Methodology for Creating the Brazilian Government Reference Price Database

Rommel Carvalho[1], Eduardo de Paiva[1], Henrique da Rocha[1], and Gilson Mendes[1]

[1]Department of Strategic Information (DIE)
Brazilian Office of the Comptroller General (CGU)
SAS, Quadra 01, Bloco A, Edifício Darcy Ribeiro
Brasília – Distrito Federal – Brazil

{rommel.carvalho,eduardo.paiva,henrique.rocha,liborio}@cgu.gov.br

Abstract. *One of the main responsibilities of the Brazilian Office of the General Comptroller (CGU) is to identify purchases that deviate from normality. A key requirement of this process is to create and maintain a reference price database. Even though the purchases are recorded daily in a centralized system, product classification is not detailed enough to produce consistent statistics of prices per product. This paper introduces a methodology developed at CGU that relies on Data Mining techniques to address the problem of ensuring a reliable repository from a wealth of mostly unreliable data. The generation of key statistical parameters and some preliminary conclusions are presented as a means to illustrate the research results.*

1. Introduction

This paper addresses the methodology for creating a database of average price paid per product by the Brazilian Federal Government. The main goal is to have a price reference so that auditors as well as regular citizens can assess whether purchases made by the Government are overpriced or not. This is an essential tool to provide both accountability as well as transparency of the daily purchases made by the Brazilian Government.

The information used in order to compute the average price per product comes from the note of purchase commitment, which are available as Open Government Data (OGD) at the Brazilian Transparency Portal. The note of purchase commitment, first stage in executing a public purchase, has a list of items that specify every product or service that is being committed. Every item has its description, price, quantity, among other information. This information was chosen because it is the most detailed information about a purchase available in the Brazilian Government database and also because every purchase made by the Federal Government must have this data entered into the SIASG system[1] in order to be able to get the money to pay the supplier. An example of a note of purchase can be found at `http://www.portaltransparencia.gov.br/despesasdiarias/`

[1]The Integrated Administration and General Services System (SIASG) enables automated control actions and management of government procurement, which provided a significant improvement in the management of spending on cost and streamlining of procedures [Moreira 2010].

empenho?documento=158457264082012NE800081. This is one of the purchases that will be analyzed in this paper (500 ml water bottles). As explained, every purchase analyzed came from this website (the Brazilian Transparency Portal), thus can be accessed by any person free of charge.

The main challenge in defining a reference price per product is how to identify which product is being described in the note of purchase commitment. The problem is that most of the relevant information is in the description text field. Fortunately, the description is semi-structured, which allows the extraction of a few extra information. The most relevant information which can be extracted is the category code.

The category code is a code used by the SIASG system to identify the product or service which is being acquired. Although important, this information is not precise enough. For instance, the code 21806 is used to identify simple batteries. However, there is no structured information which describes what type of battery it is (AA, AAA, C, D, etc).

The paper is structured as follows. Section 2 describes the proposed methodology which relies on Data Mining techniques to identify the products in each purchase in order to obtain a reliable and precise reference price. Section 3 presents some statistical results obtained from the products analyzed, including confidence intervals of price average. Finally, Section 4 draws some conclusions, describes the deployment plan of the constructed database, and presents some future work.

2. Methodology

As explained in Section 1 the main challenge when trying to define a reference price per product is to be able to correctly classify the products. Although a category code is available, in most cases, it is too broad thus not enough to pinpoint a specific product.

Although the category code defines both products and services, the only codes which will be considered when constructing the price database are of products. The reason for ignoring services is that the same service (e.g., software development) can vary too much in its details, which will reflect in totally different prices (e.g., a simple agency web page when compared to a system for allowing electronic votes for presidency). Therefore, it is not reasonable nor useful to compute averages in these situations.

