Trabalho_Final_NLP

February 13, 2020

[1]:
```
import nltk
#nltk.download()
```

[2]:
```
import os
workdir_path = r'./'
os.chdir(workdir_path)
```

[3]:
```
import pandas as pd

data = pd.read_csv('base_reclamacoes.csv', sep= ";", encoding = "ISO-8859-1")
data.head()
```

[3]:

	Regiao	estado	empresa	subsidiaria	area	serviço	problema	faixa etarea
0	Sudeste	SP	ITAÚ UNIBANCO S/A	BANCO ITAÚ/BANCO UNIBANCO	BANCOS MÚLTIPLOS, COM CARTEIRA COMERCIAL	Outros Contratos	Contrato (não cumprimento, alteração, transfer...	entre 31 a 40 anos
1	Sudeste	SP	CLARO S/A	NET	NaN	Telefonia Fixa (Plano de Expansão / Compra e ...	Contrato - Rescisão/alteração unilateral	mais de 70 anos
2	Sudeste	SP	ELETROPAULO METROPOLITANA ELETRICIDADE DE S PAULO	ELETROPAULO METROPOLITANA	DISTRIBUIÇÃO DE ENERGIA ELÉTRICA	Energia Elétrica	PID - Pedido de Indenização por Danos Morais	entre 31 a 40 anos
3	Sudeste	SP	GNN GARAGENS LTDA - EPP	NETPARK.COM.BR	ESTACIONAMENTO DE VEÍCULOS	Estacionamento (Particular, Supermercado, Sho...	Vicio de qualidade (mal executado, inadequado,...	entre 31 a 40 anos
4	Sudeste	SP	CLARO S/A	CLARO / NET / EMBRATEL / CLAROTV	TELEFONIA MÓVEL CELULAR	Telefonia Fixa (Plano de Expansão / Compra e ...		

```
4  Vicio de qualidade (mal executado, inadequado,...  entre 41 a 50 anos
```

1 Pré-processamento dos dados

[4]:
```
len(data)
```

[4]: 42307

[5]:
```
print(data.isnull().any())
```

```
Regiao           False
estado           False
empresa           True
subsidiaria       True
area              True
serviço          False
problema          True
faixa etarea     False
dtype: bool
```

[6]:
```
#removendo os nulls
data = data.dropna(how='any',axis=0)

print(data.isnull().any())
```

```
Regiao           False
estado           False
empresa          False
subsidiaria      False
area             False
serviço          False
problema         False
faixa etarea     False
dtype: bool
```

[7]:
```
#removendo todas as palavras com números e tornando as palavras minúsculas
pd.options.mode.chained_assignment = None
import re

lower_alpha = lambda x: re.sub(r"""\w*\d\w*""", ' ', x.lower())
data['problema'] = data.problema.map(lower_alpha)
data['empresa'] = data.empresa.map(lower_alpha)
data['estado'] = data.estado.map(lower_alpha)
data['serviço'] = data.serviço.map(lower_alpha)
```

```
data.head()
```

[7]:
```
   Regiao  estado                                          empresa  \
0  Sudeste     sp                                  itaú unibanco s/a
2  Sudeste     sp  eletropaulo metropolitana eletricidade de s paulo
3  Sudeste     sp                           gnn garagens ltda - epp
4  Sudeste     sp                                       claro s/a
5  Sudeste     sp             aerovias del continente americano s/a

                    subsidiaria                                        area   \
0         BANCO ITAÚ/BANCO UNIBANCO  BANCOS MÚLTIPLOS, COM CARTEIRA COMERCIAL
2         ELETROPAULO METROPOLITANA          DISTRIBUIÇÃO DE ENERGIA ELÉTRICA
3                   NETPARK.COM.BR                ESTACIONAMENTO DE VEÍCULOS
4  CLARO / NET / EMBRATEL / CLAROTV                   TELEFONIA MÓVEL CELULAR
5             AVIANCA INTERNACIONAL   TRANSPORTE AÉREO DE PASSAGEIROS REGULAR

                                        serviço  \
0                                outros contratos
2                                energia elétrica
3  estacionamento ( particular, supermercado, sho...
4  telefonia fixa ( plano de expansão / compra e ...
5  agências e operadoras de viagens (pacotes turí...

                                       problema        faixa etarea
0  contrato (não cumprimento, alteração, transfer...  entre 31 a 40 anos
2       pid - pedido de indenização por danos morais  entre 31 a 40 anos
3  vicio de qualidade (mal executado, inadequado,...  entre 31 a 40 anos
4  vicio de qualidade (mal executado, inadequado,...  entre 41 a 50 anos
5  desistência do serviço (artigo   - descumprime...  entre 21 a 30 anos
```

