SAS2257-2018

Invoiced: Using SAS® Contextual Analysis to Calculate Final Weighted Average Consumer Price

Alexandre Carvalho, SAS Institute Inc.

ABSTRACT

SAS® Contextual Analysis brings advantages to the analysis of the millions of Electronic Tax Invoices (Nota Fiscal Electrônica) issued by industries and improves the validation of taxes applied. Tax calculation is one of the analytical challenges for government finance secretaries in Brazil. This paper highlights two items of interest in the public sector: tax collection efficiency and the calculation of the final weighted average consumer price. The features in SAS® Contextual Analysis enable the implementation of a tax taxonomy that analyzes the contents of invoices, automatically categorizes the product, and calculates a reference value of the prices charged in the market. The first use case is an analysis of compliance between the official tax rate—as specified by the Mercosul Common Nomenclature (NCM)—and the description on the electronic invoice. (The NCM code was adopted in January 1995 by Argentina, Brazil, Paraguay, and Uruguay for product classification.) The second use case is the calculation of the final weighted average consumer price (PMPF). Generally, this calculation is done through sampling performed by public agencies. The benefits of a solution such as SAS Contextual Analysis are automatic categorization of all invoices and NCM code validation. The text analysis and the generated results contribute to tax collection efficiency and result in a more adequate reference value for use in the calculation of taxes on the circulation of goods and services.

INTRODUCTION

This paper focuses on the analytical challenges of government finance secretaries in Brazil, including the following:

- categorize the contents of the Electronic Tax Invoices
- improve the accuracy of the calculation of the final weighted average consumer price
- build an analytical base table that can be used as the basis for the calculation of the final weighted average consumer price

Business analysts and IT professionals are looking for solutions that are easy to use and easy to integrate into their existing systems, and that improve their analytics and their outcomes to challenges. SAS Contextual Analysis has benefits that combine machine learning and text mining with linguistic rules.

Some of these features can be directly monetized to help provide a fast return, such as the following:

- filtering documents
- predefined concepts
- ability to create and improving rules to concepts and categories
- exploring for new topics
- categorizing unstructured textual data and collections of documents

These and other features are found in SAS Contextual Analysis through a single integrated system. You can update and customize rules as needed.

DATA SOURCES FOR THIS DEMO

The data source was provided by and its use authorized by Secretaria de Estado de Fazenda de Minas Gerais (SEFA MG), Brazil. In May 2017, the data source was utilized in Proof of Concept (POC) for categorizing invoice issues. The results were reduced classification time, improved accuracy in product identification, and help with identifying anomalies in invoices and taxes.

Display 1 shows a sample of the data source with 9,955 rows and 6 variables (including descriptive text about the invoices and the NCM code). The sample contains grouped information about Electronic Tax Invoices issued to taxpayers (that is, industries). The Electronic Tax Invoices issued are a selection of the invoices issued in May 2017, and the source does not contain confidential information about taxpayers.

	DESCRIPITION_INVOICES	NCM_CHAPTER	NCM_POSITION	NCM_SUB_POSITION	NCM_ITEM	NCM_SUB_ITEM
1362	BAVARIA LATA 350ML/12	22	2203	220300	2203000	22030000
1363	ANTARCTICA SUBZERO LATA 350ML SH ...	22	2203	220300	2203000	22030000
1364	BRAHMA CHOPP LT 473ML SH C 12 NPAL	22	2203	220300	2203000	22030000
1365	BRAHMA EXTRA LONG NECK 355ML SIX-...	22	2203	220300	2203000	22030000
1366	SKOL LATA 350ML SH C/12 NPAL ...	22	2203	220300	2203000	22030000
1367	ORIGINAL 600ML 60,7915	22	2203	220300	2203000	22030000
1368	HEINEKEN VNR 355ML 1X1	22	2203	220300	2203000	22030000
1369	MILLER LN 355ML	22	2203	220300	2203000	22030000
1370	MALZBIER BRAHMA LONG NECK 355ML S...	22	2203	220300	2203000	22030000
1371	CERV SCHIN PILS 0,269LT 15 UN PBR	22	2203	220300	2203000	22030000
1372	BRAHMA CHOPP GFA VD 300ML CX C/23 ...	22	2203	220300	2203000	22030000
1373	KAISER PILSEN LATA 350 ML	22	2203	220300	2203000	22030000
1374	SKOL LATA 350ML SH C 12 NPAL	22	2203	220300	2203000	22030000
1375	BUDWEISER LN 343ML SIXPACK CARTAO...	22	2203	220300	2203000	22030000
1376	0101 - CERVEJA NOVA SCHIN 600ML	22	2203	220300	2203000	22030003
1377	CHAMP CHANDON 187ML BABY BRUT RO...	22	2204	220410	2204101	22041010
1378	ESPUMANTE	22	2204	220410	2204101	22041010
1379	CHAMP CHUVA PRATA BRANCO 660ML	22	2204	220410	2204101	22041010
1380	VINHO NAC PERGOLA 1L TINTO SUAVE	22	2204	220410	2204101	22041010

Display 1. Data Source from SEFA-MG, 2017

UNDERSTANDING THE ICMS TAX AND THE CONTENT OF THE INVOICE DESCRIPTIONS

ICMS is the tax levied on the circulation of products such as food, beverages, household appliances, communication services, transportation, and some imported products, and became law in 1997 (also known as the Lei Kandir law). In Brazil, ICMS is one of the largest sources of financial revenue. Because it is established by each state (for example, Minas Gerais, Rio de Janeiro, or São Paulo), it changes from one place to another. Tax collections can be routed to various functions (for example, health, education, payment of civil servants, and so on).

At each stage of the collection cycle, it is always necessary to issue an invoice or tax coupon, which is calculated by the taxpayer and collected by the State. There are two types of Electronic Tax Invoices: invoices issued at the industry level (electronic invoices issued by the beer, refrigerator, or fuel industries) and invoices issued at the consumer level (electronic invoices issued by restaurants to final consumers).

In Display 2, line 1375 (BUDWEISER LN 343ML SIXPACK CARTAO) provides us with the following information: Product (Budweiser), Type (LN means Long Neck), Volume (343ML), and Quantity (SIXPACK CARTAO SH C/4).

	DESCRIPITION_INVOICES
1373	KAISER PILSEN LATA 350 ML
1374	SKOL LATA 350ML SH C 12 NPAL
1375	BUDWEISER LN 343ML SIXPACK CARTAO SH C/4 68,1700
1376	0101 - CERVEJA NOVA SCHIN 600ML
1377	CHAMP CHANDON 187ML BABY BRUT ROSE(E)

Display 2. Data Source Content

WHAT IS THE MERCOSUL COMMON NOMENCLATURE (NCM CODE) FOR PRODUCT CLASSIFICATION?

The classification system for invoices follows the Mercosul Common Nomenclature (Nomenclatura Comum do Mercosul, or NCM) and was adopted in January 1995 by Argentina, Brazil, Paraguay, and Uruguay for product classification. Any merchandise, imported or purchased in Brazil, must have an NCM code in its legal documentation (invoices, legal books, and so on), whose objective is to classify the items according to the Mercosul regulation.

Display 3 shows examples of the content of Electronic Tax Invoices according to the NCM code by chapter, position, sub-position, item, and sub-item.

	DESCRIPITION_INVOICES	NCM_CHAPTER	NCM_POSITION	NCM_SUB_POSITION	NCM_ITEM	NCM_SUB_ITEM
1373	KAISER PILSEN LATA 350 ML	22	2203	220300	2203000	22030000
1374	SKOL LATA 350ML SH C 12 NPAL	22	2203	220300	2203000	22030000
1375	BUDWEISER LN 343ML SIXPACK CARTAO SH C/4 68,1700	22	2203	220300	2203000	22030000
1376	0101 - CERVEJA NOVA SCHIN 600ML	22	2203	220300	2203000	22030003
1377	CHAMP CHANDON 187ML BABY BRUT ROSE(E)	22	2204	220410	2204101	22041010

Display 3. Mercosul Common Nomenclature Content

IMPROVING CATEGORIZATION EFFICIENCY WITH SAS CONTEXTUAL ANALYSIS

The use of unstructured data is growing exponentially in government agencies. In January 2018, according to the Brazilian Federal Revenue Agency (Receita Federal Brasileira), approximately 18 billion Electronic Tax Invoices were identified, and the number of issuers was approximately 1.4 million.

