a versar em IRPJ por *clusters*

Na tabela a seguir, apresentam-se a 10.162 acórdão do CARF a versar sobre o IRPJ classificados por temática.

Tabela 1. *Clusters* por temática e peso percentual dos acórdãos no período de estudo

Cluster	Tema	Quant. acórdãos	%
0	Pedidos de restituição e compensação de pagamento indevido	1.362	13,4%
1	Presunção de omissão de receitas decorrente de depósitos bancários de origem não comprovada	1.161	11,4%
2	Pedidos de compensação decorrente de erros de preenchimento de DARF e/ ou DCTF	91	0,9%
3	Lançamento de ofício em geral, tributação reflexa do IRPJ	3.787	37,3%
4	Coeficientes de presunção de lucro em serviços hospitalares ou de construção civil	378	3,7%
5	Créditos de PIS/PASEP – Conceito de insumos	188	1,9%
6	Amortização de ágio	373	3,7%
7	Preço de transferência	208	2,0%
8	Compensação de tributos pagos a maior face a prova	232	2,3%
9	Restituição e compensação de pagamento indevido por comprovação deficiente	98	1,0%
10	Multa isolada por falta de recolhimento das estimativas de IRPJ – Concomitância de multas	567	5,6%

Tabela 1. *Clusters* por temática e peso percentual dos acórdãos no período de estudo

Cluster	Tema	Quant. acórdãos	%
11	Restituição e compensação de saldo negativo de IRPJ	1.150	11,3%
12	Restituição e compensação por pagamento indevido, incluindo saldo negativo de IRPJ, gerado pelo pagamento de estimativa	567	5,6%
	Total de acórdãos	10.162	

Fonte: Elaborada pelos autores.

4.1 *Clusters* de acórdãos de julgamento a versar em IRPJ

Em seguida, buscou-se identificar as principais matérias em cada agrupamento, sendo este o objetivo principal deste estudo. Para tal identificação, a primeira técnica empregada foi a nuvem de palavras, também conhecida como "*word cloud*". Essa representação gráfica exibe os termos mais frequentes em cada agrupamento, permitindo uma rápida visualização da relevância de cada termo, graças ao tamanho com que cada palavra é apresentada na nuvem. O pacote *wordcloud* do Python foi utilizado para criar essas nuvens.

Decidiu-se construir as *word clouds* sem as *stopwords*, utilizando a variável "*ementa_clean*" para evitar a contaminação das nuvens com termos desnecessários. Os gráficos das *word clouds* foram analisados levando em consideração a área de negócios, ou seja, com base em conhecimentos específicos sobre a legislação tributária federal e contabilidade. Essa análise possibilitou captar o sentido das expressões no contexto do PAF (Processo Administrativo Fiscal), facilitando as conexões entre essas expressões e o reconhecimento das matérias abordadas. A seguir, as nuvens criadas pelo código, uma para cada *cluster*:

4.1.1 *Cluster* 0 – Pedidos de restituição e compensação de tributos pagos a maior (pagamento indevido) – Instrução Normativa (IN) RFB nº 1.717/2017

Principais palavras ou expressões caracterizadoras do *cluster*: “compensação”, “crédito”, “DCOMP”, “retificação”, “DCTF”, “pagamento indevido”, “liquidez”, “certeza”, “erro”, “preenchimento”, “despacho decisório”.

Figura 3. *Word cloud* do *cluster* 0

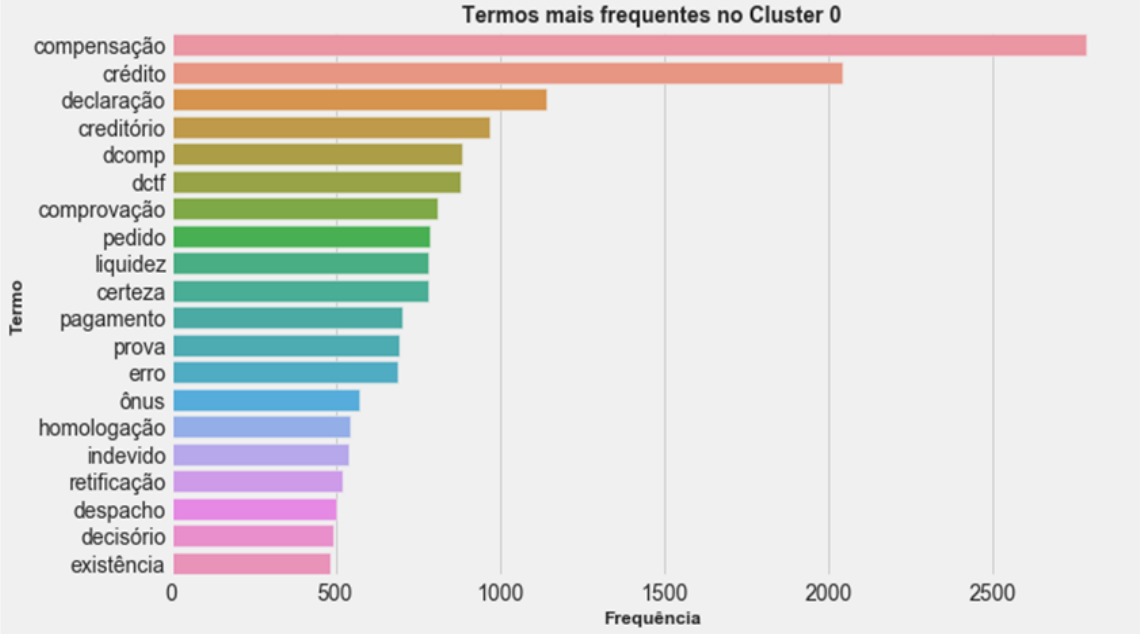

Figura 4. *Cluster* 0 – Pedidos de restituição e compensação de pagamento indevido

4.1.2 *Cluster* 1 – Lançamentos de ofício decorrentes de omissão de receitas apuradas com base em depósitos bancários de origem não comprovada (art. 42 da Lei nº 9.430/96)

Principais palavras ou expressões caracterizadoras do *cluster*: "omissão", "receita", "depósitos bancários", "origem comprovada", "multa qualificada", "interesse comum", "regularmente intimado", "presunção", "documentação hábil", "idônea, "cerceamento", "defesa", "fraude", "simulação" e "arbitramento".

Figura 5. *Word cloud* do *cluster* 1

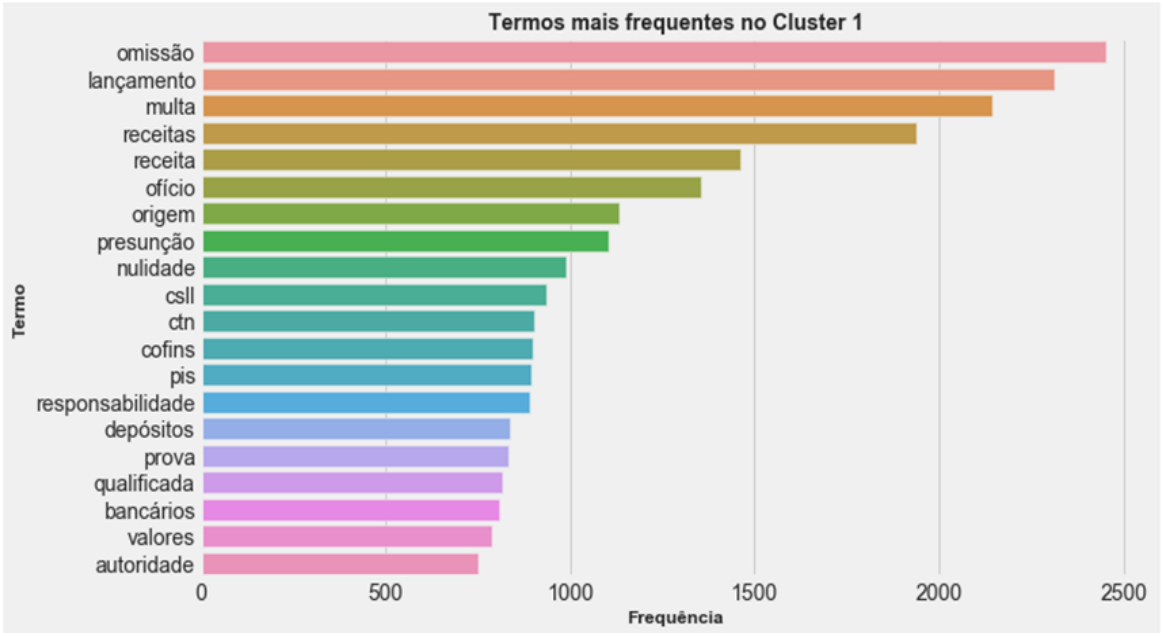

Figura 6. *Cluster* 1 – Presunção de omissão de receitas

4.1.3 *Cluster* 2 – Pedidos de compensação decorrentes de erro de preenchimento de DARF (Documento de Arrecadação de Receitas Federais) e/ou DCTF (Declaração de Débitos e Créditos Tributários Federais)

Principais palavras ou expressões caracterizadoras do *cluster*: "DARF", "DCTF", "compensação", "erro", "erros", "preenchimentos", "códigos", "pedido", "reconhecer" e "indébito".

