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

Cluster analysis of IRPJ precedents in CARF

Clusterização de precedentes de IRPJ no CARF

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

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

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.

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.

Article information
Received: May, 01st 2022
Approved: April, 27th 2023
Published: June, 02nd 2023
Responsible editor: Dr. Sílvio Hiroshi Nakao

Practical implications

Cluster analysis benefits: i) Judging bodies, identifying litigious issues in accounting and taxation, optimizing resources; ii) Taxpayers, clarifying IRPJ accounting controversies and facilitating tax compliance; iii) Tax authorities, directing training and improving inspection and tax and accounting guidance; iv) Tax and accounting professionals, promoting qualification and specialization in relevant topics.

Corresponding author: 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 INTRODUCTION AND PROBLEM

The judgments of the Brazilian Administrative Council of Tax Appeals (CARF) related to corporate income tax (IRPJ) are administrative decisions that establish criteria for calculating and taxing companies. They consist of a relevant topic in accounting, providing valuable information on the main causes of disputes and the criteria tax authorities use to verify a company's bookkeeping. In addition, the CARF judgments also help to clarify other issues, such as the measurement of the company's actual profit and the deduction of expenses.

The study of CARF judgments regarding IRPJ provides accounting professionals with a comprehensive knowledge of tax legislation and its implications for business, enhancing their ability to offer precise and effective advice to clients and mitigate the risk of tax sanctions. In addition, these judgments provide insight into tax planning and help accounting professionals identify opportunities to maximize tax deductions.

This article aims to analyze CARF judgments related to IRPJ using unsupervised Machine Learning (ML) and clustering technique. Text clustering seeks to identify groups of similar documents in a larger set, dividing them into semantically heterogeneous parts (Serras, 2021).

The main objective of this research was to use ML algorithms to cluster similar CARF judgments, facilitating the creation of uniform decisions based on semantic precedents and collaborating for the optimization of tax management. The specific objectives include the clustering analysis to identify the main subject of each cluster and the number of judgments. In addition, the study emphasizes that the analysis of the outcomes from applying the ML model requires knowledge of accounting and taxation and that the model can significantly impact accounting, especially in corporate tax planning and internal controls and processes.

This study contributes to increasing understanding of accounting-related issues that are the subject of administrative disputes between taxpayers and the Brazilian Federal Revenue Service in CARF. Although there are academic efforts to cluster legal documents, there is still a gap to be explored regarding the application of these techniques in the procedural screening of CARF judgments, especially regarding IRPJ. Thus, this research seeks to fill this gap, contributing to accounting knowledge, ensuring legal certainty, and helping CARF in its mission.

CARF is a collegiate body linked to the Ministry of Economy. It processes appeals imposed by the Special Secretary of the Brazilian Federal Revenue Service (RFB) on individual and legal taxpayers. Formerly known as Taxpayer Council, CARF is nowadays structured into chambers and classes (Rêgo, 2020), and its processes must constantly improve to reduce average trial times and offer optimal administrative performance. In this sense, Serpa (2021) advocates that new computerized systems should automate trial activity and provide agile means for taxpayers to submit appeals.

This research is relevant to the accounting literature in several dimensions. From a practical perspective, it helps professionals and scholars specialized in fiscal issues to identify, classify, and systematically group the main issues subject to tax litigation regarding the IRPJ. The scientific contribution lies in proposing an innovative methodological approach based on a ML model to cluster documents, providing efficiency and reliability in researching tax administrative precedents. This approach represents a tool for practical analysis and future research in accounting-tax and financial areas.

The next section discusses cluster analysis and the main practical contributions of this procedure applied to CARF judgments. The subsequent section outlines the methodology and the ML model adopted. Section four shows the results obtained with the application of the developed model, identifying the main clusters of judgments for IRPJ. The fifth section presents the research implications, followed by the conclusions.

2 CLUSTER ANALYSIS OF CARF JUDGMENTS RELATED TO IRPJ

The objective of any clustering process, regardless of the algorithm used, is to structure or divide a set of unstructured objects into clusters. The process minimizes the distances between similar objects within a group and maximizes the distances between the elements in a cluster and those outside it (Calambás et al., 2015), i.e., documents within a cluster should be as similar as possible, while documents in different clusters should be quite distinct (Martins, 2018).

Statistical, computational, and analytical methods can be used to identify patterns in large data sets, particularly when apparent correlations are not readily observable. With the exponential increase in textual data, analyzing patterns in legal documents has become challenging. A challenge in the tax-legal area is to respond quickly to the growing demand for tax issues. By using clustering mechanisms, it is possible to distribute work among law clerks, considering the similarity between documents (Oliveira & Nascimento, 2021).

Keywords are essential to analyze large unstructured data sets, such as those observed in CARF. Using keywords, specialists separate the documents and distribute the cases among team members, an activity that deviates from their main work, which is preparing the draft decisions for the cases (Oliveira & Nascimento, 2022).

Machine learning algorithms improve the search processes for IRPJ cases in CARF with the support of clustering and categorization processes. Legal terms and references were identified by analyzing regular expressions corresponding to the main characteristics used as clustering parameters. Cluster formation is carried out by a division process where a general cluster is defined, and the documents contained therein are iteratively analyzed. If the document similarity is low in relation to the cluster centroid, it is separated to form a new cluster (Silva et al., 2021). Distributed data mining is contemplated in the research field that involves knowledge extraction from large volumes of information stored in publicly available CARF databases (Liu & Chen, 2017). New data analysis tools provide opportunities to perform predictions, classifications, and other tasks online in their databases located in different nodes interconnected through the internet (Rodríguez, 2015).

The cluster analysis of the topics addressed in the CARF judgments regarding IRPJ is interesting for the connection with accounting and for reasons that are particular for the different players involved, as observed below:

i) From CARF's point of view, cluster analysis identifies which issues demand more hours of work from its counselors, allowing a better allocation and distribution of cases between them, facilitating specialization and, consequently, efficiency in the analysis of accounting and taxation issues.

ii) For the taxpayers, clustering allows them to identify the issues that CARF judges the most regarding IRPJ. Therefore, they can dedicate more attention to the recurrent ones, directly impacting the companies' accounting and tax management. Such knowledge, segregated by size or field of activity when possible and obtained from clustering results, could guide taxpayers when engaging in tax planning (lawful, evidently).

iii) For the tax authority (RFB), clustering directs efforts toward a) training the staff to work with the issues that generate the most disputes in administrative litigation, also addressing related accounting aspects; b) creating specialized groups of auditors in certain issues and the accounting topics they entail.

iv) For accounting and legal professionals, cluster analysis identifies the topics most likely to generate disputes, allowing these practitioners to prepare to deal with them, specialize in certain accounting and tax issues, and build competitive advantages in the market.

Considering the importance of IRPJ in administrative litigation, federal tax collection, and corporate accounting, the scope of this study was limited to analyzing decisions regarding this specific tax (IRPJ). In general, these cases are judged by CARF's 1st section.

A consequence of limiting the analysis to IRPJ is that it will naturally be restricted to assessments carried out with legal entities, such as corporations and partnerships, rather than individuals. This implies focusing on accounting and tax issues that directly affect companies' management and tax planning, contributing to a better understanding of the implications of these taxes on corporate accounting.

3 METHODOLOGY

3.1 Data collection

CARF judgments data was retrieved directly from the Administrative Council's search page[1]. The search engine allows retrieving data by the case or judgment number, by the rapporteur's name, the taxpayer's name or tax identification number (CPF for individuals and CNPJ for companies), or keywords found in the judgment summary or in the full text of the decision. Data collection was performed using Python's Selenium library and ChromeDrive implementation[2]. XPATH was used to access the content of the pages. The CARF page has the following configuration:

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

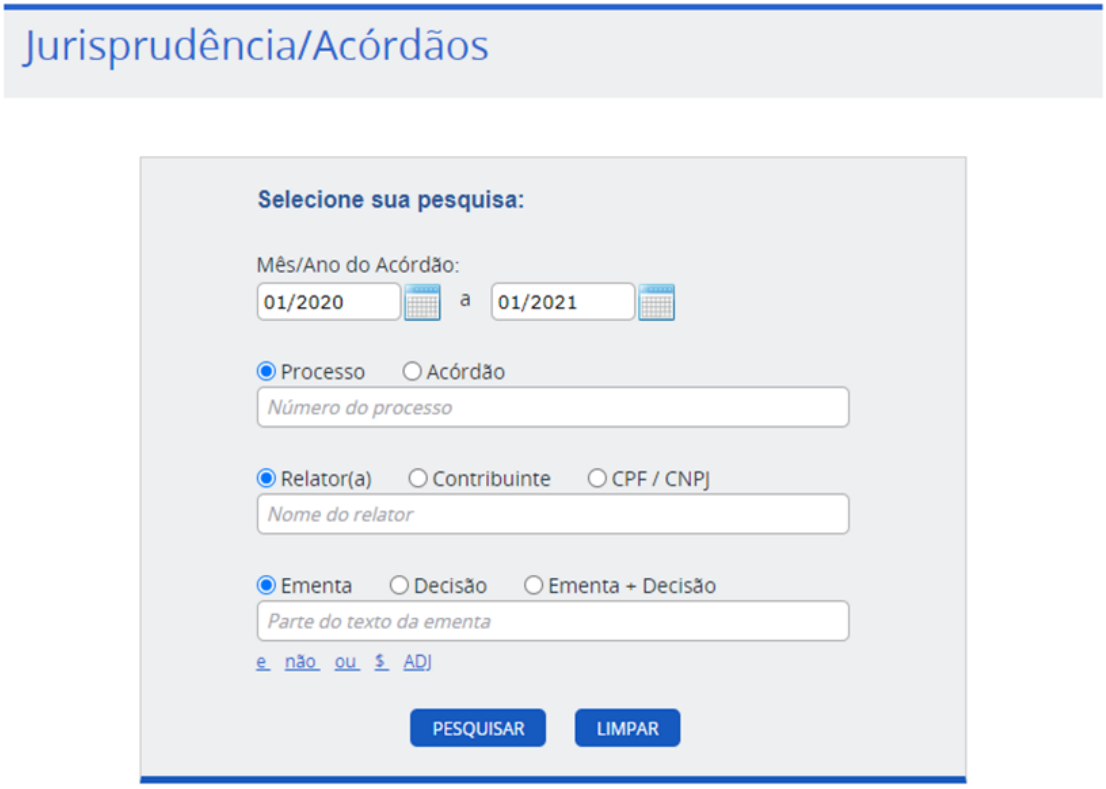

Figure 1. Data collection

Following the study's objective, the search used the term "IRPJ" (corporate income tax), identifying the decisions containing the term in the judgment summary. The summaries may address several subjects, each containing the fiscal or calendar years of the taxable event in litigation or the infraction under analysis, followed by the decision presented in the format "CASE BACKGROUND <line break> Judgment analysis." Legal theory and practice establish that the summary has two parts: the case background or preamble and the judgment analysis or decision. The first is usually written in capital letters and consists of keywords or expressions that indicate the subject discussed in the judgment. The second is the rule resulting from the judgment of the specific case, which must be objective, concise, affirmative, precise, unambiguous, coherent, and correct (Freitas, 2011).