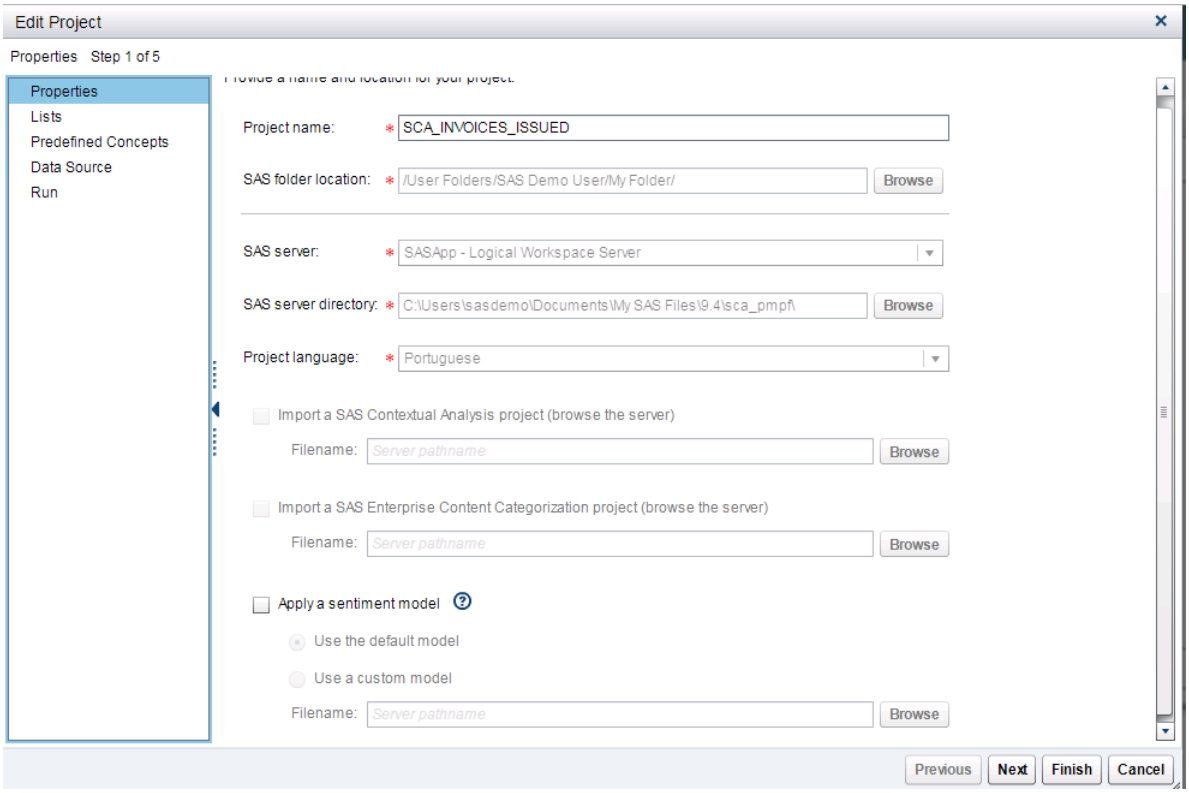
THE BENEFITS OF USING SAS CONTEXTUAL ANALYSIS

Business analysts are looking for solutions that are fast, easy to use and integrate into existing systems, and that improve their analytics and challenges. For the classification of electronic invoices, the analyst has more control with a hybrid approach. Analysts can add concepts (for example, 1LT, 500GR means quantity) and synonyms (skol, Budweiser, heinecken, brhama means beer) that specifically identify the product and its value for the tax aliquot calculation (for example, beer and 1LT the tax aliquot is 4%).

SAS Contextual Analysis combines machine learning and text mining capabilities with the ability to impose linguistic rules. SAS Contextual Analysis also enables you to filter, explore, and categorize unstructured textual data and collections of documents.Technology syntactically identifies common themes, category rules, and document sentiment, based on data. At any time, you can review and modify the results to meet your specific needs.

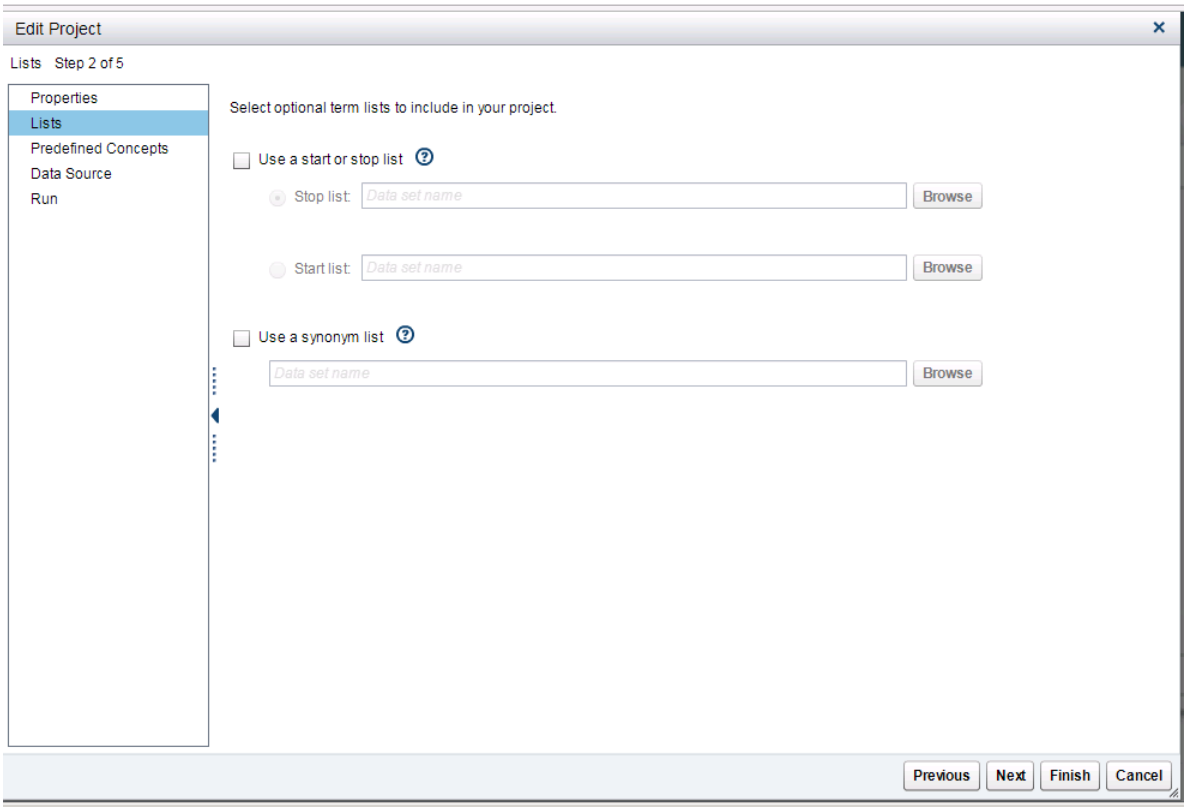
HOW TO BUILD A PROJECT IN SAS CONTEXTUAL ANALYSIS

Display 4 shows Step 1 of 5 for building a project in SAS Contextual Analysis.The analyst defines the name and location for your project, and chooses a project language. This paper doesn’t apply a sentiment model, but is possible to use either the default model or a custom model.



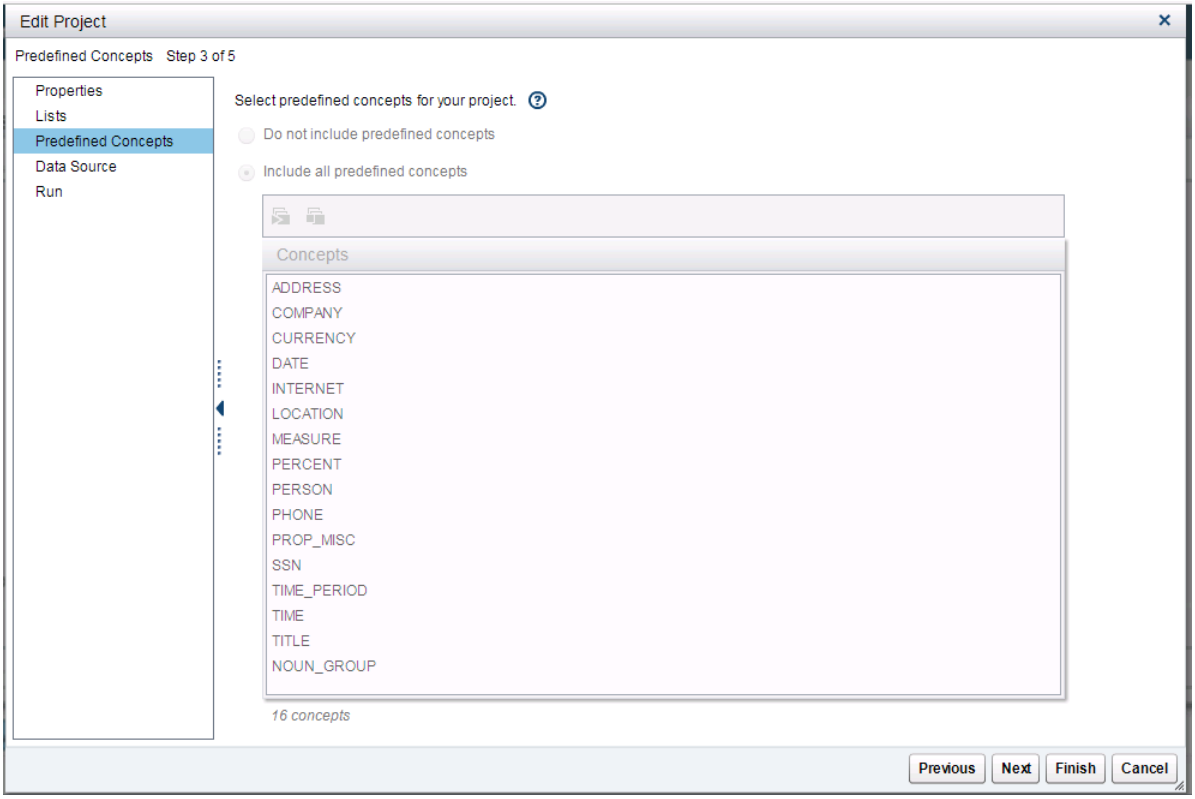
Display 4. Create a New Project: Define name, location and language for your project

