**Projeto da Disciplina BI - NPL**

Prof. Leonardo Mendonza

[prof.leonardo@ica.ele.puc-rio.br](https://meet.google.com/linkredirect?authuser=0&dest=mailto:prof.leonardo%40ica.ele.puc-rio.br)

**Componentes do Grupo:**

|  |  |  |
| --- | --- | --- |
| Nome | Matrícula | e-mail |
| Renata Regina da Fonseca Santos | 192.110.195 | renatarfsantos@zipmail.com.br |
| Sérgio Gustavo Souza Guerra | 201.110.026 | sergiog.guerra@gmail.com |
| Flávia Szczerbacki | 192.110.229 | sflavia@gmail.com |
| Mariana Gonçalves Almeida Pinto | 192.110.208 | marianagoncalvesap@gmail.com |

## Introdução

Este projeto tem por finalidade analisar todas as reclamações fundamentadas no PROCON, que foram audiência no ano de 2017, em dezessete estados brasileiros. É desejável que ao final tenhamos *WordClouds* onde seja possível analisar:

* Empresas, Serviços e Problemas mais reclamados;
* Estados com maiores números de reclamações;
* Empresas mais reclamadas por estado.

Ao longo desse documento demonstraremos os principais métodos e modelos utilizados para construção das *WordClouds*.

## Análise da Base

A base inclui dezessete estados brasileiros e se caracteriza por ter 42.306 linhas de reclamações e oito colunas, que estão distribuídas da seguinte forma:

* Região
* Estado
* Empresa
* Subsidiária
* Área (setor/segmento)
* Serviço
* Problema
* Faixa etária (do autor da reclamação)

A base recebida, por ser de campo livre, carrega consigo as reclamações, nomes de empresa, serviço sem uma normalização prévia. Diante disso, a primeira preocupação que nos moveu foi a limpeza da base, para que a aparência final da *WordCloud* viesse clara e facilitasse a compreensão correta dos dados.

* Tratamento de dados nulos

Regiao 0

estado 0

empresa 1

subsidiaria 6530

area 5329

serviço 0

problema 8936

faixa etarea 0

dtype: int64

Remoção de nulos para a criação da *WordCloud*.

# Remoção de linhas com valores nulos

data = data.dropna(subset=['empresa’, ‘problema']).astype(str)

## *WordClouds* Empresa, Serviços e Problemas

Foram seguidas as mesmas etapas de pré-processamento para a construção das *WordClouds* Empresas, Serviços e Problemas.

* 1. Abordagem inicial utilizando Expressões Regulares para Pré-processamento

Conversão do texto para letras minúsculas e criação de 2 dicionários para limpeza do texto.

dicionario\_1 = {' s/a':'',' s.a':'',' ltda':'',' s/s':'',',':'',\

'\(':'','\)':'',"-":'','\.com':'','\.br':'', *r*"^\t":'',' epp':'',' lta':'','@':''}

dicionario\_2 = {'\_sa$':'','\.':'','/':' ',' ':'\_'}

Na construção da *WordCloud* Empresa, identificamos que uma mesma empresa aparece com diferentes grafias, resultando inicialmente numa nuvem com pesos não verdadeiros.

Exemplo de nomenclatura encontrada para uma mesma empresa:

empresa

785 telefônica\_brasil

788 telefônica\_brasil

1112 telefônica\_brasil

1173 telefônica\_brasil\_\_vivo

... ...

40548 telefônica\_brasil

41022 telefônica\_brasil

41432 telefônica\_brasil\_\_vivo

41990 telefônica\_brasil\_\_vivo

42093 telefônica\_brasil\_sa\_\_vivo

144 rows × 1 columns

* 1. Abordagem final incluindo mais etapas para Pré-processamento

Incluímos a utilização de expressões regulares para remoção de números, pontuação e acentos e NLTK para Tokenização e remoção de StopWords das colunas ‘empresa’, ‘serviço’ e ‘problema’.

Para a formação dos ngrams criamos uma lista com as listas de tokens formadas em cada linha das colunas ‘empresa’, ‘serviço’ e ‘problema’. Concatenamos os tokens em ngrams e criamos um dicionário com a frequência dos ngrams.