The subjects explicitly mentioned in the judgment summary refer only to the tax or the general rules applicable generically to the specific case. They do not accurately characterize the issue under analysis. Thus, the expression "issue" was adopted to refer to the various tax-related topics, avoiding confusion with the subjects expressly indicated in the summary.

The data collection comprised 10,162 CARF judgments in the period from 2016 to 2020. Because the study aims to carry out a cluster analysis based on the judgment summary, the exploratory data analysis focused on this feature. The analyzed period did not include the years 2021 and 2022 due to changes in the CARF bylaws, which affected the dynamics of decisions. In addition, a strike by councilors representing the Brazilian Treasury affected the progress of IRPJ-related processes significantly, especially those involving large tax credits.

3.2 Data processing

This study was carried out using Python language, version 3.7.3, and the Jupyter Notebook environment of the Anaconda platform. The technique chosen to carry out the cluster analysis was the partitioning method called k-means, belonging to the Python scikit-learn library, adjusted on the sparse matrix generated by applying the tf-idf method (term frequency – inverse document frequency) in the text of the summary, with dimensionality reduction.

The tools scikit-learn[3] and nltk[4] (Natural Language Toolkit) were used, two of the main libraries for ML and NLP (Natural Language Processing) available for the Python language. For the graphic part, Matplotlib and

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

Seaborn libraries were used. After importing the libraries, the files generated in the two aforementioned notebooks were imported and integrated into a single DataFrame.

Initially, the analysis considered the summary's size (number of words) in each observation. Using the "describe" function, it was possible to verify that the smallest number of words found in a judgment summary was 9, while the largest was 1,784. The average number of words per summary was 166.6, with a standard deviation of 152.6. The median of the distribution was 114 words. Histogram plotting was performed using the Matplotlib and Seaborn libraries.

The analysis of the terms extracted from the summaries revealed the roots of the words "*compensação*" (offset), "*multa*" (fine), "*receita*" (revenue), "*provisão*" (provision), "*despesa*" (expense), "*estimativa*" (estimate), and "*omissão*" (omission), among others. Such words provide a first insight into the issues most addressed in the judgments. Some words (or rather, their roots) draw attention: "*ágio*" (premium), "*creditório*" (creditor-related), "*decadência*" (expiration), "*homologação*" (approval), "*DCOMP*" (compensation declaration), "*origem*" (origin), among others. These roots also indicate the issues most analyzed by CARF for IRPJ.

3.3 Creating the machine learning model

Machine learning (ML) aims to create algorithms capable of identifying patterns in large data sets, learning from the data with minimal human intervention. This study aimed to use an ML model to cluster CARF judgments summaries related to corporate income tax (IRPJ), identifying the issues and their relevance according to the cluster's size.

There are three types of ML: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, data is labeled, and the model learns to classify or predict values for new records (Thangaraj & Sivakami, 2018). This type of model is suitable for situations with a labeled dataset available for practice. Classification takes place in the scope of supervised learning, where the system undergoes practice and testing before classification. In unsupervised learning, labeled data is not available. Despite being a more complex process that can present performance problems, it is suitable for dealing with large volumes of data.

In this study, the unsupervised learning model was the best option since no labeled data was available. Also, this model is suitable for situations where one needs to model the data distribution to learn more about it or identify patterns and clusters.

In clustering algorithms, the analysis creates groups of similar objects, different from other groups. In this study, text document clustering techniques were used to identify, based on the summaries, the issues CARF judged. The steps taken were (Panagopoulos, 2020): i. i. Preprocess text: This step includes tokenization, stemming, and "stopword" removal. Scikit-learn performs tokenization and "stopword" removal in the k-means algorithm. ii. Represent documents as vectors: Transform the text into numerical vectors using tf-idf, considering word frequency and relevance. iii. Perform clustering: Apply the k-means algorithm, which is based on Euclidean distance and aims to minimize the sum of squared errors (SSE) of the centroids of each cluster. The steps in the k-means algorithm are as follows: specify the number of clusters (k), choose k random centroids, assign points to the nearest centroid, calculate the new centroid for each cluster, and repeat until the centroid positions no longer change, and iv. Evaluate the result: Analyze the characteristics of the clusters and determine which documents belong to each group. Visualization tools such as "word cloud" can be helpful in this analysis.

Data extracted from the CARF website and notebooks used for data extraction and processing are available at the end of the paper.

4 PRESENTATION AND ANALYSIS OF RESULTS

The clusters obtained from the judgments' summary analysis were described based on the issues addressed in each one. Based on this analysis, it was possible to associate each cluster with the main issue contained therein, as described below. The "cluster characterizing words" were those most relevant in the context and not necessarily the most frequent. Before evaluating and viewing the results, the judgments belonging to each cluster were counted.

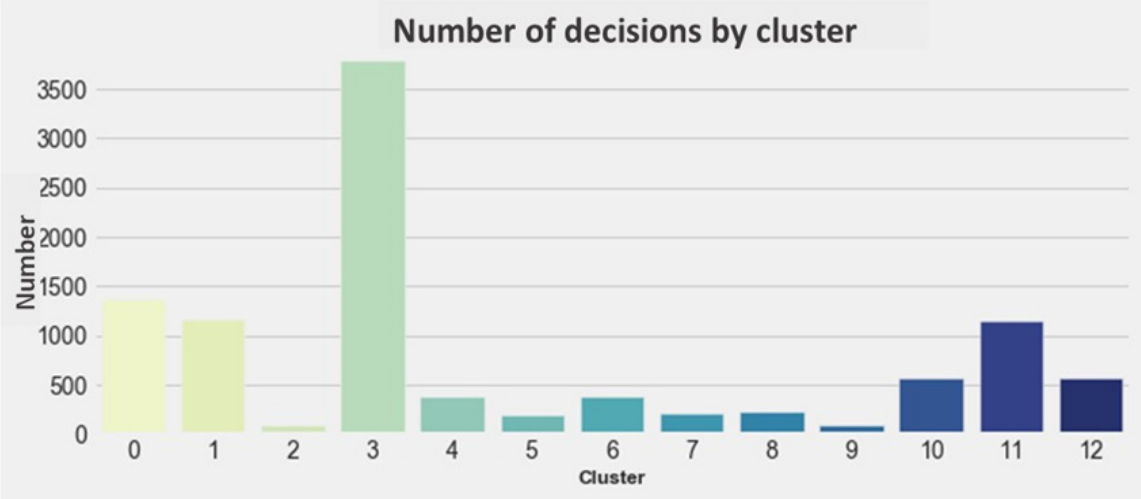

Figure 2. Distribution of the number of judgments for IRPJ by clusters

The table below shows the 10,162 CARF judgments related to IRPJ classified by theme.

Table 1. Clusters by theme and percentage of overall judgments in the analyzed period

Cluster	Theme	Number of judgments	%
0	Requests for refund and offsetting of undue payments	1,362	13.4%
1	Presumption of unreported income derived from unverified bank deposits	1,161	11.4%
2	Compensation requests due to errors in filling out the tax collection documents (DARF and/or DCTF)	91	0.9%
3	General tax assessment, indirect taxation of corporate income tax (IRPJ)	3,787	37.3%
4	Presumption of profit coefficients in hospital services or services in the construction industry	378	3.7%
5	PIS/PASEP credits - Concept of inputs	188	1.9%
6	Premium amortization	373	3.7%
7	Transfer pricing	208	2.0%
8	Offset of overpaid taxes based on evidence	232	2.3%
9	Refund and offsetting of undue payments due to insufficient evidence	98	1.0%
10	Isolated fine for failure to pay estimated IRPJ taxes – Concurrent fines	567	5.6%
11	Refund and offsetting of negative balance of IRPJ	1,150	11.3%
12	Refund and offsetting of undue payment, including negative balance of IRPJ generated by estimated tax payment	567	5.6%
	Total judgments	10,162	

Source: Elaborated by the authors.

4.1 Clusters of CARF judgments related to IRPJ

The study identified the main issues in each cluster, adopting the word cloud technique using Phyton's WordCloud package. This graphical representation displays the most frequent terms in each cluster, allowing quick visualization of the relevance of each term, thanks to the size with which each word is presented.

The word clouds were created without the "stopwords," using the variable "ementa_clean" (summary_clean) to avoid contaminating the clouds with unnecessary terms. The WordClouds charts were analyzed based on specific knowledge about federal tax legislation and accounting. This analysis captured the meaning of the expressions in the context of the PAF (Administrative Tax Process), facilitating the connections between these expressions and the recognition of the issues addressed. Below are the clouds created by the code, one for each cluster:

4.1.1 Cluster 0 – Requests for refund and offsetting of overpaid taxes (undue payment) – Normative Instruction (IN) RFB 1717/2017

Main words or expressions that characterize the cluster: "*compensação*", "*crédito*", "*DCOMP*", "*retificação*", "*DCTF*", "*pagamento indevido*", "*liquidez*", "*certeza*", "*erro*", "*preenchimento*", "*despacho decisório*" (offset, credit, compensation declaration, rectification, DCTF, undue payment, liquidity, certainty, error, completion, decision-making).