Before using any advanced techniques, experts from CGU analyzed a simple and intuitive methodology for computing the reference prices. This methodology includes 6 major steps:

1. Retrieve the notes of purchase commitment for a given period from the Brazilian Transparency Portal database.
2. For every note of purchase commitment, retrieve the category code from the SIASG's database.
3. Filter the resulting dataset by category code to retrieve only the notes of purchase commitment of a given product (e.g., code 21806 which refers to simple batteries).

4. Filter the resulting dataset by keywords in order to pinpoint a specific product (e.g., aaa).
5. Filter the resulting dataset by price range.
6. Finally, compute the reference price for the product.

Steps 1 and 2 are performed by an Extract, Transform and Load (ETL) process. Step 3 is a simple and straightforward Structured Query Language (SQL) query. In step 4, the experts define keywords that should be present in the description field of the note (e.g., aaa) but also keywords that should not be present in the description (e.g., car, if we want to make sure a car battery will not be in our result). The goal of step 4 is to pinpoint the specific product we want to compute the reference price for. Even after pinpointing a specific product, the experts realized that the price still had a large variance, which had a huge impact when computing the reference price, which resulted in an unrealistic price when compared to standard prices found in commerce. The reason for such difference was that these purchases had not only some outliers, but also some other subtleties which were not easily captured by the expert when first looking at the data. For instance, there could be purchases of pairs of batteries as well as boxes of batteries. Since these subtleties were hard to catch by hand, the experts came up with a price range which could be thought as reasonable for the type of product in the unit of measure they were thinking about plus a large margin of error in order to account for unexpected values (e.g., overprice and errors), which is the reason for step 5. Finally, step 6 was, at first, just the computation of the price average. Nevertheless, once the average was computed, the experts realized that the result was still not quite correct (too far from what it seemed to be the right one). The problem was that defining these price range was a though task and some data-points (which could be outliers) had a huge impact in the average. Therefore, they decided to use the median as the reference price.

As it can be seen, in this initial methodology, a lot of questions arise, especially in steps 4 and 5. How can we find out which keywords to use? Do we have to have experts manually searching for them? But there are hundreds of thousands of records, how many experts will be needed? How can we define the price range? Is this just an expert feeling? What if there is no expert available that knows that type of product (e.g., some specific medicine)? How can we trust the expert chose a reliable price range?

These questions led to the development of the proposed methodology in this paper, which uses well known Data Mining techniques to address these problems. Section 2.1 describes how to find similar products by clustering them based on the purchase price, which is a solid justification for price ranges used in step 5. Finally, Section 2.2 uses text mining techniques in order to classify the clusters found by using keywords from the description text field, which is an automatic way of doing step 4.

2.1. Using clustering to define reliable price range

As previously explained, one major challenge is to find a group of purchases that describe the exact same product (e.g., 500 ml water bottle) and in the same unit (e.g., box of 12 or box of 24). Intuitively, experts used price range for such tasks,

Table 1. Sample from water bottle (500 ml) dataset from years 2011 and 2012

	Description	Price (R$)	Quantity	Total Price (R$)
1	0000000001,00000 garrafa agua mineral garrafa de 500 ml de agua mineral sem gas, marca agua da pedra. marca: agua da pedra item do processo: 00003 item de material: 000009873	75.36	1.00	75.36
2	10,00000 cx agua mineral agua mineral com gas - pvc, com 24 unid de 500 ml marca: mil item do processo: 00215 item de material: 000009873	29.36	10.00	293.60
586	500,00000 garrafa agua mineral agua mineral c/ gas, garrafa c/ 500 ml marca: veragua item do processo: 00005 item de material: 000009873	0.32	500.00	160.00
587	300,00000 garrafa agua mineral agua mineral c/ gas, garrafa c/ 500 ml marca: veragua item do processo: 00005 item de material: 000009873	0.32	300.00	96.00

however, choosing the correct range and justifying the reason for that was not trivial. In this Section we will use the notes of purchase commitments for water bottles (500 ml) from 2011 and 2012 to show how clustering can reliably find these price ranges.

Table 1 presents a sample from the water bottle (500 ml) product dataset from years 2011 and 2012.