* 1. *WordClouds*
     1. *WordClouds* com as 40 empresas mais reclamadas

Podemos verificar que, sem o tratamento de acentos e a tokenização, a primeira *WordCloud* não reflete a realidade das reclamações, mostrando a empresa Claro S.A. como campeã. Ao incluirmos mais etapas de tratamento, verificamos que a empresa Telefónica Brasil tem mais reclamações.

[telefonica, brasil] 1158

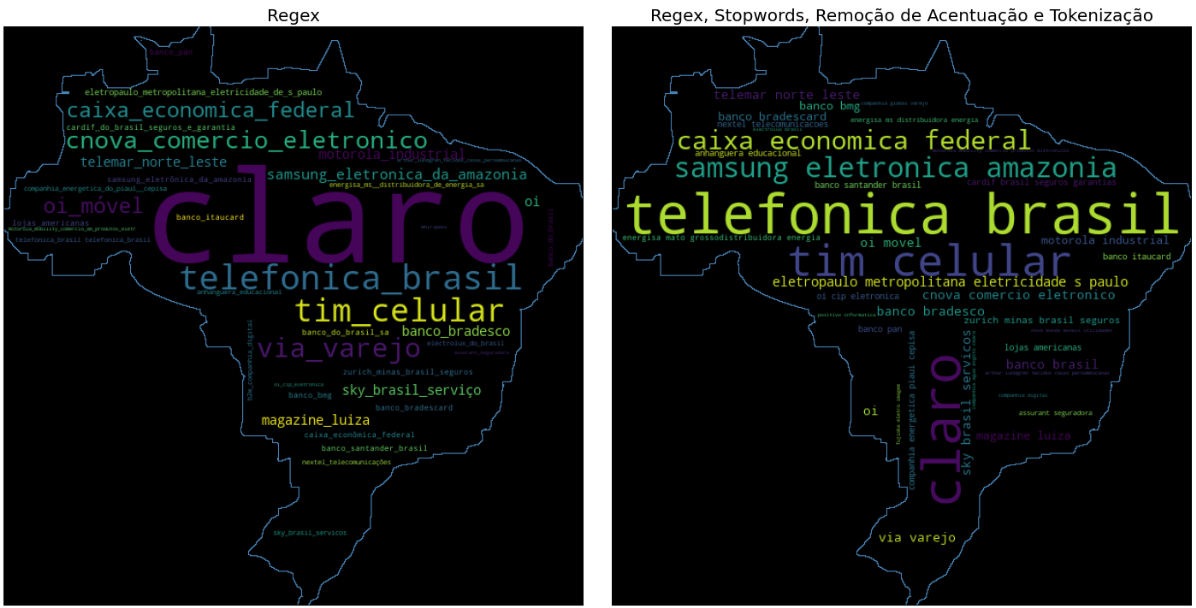
[claro] 1039

[tim, celular] 684

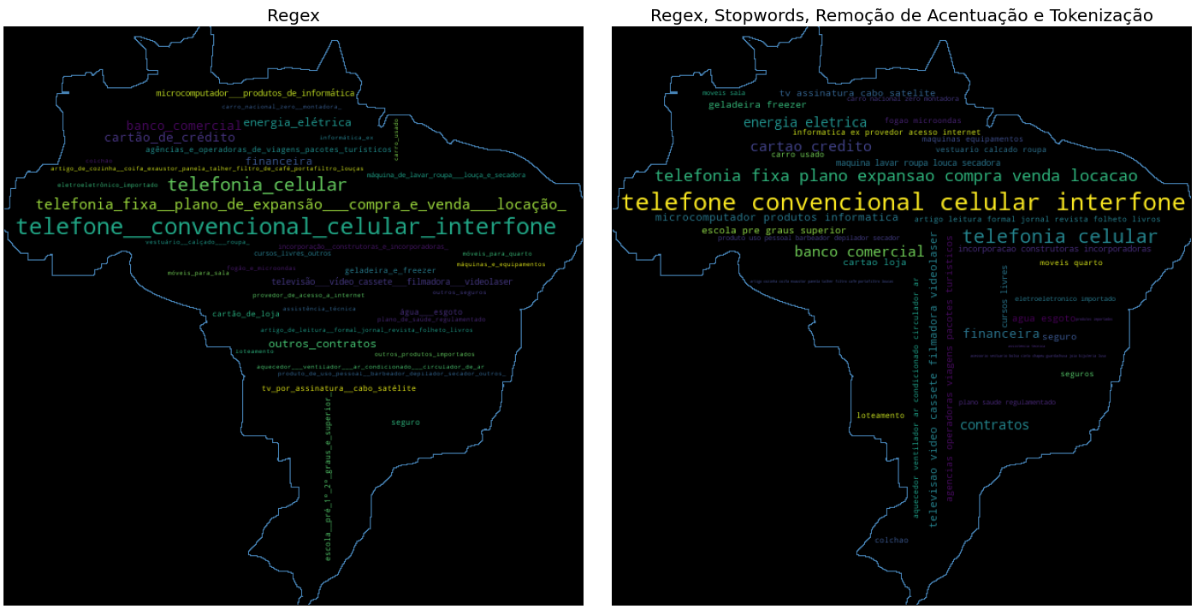
[samsung, eletronica, amazonia] 518