[8]:
```
#removendo toda a pontuação
import string

punc_re = lambda x: re.sub('[%s]' % re.escape(string.punctuation), ' ', x)
data['problema'] = data.problema.map(punc_re)
data['empresa'] = data.empresa.map(punc_re)
data['estado'] = data.estado.map(punc_re)
data['serviço'] = data.serviço.map(punc_re)



data.head()
```

[8]:
```
   Regiao  estado                                          empresa  \
0  Sudeste     sp                                  itaú unibanco s a
2  Sudeste     sp  eletropaulo metropolitana eletricidade de s paulo
3  Sudeste     sp                           gnn garagens ltda   epp
4  Sudeste     sp                                       claro s a
```

```
5  Sudeste     sp              aerovias del continente americano s a

                     subsidiaria                                          area   \
0         BANCO ITAÚ/BANCO UNIBANCO  BANCOS MÚLTIPLOS, COM CARTEIRA COMERCIAL
2         ELETROPAULO METROPOLITANA          DISTRIBUIÇÃO DE ENERGIA ELÉTRICA
3                   NETPARK.COM.BR                ESTACIONAMENTO DE VEÍCULOS
4  CLARO / NET / EMBRATEL / CLAROTV                   TELEFONIA MÓVEL CELULAR
5             AVIANCA INTERNACIONAL   TRANSPORTE AÉREO DE PASSAGEIROS REGULAR

                                        serviço  \
0                                outros contratos
2                                energia elétrica
3  estacionamento   particular  supermercado  sho...
4  telefonia fixa   plano de expansão   compra e ...
5  agências e operadoras de viagens  pacotes turí...

                                      problema        faixa etarea
0  contrato  não cumprimento  alteração  transfer...  entre 31 a 40 anos
2       pid   pedido de indenização por danos morais  entre 31 a 40 anos
3  vicio de qualidade  mal executado  inadequado ...  entre 31 a 40 anos
4  vicio de qualidade  mal executado  inadequado ...  entre 41 a 50 anos
5  desistência do serviço  artigo     descumprime...  entre 21 a 30 anos
```

[9]:
```
#removendo os "S/A" das empresas

remove_s_a = lambda x: re.sub(r"""s a""", ' ', x.lower())
remove_s = lambda x: re.sub(r""" s """, ' ', x.lower())
remove_sa = lambda x: re.sub(r""" sa """, ' ', x.lower())
data['empresa'] = data.empresa.map(remove_s_a)
data['empresa'] = data.empresa.map(remove_s)
data['empresa'] = data.empresa.map(remove_sa)

data.head()
```

[9]:
```
   Regiao  estado                                        empresa  \
0  Sudeste     sp                                  itaú unibanco
2  Sudeste     sp  eletropaulo metropolitana eletricidade de paulo
3  Sudeste     sp                          gnn garagens ltda   epp
4  Sudeste     sp                                        claro
5  Sudeste     sp              aerovias del continente americano

                     subsidiaria                                          area   \
0         BANCO ITAÚ/BANCO UNIBANCO  BANCOS MÚLTIPLOS, COM CARTEIRA COMERCIAL
2         ELETROPAULO METROPOLITANA          DISTRIBUIÇÃO DE ENERGIA ELÉTRICA
3                   NETPARK.COM.BR                ESTACIONAMENTO DE VEÍCULOS
4  CLARO / NET / EMBRATEL / CLAROTV                   TELEFONIA MÓVEL CELULAR
5             AVIANCA INTERNACIONAL   TRANSPORTE AÉREO DE PASSAGEIROS REGULAR
```

```
                                          serviço  \
0                                 outros contratos
2                                 energia elétrica
3  estacionamento   particular  supermercado  sho...
4  telefonia fixa   plano de expansão   compra e ...
5  agências e operadoras de viagens  pacotes turí...

                                          problema        faixa etarea
0  contrato  não cumprimento  alteração  transfer...  entre 31 a 40 anos
2       pid   pedido de indenização por danos morais  entre 31 a 40 anos
3  vicio de qualidade  mal executado  inadequado ...  entre 31 a 40 anos
4  vicio de qualidade  mal executado  inadequado ...  entre 41 a 50 anos
5  desistência do serviço  artigo     descumprime...  entre 21 a 30 anos
```

[10]:
```
#removendo os "ltda" das empresas

remove_ltda = lambda x: re.sub(r""" ltda """, ' ', x.lower())

data['empresa'] = data.empresa.map(remove_ltda)
```

[11]:
```
#removendo o "etc" dos problemas

remove_etc = lambda x: re.sub(r""" etc """, "", x.lower())
data['problema'] = data.problema.map(remove_etc)
data['serviço'] = data.serviço.map(remove_etc)
```