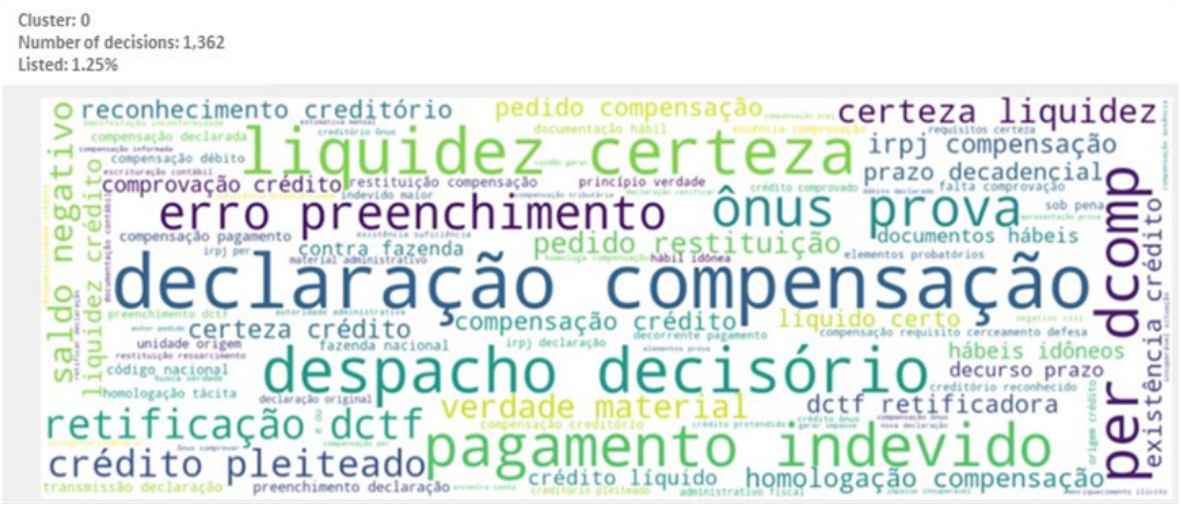

Figure 3. Word cloud of cluster 0

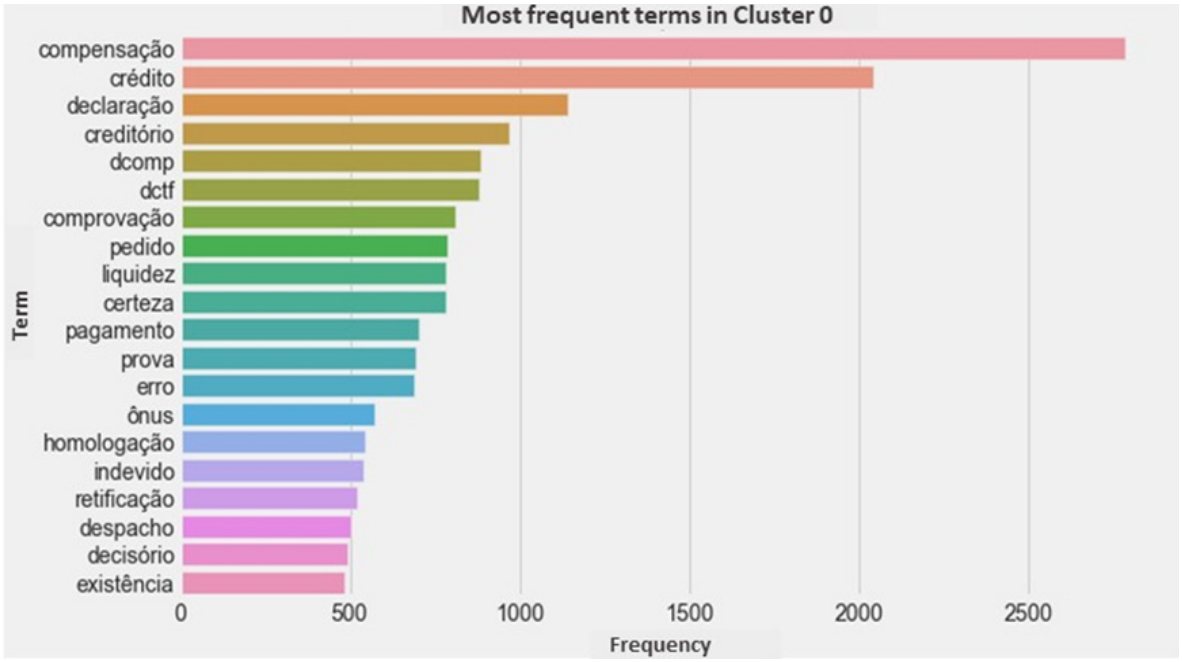

Figure 4. Cluster 0 – Requests for refund and offsetting of undue payments

4.1.2 Cluster 1 – Ex-officio entries arising from the omission of revenue calculated based on bank deposits of unproven origin (Article 42 of Law 9430/96)

Main words or expressions that characterize the cluster: "*omissão*", "*receita*", "*depósitos bancários*", "*origem comprovada*", "*multa qualificada*", "*interesse comum*", "*regularmente intimado*", "*presunção*", "*documentação hábil*", "*idônea*", "*cerceamento*", "*defesa*", "*fraude*", "*simulação*", "*arbitramento*" (omission, revenue, bank deposits, proven origin, qualified fine, common interest, regularly subpoenaed, presumption, valid documentation, reputable, restriction, defense, fraud, simulation, arbitration).

Figure 5. Word cloud of cluster 1

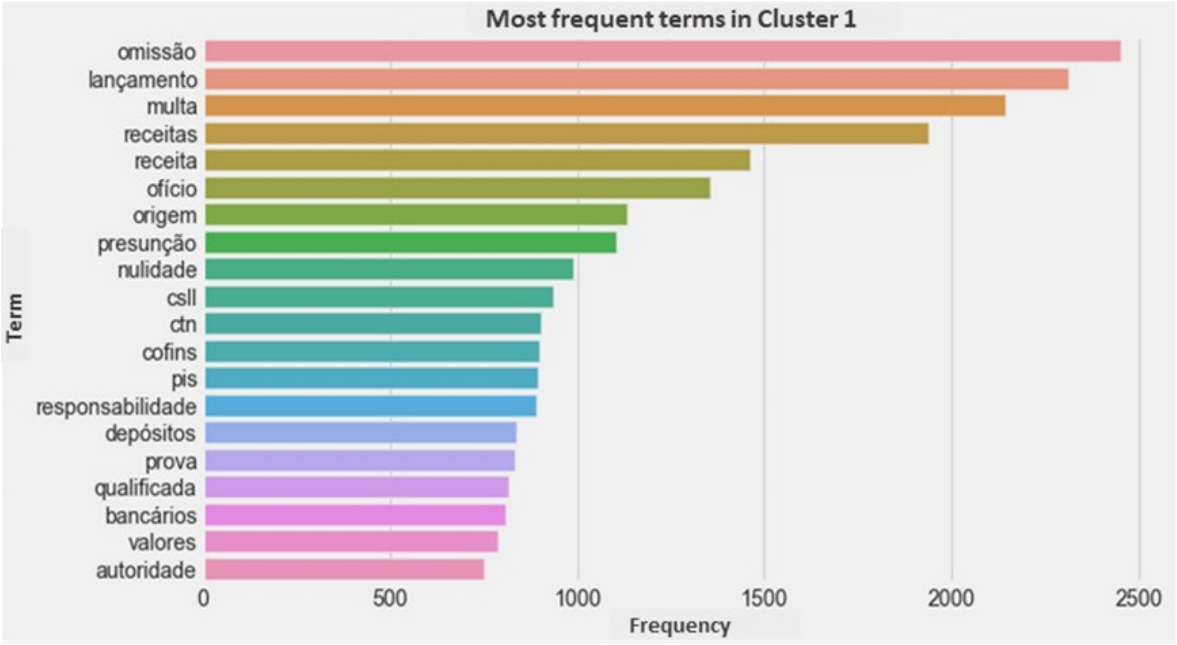

Figure 6. Cluster 1 – Presumption of unreported income

4.1.3 Cluster 2 – Compensation requests due to errors in filling out the DARF (Federal Revenue Collection Document) and/or DCTF (Declaration of Federal Tax Debts and Credits)

Main words or expressions that characterize the cluster: "*DARF*", "*DCTF*", "*compensação*", "*erro*", "*erros*", "*preenchimentos*", "*códigos*", "*pedido*", "*reconhecer*", "*indébito*" (DARF, DCTF, offset, error, errors, completions, codes, request, acknowledge, indebt).

Figure 7. Word cloud of cluster 2

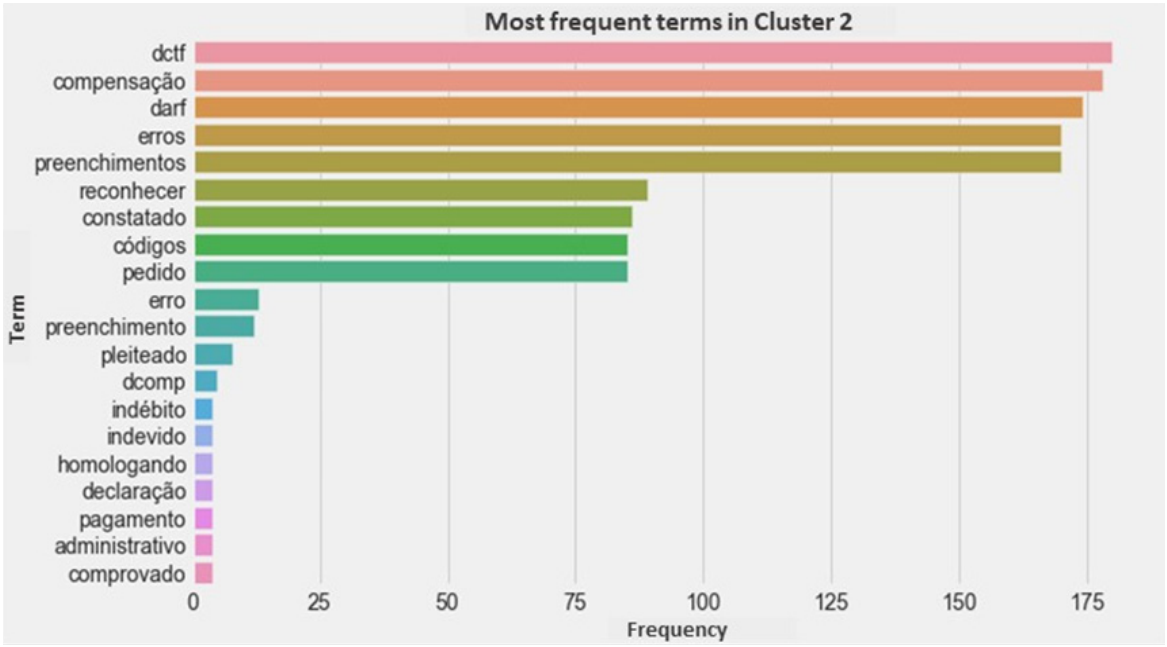

Figure 8. Cluster 2 – Compensation requests due to errors in filling out the tax collection documents (DARF and/or DCTF)

4.1.4 Cluster 3 – General tax assessment

Main words or expressions that characterize the cluster: "*lançamento*", "*ofício*", "*multa*", "*juros*", "*CSLL*", "*fiscalização*", "*cerceamento defesa*", "*regime competência*", "*glosa despesa*", "*ganho capital*", "*capital próprio*", "*prazo decadencial*", "*PIS*", "*COFINS*" (entry, ex-officio, fine, interest, tax on net profit, inspection, defense restriction, accrual regime, canceled expense, capital gain, equity, decay period, PIS, COFINS).