[caixa, economica, federal] 507

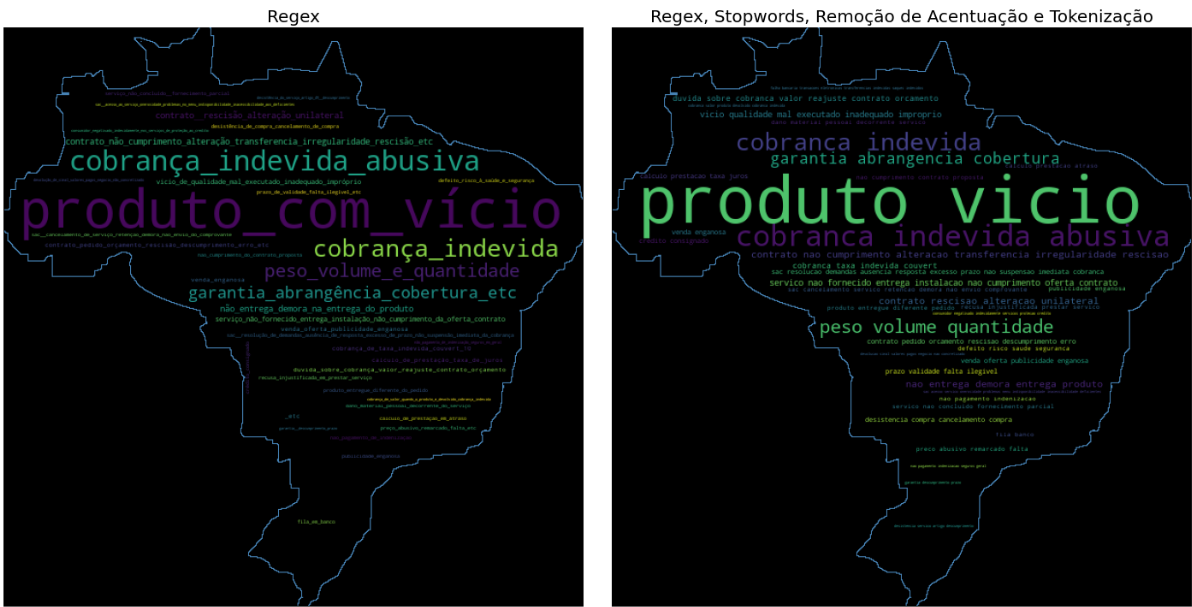
Name: empresa, Length: 7594, dtype: int64



* + 1. *WordClouds* com os 40 serviços mais reclamados



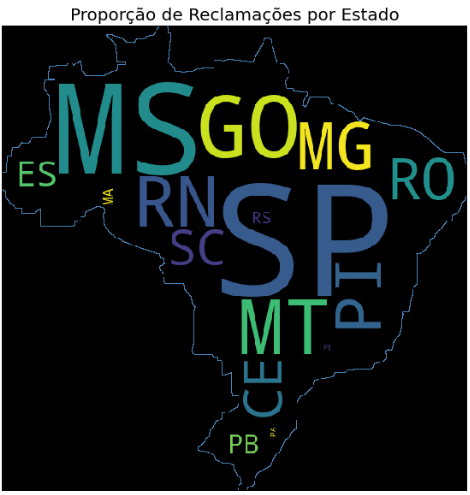
* + 1. *WordClouds* com os 40 Problemas mais reclamados



## *WordClouds* Estados e Empresas por Estado

* 1. *WordCloud* Estados

Para a criação da wordcloud Estados não foi preciso fazer nenhum pré-processamento. Criamos somente uma string única com todas as ocorrências dos estados.