2 Tokenização

[12]:
```
import nltk
from nltk.tokenize import word_tokenize

data['tokens_problema'] = data.problema.map(word_tokenize)
data['tokens_empresa'] = data.empresa.map(word_tokenize)
data['tokens_estado'] = data.estado.map(word_tokenize)
data['tokens_serviço'] = data.serviço.map(word_tokenize)


data.head()
```

[12]:
```
   Regiao  estado                                      empresa  \
0  Sudeste     sp                                itaú unibanco
2  Sudeste     sp  eletropaulo metropolitana eletricidade de paulo
3  Sudeste     sp                            gnn garagens   epp
4  Sudeste     sp                                        claro
5  Sudeste     sp          aerovias del continente americano

                    subsidiaria                                      area   \
```

```
0         BANCO ITAÚ/BANCO UNIBANCO  BANCOS MÚLTIPLOS, COM CARTEIRA COMERCIAL
2         ELETROPAULO METROPOLITANA          DISTRIBUIÇÃO DE ENERGIA ELÉTRICA
3                   NETPARK.COM.BR                ESTACIONAMENTO DE VEÍCULOS
4  CLARO / NET / EMBRATEL / CLAROTV                   TELEFONIA MÓVEL CELULAR
5             AVIANCA INTERNACIONAL   TRANSPORTE AÉREO DE PASSAGEIROS REGULAR

                                          serviço  \
0                                outros contratos
2                                energia elétrica
3  estacionamento   particular  supermercado  sho...
4  telefonia fixa   plano de expansão   compra e ...
5  agências e operadoras de viagens  pacotes turí...

                                        problema        faixa etarea  \
0  contrato  não cumprimento  alteração  transfer...  entre 31 a 40 anos
2      pid   pedido de indenização por danos morais  entre 31 a 40 anos
3  vicio de qualidade  mal executado  inadequado ...  entre 31 a 40 anos
4  vicio de qualidade  mal executado  inadequado ...  entre 41 a 50 anos
5  desistência do serviço  artigo     descumprime...  entre 21 a 30 anos

                                    tokens_problema  \
0  [contrato, não, cumprimento, alteração, transf...
2  [pid, pedido, de, indenização, por, danos, mor...
3  [vicio, de, qualidade, mal, executado, inadequ...
4  [vicio, de, qualidade, mal, executado, inadequ...
5  [desistência, do, serviço, artigo, descumprime...

                                    tokens_empresa tokens_estado  \
0                                  [itaú, unibanco]          [sp]
2  [eletropaulo, metropolitana, eletricidade, de,...          [sp]
3                              [gnn, garagens, epp]          [sp]
4                                          [claro]          [sp]
5             [aerovias, del, continente, americano]          [sp]

                                    tokens_serviço
0                                [outros, contratos]
2                                [energia, elétrica]
3  [estacionamento, particular, supermercado, sho...
4  [telefonia, fixa, plano, de, expansão, compra,...
5  [agências, e, operadoras, de, viagens, pacotes...
```

[13]:
```
#lista combinando todos os valores de tokens
word_list_problema = sum(data.tokens_problema.tolist(), [])
word_list_empresa = sum(data.tokens_empresa.tolist(), [])
word_list_estado = sum(data.tokens_estado.tolist(), [])
word_list_serviço = sum(data.tokens_serviço.tolist(), [])
```

[14]:
```
#word_list_problema[:10]
#word_list_empresa[:10]
word_list_estado[:10]
#word_list_serviço[:10]
```

[14]:
```
['sp', 'sp', 'sp', 'sp', 'sp', 'sp', 'sp', 'sp', 'sp', 'sp']
```

3 Stopwords

[15]:
```
from nltk.corpus import stopwords

stop_words = stopwords.words('portuguese')

stop_lambda = lambda x: [y for y in x if y not in stop_words]
data['tokens_stop_problema'] = data.tokens_problema.apply(stop_lambda)
data['tokens_stop_empresa'] = data.tokens_empresa.apply(stop_lambda)
data['tokens_stop_estado'] = data.tokens_estado.apply(stop_lambda)
data['tokens_stop_serviço'] = data.tokens_serviço.apply(stop_lambda)

data.head()
```

[15]:

	Regiao	estado	empresa	subsidiaria	area	serviço	problema	faixa etarea	\
0	Sudeste	sp	itaú unibanco	BANCO ITAÚ/BANCO UNIBANCO	BANCOS MÚLTIPLOS, COM CARTEIRA COMERCIAL	outros contratos	contrato não cumprimento alteração transfer...	entre 31 a 40 anos	
2	Sudeste	sp	eletropaulo metropolitana eletricidade de paulo	ELETROPAULO METROPOLITANA	DISTRIBUIÇÃO DE ENERGIA ELÉTRICA	energia elétrica	pid pedido de indenização por danos morais	entre 31 a 40 anos	
3	Sudeste	sp	gnn garagens epp	NETPARK.COM.BR	ESTACIONAMENTO DE VEÍCULOS	estacionamento particular supermercado sho...	vicio de qualidade mal executado inadequado ...	entre 31 a 40 anos	
4	Sudeste	sp	claro	CLARO / NET / EMBRATEL / CLAROTV	TELEFONIA MÓVEL CELULAR	telefonia fixa plano de expansão compra e ...	vicio de qualidade mal executado inadequado ...	entre 41 a 50 anos	
5	Sudeste	sp	aerovias del continente americano	AVIANCA INTERNACIONAL	TRANSPORTE AÉREO DE PASSAGEIROS REGULAR	agências e operadoras de viagens pacotes turí...			

```
5  desistência do serviço  artigo     descumprime...  entre 21 a 30 anos

                                    tokens_problema  \
0  [contrato, não, cumprimento, alteração, transf...
2  [pid, pedido, de, indenização, por, danos, mor...
3  [vicio, de, qualidade, mal, executado, inadequ...
4  [vicio, de, qualidade, mal, executado, inadequ...
5  [desistência, do, serviço, artigo, descumprime...

                                     tokens_empresa tokens_estado  \
0                                    [itaú, unibanco]          [sp]
2  [eletropaulo, metropolitana, eletricidade, de,...          [sp]
3                                [gnn, garagens, epp]          [sp]
4                                             [claro]          [sp]
5             [aerovias, del, continente, americano]          [sp]

                                     tokens_serviço  \
0                                [outros, contratos]
2                                [energia, elétrica]
3  [estacionamento, particular, supermercado, sho...
4  [telefonia, fixa, plano, de, expansão, compra,...
5  [agências, e, operadoras, de, viagens, pacotes...

                                tokens_stop_problema  \
0  [contrato, cumprimento, alteração, transferenc...
2          [pid, pedido, indenização, danos, morais]
3  [vicio, qualidade, mal, executado, inadequado,...
4  [vicio, qualidade, mal, executado, inadequado,...
5     [desistência, serviço, artigo, descumprimento]

                                 tokens_stop_empresa tokens_stop_estado  \
0                                    [itaú, unibanco]               [sp]
2  [eletropaulo, metropolitana, eletricidade, paulo]               [sp]
3                                [gnn, garagens, epp]               [sp]
4                                             [claro]               [sp]
5             [aerovias, del, continente, americano]               [sp]

                                 tokens_stop_serviço
0                                [outros, contratos]
2                                [energia, elétrica]
3  [estacionamento, particular, supermercado, sho...
4  [telefonia, fixa, plano, expansão, compra, ven...
5  [agências, operadoras, viagens, pacotes, turís...
```

4 Stemming

[16]:
```
from nltk.stem import SnowballStemmer

stemmer = SnowballStemmer('portuguese')
stem_lambda = lambda x: [stemmer.stem(y) for y in x]

data['tokens_stem_problema'] = data.tokens_stop_problema.apply(stem_lambda)
data['tokens_stem_empresa'] = data.tokens_stop_empresa.apply(stem_lambda)
data['tokens_stem_serviço'] = data.tokens_stop_serviço.apply(stem_lambda)

data.head()
```