Note: This is a more generic cluster. Thus, it is the largest of them all, with more than 3,700 observations. It brings together several different issues.

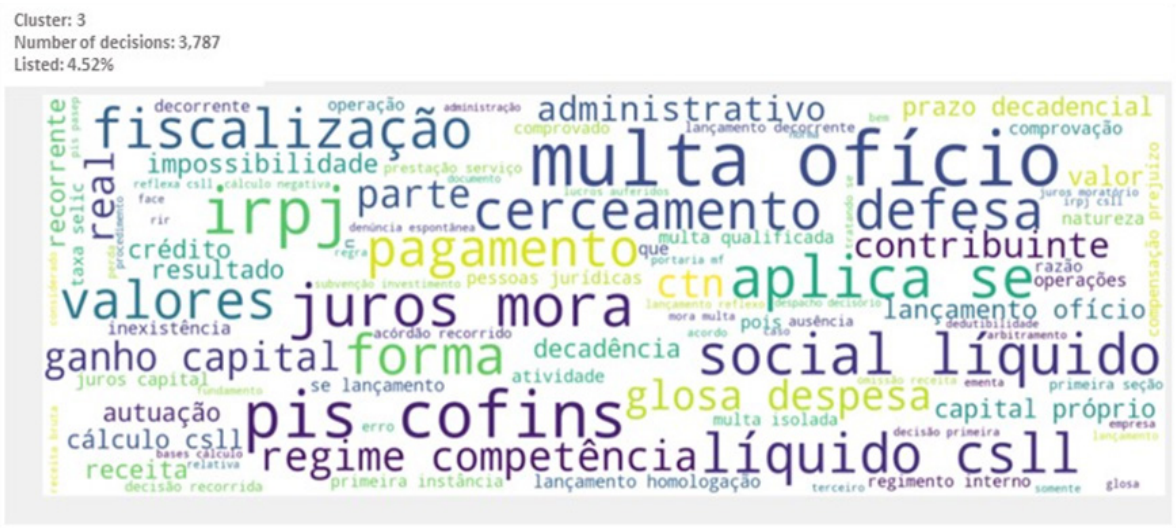

Figure 9. Word cloud of cluster 3

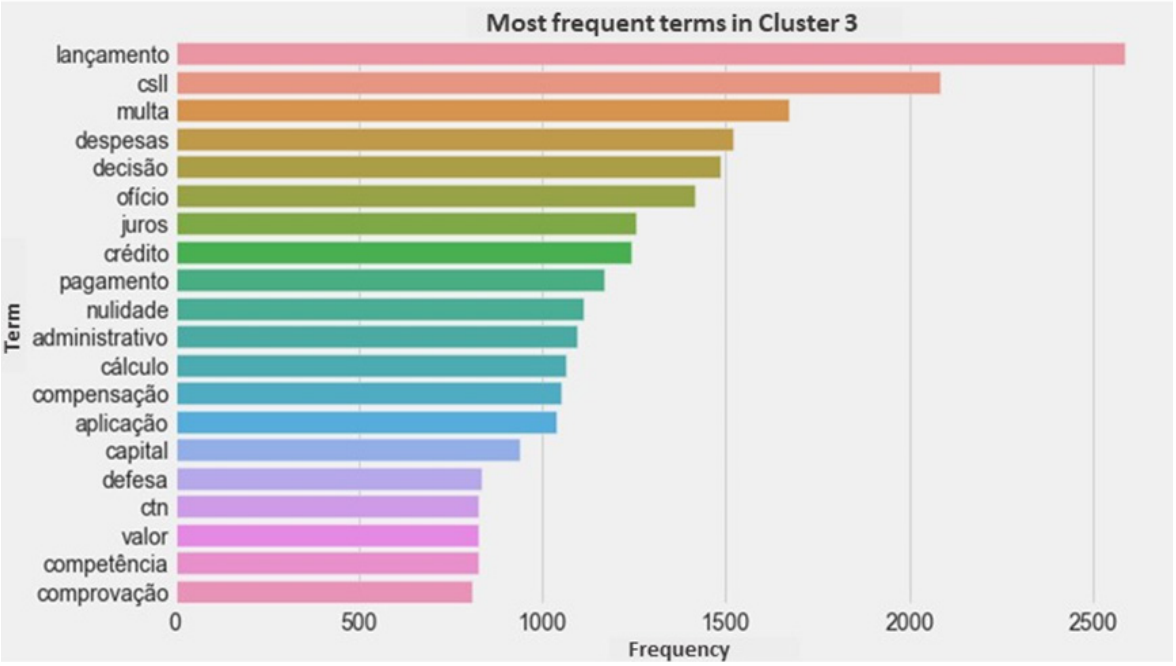

Figure 10. Cluster 3 – General tax assessment, reflex taxation

4.1.5 Cluster 4 – Presumption of profit coefficient (or percentage), encompassing revenues from hospital services and services in the construction industry, including or not the supply of materials (Article 15, Law 9249/95)

Main words or expressions that characterize the cluster: "*percentual*", "*coeficiente*", "*serviços hospitalares*", "*presumido*", "*repetitivo*", "*STJ*", "*construção civil*", "*empreitada*", "*fornecimento*" (percentage, coefficient, hospital services, presumed, repetitive, Superior Court of Justice, civil construction, contract work, supply).

Obs.: Following systematic repetitive appeals, jurisprudence is formed on the issue in the Superior Court of Justice (STJ).

Figure 11. Word cloud of cluster 4

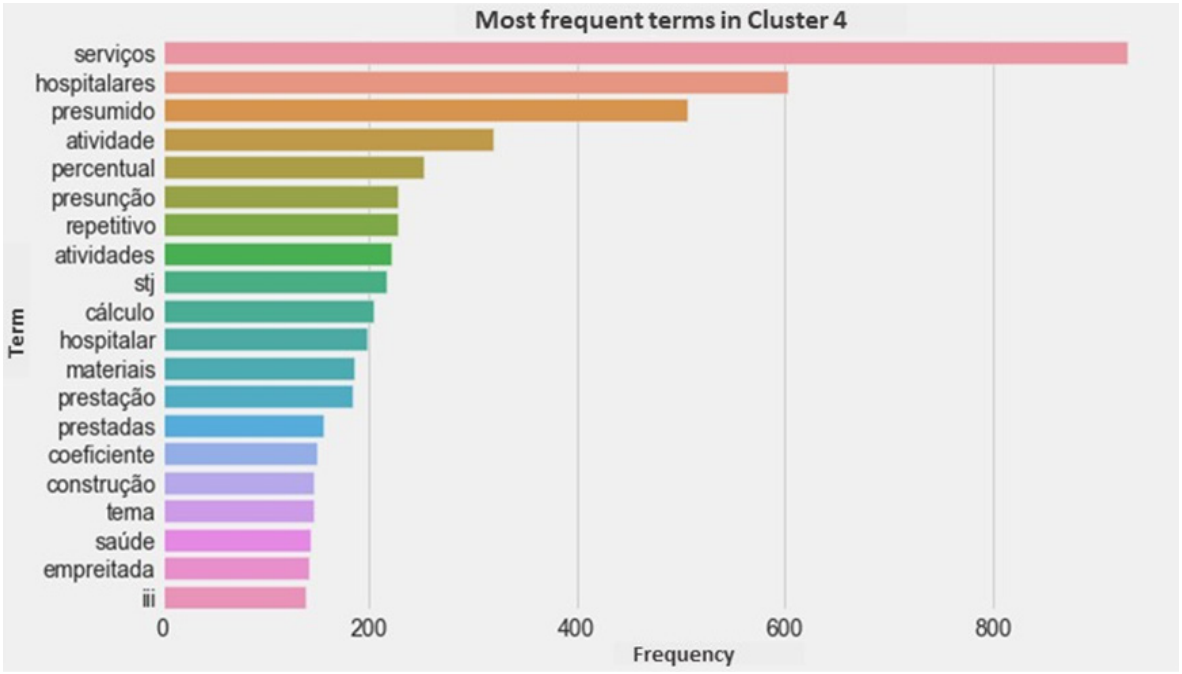

Figure 12. Cluster 4 – Presumption of profit coefficients in hospital services or services in the construction industry

4.1.6 Cluster 5 – Credits to PIS/PASEP and COFINS – Concept of inputs used in the production of goods intended for sale or provision of services (Articles 3 of Laws 10637/02 and 10833/03)

Main words or expressions that characterize the cluster: "*PIS*", "*PASEP*", "*COFINS*", "*créditos*", "*conceito*", "*insumos*", "*produtos*", "*bens*", "*serviços*", "*destinados*", "*venda*" (PIS, PASEP, COFINS, credits, concept, inputs, products, goods, services, destined, sale).

Note: These are not issues related to IRPJ. They may be processes resulting from mixed inspections, related to IRPJ/CSLL and PIS/COFINS, concomitantly. This would explain the fact that the CARF system included them among the results of the search carried out for this study.