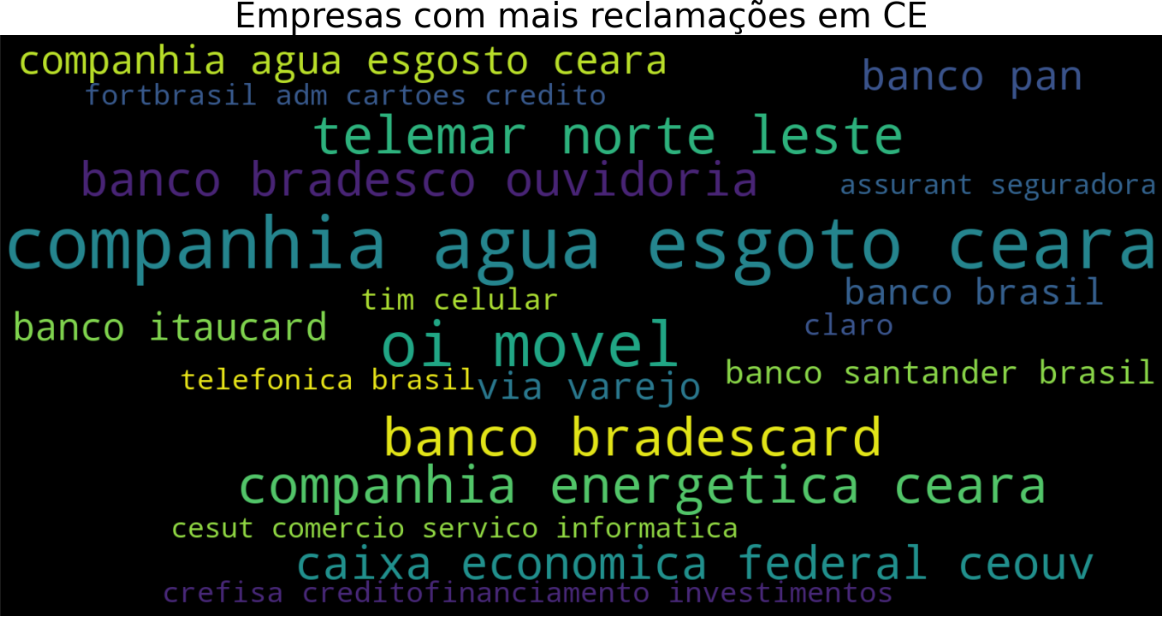


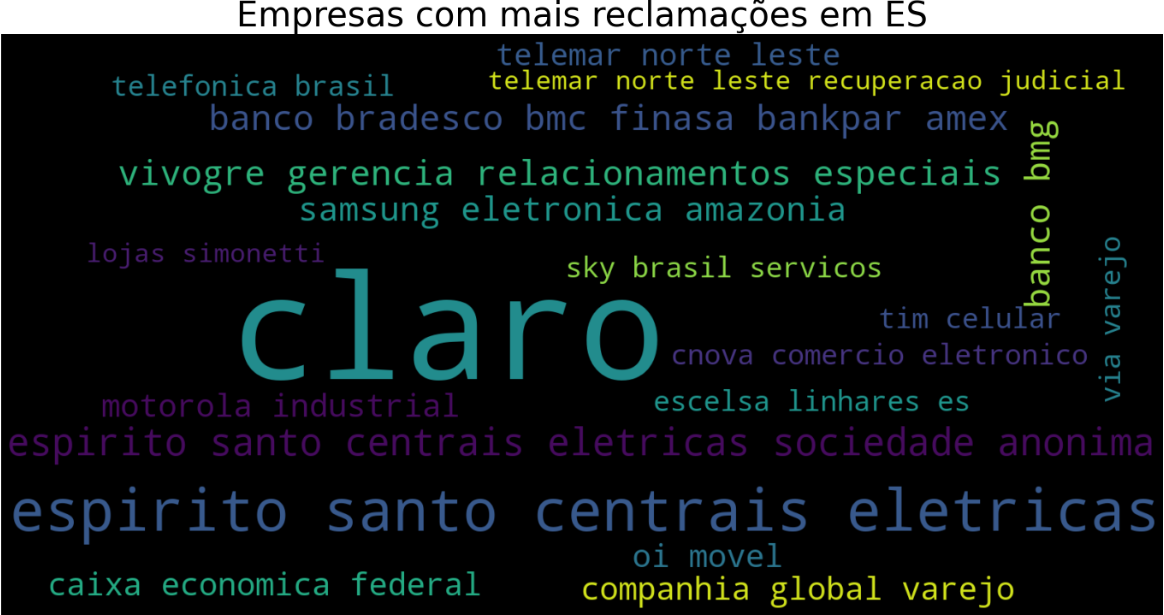
* 1. *WordCloud* Empresas por Estado

Para a *WordCloud* Empresas por Estado foram seguidas as mesmas etapas de pré-processamento utilizadas nas *WordClouds* Empresas, Serviços e Problemas.

Através da criação das *WordClouds* por estado podemos verificar que há um comportamento diferenciado no tipo de reclamação nas regiões do país. Nas regiões Sul e Sudeste constata-se claramente que os serviços de telecomunicações são mais reclamados. Já para o Norte do país existe um número maior para serviços básicos da economia como, por exemplo, água, esgoto e energia elétrica, serviços em regra prestados pelo Estado.

Abaixo seguem dois exemplos de *WordClouds* Empresas por Estado, as demais encontram-se no notebook “Procon Empresas Estados – WordCloud”.

