Metodologia Cientifica

Jhonatan Batista, Hebert, Ronnyldo e Wesley Roseno Saraiva

11 de junho de 2019

### Apresentação do problema

Existem sites dedicados a fazerem registros de reclamações de cliente em relação ao serviço oferecido por empresas, autarquias, ou por agências do próprio governo federal. Portanto, existe uma necessidade estabelecida pelo mercado, de se fazerem análises para se ter uma melhor classificação dos registros de insatisfação dos clientes e consumidores dos serviços. Desta forma, é de interesse do governo federal e de empresas, desenvolver métodos automatizados para análise de insatisfação presente nas reclamações devido ao grande número de reclamações. Estas análises podem ser úteis para otimizar a qualidade dos serviços de reclamações responsáveis por ouvir a população, bem como definir prioridades para empresas e governo nas áreas de atuação das correções dos problemas reportados pelos clientes.

Este trabalho se propõe a fazer uma análise entre o texto relativo a insatisfação do cliente, a nota de insatisfação do cliente e uma nota dada pela análise feita por dois algoritmos de análise de sentimentos. Os algoritmos utilizados para fazer a avaliação léxica são da biblioteca lexiconpt (n.d.).

A biblioteca pode ser utilizada em R e ela tem a função de fazer uma análise léxica de textos escritos em português, fazendo uma análise geral com o intuito de pontuar o texto em sua totalidade. A biblioteca então analisa palavra por palavra, pontuando o texto positivamente se houver incidência de palavras com conotação positiva (E.G. Obrigado, Feliz, Bom, Dia, etc) e pontua negativamente se houvr incidência de palavras com conotação negativa (E.G. Raiva, palavrões, etc). No fim ela faz um somatório do texto e retorna um valor sentimental do texto que pode variar de infinito negativo até infinito positivo.

Desta forma, nossa variável resposta seria o grau de insatisfação e as variáves explicativas seriam a quantidade de caracteres e as notas dada pela biblioteca lexiconpt.

###LexiconPt LexiconPt é uma biblioteca de R que traz um ferramentário para fazer análise léxica de textos em português. Ela foi feita para facilitar o trabalho de analistas na facilitação da mineração textual. Ela é capaz de fazer análise léxica de palavras trazendo o sentimento dela.(Sillasgonzaga 2018,@lexiconPt2,@lexiconDiscuss).

A ideia de mineração de opinião baseada em léxicos de palavras é saber que existem palavras com polaridade positivas (e.g. bom ou bem) e palavras que são indesejadas com polaridade negativa (e.g. ruim ou terrível). Apesar de polaridade de opinião ser aplicadas a adjetivos e advérbios, existem vebos e pronomes que possuem opinião, também.

Existem duas classes de léxicos (positivo/negativo) que se subclassificam em 5 outras classes:

The lexicon contains two classes (positive/negative) and 10 specific lists.

Tipos positivos:

1. Substantivo positivo
2. Adjetivo Positivo
3. Advérbio Positivo
4. Verbo Positivo
5. Frase positiva

Tipos negativos:

1. Substantivo negativo
2. Adjetivo negativo
3. Advérbio negativo
4. Verbo negativo
5. Frase negativo

Exemplo de uso da Biblioteca:

library("lexiconPT")  
get\_word\_sentiment("gentileza")[3]

## $sentilex  
## term grammar\_category polarity polarity\_target  
## 3608 gentileza N 1 N0  
## polarity\_classification  
## 3608 MAN

get\_word\_sentiment("mau")[3]

## $sentilex  
## term grammar\_category polarity polarity\_target  
## 4830 mau Adj -1 N0  
## polarity\_classification  
## 4830 MAN

### Descrição da Base de Dados

A base de dados foi extraída das reclamações recebidas pela ANS (Agência Nacional de Saúde Suplementar) e pela ANATEL (Agência Nacional de Telecomunicações). Onde após o recolhimento das reclamações, foi feito uma avaliação por humanos, da insatisfação presente em cada reclamação, dividida em 5 níveis de 1 a 5. Sendo 5 insatisfação total, 4 insatisfação bastante alta, 3 insatisfação média, 2 insatisfação baixa e 1 insatisfação mínima com o serviço. Com um total de 60 reclamações.