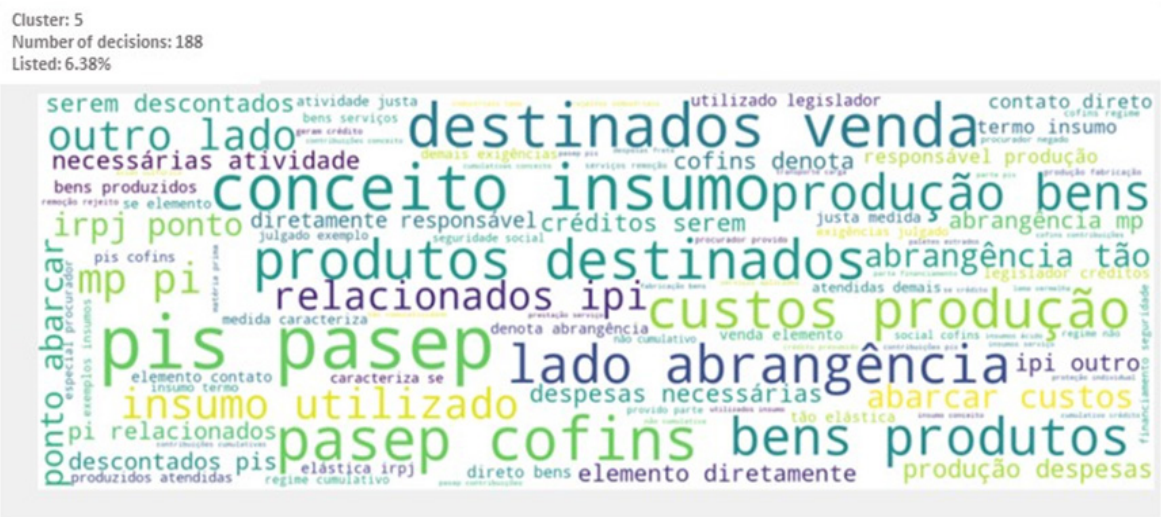

Figure 13. Word cloud of cluster 5

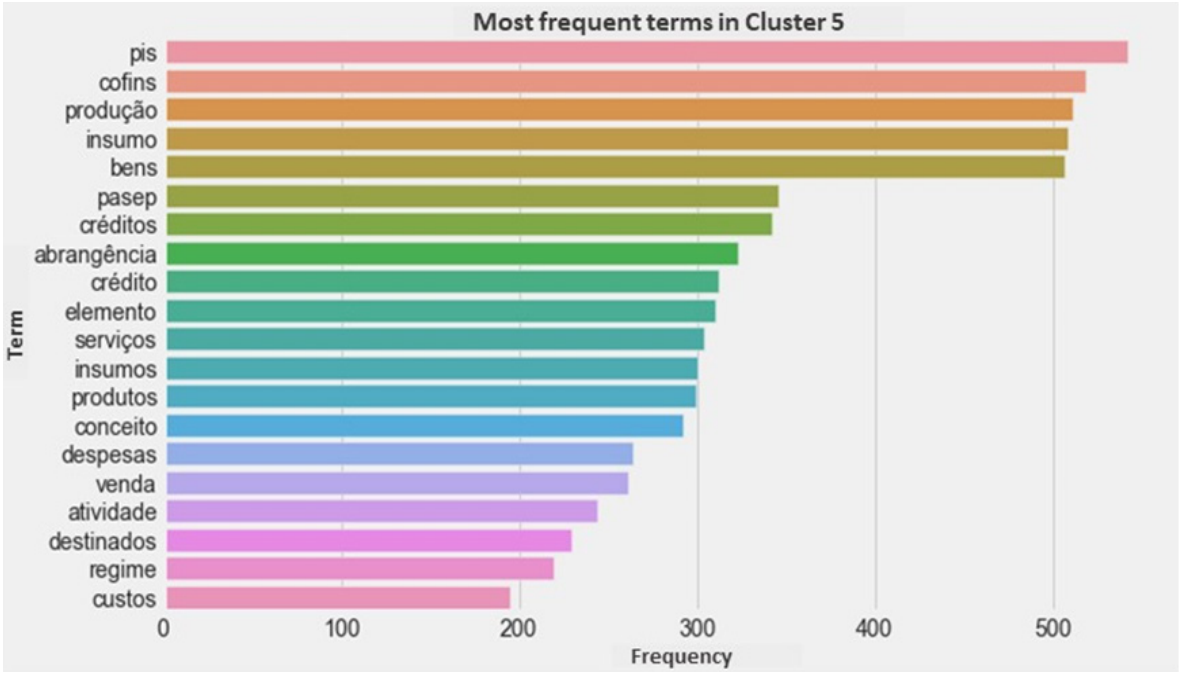

Figure 14. Cluster 5 – PIS/PASEP credits – Concept of inputs

4.1.7 Cluster 6 – Premium amortization derived from expected future profitability (Articles 385 and 386 of RIR/99 – Income Tax Regulation, Decree 3000/99, and/or supervening legislation)

Main words or expressions that characterize the 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*" (amortization, premium, future profitability, investor, investee, acquisition, investment, economic group, internal premium, vehicle, premium transfer, shareholding, effective payment, business purpose, economic substance, fine, ex-officio, isolated).

Figure 15. Word cloud of cluster 6

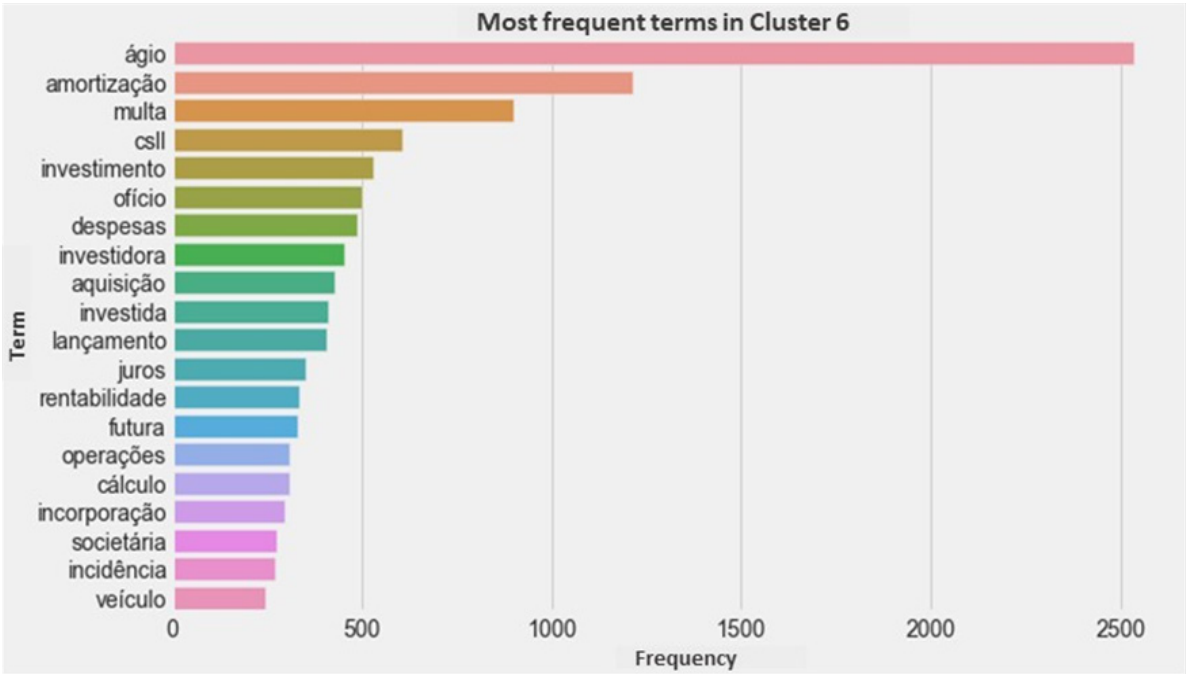

Figure 16. Cluster 6 – Premium amortization

4.1.8 Cluster 7 – Import transfer pricing, determined by the Resale Price Less Profit Method (PRL) – IN SRF 243/2002, Article 12, and IN RFB 1312/2012, Article 12

Main words or expressions that characterize the cluster: "*preço*", "*preços*", "*método*", "*transferência*", "*método PRL*", "*preço parâmetro*", "*instrução normativa*", "*bem importado*", "*frete*", "*seguro*", "*valor agregado*" (price, prices, method, transfer, PRL method, parameter price, normative instruction, imported good, freight, insurance, added value).

Figure 17. Word cloud of cluster 7

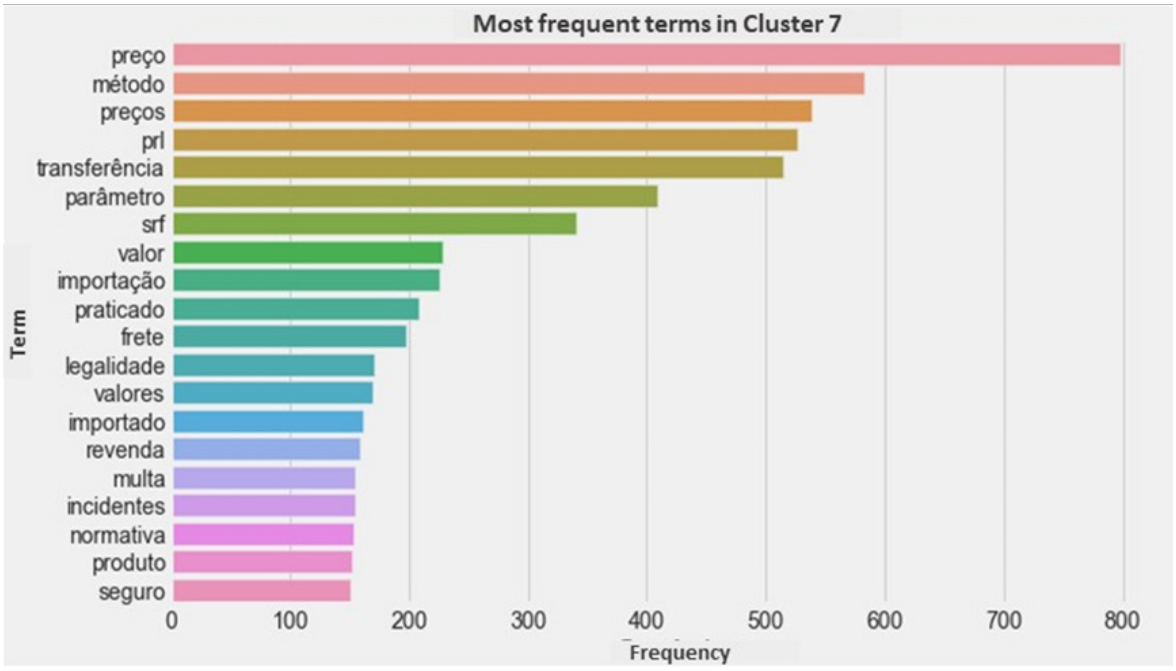

Figure 18. Cluster 7 – Transfer pricing

4.1.9 Cluster 8 – Offset of overpaid taxes (undue payment) in which the administrative authority questions the existence of credits – Questions relating to proof, as well as the certainty and liquidity of the claim

Main words or expressions that characterize the cluster: "*existência*", "*crédito*", "*compensação*", "*liquidez*", "*certeza*", "*prova*", "*provas hábeis*", "*comprovar*", "*autoridade*", "*administrativa*" (existence, credit, offset, liquidity, certainty, proof, valid evidence, proof, authority, administrative).