Display 5 shows Step 2 of 5 for building a project in SAS Contextual Analysis. When analyzing text, it is common to disregard some terms already known to analysts that would not add value to the analysis or select a list of terms for research. For example, we can use the stop list (for name Brazil, SEFA-MG) or start list (skol, brahma, or budweiser). Another important feature is to use a list of synonyms whose terms would have the same meaning across the business (LT, GR, and KG all indicate quantity).



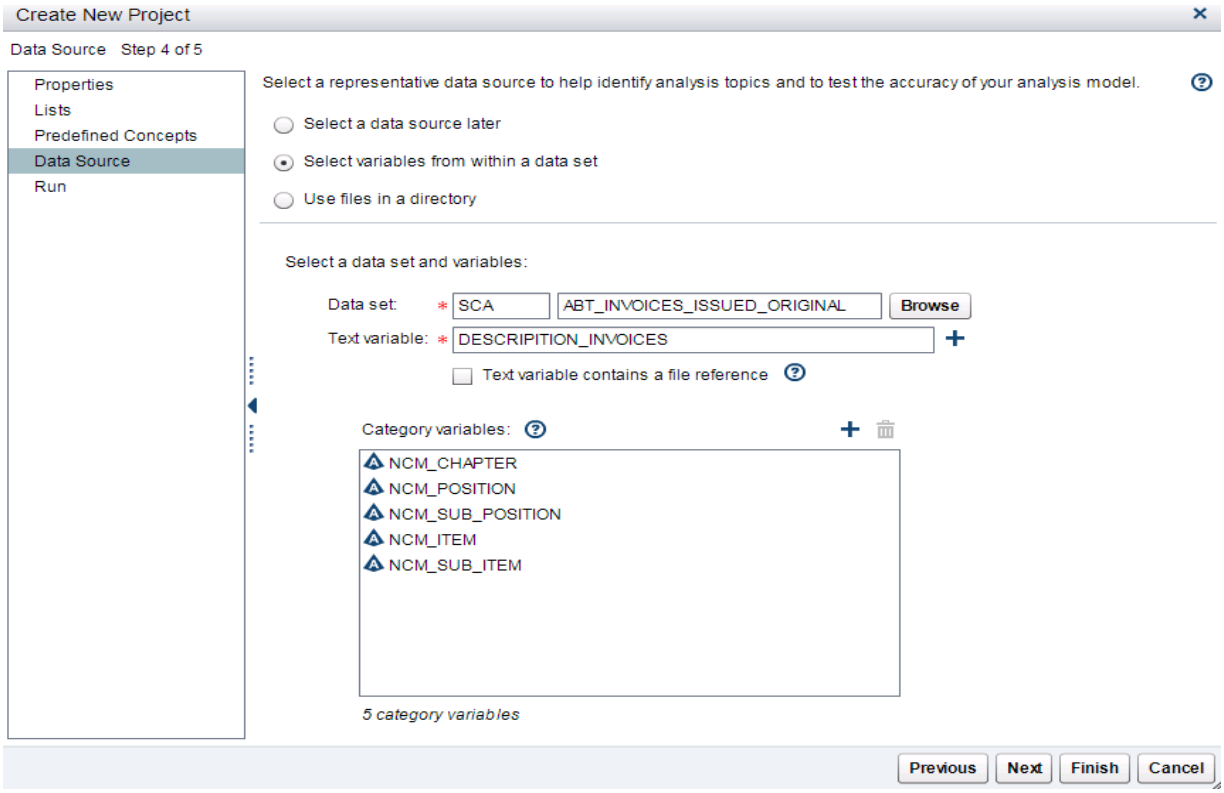
Display 5. Create a New Project: Define start list, stop list or synonyms list

Display 6 shows predefined concepts for your analysis and how SAS Contextual Analysis automatically identifies concepts such as location, currency, company, address, and so on.



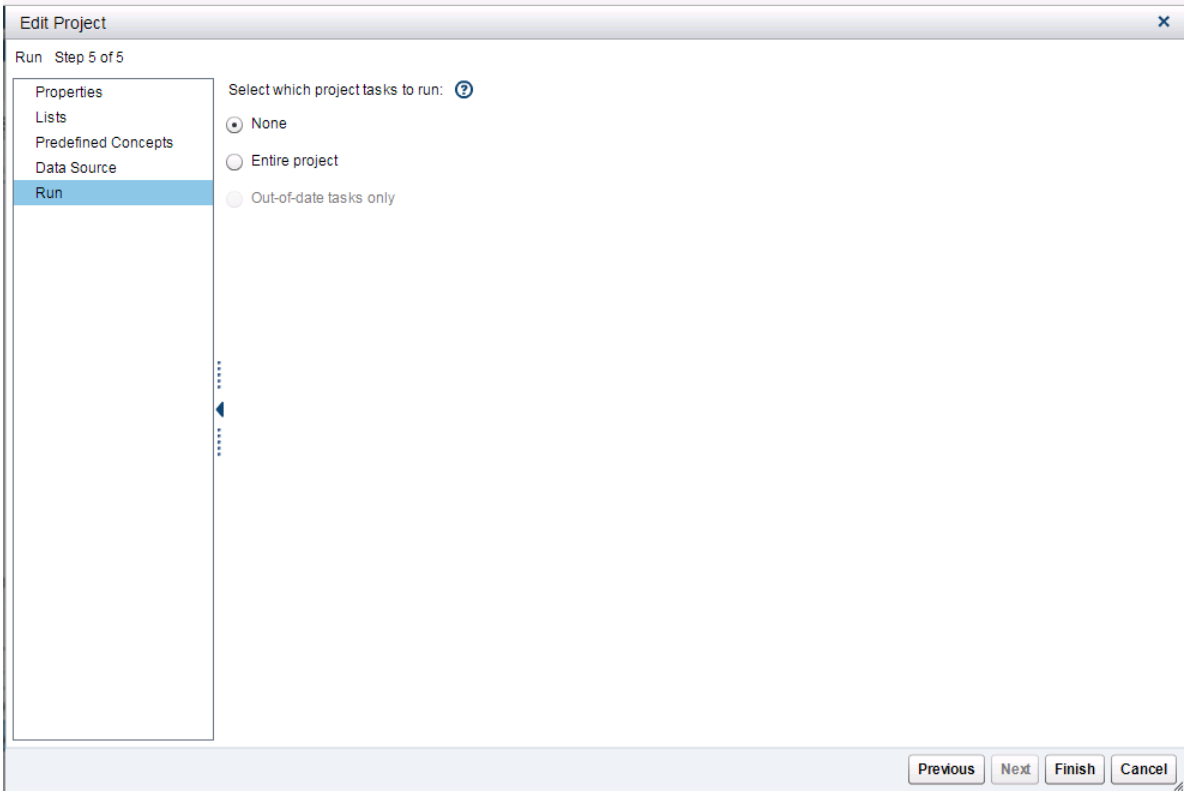
Display 6. Create a New Project: Predefined Concepts

Display 7 shows Step 4 of 5, which is when you select a SAS data set (ABT_INVOICES_ISSUED_ORIGINAL). The variable DESCRIPTION_INVOICES contains the invoice description, and text mining is used. On the other hand, NCM code information is used for categorization.