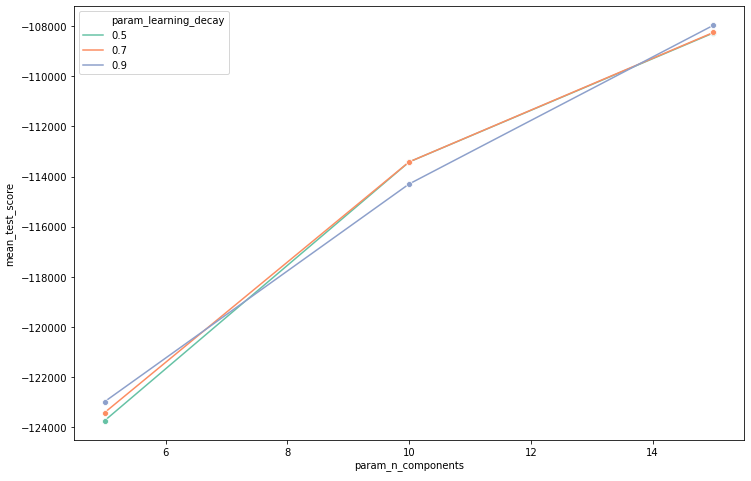




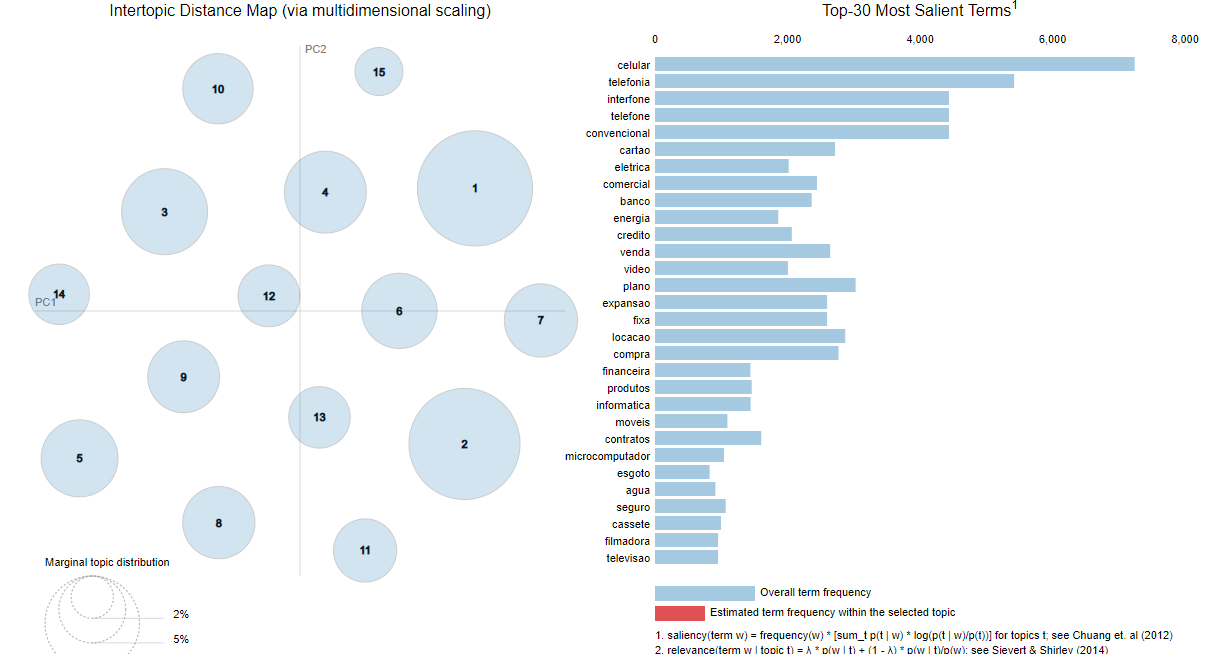
## Análise de Tópicos - Serviços

Criamos um script para analisar os tópicos existentes dentro dos serviços reclamados, considerando cada linha da coluna ‘serviço’ como um documento. A coluna ‘serviço’ foi escolhida, pois apresenta maior diversidade de informações. Seguimos as etapas abaixo:

* Pré-processamento do texto: tokenização e remoção de pontuação utilizando Gensim e remoção de StopWords utilizando NLTK;
* Criação de matriz Documento-Palavra utilizando Sklearn CountVectorizer;
* Utilização de Sklearn GridSearch para encontrar melhor modelo LDA com 'n\_components': [5, 10, 15] (número de tópicos), 'learning\_decay': [.5, .7, .9] (taxa de aprendizado);



* Treinamento do Modelo LDA com Sklearn utilizando os parâmetros escolhidos GridSearch - {'learning\_decay': 0.9, 'n\_components': 15}.



O resultado da análise de tópicos se mostrou inconclusiva, pois os dados fornecidos apresentavam sentenças muito curtas e o modelo agrupou palavras repetidas em vários tópicos, não nos permitindo identificá-los facilmente. A visualização dos tópicos encontra-se no arquivo “lda.html”.



Foram geradas *WordClouds* para cada tópico. Elas se encontram no notebook “Procon Serviços Total – LDA”.

## Conclusão

Durante o desenvolvimento das análises, aplicamos técnicas de pré-processamento, como expressões regulares, remoção *stopwords*, tokenização e Ngrams, com o objetivo de compreender e identificar de forma clara os serviços, problemas e empresas mais reclamados no Procon no ano de 2017.

Fizemos uma análise de empresas por estado e foi possível identificar que no Norte e no Nordeste se destacam mais os serviços de saneamento básico ofertados pelo estado, no Centro-Oeste há o destaque para o banco Caixa, em função de financiamento do Agronegócio, e no Sul e Sudeste observamos mais reclamações em Telecomunicações.

Desenvolvemos uma análise de tópicos para serviços, utilizando a Latent Dirichlet Allocation (**LDA**) que não se mostrou efetiva em consequência da natureza dos dados fornecidos.

Utilizando os conteúdos abordados na matéria, foi possível compreender na prática o que foi apresentado em aula. Com base nisso, concluímos que o trabalho nos possibilitou aplicar metodologias em um banco de dados real, nos mostrando como poderemos utilizar essa “caixa de ferramentas” em nossa vida profissional.

## Arquivos

Os arquivos em anexo apresentam o desenvolvimento do trabalho em Python:    