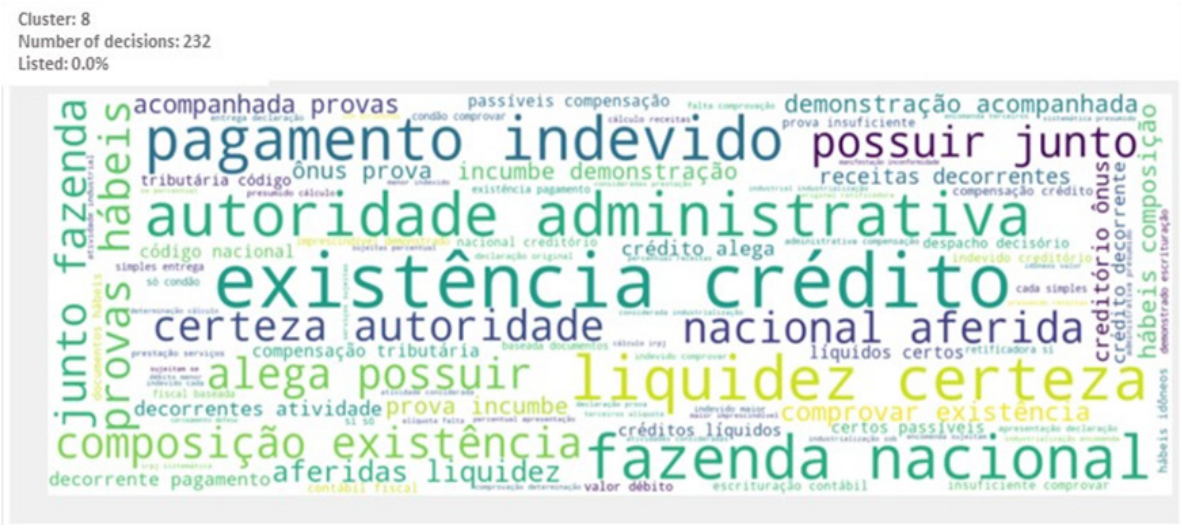

Figure 19. Word cloud of cluster 8

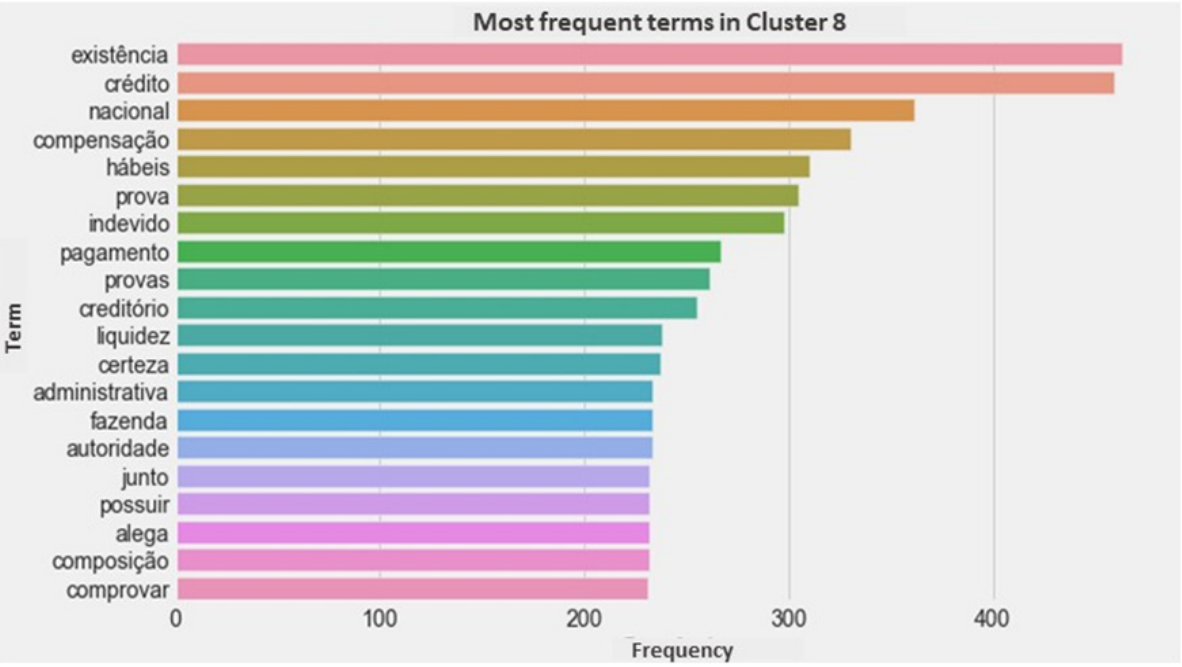

Figure 20. Cluster 8 – Offset of overpaid taxes based on evidence

4.1.10 Cluster 9 – Refund and offsetting of undue payments (indebt) due to insufficient evidence

Main words or expressions that characterize the cluster: "*indébito*", "*comprovação*", "*líquida*", "*certa*", "*compensação*", "*restituição*", "*comprovado*", "*deficiente*" (indebt, evidence, net, correct, offset, refund, evidenced, insufficient).

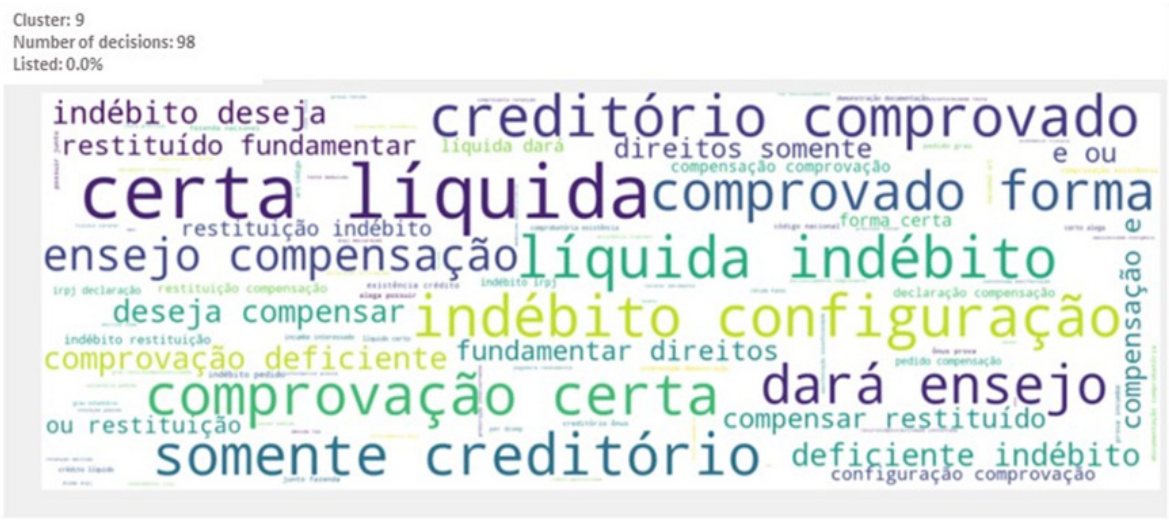

Figure 21. Word cloud of cluster 9

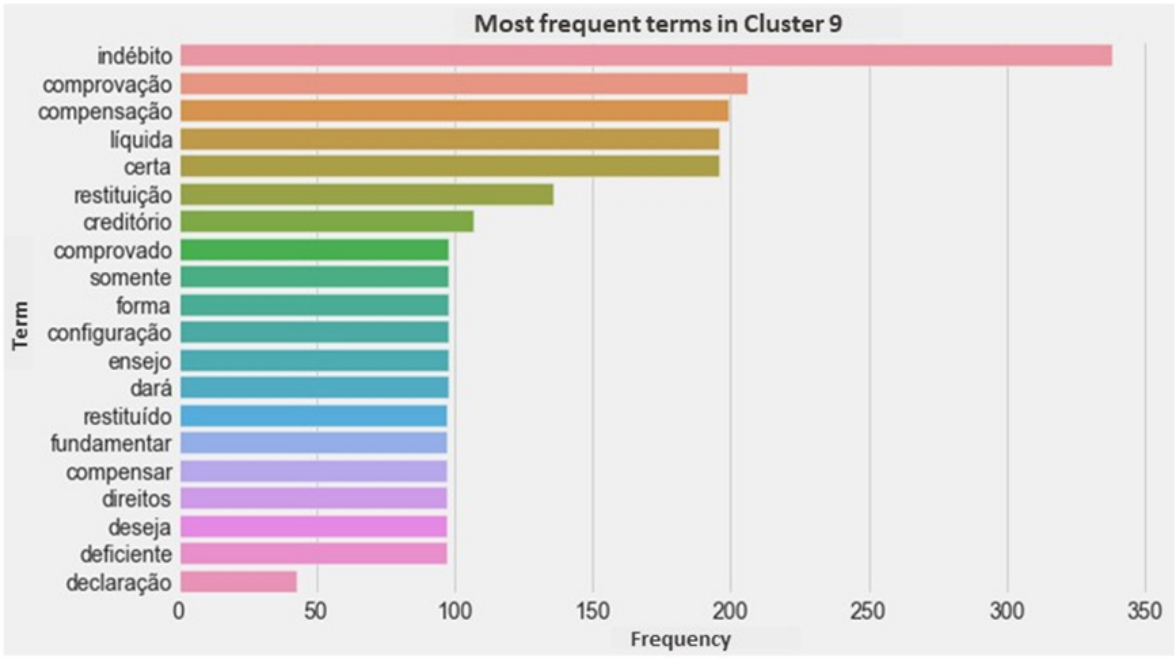

Figure 22. Cluster 9 – Refund and offsetting of undue payments due to insufficient evidence

4.1.11 Cluster 10 – Isolated fine for failure to pay estimated IRPJ and CSLL (tax on net profit) – Concurrent isolated fine and the ex-officio fine Articles 2 and 44, item II, of Law 9430/96)

Main words or expressions that characterize the cluster: "*multa isolada*", "*multa*", "*ofício*", "*juros*", "*falta recolhimento*", "*falta pagamento*", "*recolhimento estimativa*", "*estimativas mensais*", "*ajuste anual*", "*concomitância*", "*incidência*", "*IRPJ*", "*CSLL*" (isolated fine, fine, ex-officio, interests, missing collection, missing payment, estimated collection, monthly estimates, annual adjustment, concomitance, incidence, IRPJ, tax on net profit).