Figura 7. *Word cloud* do *cluster* 2

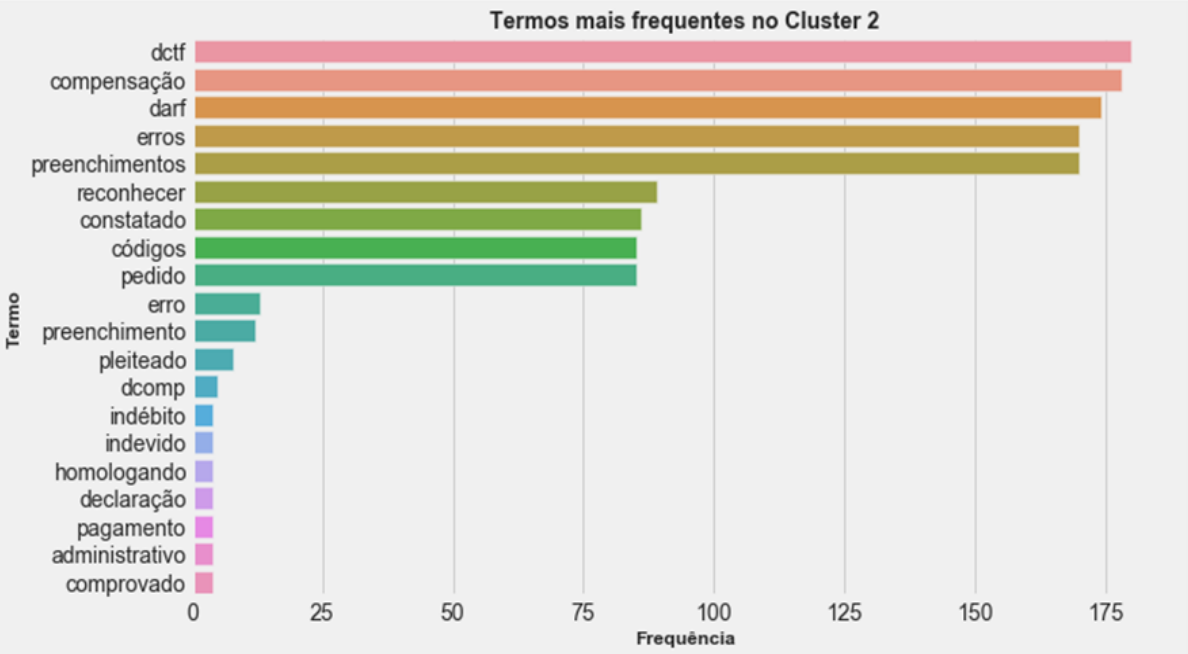

Figura 8. *Cluster* 2 – Pedidos de compensação decorrente de erros de preenchimento de DARF e/ou DCTF

4.1.4 *Cluster* 3 – Lançamentos de ofício em geral

Principais palavras ou expressões caracterizadoras do *cluster*: “lançamento”, “ofício”, “multa”, “juros”, “CSLL”, “fiscalização”, “cerceamento defesa”, “regime competência”, “glosa despesa”, “ganho capital”, “capital próprio”, “prazo decadencial”, “PIS”, “COFINS”.

Obs.: Trata-se de um *cluster* de caráter mais genérico. Não à toa, é o maior deles, com mais de 3.700 observações. Aparentemente, reúne diversas matérias distintas.

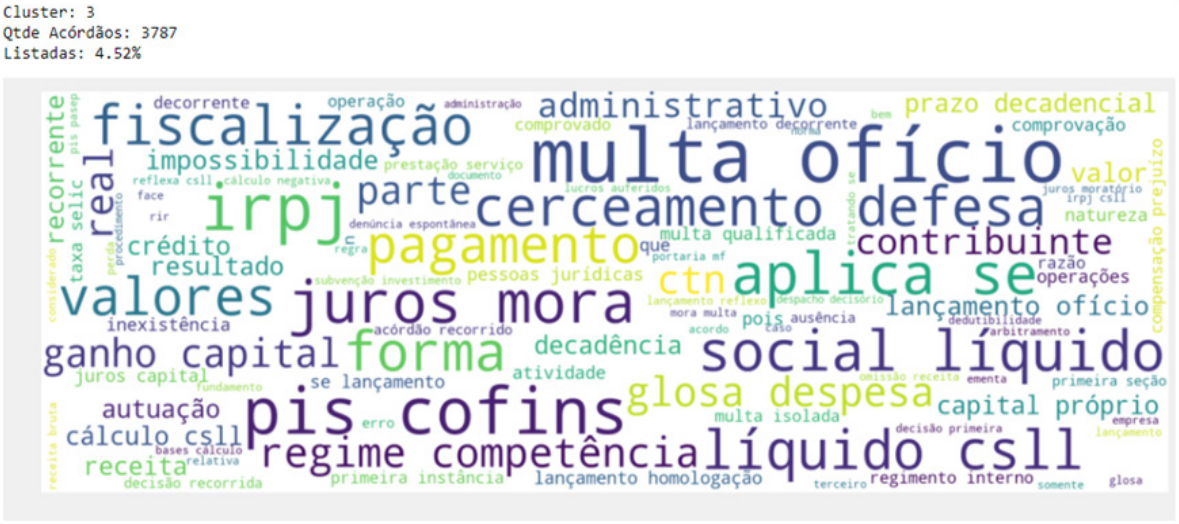

Figura 9. *Word cloud* do *cluster* 3

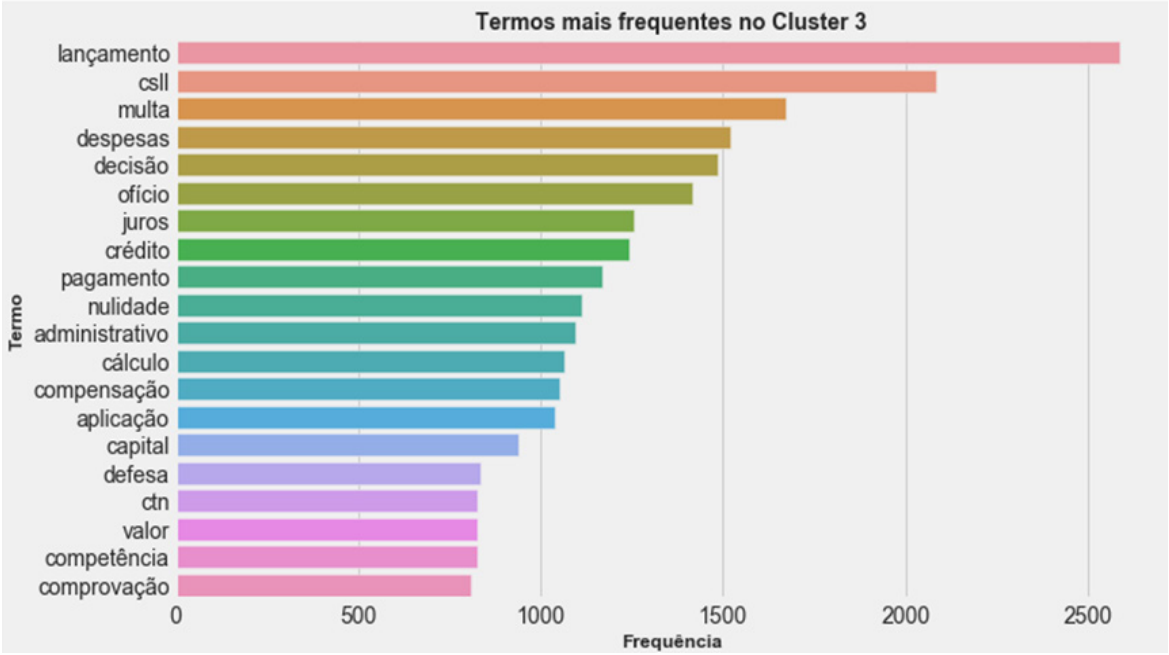

Figura 10. *Cluster* 3 – Lançamento de ofício em geral, tributação reflexa

4.1.5 *Cluster* 4 – Coeficiente (ou percentual) de presunção na apuração do lucro presumido, envolvendo receitas de serviços hospitalares e de serviços de construção civil, com ou sem o fornecimento de materiais (art. 15 da Lei nº 9.249/95)

Principais palavras ou expressões caracterizadoras do *cluster*: "percentual", "coeficiente", "serviços hospitalares", "presumido", "repetitivo", "STJ", "construção civil", "empreitada", "fornecimento".