Display 7. Create a New Project: Select Data Set and Variables

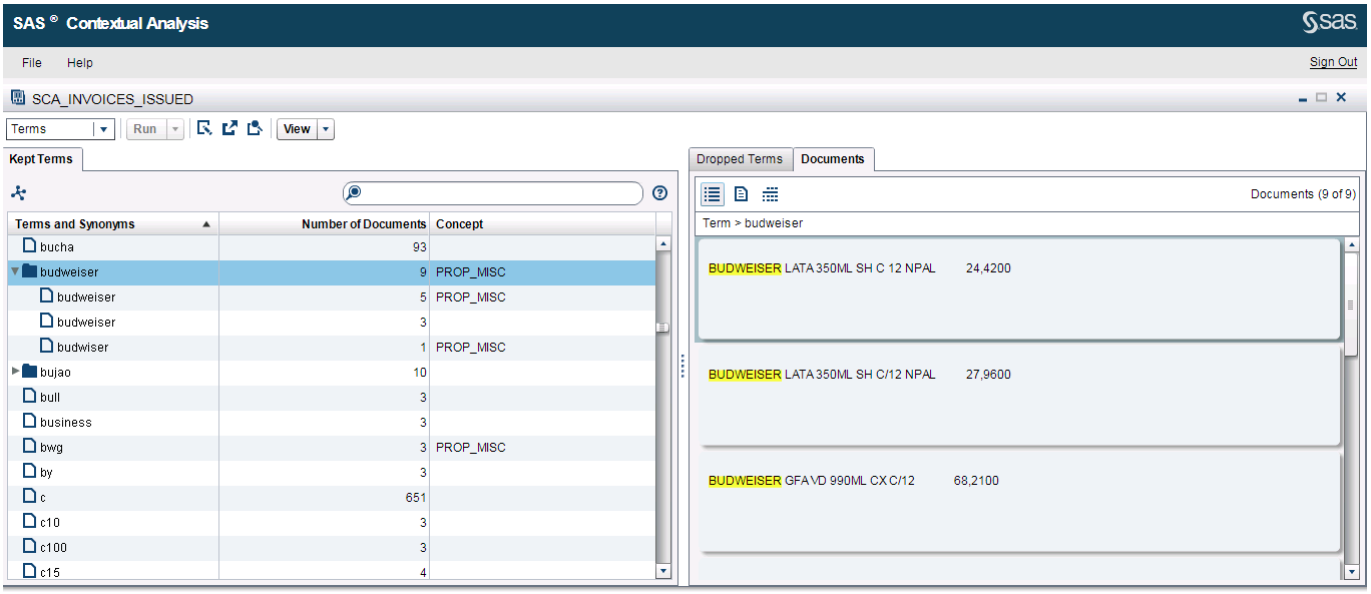
And finally, you are ready to run the project (Display 8).



Display 8. Create a New Project: Run the entire project

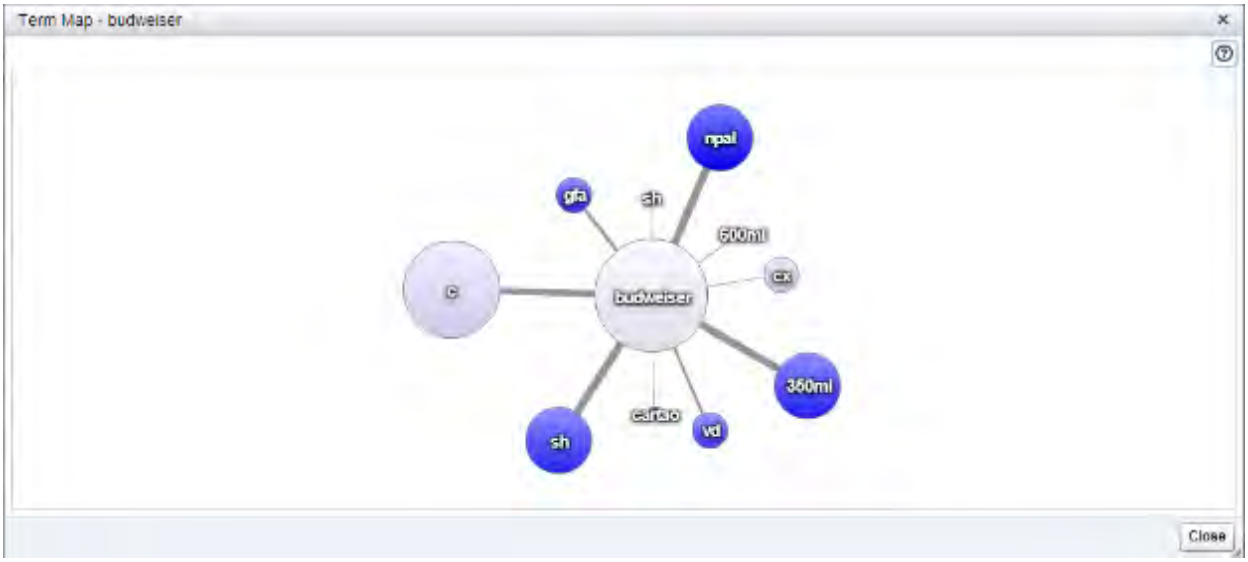
IDENTIFY TERMS: NAME, TYPE, AND PRODUCT QUANTITY

Display 9 focuses on the term “budweiser”. In this case, you can see the stemming for the term “budweiser”, including the three forms it takes and the few rare misspellings that have occurred in the documents (for example, “budwiser”). In this example, "budweiser" is the description of a type of beer (product name).



Display 9. Create a New Project: Identifying Terms

In the term map shown in Display 10, you can see that there is additional information about the product type (for example, “gf” and “cx” mean “bottle”) and volume (350ml or 600ml). The term map can help you refine your terms list and create rules for classification.