[16]:
```
   Regiao  estado                                          empresa  \
0  Sudeste     sp                                   itaú unibanco
2  Sudeste     sp  eletropaulo metropolitana eletricidade de paulo
3  Sudeste     sp                                gnn garagens   epp
4  Sudeste     sp                                           claro
5  Sudeste     sp          aerovias del continente americano

                     subsidiaria                                          area   \
0         BANCO ITAÚ/BANCO UNIBANCO  BANCOS MÚLTIPLOS, COM CARTEIRA COMERCIAL
2         ELETROPAULO METROPOLITANA          DISTRIBUIÇÃO DE ENERGIA ELÉTRICA
3                   NETPARK.COM.BR                ESTACIONAMENTO DE VEÍCULOS
4  CLARO / NET / EMBRATEL / CLAROTV                   TELEFONIA MÓVEL CELULAR
5             AVIANCA INTERNACIONAL   TRANSPORTE AÉREO DE PASSAGEIROS REGULAR

                                       serviço  \
0                              outros contratos
2                              energia elétrica
3  estacionamento   particular  supermercado  sho...
4  telefonia fixa   plano de expansão   compra e ...
5  agências e operadoras de viagens  pacotes turí...

                                      problema        faixa etarea  \
0  contrato  não cumprimento  alteração  transfer...  entre 31 a 40 anos
2      pid   pedido de indenização por danos morais  entre 31 a 40 anos
3  vicio de qualidade  mal executado  inadequado ...  entre 31 a 40 anos
4  vicio de qualidade  mal executado  inadequado ...  entre 41 a 50 anos
5  desistência do serviço  artigo     descumprime...  entre 21 a 30 anos

                                  tokens_problema  \
0  [contrato, não, cumprimento, alteração, transf...
2  [pid, pedido, de, indenização, por, danos, mor...
3  [vicio, de, qualidade, mal, executado, inadequ...
4  [vicio, de, qualidade, mal, executado, inadequ...
5  [desistência, do, serviço, artigo, descumprime...
```

```
                                    tokens_empresa tokens_estado  \
0                                   [itaú, unibanco]          [sp]
2  [eletropaulo, metropolitana, eletricidade, de,...          [sp]
3                                [gnn, garagens, epp]          [sp]
4                                             [claro]          [sp]
5             [aerovias, del, continente, americano]          [sp]

                                     tokens_serviço  \
0                                [outros, contratos]
2                                [energia, elétrica]
3  [estacionamento, particular, supermercado, sho...
4  [telefonia, fixa, plano, de, expansão, compra,...
5  [agências, e, operadoras, de, viagens, pacotes...

                                tokens_stop_problema  \
0  [contrato, cumprimento, alteração, transferenc...
2          [pid, pedido, indenização, danos, morais]
3  [vicio, qualidade, mal, executado, inadequado,...
4  [vicio, qualidade, mal, executado, inadequado,...
5     [desistência, serviço, artigo, descumprimento]

                                 tokens_stop_empresa tokens_stop_estado  \
0                                   [itaú, unibanco]               [sp]
2  [eletropaulo, metropolitana, eletricidade, paulo]               [sp]
3                                [gnn, garagens, epp]               [sp]
4                                             [claro]               [sp]
5             [aerovias, del, continente, americano]               [sp]

                                  tokens_stop_serviço  \
0                                [outros, contratos]
2                                [energia, elétrica]
3  [estacionamento, particular, supermercado, sho...
4  [telefonia, fixa, plano, expansão, compra, ven...
5  [agências, operadoras, viagens, pacotes, turís...

                                tokens_stem_problema  \
0  [contrat, cumpriment, alter, transferenc, irre...
2                     [pid, ped, indeniz, dan, mor]
3   [vici, qualidad, mal, execut, inadequ, impr0pri]
4   [vici, qualidad, mal, execut, inadequ, impr0pri]
5               [desistent, servic, artig, descumpr]

                      tokens_stem_empresa  \
0                          [itaú, unibanc]
2  [eletropaul, metropolitan, eletr, paul]
3                     [gnn, garagens, epp]
```

```
4                                  [clar]
5        [aerov, del, continent, american]

                                tokens_stem_serviço
0                                   [outr, contrat]
2                                    [energ, elétr]
3           [estacion, particul, supermerc, shopping]
4  [telefon, fix, plan, expansã, compr, vend, loc...
5              [agênc, oper, viagens, pacot, turíst]
```

5 Part-of-speech tagging

[17]:
```
from nltk.tag import pos_tag

pos_lambda = lambda x: nltk.pos_tag(x)
data['tokens_pos_problema'] = (data.tokens_stop_problema.apply(pos_lambda))
data['tokens_pos_empresa'] = (data.tokens_stop_empresa.apply(pos_lambda))
data['tokens_pos_serviço'] = (data.tokens_stop_serviço.apply(pos_lambda))

data.head()
```