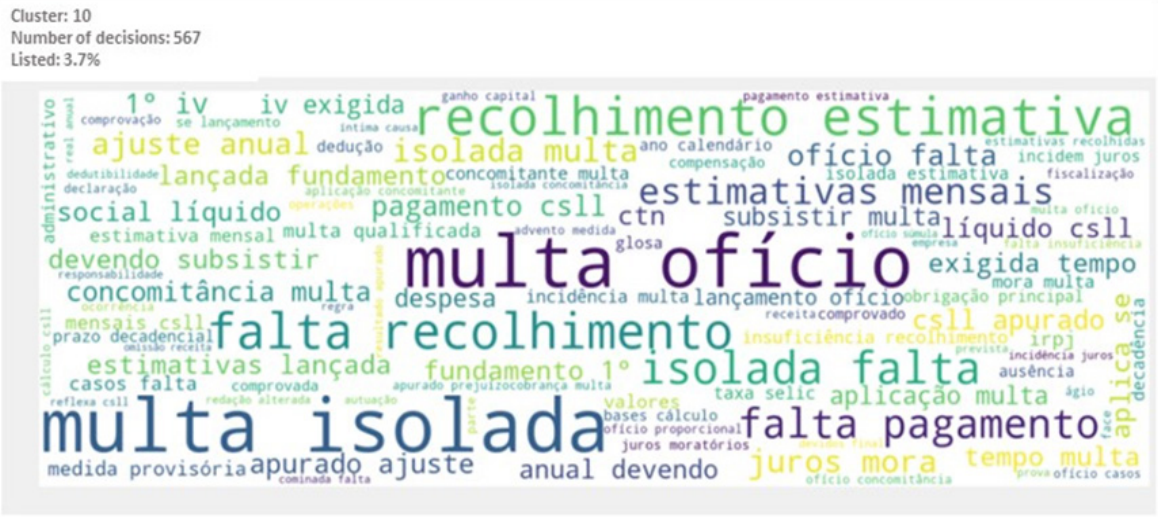

Figure 23. Word cloud of cluster 10

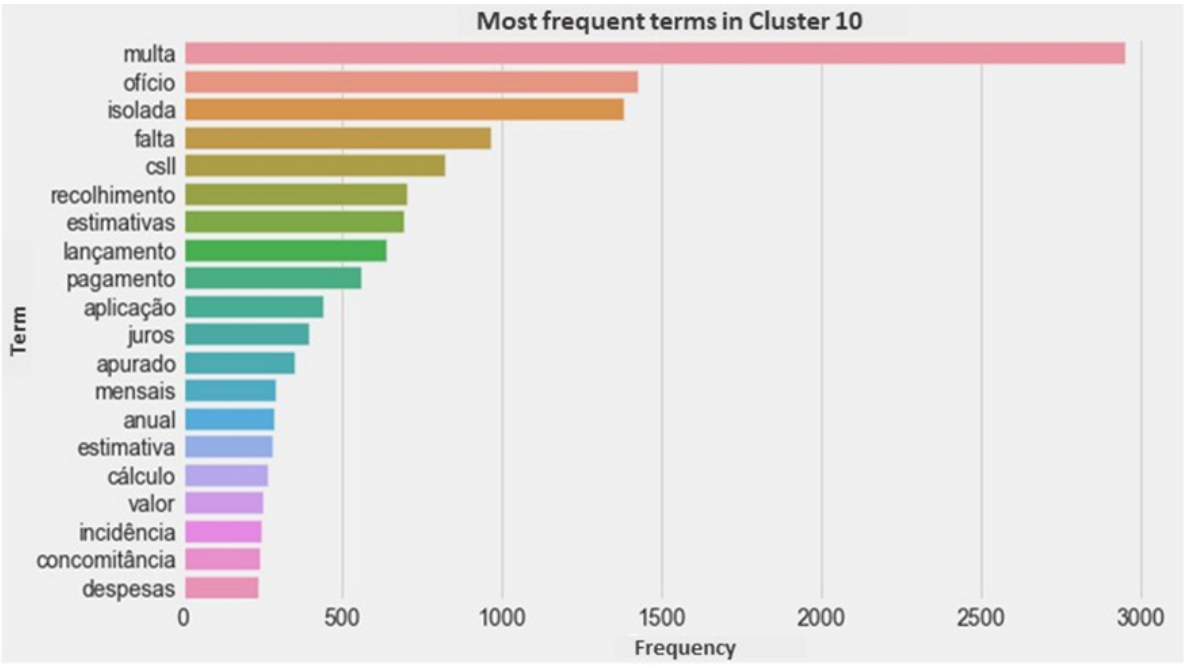

Figure 24. Cluster 10 – Isolated fine for failure to pay estimated IRPJ taxes – Concurrent fines

4.1.12 Cluster 11 – Refund and offsetting of negative balance of IRPJ (IN RFB 1717/2017)

Main words or expressions that characterize the cluster: "*saldo negativo*", "*PER*", "*DCOMP*", "*compensação*", "*estimativas compensadas*", "*retido*", "*fonte*", "*retenção*", "*IRRF*", "*creditório*", "*declaração compensação*", "*fonte pagadora*", "*homologação*" (negative balance, electronic request of refund, compensation declaration, offset, compensated estimates, withheld, provider, withholding, income tax withholding, document backing a credit, compensation declaration, payroll provider, homologation).

Figure 25. Word cloud of cluster 11

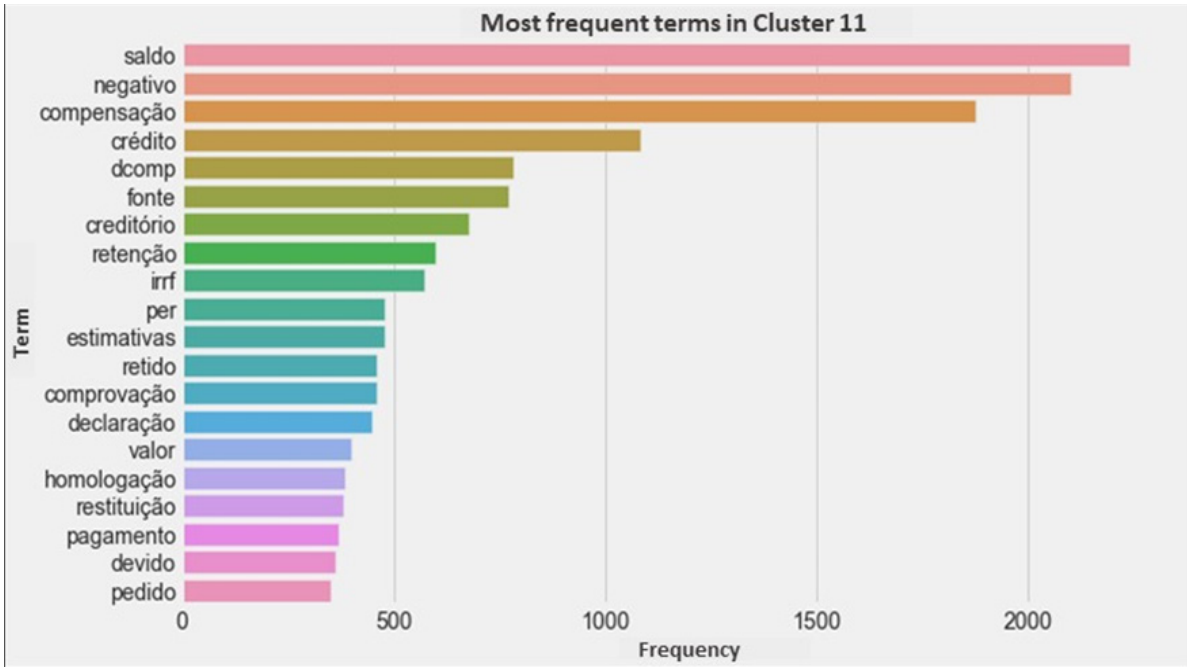

Figure 26. Cluster 11 – Refund and offsetting of negative balance of IRPJ

4.1.13 Cluster 12 – Refund and offsetting of undue payment, including negative balance of IRPJ generated by estimated tax payment – IN RFB 1717/2017

Main words or expressions that characterize the cluster: "*pagamento indevido*", "*compensação*", "*restituição*", "*estimativa*", "*pagamento*", "*indevido*", "*indébito*", "*crédito*", "*DCOMP*", "*PER*", "*recolhimento*", "*saldo negativo*" (undue payment, offset, refund, estimate, payment, undue, indebt, credit, compensation declaration, electronic request of refund, collection, negative balance).