The main assumption that must hold when performing this clustering analysis is that most of the purchases of the same product in the same unit (e.g., box of 12 water bottles) have similar price while purchases of the same product in different unit (e.g., box of 24 500 ml water bottles) have significantly different price. The same must be true if different products (e.g., aa vs D batteries) are present in the same dataset which is being clustered. This is a reasonable assumption, since error and/or fraud is usually the exception, not the rule.

This is exactly what cluster analysis do. Clustering has the objective of grouping a set of objects in order to maximize their similarity inside the same group (called cluster) and minimize the similarity they have to objects in other groups (clusters) [Jain et al. 1999].

There are several clustering algorithms. One of the most common is the k-means [Hartigan and Wong 1979], which represents each cluster by a single mean vector.

Every analysis made on the note of purchase commitment dataset was done using the R software[2], including the clustering analysis.

Although the k-means algorithm was used during our first analysis, we switched to fixed point clusters (FPC) [Hennig 2002, Hennig 2003] algorithm available as an R package[3]. The main reason for using FPC is that it is not necessary to define the number of clusters in advance (before running the clustering algorithm) as in k-means clustering algorithm. FPC computes the number of clusters automatically via bootstrap as explained in [Fang and Wang 2012]. This automatic process is crucial, since we want to analyze several products and not every expert would be able to identify the correct number of clusters without some additional training.

Figure 1 shows the result of clustering the water bottle (500 ml) product dataset from years 2011 and 2012. At this point, all we can say is that there are price ranges which seem to group similar products. But, what do these groups mean? Do they really represent similar products? Just by looking at their value,

[2]http://www.r-project.org/

[3]R: Linear Regression Fixed Point Clusters at `http://rss.acs.unt.edu/Rdoc/library/fpc/html/fixreg.html`

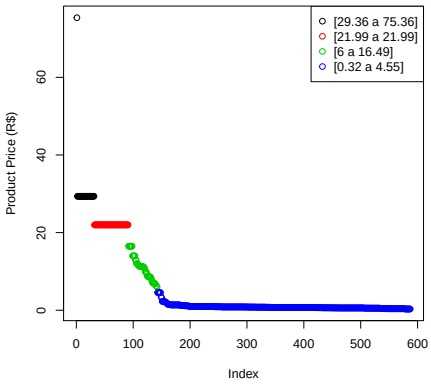

Figure 1. Result of clustering the water bottle (500 ml) product dataset from years 2011 and 2012

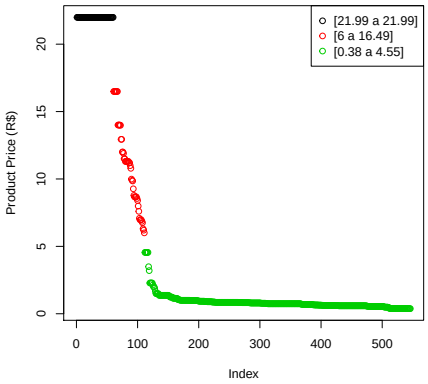

Figure 2. Result of clustering the water bottle (500 ml) product dataset from years 2011 and 2012 without outliers

we are not able to validate these cluster. We need to understand what they have in common, i.e., which type of product they represent. Section 2.2 presents how we use text mining techniques in the description text field in order to classify each cluster.

2.2. Using text mining for classifying product clusters

Text mining refers to the process of discovering useful information from text. This useful information is usually found via statistical pattern learning. Text mining usually involves the process of structuring the text, finding patterns using the structured data found, and finally analyzing its result [Tan et al. 2005].

There are several tasks involved in text mining: text categorization and text clustering [Srivastava and Sahami 2009, Sebastiani 2002], concept/entity extraction [Al Fawareh et al. 2008], production of granular taxonomies [Feldman et al. 1998], sentiment analysis [Pang and Lee 2008, Gamon et al. 2005], document summarization [Larocca Neto et al. , Larsen and Aone 1999, Hu and Liu 2004], and entity relation modeling [Cohen and Hersh 2005, Kao and Poteet 2007]. The overall goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) [Manning and Schütze 1999] and analytical methods.