Obs.: Aparentemente há jurisprudência formada sobre a matéria no âmbito do Superior Tribunal de Justiça (STJ), conforme sistemática dos recursos repetitivos.

```
Cluster: 4
Qtde Acórdãos: 378
Listadas: 0.0%
```

Figura 11. *Word cloud* do *cluster* 4

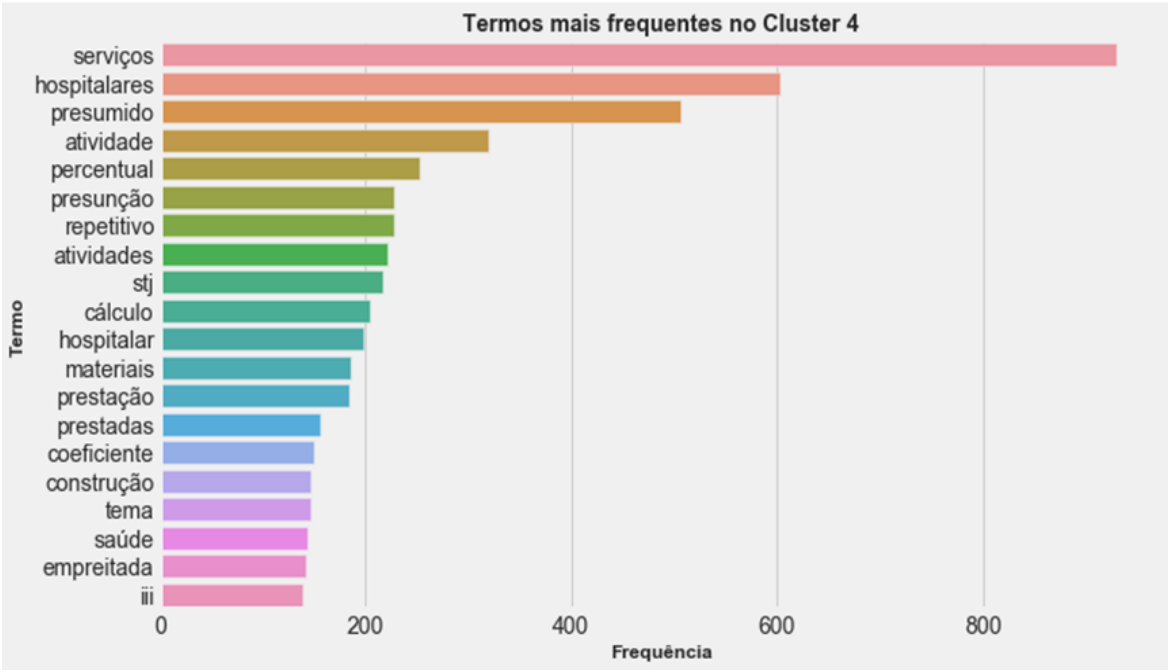

Figura 12. *Cluster* 4 – Coeficientes de presunção de lucro presumido em serviços hospitalares ou de construção civil

4.1.6 *Cluster* 5 – Créditos de Pis/Pasep e Cofins: Conceito de insumos utilizados na produção de bens ou produtos destinados à venda ou na prestação de serviços (Arts. 3º das Leis nº 10.637/02 e 10.833/03)

Principais palavras ou expressões caracterizadoras do *cluster*: “Pis”, “Pasep”, “Cofins”, “créditos”, “conceito”, “insumos”, “produtos”, “bens”, “serviços”, “destinados”, “venda”.

Obs: Aparentemente, não se trata de matéria relacionada ao IRPJ. Podem ser processos decorrentes de fiscalizações mistas, relativas ao IRPJ/CSLL e ao Pis/Cofins, concomitantemente. Isso explicaria o fato de o sistema do CARF ter retornado esses acórdãos quando da coleta.

Figura 13. *Word cloud* do *cluster* 5

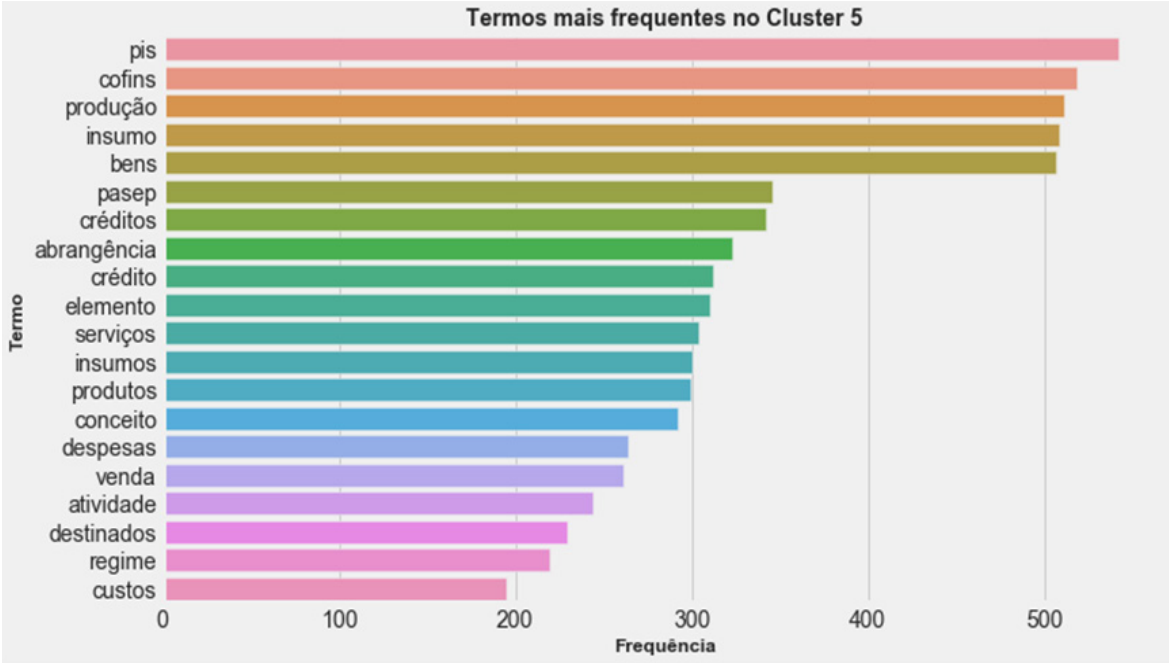

Figura 14. *Cluster* 5 – Créditos de PIS/PASEP – Conceito de insumos

4.1.7 *Cluster* 6 – Amortização de ágio derivado da expectativa de rentabilidade futura (arts. 385 e 386 do RIR/99 – Regulamento do Imposto de Renda, Decreto nº 3.000/99, e/ou legislação superveniente)

Principais palavras ou expressões caracterizadoras do *cluster*: "amortização", "ágio", "rentabilidade futura", "investidora", "investida", "aquisição", "investimento", "grupo econômico", "ágio interno", "veículo", "transferência ágio", "participação societária", "efetivo pagamento", "propósito negocial", "substância econômica", "multa", "ofício", "isolada".