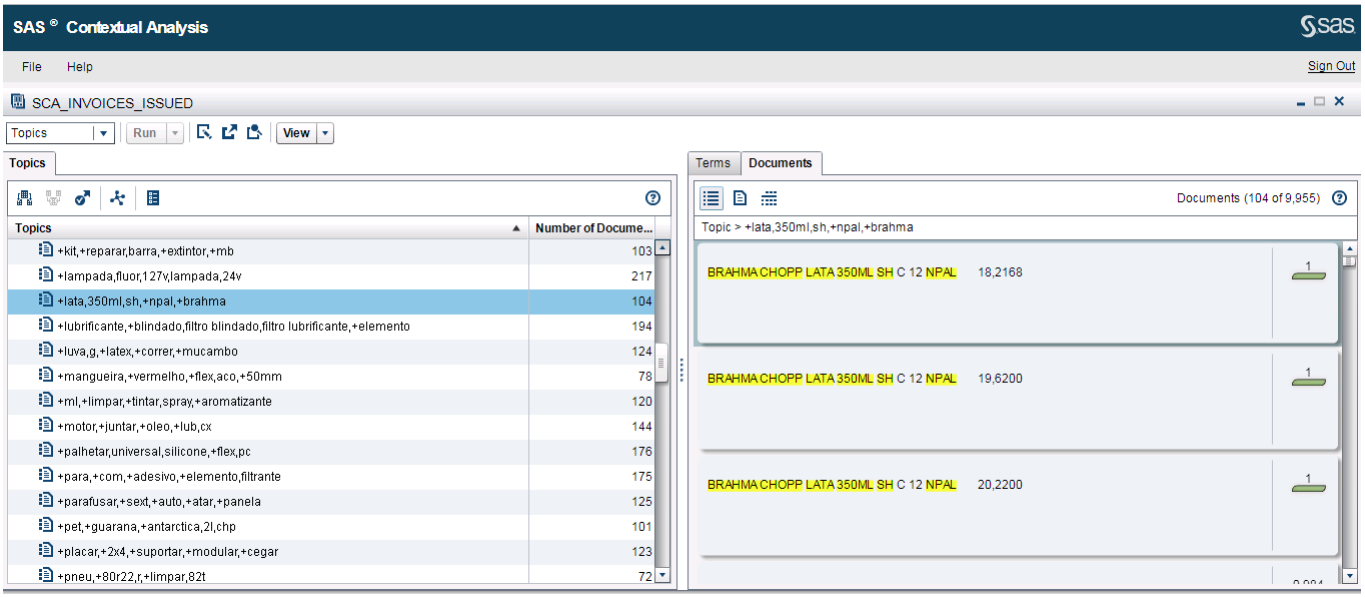


Display 10. Create New Project: Term Map

DISCOVERING TOPICS FOR THE ELECTRONIC TAX INVOICES

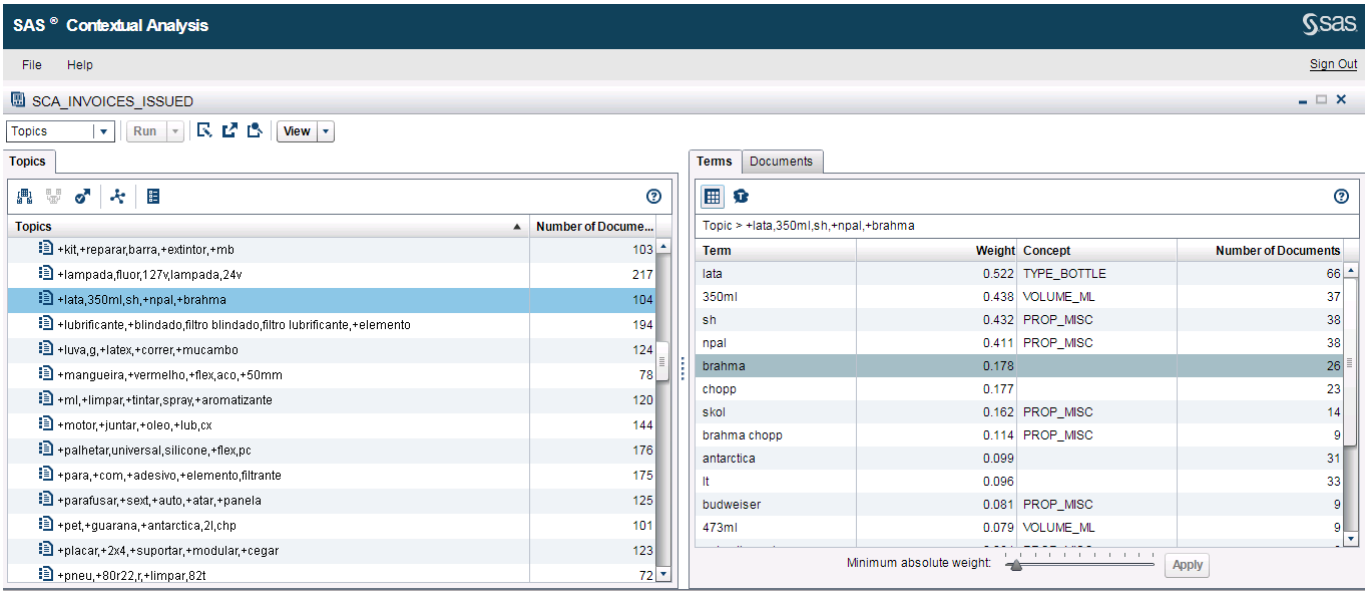
In particular, the Topics functionality in SAS Contextual Analysis can help you to automatically identify the contents of of your documents, which are in this case Electronic Tax Invoices.

Display 11 shows the documents for the topic **+lata+350ml,sh,+npal+brhama** . On the right side of the window, you can see a set of tax invoices that identify as a type of beer.

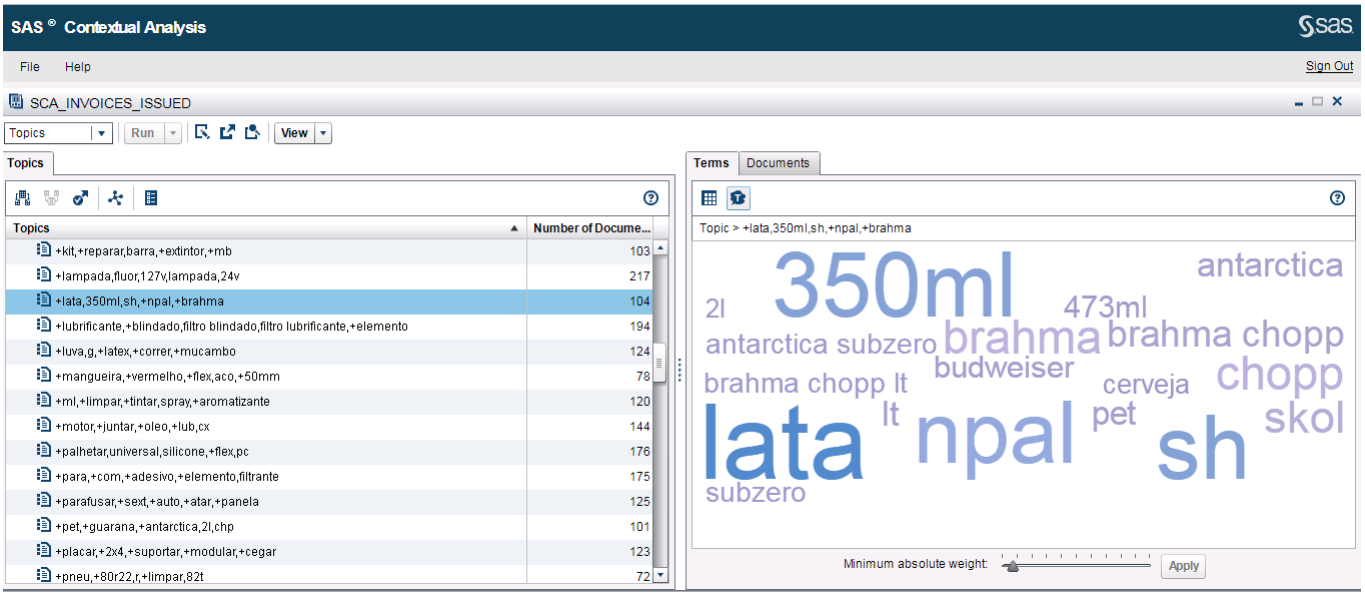


Display 11. Identify Emerging Issues

In Display 12 and Display 13, you see the **Terms** tab, on which you can choose from two different views of the terms that constitute the topics. You can also choose different views of the documents that are associated with the topics.