Section 2.1 presented how we use clustering to define groups of purchases that fall in the same price range. However, in order to make sure these purchases are indeed talking about the same product, we need to extract its semantics from the description text field. Therefore, our main goal in this Section is to use text mining for classifying each cluster to be able to make such assessment.

After applying some text mining techniques, like creating the corpus, which represents a collection of text documents, preprocessing the corpus (e.g., stripping white spaces, removing stop-words), and creating the term-document matrix, we were able to find the most frequent words in each cluster. Finally, with these most frequent words per cluster, we were able to compute the words that better define the cluster, which we interpret as the cluster classification. These words are the set of

Table 2. Result of the text mining classification of each cluster found in the water bottle (500 ml) product dataset

	Most Frequent Words in the Cluster	Words that Better Define the Cluster
1	000009873 / 00215 / 24 / 500 / agua / cx / gas / item / marca: / material: / mil / mineral / ml / processo: / pvc, / unid	00215 / 24 / cx / gas / mil / ml / pvc, / unid
2	000009873 / 00214 / 24 / 500 / agua / cx / item / marca: / material: / mil / mineral / ml / processo: / pvc, / unid	00214 / 24 / cx / mil / ml / pvc, / unid
3	000009873 / 12 / 500 / agua / gas, / item / marca: / material: / mineral / processo:	12 / gas,
4	000009873 / 500 / agua / garrafa / gas, / item / marca: / material: / mineral / ml / processo:	garrafa / gas, / ml

most frequent words minus the intersection of the sets of most frequent words from all clusters but the one being analyzed. All text mining analysis was done using the Text Mining Infrastructure in R (TM) package[4] [Feinerer et al. 2008, Feinerer 2013].

Table 2 presents the text mining classification result for each cluster found for the water bottle (500 ml) product dataset from years 2011 and 2012. As it can be seen, the first and second clusters represent purchases of boxes of 24 bottles (from the keyword *cx*, which is short for *caixa* - box in Portuguese - and keyword 24). The third cluster represents purchases of boxes of 12 bottles (from the keyword 12). Finally, the fourth cluster represents purchases of single bottles (from the keyword *garrafa*, which is bottle in Portuguese).

Although a human being has to verify the sets of keywords in order to better classify each cluster, it is a lot faster, easier, and less error-prone than going through every single purchase manually. This product alone, which is one of the products with fewer purchases, has over 500 purchases just in 2011 and 2012. It would be unrealistic to ask an expert to look at all these purchases for hundreds of products. However, with this methodology, this is actually feasible. In fact, it is quite fast. It took an expert, in average, less than 10 minutes to generate the report with this classification information and assess what each cluster means in a more human understandable way. This is why this methodology is so useful.

One question that may arise from the classification found by the text mining algorithm is why the first two groups were not identified as just one. Since the words that better define both groups are the same.

The problem is that the highest data-point is way higher than all the other values (see Figure 1). In fact, this is a clear outlier. When finding the clusters for this product, this outlier ends up causing the creation of a new cluster.

After some exploratory analysis we were able to identify that for most of the products we analyzed (about 50 products), at least 1% of the data-points are outliers. Therefore, for the purpose of better identifying the clusters, we may ignore these outliers by filtering the 1% of the data-points with the highest and lowest values (0.5% of the highest and 0.5% of the lowest).

Figure 2 shows the result of clustering the water bottle (500 ml) product dataset from years 2011 and 2012, but now removing the supposed outliers. Notice that now the number of clusters found was 3, as expected, not 4. As a result, our text mining now is able to correctly classify all three clusters.