Figura 15. *Word cloud* do *cluster* 6

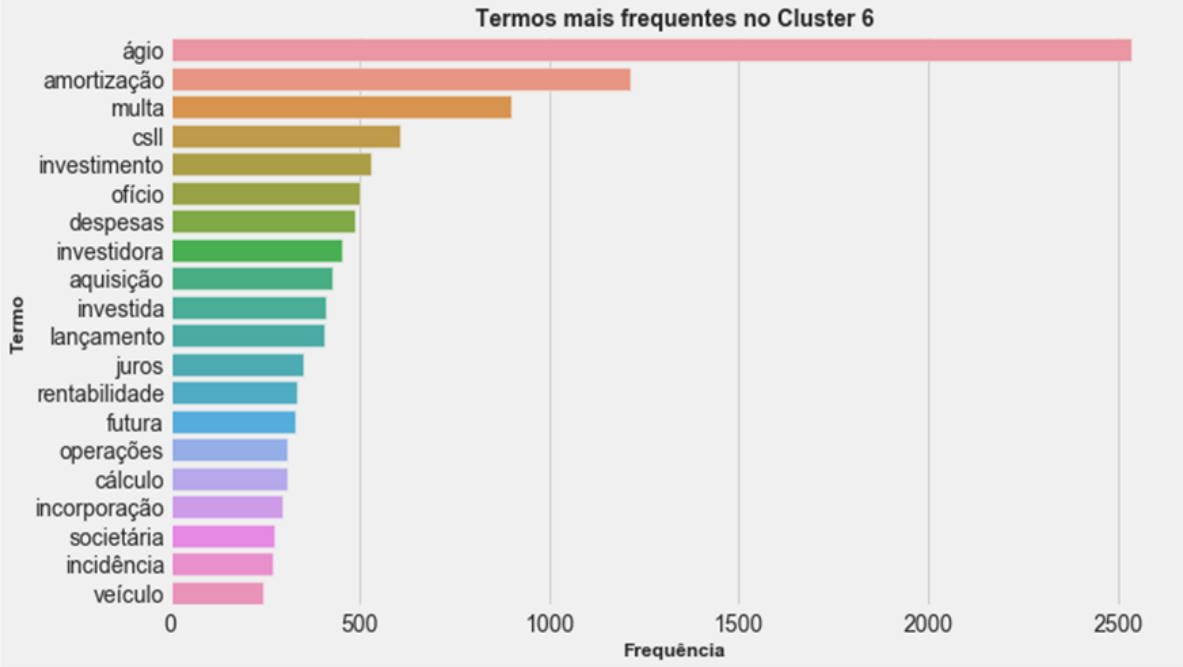

Figura 16. *Cluster* 6 – Amortização de ágio

4.1.8 *Cluster* 7 – Preços de transferência na importação, apurados pelo Método do Preço de Revenda menos Lucro (PRL) – IN SRF nº 243/2002, art. 12, e IN RFB nº 1.312/2012, art. 12

Principais palavras ou expressões caracterizadoras do *cluster*: "preço", "preços", "método", "transferência", "método PRL", "preço parâmetro", "instrução normativa", "bem importado", "frete", "seguro", "valor agregado".

Figura 17. *Word cloud* do *cluster* 7

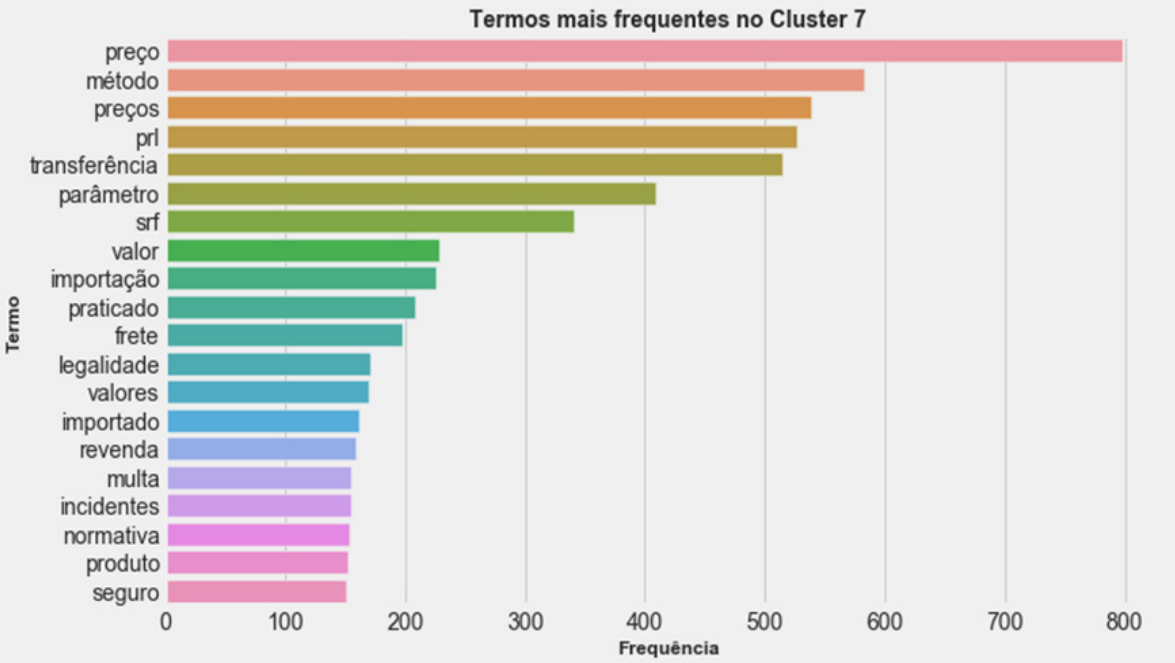

Figura 18. *Cluster* 7 – Preço de transferência

4.1.9 *Cluster* 8 – Compensação de tributos pagos a maior (pagamento indevido) em que a autoridade administrativa questiona a existência dos créditos – Questões relativas à prova, bem como à certeza e liquidez do crédito

Principais palavras ou expressões caracterizadoras do *cluster*: “existência”, “crédito”, “compensação”, “liquidez”, “certeza”, “prova”, “provas hábeis”, “comprovar”, “autoridade”, “administrativa”.

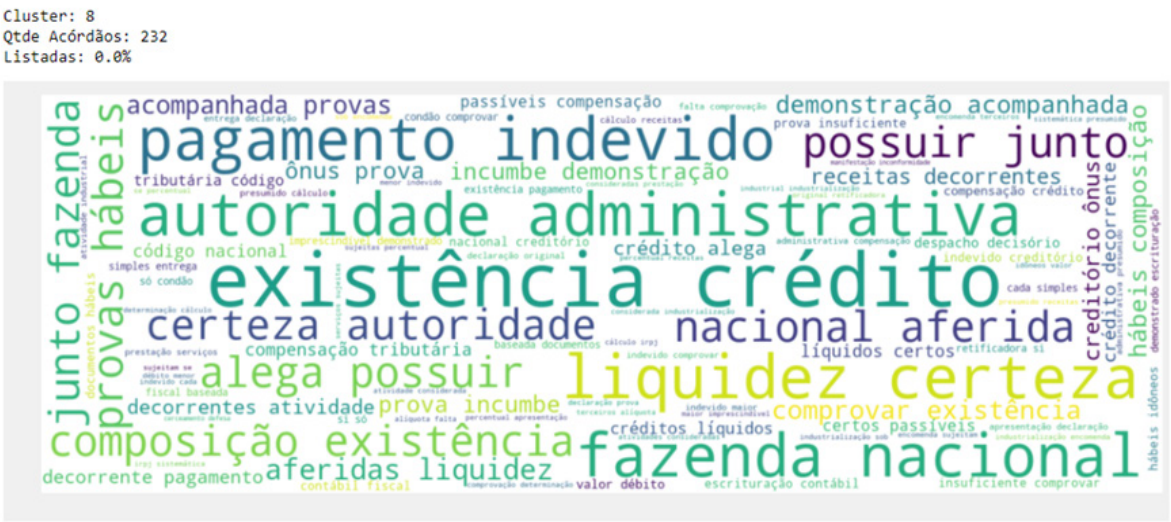

Figura 19. *Word cloud* do *cluster* 8

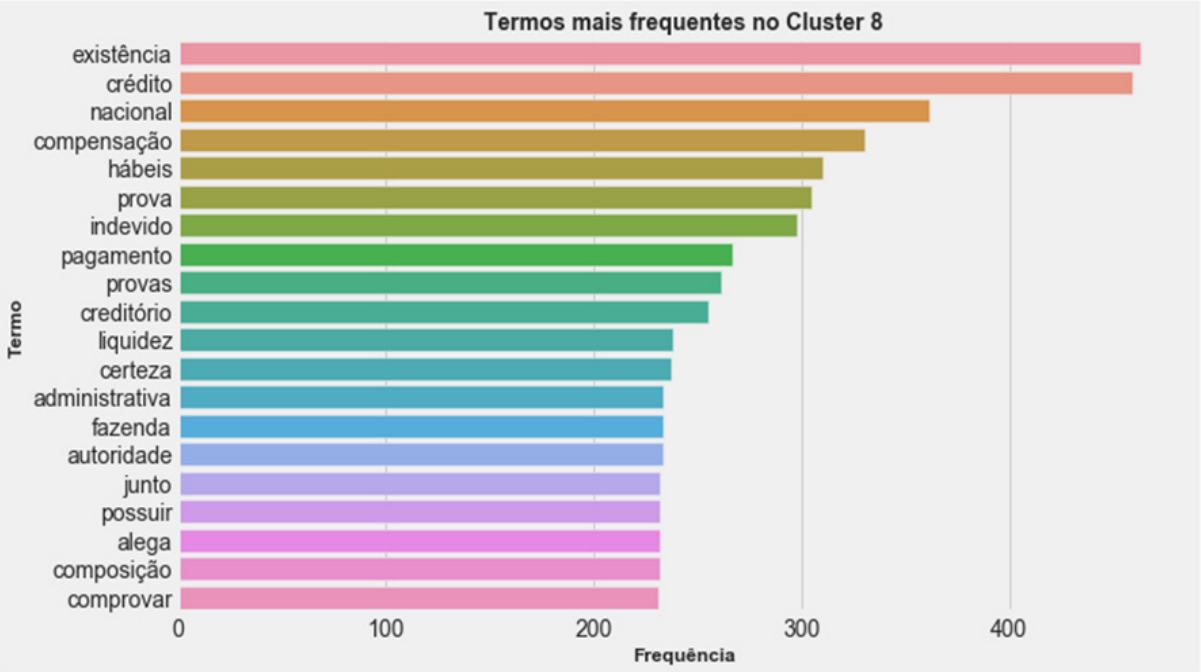

Figura 20. *Cluster* 8 – Compensação de tributos pagos a maior face a prova

4.1.10 *Cluster* 9 – Restituição e compensação referentes a pagamento indevido (indébito) com comprovação deficiente