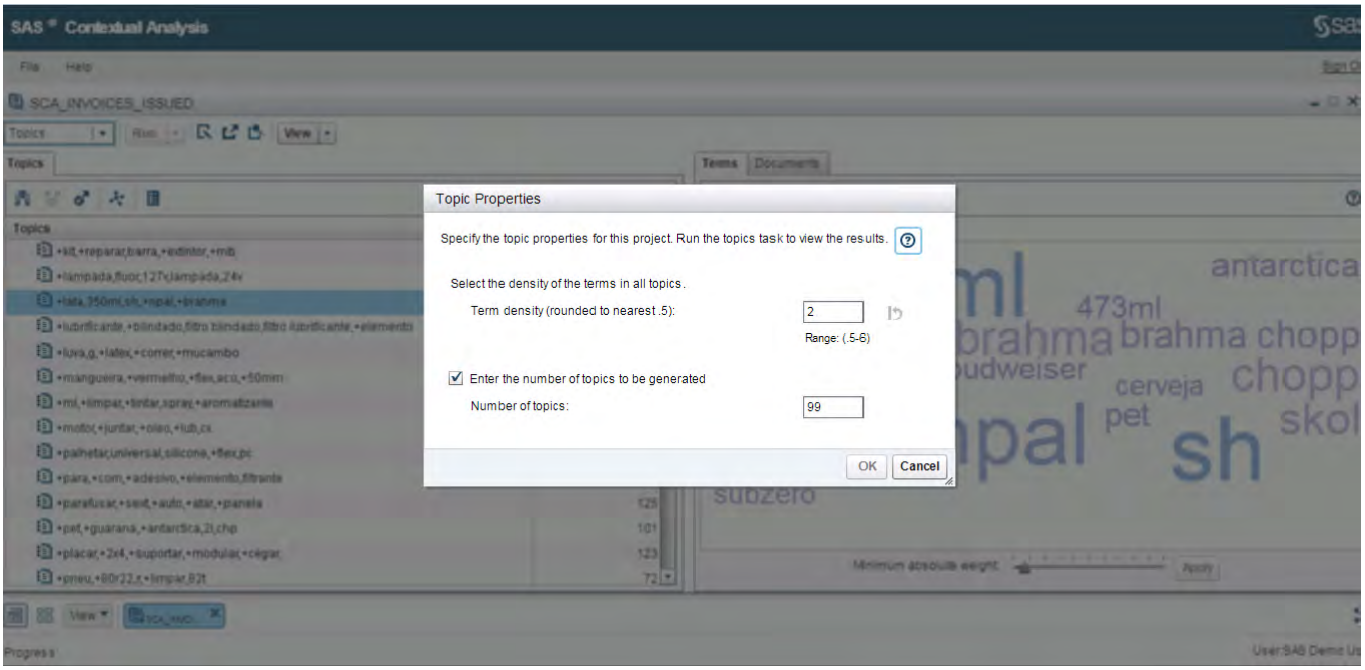


Display 12. The Terms Tab: View Tabular Form



Display 13. The Terms Tab: View Graphic Form

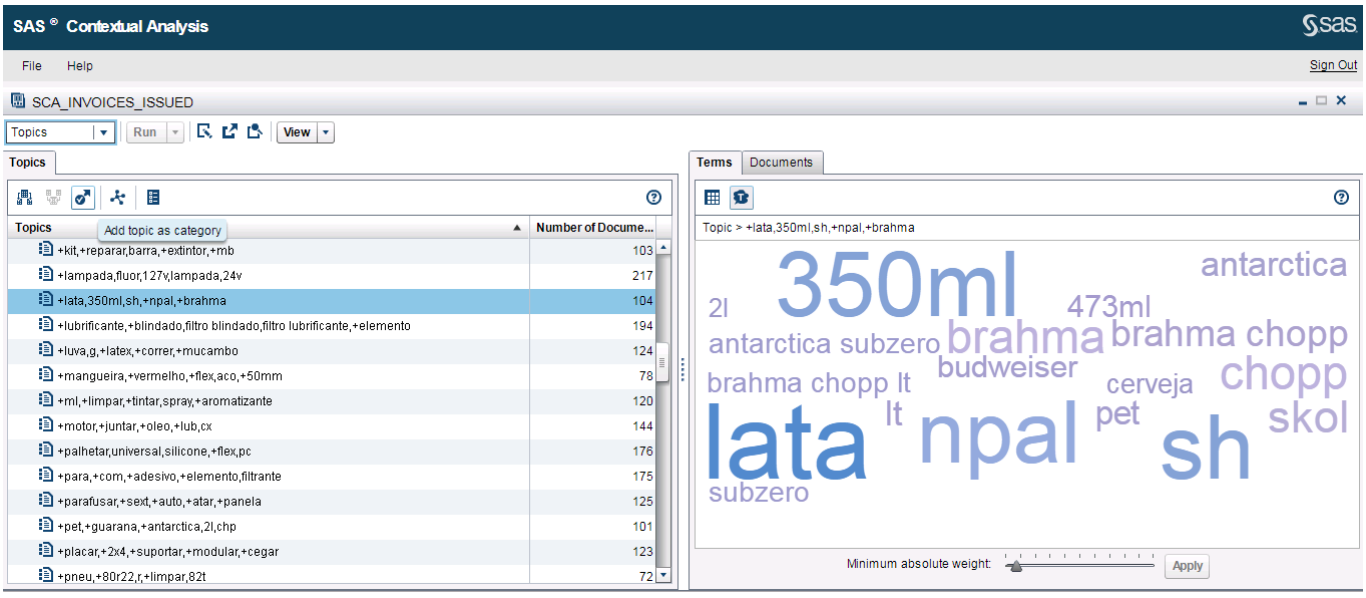
In some situations, the analyst needs to define a specific number of topics because of the structure of their challenges. In Display 14, we change the number of the topics to 99.



Display 14. Topic Properties

HOW TO TRANSFORM TOPICS INTO CATEGORIES

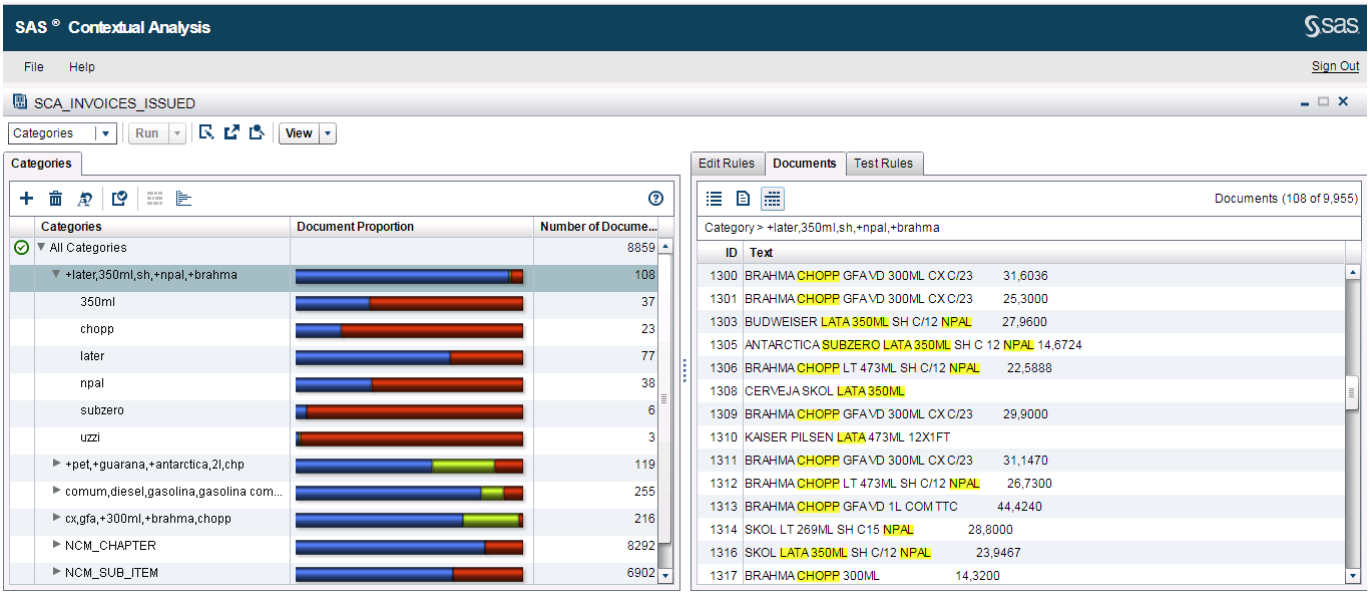
After the topics are validated, you can create categories. Let's continue with the topic that identifies drinks, and promote some topics to be categories. First, you choose a topic and click the Add Topic icon, as shown in Display 15.



Display 15. Promote Topics to Categories

SAS Contextual Analysis suggests possible rules for classifying newly issued invoices. In this example, we transform the topic, which is the type of drinks, into a category that is defined as BEERS (see Display 16). On the **Documents** tab, you can see that out of 9,955 documents, 108 were categorized belonging to the BEERS category

This analysis evaluates how well the displayed linguistic definitions of the categories approximate the underlying machine learning definitions. This is important because you will use the linguistic definitions to score other documents.