Foi gerado um CSV contendo id,orgao,data,titulo,texto,link,Grupo que vai avaliar,insatisfacao avaliadores,range.avaliacoes,Numero de caracteres, Sentimento da Biblioteca. Como existe muita informação, foi feita uma versão resumida, com apenas um exemplo, para melhor entendimento dos dados que estão sendo trabalhados.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | id | orgao | data | titulo | texto | link | insatisfacao | qtdCaracteres | Sentimento.Biblioteca |
| 2 | 2 | ANS | Terça-feira, 26 de Fevereiro de 2019 - 08:49 | Cabesp familia mensalidade absurda | Cabesp familia teve reajuste de 30% em 2 anos | <https://cidadao.reclameaqui.com.br/406886/ans-agencia-nacional-de-saude-suplementar/cabesp-familia-mensalidade-absurda/> | 1 | 45 | 0 |

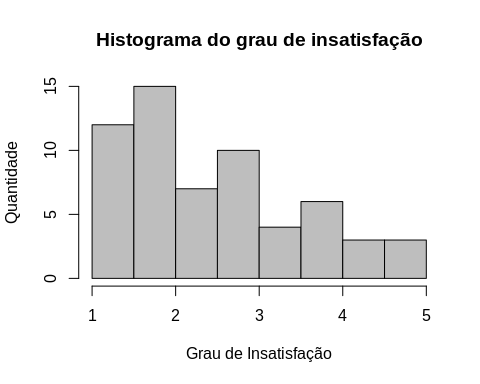
Tendo essa base para análise, foi estudada a viabilidade de automatização da avaliação do nível de insatisfação, através da análise de duas variáveis. A primeira variável sendo a quantidade de caracteres presentes na reclamação, onde se parte do pressuposto de haver uma correlação do tamanho da reclamação com o quanto de insatisfação presente na mesma. E a segunda sendo uma análise léxica através de uma biblioteca especializada disponíveis para a linguagem R, onde se tem o nível de polaridade do sentimento presente no texto.

A ideia é fazer um simples estudo relacionando tais variáveis. Será que a quantidade de texto escrita pelo cliente pode ser um indicador da insatisfação dada por ele nos sites de reclamações? Ou seja, o esforço para escrever o texto de reclamação existe alguma relação com a insatisfação do cliente? E com a análise léxica feita pela biblioteca, será que existe alguma relação?

### Histograma

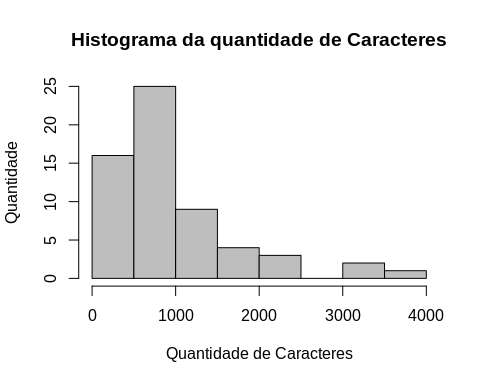
Histogramas possuem a capacidade de trazer uma visualização da distribuição de frequências de uma certa característica das observações. A base representa a classe dos dados, enquanto a altura representa a frequência observada daquela determinada classe. Ele pode servir para melhor entender a distribuição dos nossos dados, indicando se a distribuição dos dados é normal ou não, por exemplo, bem como indicar a possibilidade de haver misturas de populações(Frost 2019). Nos nossos estudos desenvolvemos histograma das nossas três variáveis.

hist(insatisfacao,  
 main="Histograma do grau de insatisfação",  
 xlab="Grau de Insatisfação",  
 ylab="Quantidade",  
 xlim=c(1,5),  
 col="gray",  
 freq=TRUE  
);