Principais palavras ou expressões caracterizadoras do *cluster*: "indébito", "comprovação", "líquida", "certa", "compensação", "restituição", "comprovado", "deficiente".

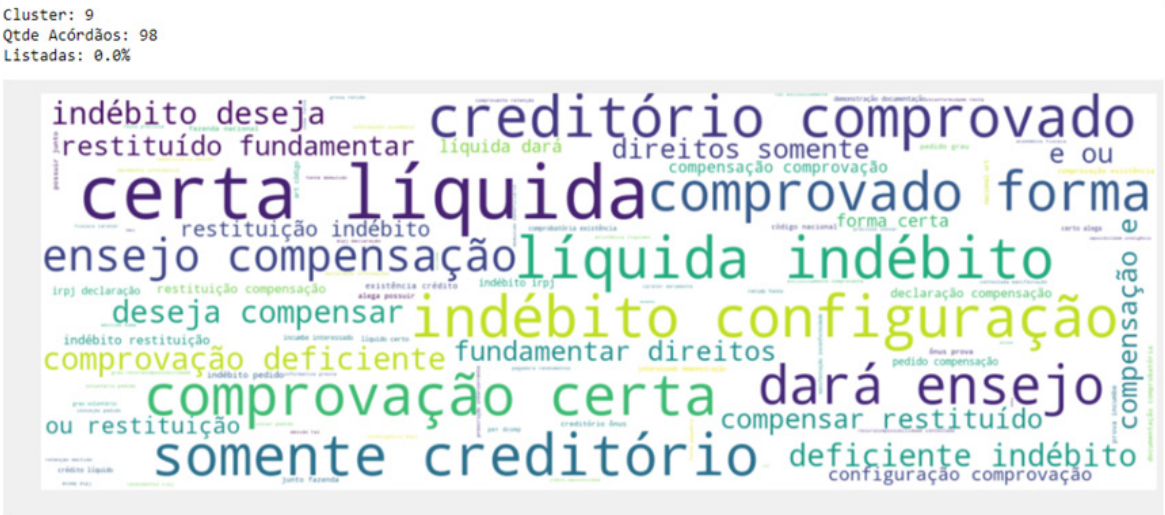

Figura 21. *Word cloud* do *cluster* 9

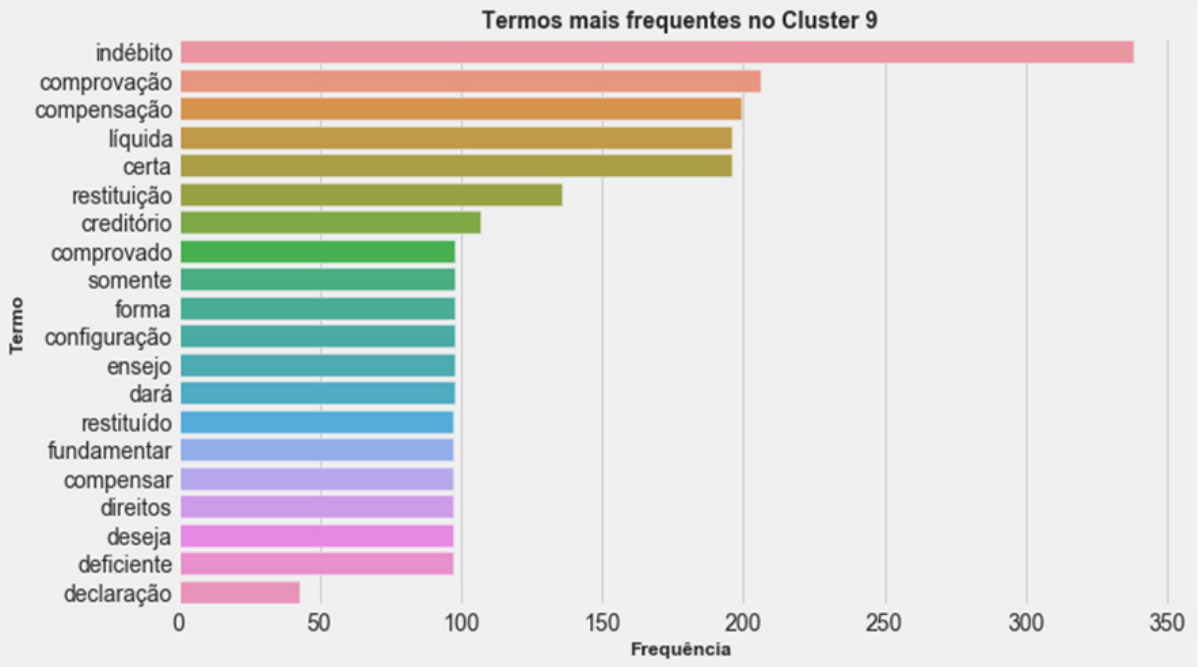

Figura 22. *Cluster* 9 – Restituição e compensação de pagamento indevido por comprovação deficiente

4.1.11 *Cluster* 10 – Multa isolada por falta de recolhimento das estimativas de IRPJ e CSLL – Concomitância da multa isolada com a multa de ofício (arts. 2º e 44, inciso II, da Lei nº 9.430/96)

Principais palavras ou expressões caracterizadoras do *cluster*: "multa isolada", "multa", "ofício", "juros", "falta recolhimento", "falta pagamento", "recolhimento estimativa", "estimativas mensais", "ajuste anual", "concomitância, "incidência", "IRPJ", "CSLL".