[17]:
```
   Regiao  estado                                        empresa  \
0  Sudeste     sp                                 itaú unibanco
2  Sudeste     sp  eletropaulo metropolitana eletricidade de paulo
3  Sudeste     sp                              gnn garagens   epp
4  Sudeste     sp                                          claro
5  Sudeste     sp          aerovias del continente americano

                     subsidiaria                                          area   \
0        BANCO ITAÚ/BANCO UNIBANCO  BANCOS MÚLTIPLOS, COM CARTEIRA COMERCIAL
2        ELETROPAULO METROPOLITANA          DISTRIBUIÇÃO DE ENERGIA ELÉTRICA
3                  NETPARK.COM.BR                ESTACIONAMENTO DE VEÍCULOS
4  CLARO / NET / EMBRATEL / CLAROTV                   TELEFONIA MÓVEL CELULAR
5           AVIANCA INTERNACIONAL   TRANSPORTE AÉREO DE PASSAGEIROS REGULAR

                                      serviço  \
0                             outros contratos
2                             energia elétrica
3  estacionamento   particular  supermercado  sho...
4  telefonia fixa   plano de expansão   compra e ...
5  agências e operadoras de viagens  pacotes turí...

                                    problema        faixa etarea  \
0  contrato  não cumprimento  alteração  transfer...  entre 31 a 40 anos
2      pid   pedido de indenização por danos morais  entre 31 a 40 anos
3  vicio de qualidade  mal executado  inadequado ...  entre 31 a 40 anos
```

```
4  vicio de qualidade  mal executado  inadequado ...  entre 41 a 50 anos
5  desistência do serviço  artigo     descumprime...  entre 21 a 30 anos

                                    tokens_problema  \
0  [contrato, não, cumprimento, alteração, transf...
2  [pid, pedido, de, indenização, por, danos, mor...
3  [vicio, de, qualidade, mal, executado, inadequ...
4  [vicio, de, qualidade, mal, executado, inadequ...
5  [desistência, do, serviço, artigo, descumprime...

                                     tokens_empresa  ...  \
0                                   [itaú, unibanco]  ...
2  [eletropaulo, metropolitana, eletricidade, de,...  ...
3                               [gnn, garagens, epp]  ...
4                                            [claro]  ...
5              [aerovias, del, continente, americano]  ...

                                tokens_stop_problema  \
0  [contrato, cumprimento, alteração, transferenc...
2          [pid, pedido, indenização, danos, morais]
3  [vicio, qualidade, mal, executado, inadequado,...
4  [vicio, qualidade, mal, executado, inadequado,...
5     [desistência, serviço, artigo, descumprimento]

                                 tokens_stop_empresa tokens_stop_estado  \
0                                   [itaú, unibanco]                [sp]
2  [eletropaulo, metropolitana, eletricidade, paulo]                [sp]
3                               [gnn, garagens, epp]                [sp]
4                                            [claro]                [sp]
5              [aerovias, del, continente, americano]                [sp]

                                 tokens_stop_serviço  \
0                                 [outros, contratos]
2                                 [energia, elétrica]
3  [estacionamento, particular, supermercado, sho...
4  [telefonia, fixa, plano, expansão, compra, ven...
5  [agências, operadoras, viagens, pacotes, turís...

                                tokens_stem_problema  \
0  [contrat, cumpriment, alter, transferenc, irre...
2                      [pid, ped, indeniz, dan, mor]
3   [vici, qualidad, mal, execut, inadequ, imprópri]
4   [vici, qualidad, mal, execut, inadequ, imprópri]
5               [desistent, servic, artig, descumpr]

                       tokens_stem_empresa  \
0                          [itaú, unibanc]
```

```
2  [eletropaul, metropolitan, eletr, paul]
3                     [gnn, garagens, epp]
4                                  [clar]
5       [aerov, del, continent, american]

                                tokens_stem_serviço  \
0                                   [outr, contrat]
2                                    [energ, elétr]
3         [estacion, particul, supermerc, shopping]
4  [telefon, fix, plan, expansã, compr, vend, loc...
5             [agênc, oper, viagens, pacot, turíst]

                                tokens_pos_problema  \
0  [(contrato, NN), (cumprimento, NN), (alteração...
2  [(pid, JJ), (pedido, NN), (indenização, NN), (...
3  [(vicio, NN), (qualidade, NN), (mal, JJ), (exe...
4  [(vicio, NN), (qualidade, NN), (mal, JJ), (exe...
5  [(desistência, NN), (serviço, NN), (artigo, NN...

                                 tokens_pos_empresa  \
0                      [(itaú, NN), (unibanco, NN)]
2  [(eletropaulo, NN), (metropolitana, NNS), (ele...
3          [(gnn, NN), (garagens, NNS), (epp, VBP)]
4                                   [(claro, NN)]
5  [(aerovias, JJ), (del, NN), (continente, NN), ...

                                 tokens_pos_serviço
0                 [(outros, NNS), (contratos, NNS)]
2                   [(energia, NN), (elétrica, NN)]
3  [(estacionamento, IN), (particular, JJ), (supe...
4  [(telefonia, NN), (fixa, NN), (plano, NN), (ex...
5  [(agências, JJ), (operadoras, NNS), (viagens, ...

[5 rows x 22 columns]
```

6 Lemmatization

[18]:
```python
#não ficou bom com stemming, logo, vamos usar lemmatization
from nltk.corpus import wordnet
from nltk.stem.wordnet import WordNetLemmatizer


def get_wordnet_pos(treebank_tag):
    if treebank_tag.startswith('J'):
        return wordnet.ADJ
    elif treebank_tag.startswith('V'):
```

```
        return wordnet.VERB
    elif treebank_tag.startswith('N'):
        return wordnet.NOUN
    elif treebank_tag.startswith('R'):
        return wordnet.ADV
    else:
        return wordnet.NOUN

lemmatizer = WordNetLemmatizer()
lemmatizer_fun = lambda x: lemmatizer.lemmatize(*x)
data['tokens_lemma_problema'] = data.tokens_pos_problema\
                        .apply(lambda x: [(y[0], get_wordnet_pos(y[1])) for y␣
 ↪in x])\
                        .apply(lambda x: [lemmatizer_fun(y) for y in x])
data['tokens_lemma_empresa'] = data.tokens_pos_empresa\
                        .apply(lambda x: [(y[0], get_wordnet_pos(y[1])) for y␣
 ↪in x])\
                        .apply(lambda x: [lemmatizer_fun(y) for y in x])
data['tokens_lemma_serviço'] = data.tokens_pos_serviço\
                        .apply(lambda x: [(y[0], get_wordnet_pos(y[1])) for y␣
 ↪in x])\
                        .apply(lambda x: [lemmatizer_fun(y) for y in x])


data.head().T
```