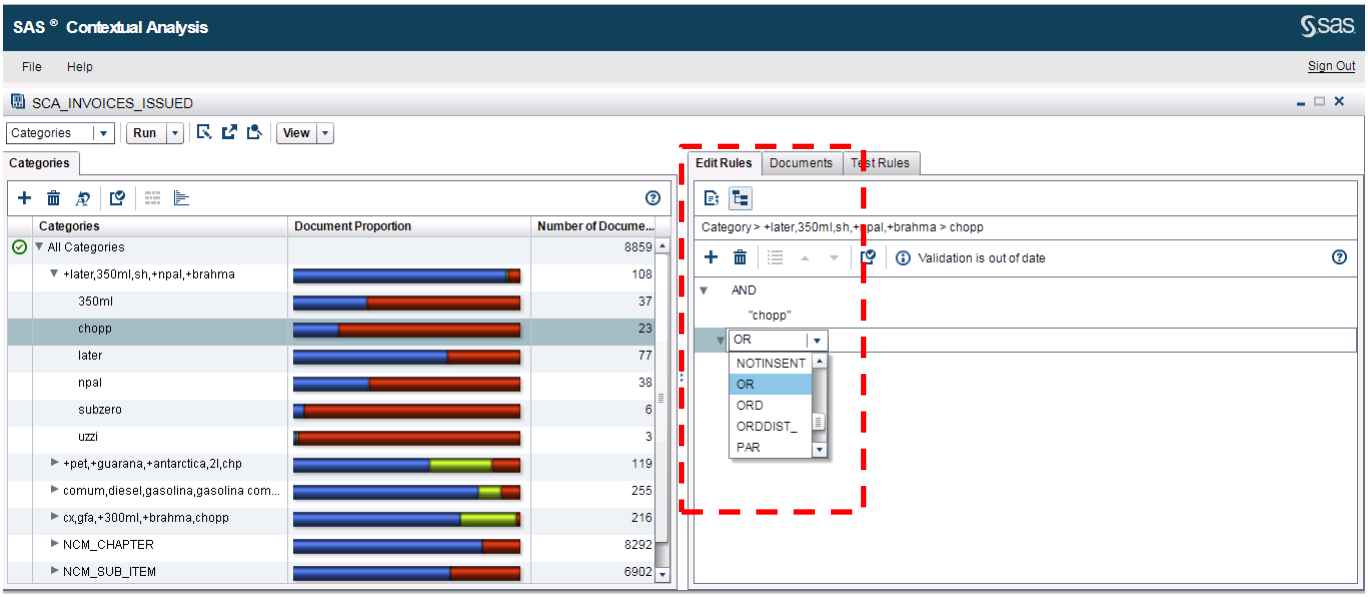


Display 16. Examples of a Category and Its Taxonomies

CATEGORIZATION: EDIT RULES AND SCORE NEW DOCUMENTS

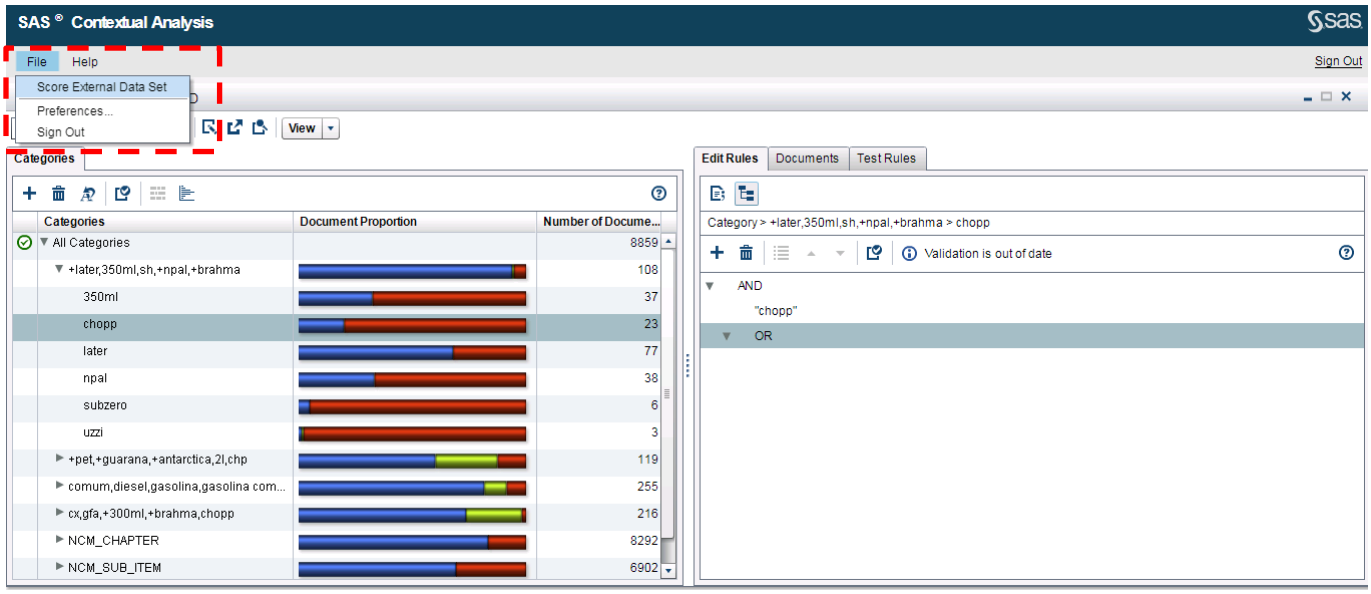
One of the first challenges for the business analyst is to develop a taxonomy that automatically categorizes invoice issues and that is updated in a recurring and more accurate manner according to the NCM. The results and benefits of accomplishing this are immediate, such as properly identifying the tax rate (for example, ICMS) and identifying possible anomalies in the application of the tax rate.

Display 17 shows the new category available in the Categories section. At this point, the analyst can improve the categorization process with the inclusion of his business knowledge on the **Edit Rules** tab.



Display 17. Examples of Categories and Their Taxonomies

You can also use models developed in SAS Contextual Analysis to score additional text data. Select **File>Score External Data Set** (see Display 18). A window appears in which you can identify the textual data that you want to score. Additionally, you can view and export the DS2 macro code used to define concepts, sentiment, and categories for use anywhere you can run SAS.



Display 18. Score External Data Set

Display 19 shows the results after categorization. The variable *document_id* is the ID of the invoices; the variable *name* is the name of the category, and the text with the description of the notes is in the Description_Invoices column.

	document_id	DESCRIPITION_INVOICES	name	term	column_na...
49	1278	FUSION LATA 250ML SIX-PACK	Top/+later,350ml,sh,+npal,+brahma	LATA	c_994
50	1284	BUDWEISER LATA 350ML SH C 12 NPAL 24,4200	Top/+later,350ml,sh,+npal,+brahma	350ML	c_994
51	1284	BUDWEISER LATA 350ML SH C 12 NPAL 24,4200	Top/+later,350ml,sh,+npal,+brahma	LATA	c_994
52	1284	BUDWEISER LATA 350ML SH C 12 NPAL 24,4200	Top/+later,350ml,sh,+npal,+brahma	NPAL	c_994
53	1285	BRAHMA CHOPP LATA 350ML SH C 12 NPAL 19,6200	Top/+later,350ml,sh,+npal,+brahma	350ML	c_994
54	1285	BRAHMA CHOPP LATA 350ML SH C 12 NPAL 19,6200	Top/+later,350ml,sh,+npal,+brahma	CHOPP	c_994
55	1285	BRAHMA CHOPP LATA 350ML SH C 12 NPAL 19,6200	Top/+later,350ml,sh,+npal,+brahma	LATA	c_994
56	1285	BRAHMA CHOPP LATA 350ML SH C 12 NPAL 19,6200	Top/+later,350ml,sh,+npal,+brahma	NPAL	c_994
57	1287	SKOL LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	350ML	c_994
58	1287	SKOL LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	LATA	c_994
59	1287	SKOL LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	NPAL	c_994

Display 19. Categorization Result

INPUTS FOR CALCULATING THE FINAL WEIGHTED AVERAGE CONSUMER PRICE

The calculation of the final weighted consumer average price is updated frequently, and the values for some products rise more than others. In Brazil, the most common products for which the ICMS is calculated based on the final weighted average consumer price are fuels, drinks, and cosmetics, among other goods.

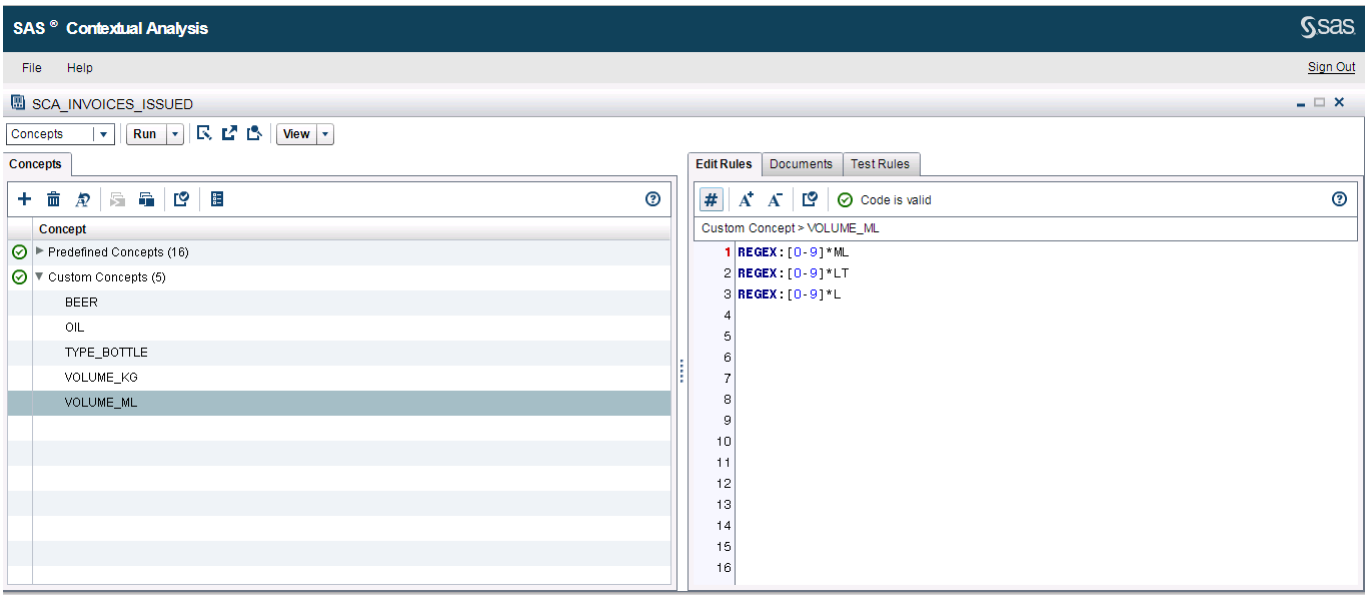
The taxpayer needs to be aware of this calculation and determine whether they are subject it. Otherwise, taxpayers might end up doing their ICMS calculations erroneously.

For this reason, there is a need to extract concepts like volume, type, quantity, and product name from the thousands or millions of Electronic Tax Invoices for inclusion in the calculation of the final weighted average consumer price.