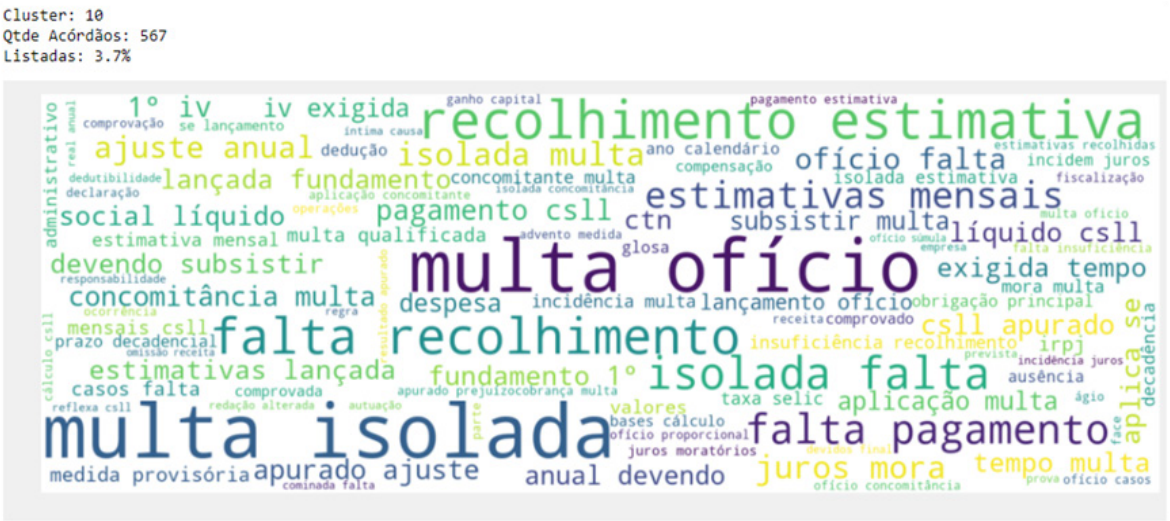

Figura 23. *Word cloud* do *cluster* 10

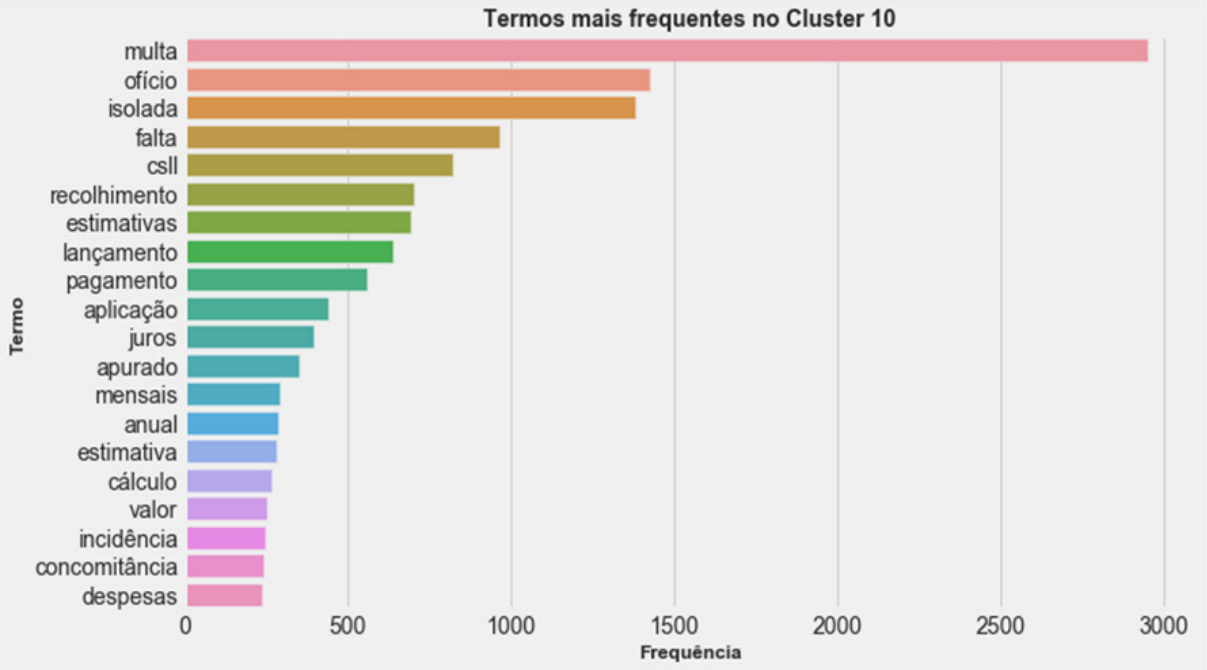

Figura 24. *Cluster* 10 – Multa isolada por falta de recolhimento das estimativas de IRPJ – Concomitância de multas

4.1.12 *Cluster* 11 – Restituição e compensação de saldo negativo de IRPJ (IN RFB nº 1.717/2017)

Principais palavras ou expressões caracterizadoras do *cluster*: “saldo negativo”, “PER”, “DCOMP”, “compensação”, “estimativas compensadas”, “retido”, “fonte”, “retenção”, “IRRF”, “creditório”, “declaração compensação”, “fonte pagadora”, “homologação”.

Figura 25. *Word cloud* do *cluster* 11

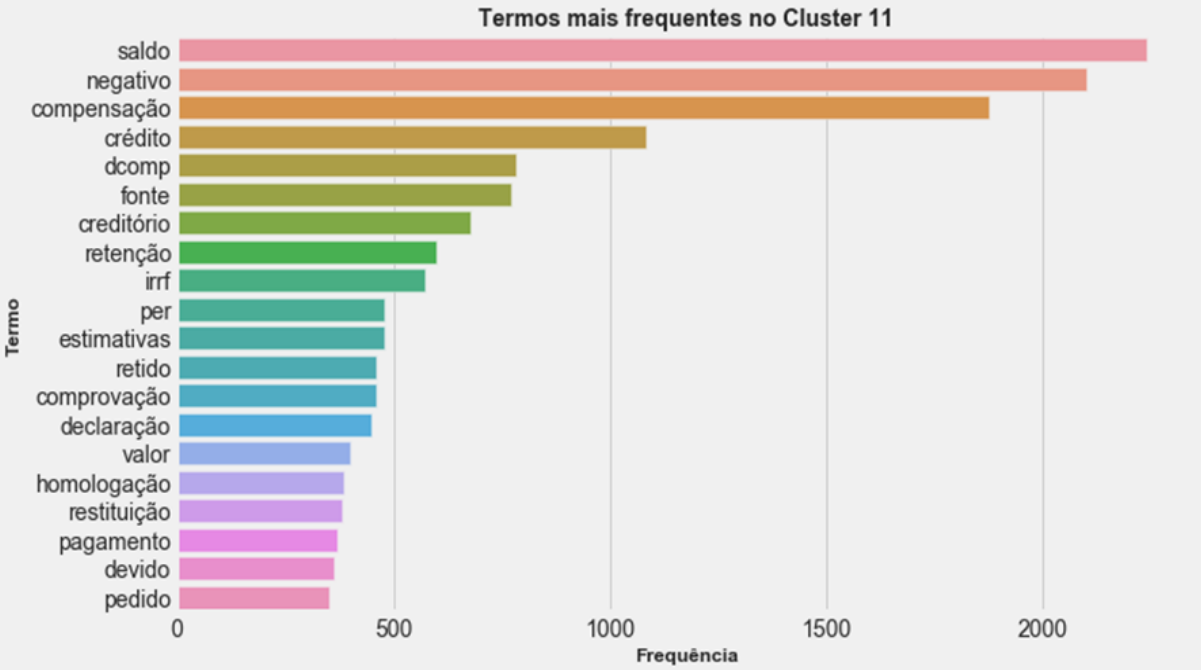

Figura 26. *Cluster* 11 – Restituição e compensação de saldo negativo de IRPJ

4.1.13 *Cluster* 12 – Restituição e compensação por pagamento indevido (indébito), incluindo o saldo negativo de IRPJ gerado pelo pagamento de estimativas – IN RFB nº 1.717/2017

Principais palavras ou expressões caracterizadoras do *cluster*: "pagamento indevido", "compensação", restituição", "estimativa", "pagamento", "indevido", "indébito", "crédito", "DCOMP", "PER", "recolhimento", "saldo negativo".