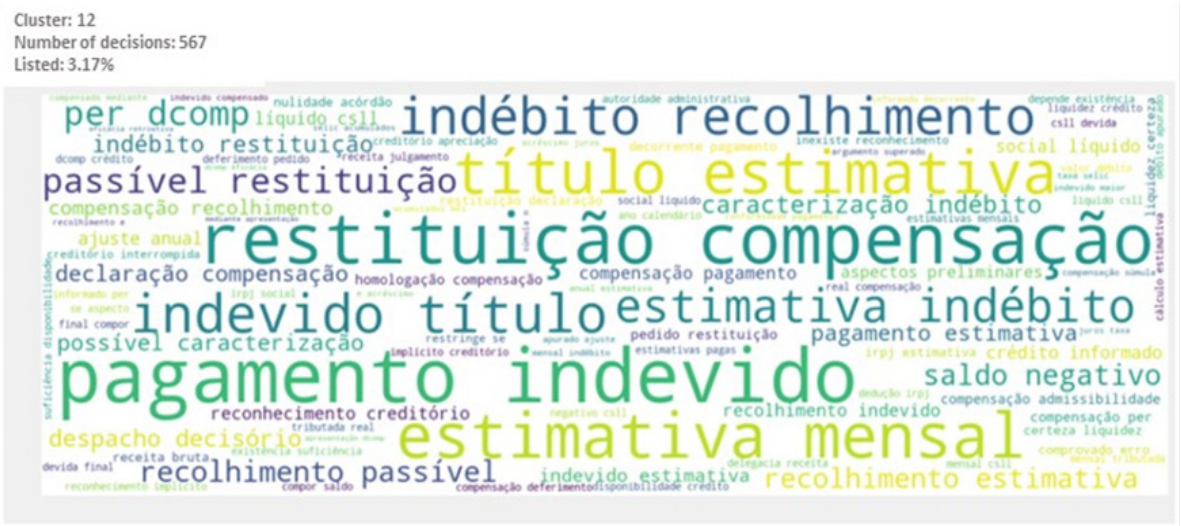

Figure 27. Word cloud of cluster 12

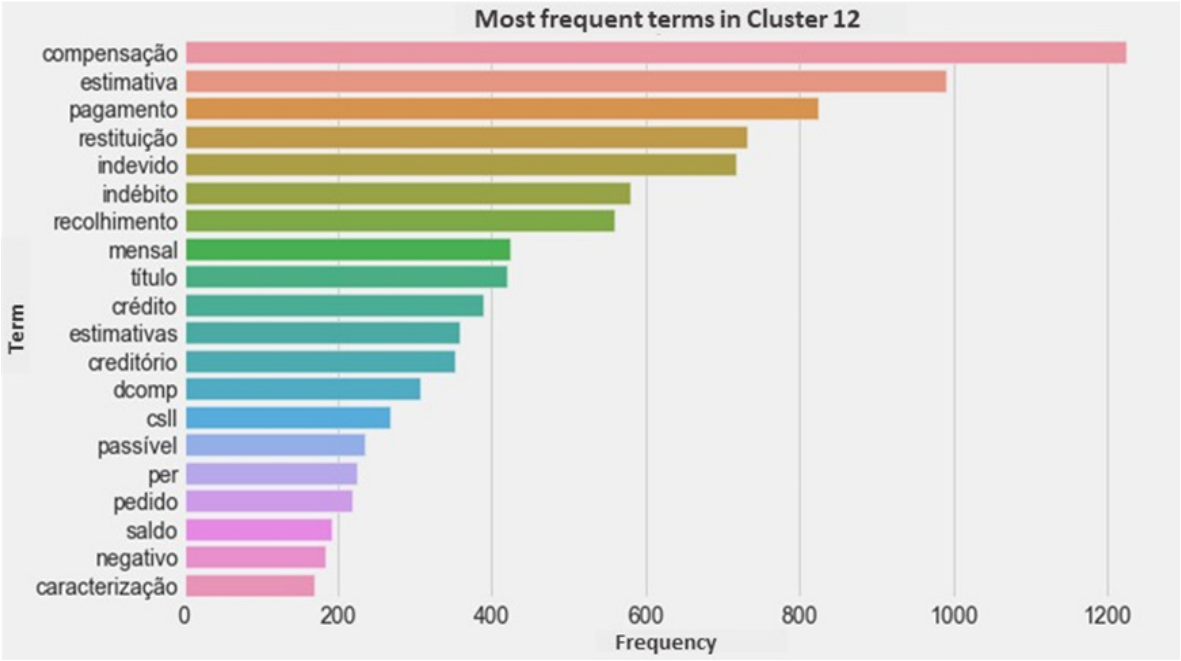

Figure 28. Cluster 12 – Refund and offsetting of undue payment, including negative balance of IRPJ generated by estimated tax payment

4.2 Discussion of results

The cluster analysis of 10,162 CARF judgments from 2016 to 2020 related to IRPJ revealed a wide variety of decisions involving various issues, including refund/offset resulting from undue payment, refund/offset of negative IRPJ balance, omission of revenues based on unproven bank deposits, premium amortization, and transfer pricing, among others.

Clusters 0, 2, 8, and 9 address similar issues, specifically refund and offset for undue payments or amounts exceeding the actual obligation. Similarly, clusters 11 and 12 pertain to refunds and offsets of negative IRPJ balances. In theory, these six clusters could be combined into two, reducing the total number of clusters to nine. However, the nuances within each cluster indicate the possibility of forming distinct groups.

Additionally, cluster 3 contains significantly more observations than the other clusters. The concentration of cases in this cluster can be explained by the nature of ex-officio entries related to IRPJ, which also encompass other taxes such as PIS/COFINS and CSLL. While the judgments in this cluster share a factual basis related to IRPJ assessments, the specific aspects of PIS/COFINS and CSLL legislation are also considered. Cluster 5, on the

other hand, addresses controversies regarding PIS/PASEP and COFINS credits, where the concept of input and production is highly debated, indicating recurring issues in CARF judgments.

Among the themes primarily accounting-related, clusters stand out that address aspects related to the presumption of profit coefficients, premium amortization, and transfer pricing. Cluster 4 focuses on the controversy surrounding the presumption of profit coefficients in hospital services and services in the construction industry, with or without the supply of materials. Cluster 6 examines the accounting treatment of premiums on equity interest acquisitions, which refers to the positive difference between the market value and book value of assets in the investee company. Cluster 7 deals with import transfer pricing, specifically the prices established between related companies for the transfer of goods or services. The Resale Price Less Profit (PRL) Method is used to determine these prices, and there is debate regarding the inclusion of amounts related to freight, insurance, and taxes in the price charged for applying PRL.

Cluster 1 focuses on judgments related to the presumption of revenue omission resulting from unproven bank deposits. This occurs when the account holder lacks proof of the origin of the deposited amounts. In such cases, it is presumed, based on the law, that these amounts represent undeclared revenue or billing and are subject to corporate income tax (IRPJ).

Finally, cluster 10 addresses the discussion of simultaneously applying isolated and ex-officio fines for non-payment of IRPJ and tax on net profit (CSLL). The isolated fine is imposed when monthly estimates of IRPJ and CSLL are not collected, while the ex-officio fine is applied when the taxpayer fails to pay the IRPJ and CSLL determined in the annual adjustment.

5 IMPLICATIONS FOR ACCOUNTING

In view of the distribution of judgments by clusters, it appears that the discussion on tax refunds and offset is of paramount importance for accounting professionals as it enables the refund of unduly paid amounts. When excess taxes are paid, the refund consists of returns of the amounts paid in excess, while the offset represents the use of this amount to settle future taxes. Accounting professionals need in-depth knowledge of tax laws and regulations to properly execute tax refunds or offsets, including understanding the rules regarding the extinction of tax obligations.

In addition, they must be aware of the processes for requesting tax refunds or offsets, such as using the Request for Refund, Reimbursement, or Reimbursement and Offset Statement (PER/DCOMP) program. In this context, the importance of internal control stands out, as it helps to prevent errors and fraud, ensuring compliance with tax rules and the correct application of laws. An efficient internal control system contributes to the accuracy of accounting and tax information, minimizing risks and avoiding undue or overpayments, which directly impact the companies' financial health and reputation, as well as reducing tax disputes.

Robust internal controls are essential to avoid entries resulting from the presumption of revenue omission from bank deposits of unproven origin and the incidence of isolated and ex-officio fines for failure to pay taxes at the appropriate time. Internal control systems are also relevant to determine the company's category and profit ratio. An efficient accounting policy can significantly reduce litigation, avoid unnecessary fines and penalties, and ensure the accuracy of financial and tax information.

The "concept of inputs" refers to the non-cumulative nature of PIS and COFINS. Inputs are determined considering their essentiality or relevance in economic activity. Accounting aspects are crucial to correctly apply this concept, classifying goods and services and ensuring compliance with tax legislation.

Proper tax planning is an important part of accounting and can help companies reduce risk and avoid litigation related to tax matters. With regard to transfer pricing, accounting must consider applicable rules and regulations to avoid tax problems and ensure compliance with tax laws. The premium amortization must follow the legislation and be adequately addressed to ensure the correct tax calculation and collection.

Tax planning is a fundamental tool for business management and must be aligned with managerial accounting. The high number of judgments addressing transfer pricing and premium amortization confirm the importance of adequate and effective tax planning. In summary, accounting and tax planning are intrinsically intertwined, and the clustered analysis of judgments highlights the need for a meticulous and well-founded approach to ensure tax compliance and minimize risks.

6 CONCLUSION

The 10,162 CARF judgments related to IRPJ analyzed using a machine learning (ML) algorithm formed 13 clusters. The analysis focused on identifying the main subject of each cluster and the number of judgments issued, facilitating the collection of valuable information about the progress of these processes to support accountants and tax lawyers in decision-making.

Furthermore, data analysts must know accounting and taxation to analyze the results of the ML model. Conversely, contemporary challenges have shown that data science is crucial for professionals in all areas of knowledge, particularly accounting and tax law.

The adoption of ML may have significant impacts when incorporated into accounting, especially in optimizing companies' tax planning and improving their internal controls and processes. For example, identifying trends and patterns in the authorities' decisions in litigations regarding corporate income tax can help companies develop more efficient tax planning strategies and improve internal controls. As for improving processes, insights from such an analysis of judgments may guide a review of tax procedures and adaptations to comply with changes in tax legislation.

Despite its virtues, this research has some limitations, such as the difficulties in data collection and the need to adapt notebooks to perform it without interruptions – aspects that can be addressed in future works. Another limitation was the concentration of many decisions in a single group (cluster 3), making it difficult to define the main subject addressed. Future research can address this difficulty by exploring the nature of the financial entries included in this cluster.

Future studies may expand the scope of the analysis to other taxes – for example, personal income tax (IRPF), contributions to programs that finance unemployment and other work-related benefits (PIS/PASEP), the contribution for social security financing (COFINS), and taxes on foreign trade – and the period of the judgments analyzed, covering, for example, those issued after 2010. This approach would allow for a classification of relevant themes for different taxes. In the methodological aspect, another possibility would be the application of clustering algorithms using other database characteristics, such as the type of judgment, the year of the inspection, the amount of the tax credit in dispute, and the entire text of the judgment (instead of only the judgment summary).

DATABASE AND RESEARCH NOTEBOOKS

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

REFERENCES

Borcan, M. (2020, june 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 Brasília]. 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

How to cite this paper

Costa, F. C. L., Martinez, A. L., & Klann, R. C. (2023). Cluster analysis of IRPJ precedents in CARF. *Revista de Contabilidade e Organizações*, 17:e197181. DOI: http://dx.doi.org/10.11606/issn.1982-6486.rco.2023.197181