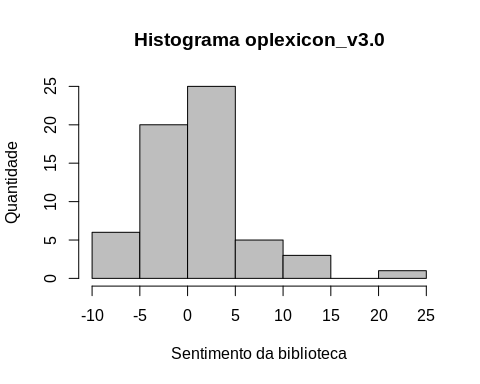


Como podemos ver acima, grande parte do grau de insatisfação se encontra distorcido a direita e possui mais de um pico, o que pode indicar que a distribuição do grau de insatisfação dos clientes não é normal. Da mesma forma temos os histograma da quantidade de caracteres. Que também estão distorcido à direita dando indícios que não possuem uma distribuição normal.

hist(qtdCaracteres,  
 main="Histograma da quantidade de Caracteres",  
 xlab="Quantidade de Caracteres",  
 ylab="Quantidade",  
 col="gray",  
 freq=TRUE  
);

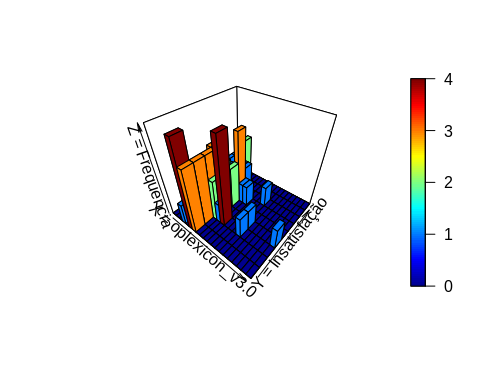


hist(robo,  
 main="Histograma oplexicon\_v3.0",  
 xlab="Sentimento da biblioteca",  
 ylab="Quantidade",  
 col="gray",  
 freq=TRUE  
);



Foram desenvolvidos, também histogramas tridimensionais para melhor entendimento da distribuição das classes se relacionando entre si. A lógica é a mesma: A base e a largura representam as classes e a altura representa a frequencia de ocorrencia das classes combinadas.

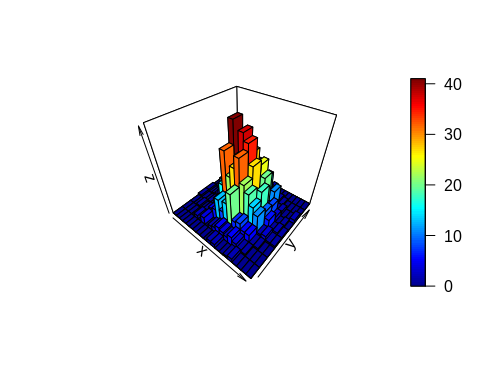
## Calculate joint counts at cut levels:  
z <- table(cut(robo, 20), cut(insatisfacao, 9))  
hist3D(z=z, border="black", zlab="Z = Frequencia", ylab="Y = Insatisfação",xlab="X = oplexicon\_v3.0")



Acima é possível ver a distribuição das classes. O gráfico é um pouco mais difícil de entender, pois estamos usando duas classes simultaneamente para demonstrar a frequencia de ocorrência, mas cada prédio representa uma combinação de classes do grau de insatisfação do cliente e do sentimento da biblioteca utilizada simultaneamente. De tal forma que quanto mais a classe se distancia da origem do eixo X maior é o sentimento da biblioteca; e quanto mais a classe se distancia da origem do eixo y maior é o grau de insatisfação registrado.