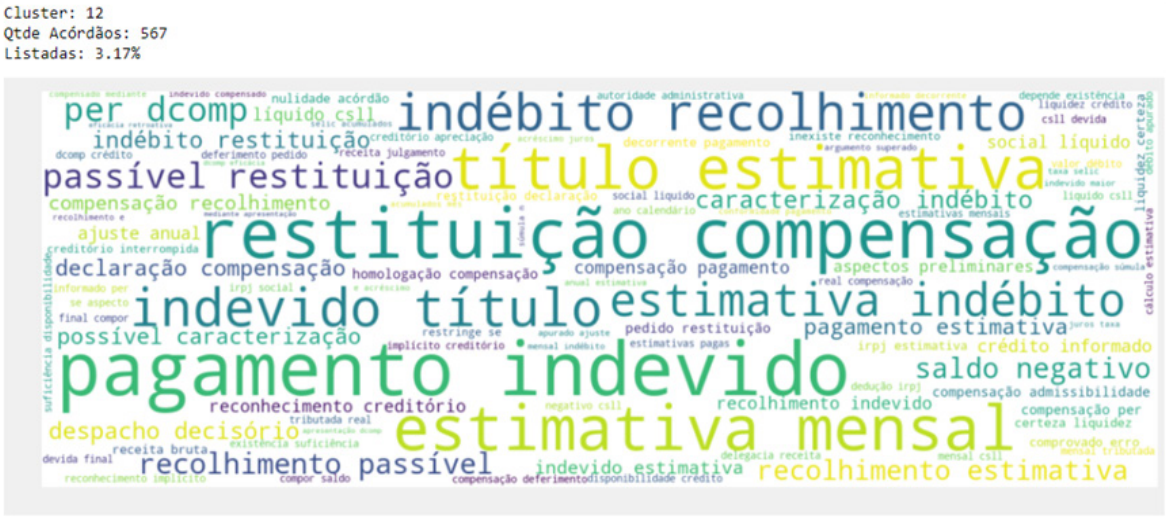

Figura 27. *Word cloud* do *cluster* 12

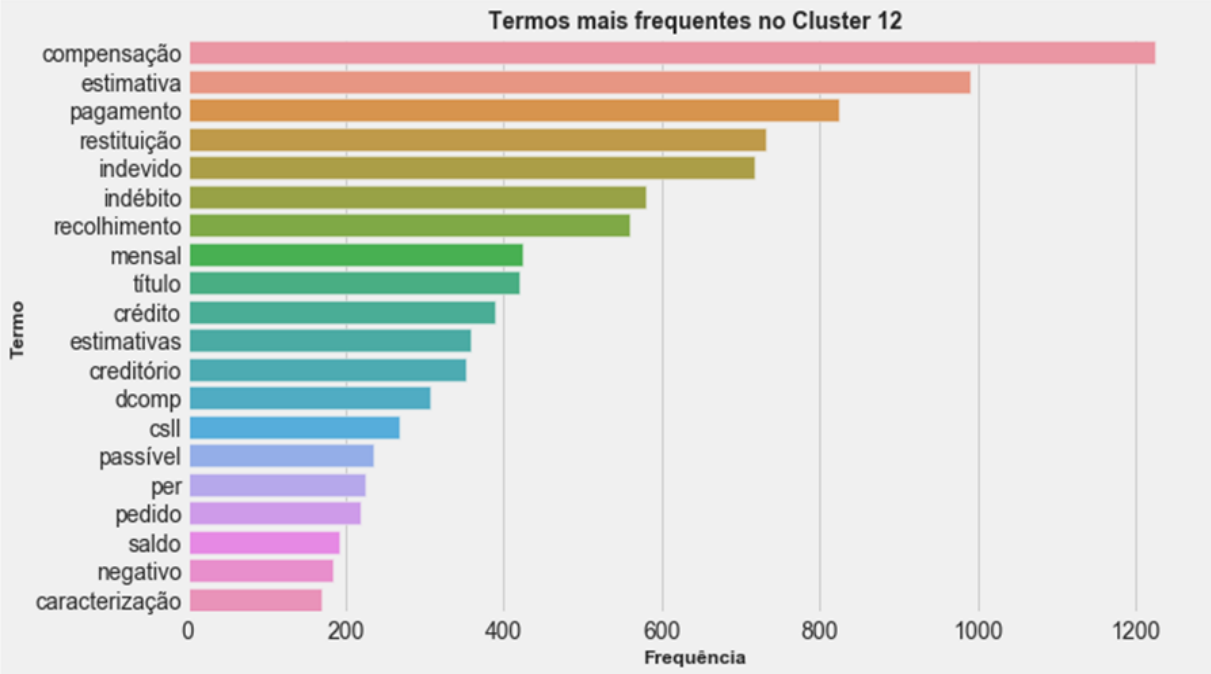

Figura 28. *Cluster* 12 – Restituição e compensação por pagamento indevido, incluindo saldo negativo de IRPJ, gerado pelo pagamento de estimativa

4.2 Discussão dos resultados

A análise clusterizada de 10.162 acórdãos, proferidos no CARF no período de 2016 a 2020, que versaram sobre IRPJ, mostrou uma grande variedade de decisões, envolvendo matérias como restituição/compensação decorrentes de pagamento indevido, restituição/compensação de saldo negativo de IRPJ, omissão de receitas com base em depósitos bancários de origem não comprovada, amortização de ágio e preço de transferência, entre outros.

Note-se que os *clusters* 0, 2, 8 e 9 parecem tratar de matéria afins, quer seja, a restituição e compensação por pagamento indevido ou maior que o devido. O mesmo acontece para os *clusters* 11 e 12, cuja matéria parece ser a restituição e compensação de saldo negativo de IRPJ. Ou seja, esses 6 *clusters* poderiam, em tese, ser reduzidos a apenas 2, o que reduziria o número total de *clusters* a 9, entretanto se encontram nuances em cada um deles que permitem uma diferenciação entre si.

Adicionalmente, percebe-se que o *cluster* 3 possui uma quantidade de observações muito maior do que os demais. É possível que a análise das características das decisões agrupadas nesse *cluster* possa oferecer uma explicação para essa concentração. Entretanto, em linhas gerais, esse *cluster* representa os lançamentos de ofício relativo a IRPJ, captando também Pis/Cofins e CSLL. Isso decorre do fato de que o que se decide em IRPJ é aplicável aos autos de infração reflexos. Os acórdãos nesse *cluster* certamente recaem sobre a mesma base fática de interesse em lançamentos de IRPJ, porém foram ressalvados os aspectos específicos inerentes à legislação do PIS/Cofins e CSLL. Aliás os efeitos de tributação reflexa do IRPJ também se fazem sentir nos *cluster* 5, que faz referência a controvérsias em torno de créditos de PIS/PASEP e COFINS, onde a preocupação com o conceito de insumo e produção se faz presente, revelando a que o tema é controverso, e define um tipo de matéria recorrente em acórdãos do CARF.

Dentre as temáticas de natureza primordialmente contábil, destacam-se os agrupamentos que abordam, respectivamente, aspectos relacionados aos coeficientes de presunção do lucro, amortização de ágio e preço de transferência. No *cluster* 4, a controvérsia se concentra no coeficiente (ou percentual) de presunção na apuração do lucro presumido, envolvendo receitas de serviços hospitalares e de serviços de construção civil, com ou sem fornecimento de materiais. O *cluster* 6 discute a contabilização do ágio na aquisição de participações societárias. O ágio é a diferença positiva entre o valor de mercado dos bens do ativo e o valor contábil desses mesmos bens na sociedade investida. Por sua vez, o *cluster* 7 aborda os preços de transferência na importação, que são preços estabelecidos entre empresas vinculadas para transferência de bens ou serviços. O Método do Preço de Revenda menos Lucro (PRL) é uma das formas de apurar esses preços. Há controvérsia sobre a consideração dos valores referentes a frete, seguro e tributos no preço praticado para fins de aplicação do PRL.

Entre os acórdãos, destaca-se também o *cluster* 1, relacionado à presunção de omissão de receitas a partir de depósitos bancários não comprovados, que ocorre quando os valores depositados em uma conta corrente não têm origem comprovada pelo titular da conta. Nesses casos, presume-se, com base na lei, que tais valores representam receita ou faturamento não declarado, e assim, cobrar IRPJ sobre esses montantes.

Por fim, existe ainda o *cluster* 10, referente à discussão sobre a possibilidade de aplicação da multa isolada e da multa de ofício ao mesmo tempo, em relação ao não pagamento do IRPJ e da CSLL. A multa isolada é aplicada quando há falta de recolhimento das estimativas mensais de IRPJ e CSLL. Por outro lado, a multa de ofício é aplicada quando o contribuinte não paga o IRPJ e a CSLL apurados no ajuste anual.

5 IMPLICAÇÕES PARA CONTABILIDADE

Diante da distribuição dos acórdãos por *clusters*, constata-se que a discussão sobre restituição e compensação tributária é de suma importância para profissionais contábeis, uma vez que possibilita a recuperação de valores pagos indevidamente. Quando ocorre o pagamento de impostos em excesso, a restituição consiste na devolução em espécie dos valores pagos a maior, enquanto a compensação representa a utilização desse montante para quitar tributos futuros. Os profissionais contábeis necessitam possuir conhecimentos aprofundados sobre as leis e regulamentos tributários a fim de executar adequadamente a restituição ou compensação tributária, incluindo a compreensão das normas referentes à extinção das obrigações tributárias.

Ademais, é imprescindível que estejam cientes dos processos para solicitar a restituição ou compensação tributária, como o uso do programa Pedido de Restituição, Ressarcimento ou Reembolso e Declaração de Compensação (PER/DCOMP). Neste contexto, a importância do controle interno se destaca, visto que auxilia na prevenção de erros e fraudes, garantindo a conformidade com as normas tributárias e a correta aplicação das leis. Um eficiente sistema de controle interno contribui para a precisão das informações contábeis e fiscais, minimizando riscos e evitando pagamentos indevidos ou a maior, o que impacta diretamente na saúde financeira e na reputação das empresas, bem como na redução dos litígios tributários.

Controles internos robustos são fundamentais para evitar lançamentos decorrentes de presunção de omissão de receitas a partir de depósitos bancários de origem não comprovada, assim como a incidência de multas isoladas e de ofício por falta de recolhimento de tributos no momento adequado. Os sistemas de controle interno também são relevantes para determinar em qual categoria e coeficiente de lucro a empresa se enquadram. Uma política contábil eficiente pode reduzir significativamente o montante de litígios, evitar multas e penalidades desnecessárias e assegurar a precisão das informações financeiras e tributárias.