[4] `http://tm.r-forge.r-project.org/`

During the process of removing outliers, we might end up removing data that was not necessarily an outlier. Nevertheless, even by removing some of these data, we still have enough data-points to compute with a reasonable confidence the average price per product. After all, we still have 99% of the data-points to make that estimate. Furthermore, as we will see in Section 3, we just compute confidence intervals for large datasets (more than 30 data-points). For clusters with less than 30 data-points, we simply do not add to our price database, since we are not confident in its accuracy.

Finally, we also generate an SQL query in order to allow the recovery of all data-points that contain the words that better describe each cluster. This way, we might get some purchases that were classified as part of one cluster (e.g., box of 24 bottles), when in fact they could be part of a different cluster (e.g., single bottle). In fact, this is the case for the purchase with the highest price. Although it was classified as a purchase of a box of 24 bottles, it is in fact a purchase of a single bottle (as it can be seen in the purchase description in Table 1).

3. Results

This Section presents the last step in our methodology, which is computing the statistical results per type of product (i.e., each cluster defined previously) in order to define the reference price of the product which will be added to our reference price database.

Although we have analyzed the overall fit of the clusters found only manually and for a random sample of the purchases, we believe that it classified most of the purchases in the correct category (e.g., bottle, box with 12 bottles, and box with 24 bottles). However, a more thorough performance analysis is needed in order to confirm our initial analysis/conclusion[5].

Nevertheless, we also compared the final mean price found for each of the products and categories we evaluated with the price found in regular stores and they were similar. Since finding reference price (means) is our main goal, we believe our approach presents at least a satisfactory result. If we had too many products being classified in the wrong category, the final mean price would be much higher or lower than expected. For instance, on the one hand, if too many purchases of single bottles were misclassified as purchases of boxes of 12 bottles, we would end up having a mean price for this category much lower than what is expected. On the other hand, if too many purchases of boxes of 12 bottles were misclassified as purchases of single bottles, we would end up having a mean price for this category much higher than expected. Since all categories presented a price similar to those found in regular stores, we believe they were correctly classified in our approach.

One problem we were able to identify with our approach is in cases where a product A shares the same price range with a different product B when they both share the same category code. If this happens (and it does happen in some cases), we are not able to differentiate between the two using our approach (which

[5]This is already being done, however, due to the size of the database, and the fact that we have to manually classify each purchase we want to validate, this is taking a long time and we were not able to finish it at this point.

Table 3. Statistics summary per cluster for the price of water bottles (500 ml) in 2011 and 2012

Cluster	Range [Min,Max]	Quart.[1st,3rd]	Mean 95% Conf. Int.	Mean	Median	Size
All - per Purchase	[0.38, 21.99]	[0.62, 1.37]	[3.53, 4.71]	4.12	0.82	546
All - per Product	[0.38, 21.99]	[0.54, 0.9]	[2.66, 2.72]	2.69	0.60	158,567
1 - per Purchase	[21.99, 21.99]	[21.99, 21.99]	[21.99, 21.99]	21.99	21.99	60
1 - per Product	[21.99, 21.99]	[21.99, 21.99]	[21.99, 21.99]	21.99	21.99	11,352
2 - per Purchase	[6, 16.49]	[8.61, 12.94]	[10.06, 11.80]	10.93	11.20	51
2 - per Product	[6, 16.49]	[8.67, 11.3]	[9.64, 9.76]	9.70	8.67	7,158
3 - per Purchase	[0.38, 4.55]	[0.59, 0.89]	[0.80, 0.91]	0.86	0.74	435
3 - per Product	[0.38, 4.55]	[0.39, 0.78]	[0.76, 0.77]	0.77	0.59	140,057

uses only price to differentiate the two). However, we were able to select several category codes that did not have this problem and we focused our analysis on these products. In future work we plan on incorporating other features in order to allow the differentiation in these cases.

Table 3 presents some of the summary statistics computed for the water bottle (500 ml) product. These summaries were computed for each cluster and also for the entire dataset (if there subtypes of products were not identified) for comparison. Furthermore, these summaries were computed per purchase, but also per product. Each purchase has a quantity associated with it, which is used to compute the summaries per product. The reason to compute these summaries per product is that the price paid in a purchase of 100 thousand bottles should have a higher weight than the price paid in a purchase of a single bottle.