HOW SAS CONTEXTUAL ANALYSIS ENRICHES THE CALCULATION

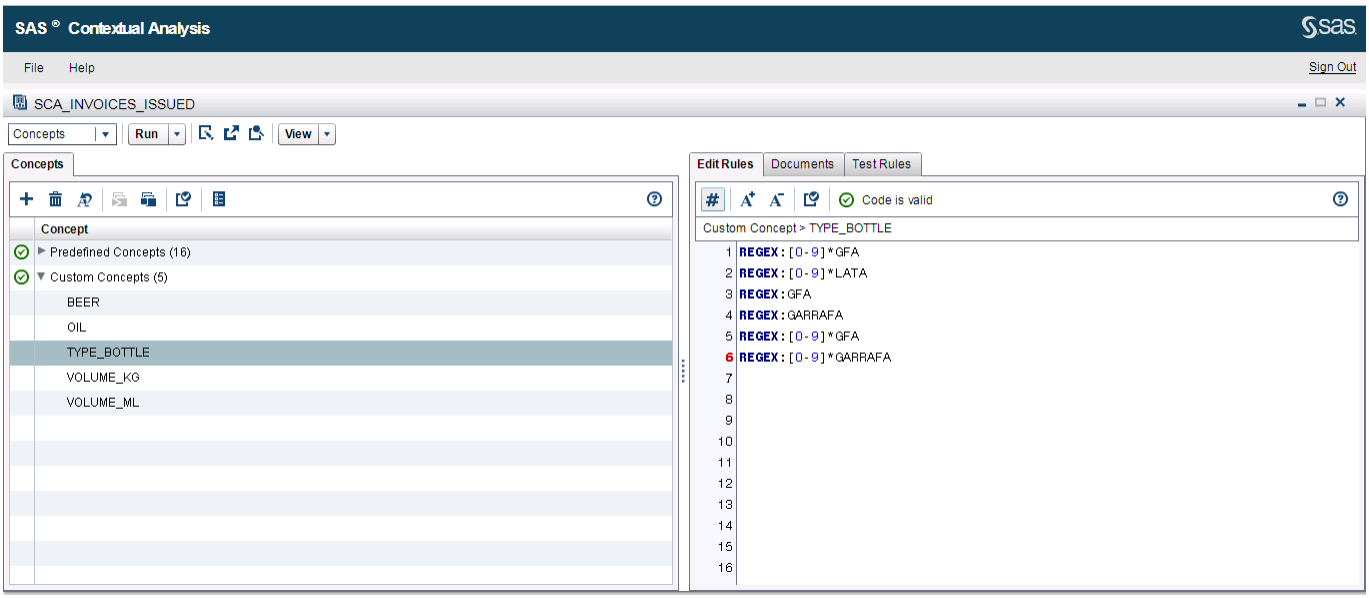
SAS Contextual Analysis uses language interpretation and text interpretation (LITI) syntax and its concept rules to recognize terms like. kg, ml, bottle, and so on, in context, so that you can extract only concepts in a document (for example, “Budweiser 355ML”) that match your rule.

In Display 20, you can see a custom concept node named VOLUME_LT and regular expressions (Regex syntax). These elements will extract all Electronic Tax Invoices in our data source that contain “LT” and all combinations that include numbers (RECEX: [0-9]*LT). The operator - is a wildcard that matches any character.

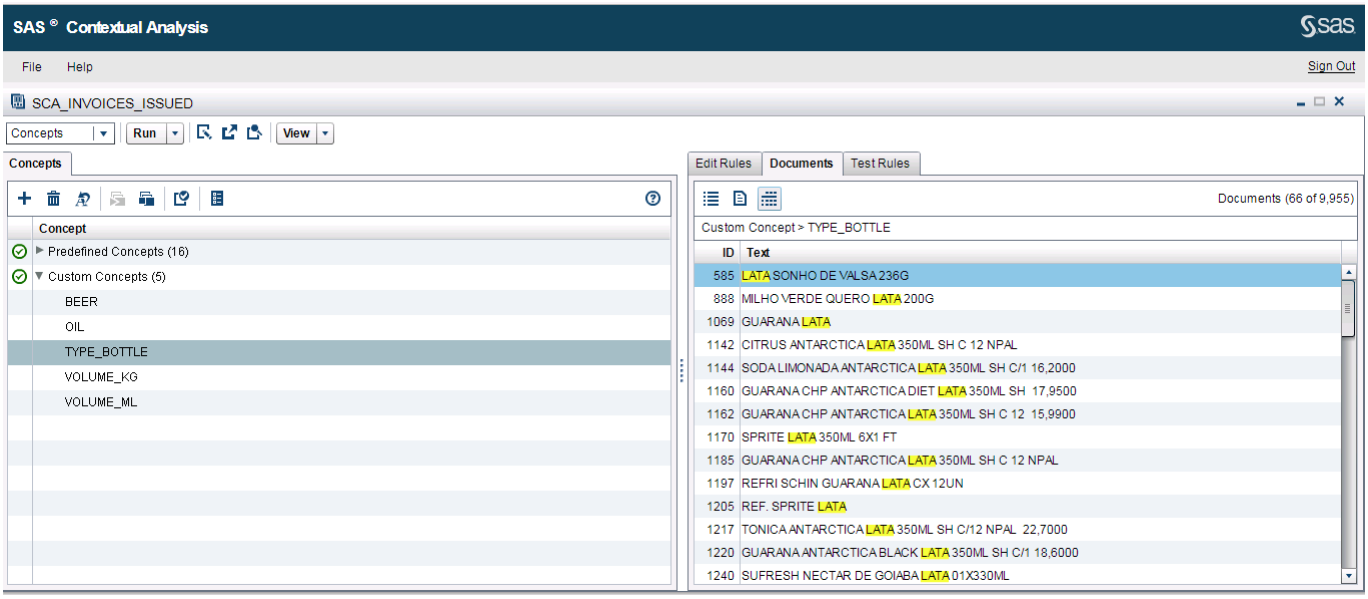


Display 20. Custom Concepts and Editing Rules for VOLUME_ML

Display 21 shows the rule for identifying all Electronic Tax Invoices for the concept node TYPE_BOTTLE that contain the terms "GFA, LATA, GARRAFA" and any number combination. Each document is evaluated separately for matches (shown in Display 22).



Display 21. Custom Concepts and Editing Rules for TYPE_BOTTLE



Display 22. Results of the Custom Rule for TYPE_BOTTLE

ANALYTICAL BASE TABLE FOR THE CALCULATION

Today, the final weighted average consumer price is typically obtained from sample surveys of final consumer prices. Such surveys can be ordered from the Finance Secretary.

Display 23 shows an example created in the SAS Contextual Analysis, which shows a possible analytical basis that can be used in the final weighted final consumer price calculation. The variable *document_id* represents the identification of the electronic invoice, DESCRIPITION_INVOICES contains the contents of the invoice, *name* is the category, and *term* is the result of extracting the electronic invoice concepts.

As an example, we could calculate the average final consumer price of all invoices classified as BEERS (*name*) and sold in cans of 355ML (*term* = “can" and name = "+ chopp + brhama + ...") . This process would already be automated, and it would be possible to generate reports in SAS Visual Analytics. This same logic would enrich the calculation for other items like food, building materials, and so on.

	document_id	DESCRIPITION_INVOICES	name	term
1	585	LATA SONHO DE VALSA 236G	Top/+later,350ml,sh,+npal,+brahma	LATA
2	746	PANETT BAUDUCCO LATA DOURADA 1KG	Top/+later,350ml,sh,+npal,+brahma	LATA
3	888	MILHO VERDE QUERO LATA 200G	Top/+later,350ml,sh,+npal,+brahma	LATA
4	1069	GUARANA LATA	Top/+later,350ml,sh,+npal,+brahma	LATA
5	1142	CITRUS ANTARCTICA LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	350ML
6	1142	CITRUS ANTARCTICA LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	LATA
7	1142	CITRUS ANTARCTICA LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	NPAL
8	1144	SODA LIMONADA ANTARCTICA LATA 350ML SH C/1 1...	Top/+later,350ml,sh,+npal,+brahma	350ML
9	1144	SODA LIMONADA ANTARCTICA LATA 350ML SH C/1 1...	Top/+later,350ml,sh,+npal,+brahma	LATA
10	1150	COCA COLA LATA ZERO 350 ML	Top/+later,350ml,sh,+npal,+brahma	LATA
11	1160	GUARANA CHP ANTARCTICA DIET LATA 350ML SH 1...	Top/+later,350ml,sh,+npal,+brahma	350ML
12	1160	GUARANA CHP ANTARCTICA DIET LATA 350ML SH 1...	Top/+later,350ml,sh,+npal,+brahma	LATA

Display 23. Calculating the Final Weighted Average Consumer Price

CONCLUSION

This paper shows how you can use SAS Contextual Analysis to automate the process of product categorization and create custom concepts, using data that supports the calculation of the tax for the Electronic Tax Invoice. This methodology can be used in other Mercosul countries to reduce analysis

time. This methodology can also improve governance, trust, and accuracy for the validation of invoice issues.

REFERENCES

COSTA, Leonardo De Andrade. 2015. “Processo Administrativo Tributário.” FGV Direito Rio. Available https://direitorio.fgv.br/sites/direitorio.fgv.br/files/u100/processo_administrativo_tributario_2015-2.pdf

SAS Contextual Analysis 14.1: Reference Help.

Website “O seu Portal Jurídico da internet-Âmbito Jurídico”. Available http://www.ambito-juridico.com.br/site/. Accessed on February 20, 2018.

Website “Portal da Nota Fiscal Electrônica”. Available http://www.nfe.fazenda.gov.br/portal/principal.aspx. Accessed on February 20, 2018.

ACKNOWLEDGMENTS

The author thanks Mauricio Fonseca Fernandino and Secretaria de Fazenda de Minas Gerais for Data Sources. The author thanks Joan Keyser and Amy Wolfe for help in reviewing the text.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Alexandre Carvalho
SAS Institute Brasil
55 21 99121-3280
Alexandre.carvalho@sas.com
https://br.linkedin.com/in/alexandrecarvalho