[18]:

	0
Regiao	Sudeste
estado	sp
empresa	itaú unibanco
subsidiaria	BANCO ITAÚ/BANCO UNIBANCO
area	BANCOS MÚLTIPLOS, COM CARTEIRA COMERCIAL
serviço	outros contratos
problema	contrato não cumprimento alteração transfer...
faixa etarea	entre 31 a 40 anos
tokens_problema	[contrato, não, cumprimento, alteração, transf...
tokens_empresa	[itaú, unibanco]
tokens_estado	[sp]
tokens_serviço	[outros, contratos]
tokens_stop_problema	[contrato, cumprimento, alteração, transferenc...
tokens_stop_empresa	[itaú, unibanco]
tokens_stop_estado	[sp]
tokens_stop_serviço	[outros, contratos]
tokens_stem_problema	[contrat, cumpriment, alter, transferenc, irre...
tokens_stem_empresa	[itaú, unibanc]
tokens_stem_serviço	[outr, contrat]
tokens_pos_problema	[(contrato, NN), (cumprimento, NN), (alteração...

```
tokens_pos_empresa                          [(itaú, NN), (unibanco, NN)]
tokens_pos_serviço                    [(outros, NNS), (contratos, NNS)]
tokens_lemma_problema  [contrato, cumprimento, alteração, transferenc...
tokens_lemma_empresa                                  [itaú, unibanco]
tokens_lemma_serviço                               [outros, contratos]

                                                                    2  \
Regiao                                                        Sudeste
estado                                                             sp
empresa                 eletropaulo metropolitana eletricidade de paulo
subsidiaria                                 ELETROPAULO METROPOLITANA
area                                DISTRIBUIÇÃO DE ENERGIA ELÉTRICA
serviço                                             energia elétrica
problema                    pid   pedido de indenização por danos morais
faixa etarea                                        entre 31 a 40 anos
tokens_problema       [pid, pedido, de, indenização, por, danos, mor...
tokens_empresa        [eletropaulo, metropolitana, eletricidade, de,...
tokens_estado                                                    [sp]
tokens_serviço                                   [energia, elétrica]
tokens_stop_problema          [pid, pedido, indenização, danos, morais]
tokens_stop_empresa   [eletropaulo, metropolitana, eletricidade, paulo]
tokens_stop_estado                                               [sp]
tokens_stop_serviço                              [energia, elétrica]
tokens_stem_problema                      [pid, ped, indeniz, dan, mor]
tokens_stem_empresa             [eletropaul, metropolitan, eletr, paul]
tokens_stem_serviço                                   [energ, elétr]
tokens_pos_problema   [(pid, JJ), (pedido, NN), (indenização, NN), (...
tokens_pos_empresa    [(eletropaulo, NN), (metropolitana, NNS), (ele...
tokens_pos_serviço                      [(energia, NN), (elétrica, NN)]
tokens_lemma_problema         [pid, pedido, indenização, danos, morais]
tokens_lemma_empresa  [eletropaulo, metropolitana, eletricidade, paulo]
tokens_lemma_serviço                             [energia, elétrica]

                                                                    3  \
Regiao                                                        Sudeste
estado                                                             sp
empresa                                           gnn garagens   epp
subsidiaria                                           NETPARK.COM.BR
area                                   ESTACIONAMENTO DE VEÍCULOS
serviço               estacionamento   particular  supermercado  sho...
problema              vicio de qualidade  mal executado  inadequado ...
faixa etarea                                        entre 31 a 40 anos
tokens_problema       [vicio, de, qualidade, mal, executado, inadequ...
tokens_empresa                                   [gnn, garagens, epp]
tokens_estado                                                    [sp]
tokens_serviço        [estacionamento, particular, supermercado, sho...
tokens_stop_problema  [vicio, qualidade, mal, executado, inadequado,...
```

```
tokens_stop_empresa                                [gnn, garagens, epp]
tokens_stop_estado                                                 [sp]
tokens_stop_serviço    [estacionamento, particular, supermercado, sho...
tokens_stem_problema    [vici, qualidad, mal, execut, inadequ, imprópri]
tokens_stem_empresa                                [gnn, garagens, epp]
tokens_stem_serviço            [estacion, particul, supermerc, shopping]
tokens_pos_problema    [(vicio, NN), (qualidade, NN), (mal, JJ), (exe...
tokens_pos_empresa              [(gnn, NN), (garagens, NNS), (epp, VBP)]