The first thing we notice is that the number of purchases of single bottles is by far larger than the purchase of boxes of bottles (about 80% of the total purchases). Therefore, the reference price for single bottles should be considered more reliable, or closer to the "real" fair price, than the reference price for boxes. This is actually reflected in the mean 95% confidence interval. The confidence interval range for single bottle purchases is a much narrower ([0.80, 0.91]) than box purchases ([10.06, 11.80] for the box with 12 bottles). This is not always true, as it can be seen for the box with 24 bottles, since all 60 purchases had the same price associated with them (21.99).

The same principle holds when comparing the mean confidence intervals of purchases and products, since the number of products bought is much higher than the number of purchases made.

It is also important to validate if these values correspond to reality. A simple way to validate the price is to compare to the price a regular citizen would pay, since this product is common and easy to find. For instance, a water bottle (500 ml) can be bought in the supermarket for a little less than R$1.00 here in Brazil. So, if the Government pays, in average, R$0.77 per bottle, than we known we have a good price for reference. Furthermore, we also have some good news in the sense that the Government is paying less than a regular citizen who is buying just a few bottles.

Nevertheless, it does not sound reasonable to expect every single purchase to be made for that computed average price every time, since the price may vary during the year, it varies from state to state, etc. Therefore, it is more reasonable to define a reference interval instead of single price per product. This is where the ranges shown in Table 3 play a key role. As it can be seen, 50% of all purchases of

single bottles paid between R$0.59 and R$0.89 and 50% of all bottles purchased cost between R$0.39 and R$0.78. A more interesting interval is shown in the mean 95% confidence interval. Although the mean of purchases of single bottles was R$0.86, its 95% confidence interval goes from R$0.80 to R$0.91. On the other hand, while the mean of the price paid per bottle was R$0.77, its 95% confidence interval goes from R$0.76 to R$0.77.

Even though it is hard to decide which interval to use for the reference price. After some discussion with the experts at CGU, it was decided that the reference price will be the interval from the minimum value between both confidence intervals (per purchase and per product) to the maximum value between both confidence intervals. Thus, the reference price for water bottle (500 ml) will be from R$0.76 to R$0.91.

It is unquestionable that, independent of which interval is chosen, having these parameters will be of great value to those responsible for the procurements, for auditors, and also for citizens.

The department responsible for all purchases at CGU (i.e., responsible for the procurements) has already shown interest in our reference price database and has asked us to compute reference price for many other products.

The department at CGU responsible for auditing and inspecting all Federal expenses has also shown interest in our database. They use the information in our database to verify if the procurements they have to analyze are not overprice. For instance, for this water bottle (500 ml) product we can identify a purchase of 1.8 thousand bottles at the price of R$3.49, a price 453% higher than the average price (R$0.77). This purchase alone incurred in a loss of almost R$5 thousand. Notice that we are describing a purchase of a product that is worth less than R$1.00. In fact, the potential loss with purchases of water bottles (500 ml) (just the purchases of single bottles, not boxes) is a bit over R$13 thousand. This value was computed as the sum over all purchases above the maximum reference price (R$0.91) of the purchase price minus R$0.91 times the quantity of bottles bought. Even greater losses might be expected for products with higher average price. This is why this reference price database is so valuable and important for a better management of public resources.

4. Conclusions

This paper presented a major problem that prohibits the creation of a reference price database in the Brazilian Government, which is being unable to categorize Government purchases precisely. The main challenge comes from the fact that the information available is not structured and the only classification available is too broad to allow the definition of a reference price.

In order to tackle this problem, CGU developed a methodology, described in Section 2, which uses Data Mining techniques to overcome the challenges presented. Section 3 demonstrated that the methodology works by showing how this methodology was used to compute the reference price of water bottles (500 ml).