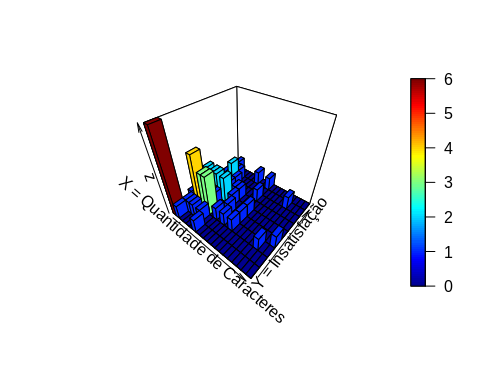
Da mesma forma, o Histograma não representa uma distribuição normal. Foi construido também um exemplo de histograma tridimensional de distribuição normal que pode ser visto a seguir:

hist3D(z=table(cut(rnorm(1000), 20), cut(rnorm(1000), 9)), border="black")



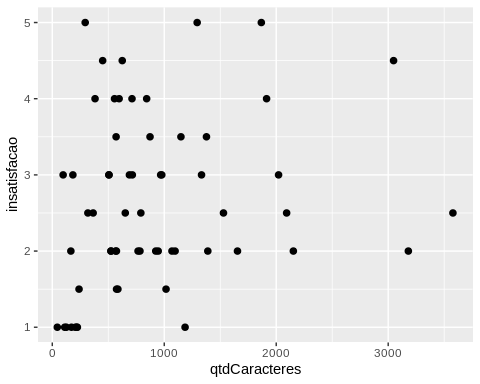
Segue, também, o histograma da quantidade de caracteres e graus de insatisfação:

z <- table(cut(qtdCaracteres, 20), cut(insatisfacao, 9))  
hist3D(z=z, border="black", ylab="Y = Insatisfação",xlab="X = Quantidade de Caracteres")



###Regressão linear simples entre as variáveis A princípio utilizamos a correlação de pearson para fazer os cálculos de nossas correlações, contudo, o resultado da correlação foi muito baixa por conta da dispersão dos nossos dados. Eles estão muito dispersos como é possível ver no gráfico a seguir.

## Warning: Removed 10 rows containing missing values (geom\_point).



Utilizando os códigos a seguir fizemos as correlações entre as variáveis para tentar verificar a existência de uma correlação entre elas, contudo todas elas retornaram valores abaixo de 0.3, indicando que existe uma correlação muito baixa entre todas as variáveis.

cor(insatisfacao,qtdCaracteres, method = "spearman")

## [1] NA

No caso do relacionamento da quantidade de caracteres com o grau de insatisfação, existe uma baixa correlação positiva (quando a insatisfação cresce, a quantidade de caracteres também cresce).

cor(insatisfacao,robo, method = "spearman")

## [1] NA

No caso do relacionamento da quantidade de caracteres com o sentimento do robô, existe uma baixa correlação negativa (quando a insatisfação cresce, o grau de sentimento do robô sobre aquele texto decresce, o que faz sentido uma vez que a biblioteca analisa o sentimento que o robô teve ao fazer a análise léxica).

cor(robo,qtdCaracteres, method = "spearman")

## [1] NA

No caso do relacionamento da quantidade de caracteres com o grau de insatisfação, existe uma baixa correlação positiva.

### Analisando os dados por Análise de Regressão Múltipla

regMult <- lm(insatisfacao ~ robo + qtdCaracteres, data=dados);  
kable(confint(regMult))

|  |  |  |
| --- | --- | --- |
|  | 2.5 % | 97.5 % |
| (Intercept) | 1.8589502 | 2.7155958 |
| robo | -0.1399118 | -0.0238675 |
| qtdCaracteres | 0.0001205 | 0.0008837 |

Segundo a apostila da disciplina e consultas externas de livros, podemos utilizar com sucesso o método da soma dos mínimos quadrados para fazer uma análise de correlação linear, de forma que buscamos representar os dados por meio de uma reta, levando em consideração uma probabilidade de erro. No caso desta análise de regressão múltipla, nossos dados se comportaram da forma:

Onde X é a insatisfação registrada pelo humano, X1 é o sentimento registrado pela biblioteca e x2 é a quantidade de caracteres da reclamação, trocando as variáveis, temos:

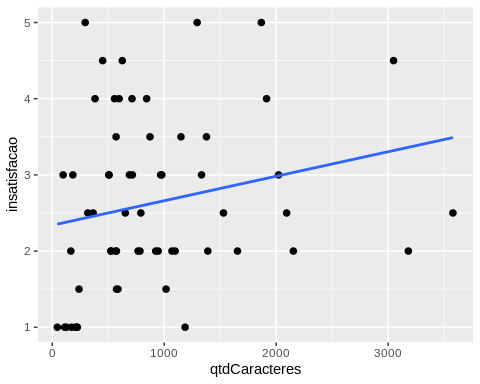
O que pode-se extrair da função dada acima é que o ponto de interceptação da reta de insatisfação no modelo é 1.90e as variáveis X1 e X2 possuem muito baixo. Neste caso X1 é uma variável que explica pouco a variável X e X2 é uma variável que não explica X1.

Esta baixa correlação é confirmada pelos gráficos de dispersão construídos na fase de regressão linear simples das variáveis.

ggplot(dados, aes(x=qtdCaracteres, y=insatisfacao)) +  
 geom\_point(size=2)+  
 geom\_smooth(method=lm,se=FALSE)

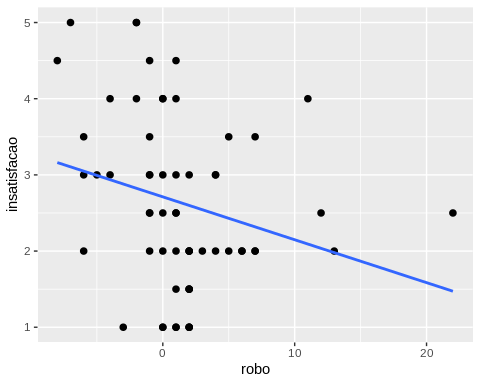
## Warning: Removed 10 rows containing non-finite values (stat\_smooth).

## Warning: Removed 10 rows containing missing values (geom\_point).



ggplot(dados, aes(x=robo, y=insatisfacao)) +  
 geom\_point(size=2)+  
 geom\_smooth(method=lm,se=FALSE)

## Warning: Removed 10 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 10 rows containing missing values (geom\_point).



Sendo assim, para buscarmos investigar essa hipótese de existir ou não uma correlação linear entre eles, dadas as especulações acima, justifica-se o uso do modelo de mínimos quadrados (MMQ).

### Verificando a adequação do modelo

Para verificar a adequação do modelo aos dados ou seja, o quão bem, esse modelo representa nossos dados, depende-se de uma medida chamada medida de adequação, ou coeficiente de determinação / explicação chamado , onde é é nosso coeficiente de correlação linear amostral. Trata-se de uma definição abordada na regressão simples, mas que pode ser expandida à análise de regressão múltipla. Foi calculado por meio do software proposto pelo professor, a nossa tabela anova, onde temos acesso a estatísticas que podem nos auxiliar no cálculo dessa medida.

kable(anova(regMult))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| robo | 1 | 4.705206 | 4.705206 | 4.253757 | 0.0437314 |
| qtdCaracteres | 1 | 7.678753 | 7.678753 | 6.942003 | 0.0108234 |
| Residuals | 57 | 63.049375 | 1.106129 | NA | NA |

summary(regMult);

##   
## Call:  
## lm(formula = insatisfacao ~ robo + qtdCaracteres, data = dados)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.8009 -0.7454 -0.1890 0.6930 2.4013   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.2872730 0.2138977 10.693 3.04e-15 \*\*\*  
## robo -0.0818897 0.0289754 -2.826 0.00648 \*\*   
## qtdCaracteres 0.0005021 0.0001906 2.635 0.01082 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.052 on 57 degrees of freedom  
## (10 observations deleted due to missingness)  
## Multiple R-squared: 0.1642, Adjusted R-squared: 0.1348   
## F-statistic: 5.598 on 2 and 57 DF, p-value: 0.00603

Sabendo que nossa Medida é a estatística que indica o quanto o modelo explica o comportamento variável resposta. pode ser calculada utilizando:

m <- plm(insatisfacao ~ robo + qtdCaracteres, data=dados, model = "pooling")