O "conceito de insumos" se refere à não cumulatividade do PIS e COFINS. Insumos são determinados considerando sua essencialidade ou relevância na atividade econômica. Aspectos contábeis são cruciais para aplicar corretamente esse conceito, classificando bens e serviços e garantindo a adequação à legislação tributária.

Um planejamento tributário adequado é uma parte importante da contabilidade e pode ajudar as empresas a reduzir riscos e evitar litígios relacionados a questões fiscais. No que se refere aos preços de transferência, a contabilidade deve considerar as regras e regulamentos aplicáveis a fim de evitar problemas fiscais e garantir a conformidade com as leis tributárias. Em relação à amortização de ágio, deve-se tratar adequadamente, e conforme a legislação, a amortização do ágio para garantir a correta apuração e recolhimento dos tributos.

O planejamento tributário é uma ferramenta crucial para a gestão empresarial e deve estar alinhado à contabilidade gerencial. A presença de diversos acórdãos abordando temas como preço de transferência e amortização de ágio ressalta a relevância de um planejamento tributário adequado e eficaz. Em síntese, a contabilidade e o planejamento tributário estão intrinsecamente interligados, e a análise clusterizada dos acórdãos evidencia a necessidade de uma abordagem meticulosa e bem fundamentada para assegurar a conformidade fiscal e minimizar riscos.

6 CONCLUSÃO

Utilizando um algoritmo de Aprendizado de Máquina (ML), 10.162 precedentes do CARF relacionados ao IRPJ foram agrupados em 13 *clusters*. A análise da clusterização concentrou-se na identificação do principal assunto de cada *cluster* e na quantidade de acórdãos proferidos. Essa abordagem objetivou classificar processos semelhantes julgados no CARF, facilitando a coleta de informações valiosas para o andamento de processos relacionados ao IRPJ e fornecendo suporte à tomada de decisão por contadores e advogados tributários.

Ademais, a análise dos resultados do modelo de aprendizado de máquina requer conhecimento do analista

de dados na área de contabilidade e tributação. Nesse sentido, pode-se enfatizar a importância do estudo da Ciência de Dados por profissionais de todas as áreas do conhecimento, especialmente contabilidade e jurídico tributárias, considerando os desafios contemporâneos.

Incorporando os efeitos na contabilidade, em especial no planejamento tributário, controles internos e processos internos, a adoção de aprendizado de máquina pode gerar impactos significativos. Por exemplo, a identificação de tendências e padrões em decisões tributárias pode auxiliar empresas a desenvolverem estratégias de planejamento tributário mais eficientes e aperfeiçoar seus controles internos. Além disso, os *insights* obtidos com a análise de acórdãos podem contribuir para a melhoria de processos internos, como a revisão de procedimentos fiscais e a adaptação às mudanças na legislação tributária. Em suma, a utilização de aprendizado de máquina na análise de acórdãos do CARF pode fornecer informações valiosas para contadores e advogados tributários, além de contribuir para a otimização do planejamento tributário, aprimoramento de controles internos e melhoria de processos internos nas empresas.

Apesar de suas virtudes, esta pesquisa apresenta algumas limitações, como as dificuldades na coleta de dados e a necessidade de adaptar os *notebooks* para realizá-la sem interrupções - aspectos que podem ser abordados em futuros trabalhos. Uma limitação aparente foi a concentração de muitas decisões em um único grupo (*cluster* 3), dificultando a definição da matéria principal tratada. Pesquisas futuras podem se dedicar a solucionar essa dificuldade, explorando em mais detalhes a natureza dos lançamentos incluídos nesse *cluster*.

Para trabalhos futuros, sugere-se ampliar o escopo da análise para outros tributos (IRPF, Pis/Pasep, Cofins, tributos sobre o comércio exterior, entre outros) e expandir o período dos acórdãos analisados, abrangendo, por exemplo, aqueles emitidos a partir de 2010. Isso permitiria uma classificação das temáticas relevantes em diversos tributos. No aspecto metodológico, outra possibilidade seria a aplicação de algoritmos de clusterização utilizando outras características do banco de dados, como o tipo de acórdão, o ano da fiscalização, o valor do crédito tributário em litígio e o texto da decisão (e não apenas o da ementa).

BASE DE DADOS E *NOTEBOOKS*

https://1drv.ms/u/s!ApYzxkx0UDRUgaEYQ_8Y1QC5I511SA?e=MgsARV

REFERÊNCIAS

Borcan, M. (2020, junho 8). *TF-IDF Explained And Python Sklearn Implementation*. Medium. https://towardsdatascience.com/tf-idf-explained-and-python-sklearn-implementation-b020c5e83275

Calambás, M. A., Ordóñez, A., Chacón, A., & Ordoñez, H. (2015). Judicial precedents search supported by natural language processing and clustering. *2015 10th Computing Colombian Conference* (*10CCC*), 372–377. https://doi.org/10.1109/ColumbianCC.2015.7333448

Oliveira, R. S., & Nascimento, E. G. S. (2021). Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches. Em *Artificial Intelligence* (Vol. 0). IntechOpen. https://doi.org/10.5772/intechopen.99875

Oliveira, R. S., & Nascimento, E. G. S. (2022). Brazilian Court Documents Clustered by Similarity Together Using Natural Language Processing Approaches with Transformers. *arXiv:2204.07182* [*cs*]. http://arxiv.org/abs/2204.07182

Dhanani, J., Mehta, R., & Rana, D. (2021). Legal document recommendation system: A cluster based pairwise similarity computation. *Journal of Intelligent & Fuzzy Systems*, 41(5), 5497–5509. https://doi.org/10.3233/JIFS-189871

Freitas, V. P. de. (2011). *Ementas de acórdãos pedem clareza e precisão*. Consultor Jurídico. http://www.conjur.com.br/2011-nov-13/segunda-leitura-ementas-acordaos-pedem-clareza-precisao

Liu, Z., & Chen, H. (2017). A predictive performance comparison of machine learning models for judicial cases. *2017 IEEE Symposium Series on Computational Intelligence* (*SSCI*), 1–6. https://doi.org/10.1109/SSCI.2017.8285436

Martins, A. D. M. (2018). *Agrupamento automático de documentos jurídicos com uso de inteligência artificial*. https://repositorio.idp.edu.br//handle/123456789/2635

Panagopoulos, D. (2020). *Clustering documents with Python*. Medium. https://towardsdatascience.com/clustering-

documents-with-python-97314ad6a78d

Rêgo, A. G. (2020). *Em que medida um tribunal administrativo tributário federal contribui para a defesa de interesses da sociedade brasileira* [Curso de Altos Estudos em Defesa (CAED)]. Escola Superior de Guerra (Campus Brasília). https://repositorio.esg.br/handle/123456789/1124

Rodríguez, Z. E. M. (2015). Aplicación de la minería de datos distribuida usando algoritmo de clustering k-means para mejorar la calidad de servicios de las organizaciones modernas caso: Poder judicial. *Repositorio de Tesis - UNMSM*. https://cybertesis.unmsm.edu.pe/handle/20.500.12672/4472

Serpa, S. de V. (2021). *Uma análise econômica do contencioso tributário brasileiro* [Dissertação de Mestrado em Economia do Setor Público, Universidade de Brasilia]. https://repositorio.unb.br/handle/10482/42310

Serras, F. R. (2021). *Algoritmos baseados em atenção neural para a automação da classicação multirrótulo de acórdãos jurídicos* [Text, Universidade de São Paulo]. https://doi.org/10.11606/D.45.2021.tde-07062021-135753

Silva, I. L. A. da, Mello, R. F., Miranda, P. B. C., Nascimento, A. C. A., Maldonado, I. W. S., & Filho, J. L. M. C. (2021). Assessment of text clustering approaches for legal documents. *Anais do Encontro Nacional de Inteligência Artificial e Computacional* (*ENIAC*), 37–48. https://doi.org/10.5753/eniac.2021.18239

Thangaraj, M., & Sivakami, M. (2018). Text Classification Techniques: A Literature Review. Interdisciplinary *Journal of Information, Knowledge, and Management*, 13, 117–135. http://dx.doi.org.ezproxy.usal.es/10.28945/4066

Yang, F., Chen, J., Huang, Y., & Li, C. (2020). Court Similar Case Recommendation Model Based on Word Embedding and Word Frequency. *2020 12th International Conference on Advanced Computational Intelligence* (*ICACI*), 165–170. https://doi.org/10.1109/ICACI49185.2020.9177720

Como citar este artigo

Costa, F. C. L., Martinez, A. L., & Klann, R. C. (2023). Clusterização de precedentes de IRPJ no CARF. *Revista de Contabilidade e Organizações*, 17:e197181. DOI: http://dx.doi.org/10.11606/issn.1982-6486.rco.2023.197181