One of the major benefits of the methodology and resulting reference price

database presented in this paper is in improving the management of public resources. As an example, recently, the Court of Auditors from the Federal District (DF) did a similar task of finding reference prices but for a specific procurement. The procurement of the DF Health Secretariat, which was initially estimated in almost R$86 millions to buy different medications, was only allowed to continue after it was re-estimated. The results were outrageous as the re-estimated price felt to less than R$13 millions. The use of reference prices was responsible for savings over 85%, about R$73 millions[6].

For future work we intend to make a more rigorous and thorough performance evaluation of the classification given by our approach as well as incorporating new features in our clustering analysis in order to differentiate between products that have similar price and share the same product category.

As described in Section 3, many different stakeholders are interested in the price reference database, including the department responsible for the procurements at CGU, the department at CGU responsible for auditing and inspecting Federal expenses, as well as citizens. Thus, CGU has made an effort to create an initial version of the database with 65 products, which is already available to all departments at CGU for querying. Moreover, since transparency and accountability are major goals of the institution, it is being discussed how to make this database available for all citizens in the Transparency Portal. A commitment was already made in Brazil's second action plan developed for the Open Government Partnership (OGP) and it can be found at `http://www.cgu.gov.br/Imprensa/Noticias/2013/noticia05413.asp`.

Besides allowing simple queries in the database, tools for improving auditing and inspection are being designed. The focus will be in allowing the identification of who is spending over the reference price, which agencies are being able to buy for less, how much the Government is spending over the reference price, among other features.

Acknowledgments.

The authors would like to thank CGU for the support in this project and for letting the authors publish these results.

References

Al Fawareh, H., Jusoh, S., and Osman, W. (2008). Ambiguity in text mining. In *International Conference on Computer and Communication Engineering, 2008. ICCCE 2008*, pages 1172 –1176.

Cohen, A. M. and Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71.

Fang, Y. and Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Comput. Stat. Data Anal.*, 56(3):468–477.

Feinerer, I. (2013). Introduction to the tm package text mining in r. Technical report.

[6] `http://www.tc.df.gov.br/web/tcdf1/noticias/-/asset_publisher/a5YM/content/valor-de-licitacao-para-compra-de-remedios-tem-reducao-de-85`

Feinerer, I., Hornik, K., and Meyer, D. (2008). Text mining infrastructure in r. *Journal of Statistical Software*, 25(5):1–54.

Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., and Zamir, O. (1998). Text mining at the term level. In Zytkow, J. and Quafafou, M., editors, *Principles of Data Mining and Knowledge Discovery*, volume 1510 of *Lecture Notes in Computer Science*, pages 65–73. Springer Berlin / Heidelberg.

Gamon, M., Aue, A., Corston-Oliver, S., and Ringger, E. (2005). Pulse: Mining customer opinions from free text. In *Advances in Intelligent Data Analysis VI*, volume 3646 of *Lecture Notes in Computer Science*, pages 741–741. Springer Berlin / Heidelberg.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 28(1):100.

Hennig, C. (2002). Fixed point clusters for linear regression: Computation and comparison. *Journal of Classification*, 19(2):249–276.

Hennig, C. (2003). Clusters, outliers, and regression: fixed point clusters. *J. Multivar. Anal.*, 86(1):183–212.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323.

Kao, A. and Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. Springer.

Larocca Neto, J., Santos, A. D., Kaestner, C. A., and Freitas, A. A. *Document Clustering and Text Summarization*. Postgraduate thesis, Pontificia Universidade Catolica do Parana.

Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 16–22, New York, NY, USA. ACM.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. BARNES & NOBLE.

Moreira, C. H. d. A. (2010). Implementation of e-procurement system in brazil. In *Proceedings: Towards Frontiers in Public Procurement*.

Pang, B. and Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

Srivastava, A. and Sahami, M., editors (2009). *Text Mining: Classification, Clustering, and Applications*. Chapman and Hall/CRC, 1 edition.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley, 1 edition.