## at least one couple (id-time) has NA in at least one index dimensionin resulting pdata.frame  
## to find out which, use e.g.table(index(your\_pdataframe), useNA = "ifany")

val.rr<- r.squared(m)  
val.r <- sqrt(val.rr)  
val.rr

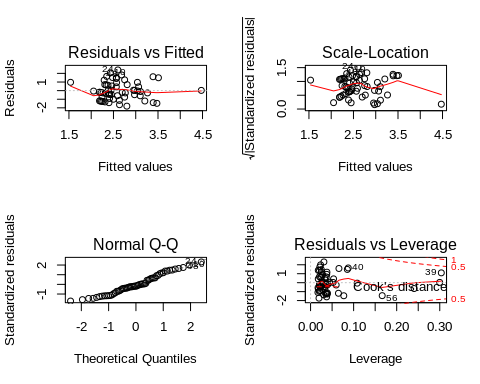
## [1] 0.1641709

E sendo assim chegamos a estatística Nesse sentido podemos afirmar por meio das estatísticas observadas que apenas cerca de % da variação total de , pode ser explicada pela reta de regressão proposta. Utilizando também a estatística de correlação linear, temos que como ela está muito baixa, há indícios de não existência de correlação entre as variáveis e o fator . Chegando a essa ideia, podemos concluir que nosso modelo talvez estivesse mais bem adequado com a adição de novas variáveis, ou que realmente não haja formas de correlação linear entre os pontos analisados, diversos pontos podem ser apontados, para chegarmos a esse resultado.

Além disso estatística F é e o p-valor é que é baixo o suficiente para rejeitar a hipótese nula de que . Corroborando mais uma vez a entender que o nosso modelo não é adequado para tentar explicar a variável de insatisfação.

###Análise de Resíduos

layout(matrix(c(1,2,3,4),2,2))  
plot(regMult)



res<- resid(regMult)  
tn <- shapiro.test(res)  
tn

##   
## Shapiro-Wilk normality test  
##   
## data: res  
## W = 0.96646, p-value = 0.09777

Verificando o modelo, os resíduos desse modelo nos mostra que pelo teste de normalidade de shapiro, foi obtido o resultado de 0.97727 com p-valor de 0.3242 que é maior que 0.05. Ou seja, não houve violação da normalidade dos resíduos. Portanto os erros possuem distribuição normal.

ncvTest(regMult)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.04211193, Df = 1, p = 0.83741

A estatística do Chi quadrado 1.270956, e o p-valor = 0.25 > 0.05 indicam que não há uma violação da homogeneidade dos resíduos.

Portanto, apesar dos resíduos indicarem um modelo com homogeneidade de resíduos e o teste de normalidade indicarem que os erros possuem uma distribuição normal, temos que o modelo não é tão bem representativo. As variáveis explicativas não explicam bem a variável resposta e um estudo complementar mais avançado é interessante de se fazer. Neste caso, seria o estudo se utilizando de modelos lineares generalizados com o intuito de fundamentar um modelo com os componentes aleatório, sistemático e de ligação para fazer um modelo mais completo e representativo dos dados. (Cordeiro and Demétrio 2008, seção 2.3)

Cordeiro, Gauss Moutinho, and Clarice GB Demétrio. 2008. “Modelos Lineares Generalizados E Extensões.” *Sao Paulo* 33.

Frost, Jim. 2019. “Using Histograms to Understand Your Data.” *Statistics by Jim*. <https://statisticsbyjim.com/basics/histograms/>.

Sillasgonzaga. 2018. “Sillasgonzaga/lexiconPT.” *GitHub*. <https://github.com/sillasgonzaga/lexiconPT>.

Souza, Marlo, Renata Vieira, Débora Busetti, Rove Chishman, and Isa Alves. 2AD. “Construction of a Portuguese Opinion Lexicon from Multiple Resources,” November, 59–66.

n.d. *lexiconPT*. <https://cran.r-project.org/web/packages/lexiconPT/README.html>.

n.d. *Department of Computer Science*. <https://www.cs.uic.edu/~lzhang3/programs/OpinionLexicon.html>.