tokens_pos_serviço     [(estacionamento, IN), (particular, JJ), (supe...
tokens_lemma_problema  [vicio, qualidade, mal, executado, inadequado,...
tokens_lemma_empresa               [gnn, garagens, epp]
tokens_lemma_serviço   [estacionamento, particular, supermercado, sho...

                                                                    4  \
Regiao                                                        Sudeste
estado                                                             sp
empresa                                                       claro
subsidiaria                              CLARO / NET / EMBRATEL / CLAROTV
area                                             TELEFONIA MÓVEL CELULAR
serviço                telefonia fixa   plano de expansão   compra e ...
problema               vicio de qualidade  mal executado  inadequado ...
faixa etarea                                       entre 41 a 50 anos
tokens_problema        [vicio, de, qualidade, mal, executado, inadequ...
tokens_empresa                                                [claro]
tokens_estado                                                   [sp]
tokens_serviço         [telefonia, fixa, plano, de, expansão, compra,...
tokens_stop_problema   [vicio, qualidade, mal, executado, inadequado,...
tokens_stop_empresa                                           [claro]
tokens_stop_estado                                              [sp]
tokens_stop_serviço    [telefonia, fixa, plano, expansão, compra, ven...
tokens_stem_problema    [vici, qualidad, mal, execut, inadequ, imprópri]
tokens_stem_empresa                                            [clar]
tokens_stem_serviço    [telefon, fix, plan, expansã, compr, vend, loc...
tokens_pos_problema    [(vicio, NN), (qualidade, NN), (mal, JJ), (exe...
tokens_pos_empresa                                      [(claro, NN)]
tokens_pos_serviço     [(telefonia, NN), (fixa, NN), (plano, NN), (ex...
tokens_lemma_problema  [vicio, qualidade, mal, executado, inadequado,...
tokens_lemma_empresa                                          [claro]
tokens_lemma_serviço   [telefonia, fixa, plano, expansão, compra, ven...

                                                                    5
Regiao                                                        Sudeste
estado                                                             sp
empresa                             aerovias del continente americano
subsidiaria                                        AVIANCA INTERNACIONAL
area                             TRANSPORTE AÉREO DE PASSAGEIROS REGULAR
serviço                agências e operadoras de viagens  pacotes turí...
```

```
problema               desistência do serviço  artigo     descumprime...
faixa etarea                                                entre 21 a 30 anos
tokens_problema        [desistência, do, serviço, artigo, descumprime...
tokens_empresa                   [aerovias, del, continente, americano]
tokens_estado                                                       [sp]
tokens_serviço         [agências, e, operadoras, de, viagens, pacotes...
tokens_stop_problema      [desistência, serviço, artigo, descumprimento]
tokens_stop_empresa              [aerovias, del, continente, americano]
tokens_stop_estado                                                  [sp]
tokens_stop_serviço    [agências, operadoras, viagens, pacotes, turís...
tokens_stem_problema                [desistent, servic, artig, descumpr]
tokens_stem_empresa                    [aerov, del, continent, american]
tokens_stem_serviço                 [agênc, oper, viagens, pacot, turíst]
tokens_pos_problema    [(desistência, NN), (serviço, NN), (artigo, NN...
tokens_pos_empresa     [(aerovias, JJ), (del, NN), (continente, NN), ...
tokens_pos_serviço     [(agências, JJ), (operadoras, NNS), (viagens, ...
tokens_lemma_problema     [desistência, serviço, artigo, descumprimento]
tokens_lemma_empresa             [aerovias, del, continente, americano]
tokens_lemma_serviço   [agências, operadoras, viagens, pacotes, turís...
```

[19]:
```
#encontrando as palavras mais comuns
word_list_clean_problema = sum(data.tokens_lemma_problema.tolist(), [])
word_list_clean_empresa = sum(data.tokens_lemma_empresa.tolist(), [])
word_list_clean_estado = sum(data.tokens_stop_estado.tolist(), [])
word_list_clean_serviço = sum(data.tokens_lemma_serviço.tolist(), [])
```

[20]:
```
#word_list_clean_problema[:10]
word_list_clean_empresa[:10]
#word_list_clean_estado[:10]
#word_list_clean_serviço[:10]
```

[20]:
```
['itaú',
 'unibanco',
 'eletropaulo',
 'metropolitana',
 'eletricidade',
 'paulo',
 'gnn',
 'garagens',
 'epp',
 'claro']
```

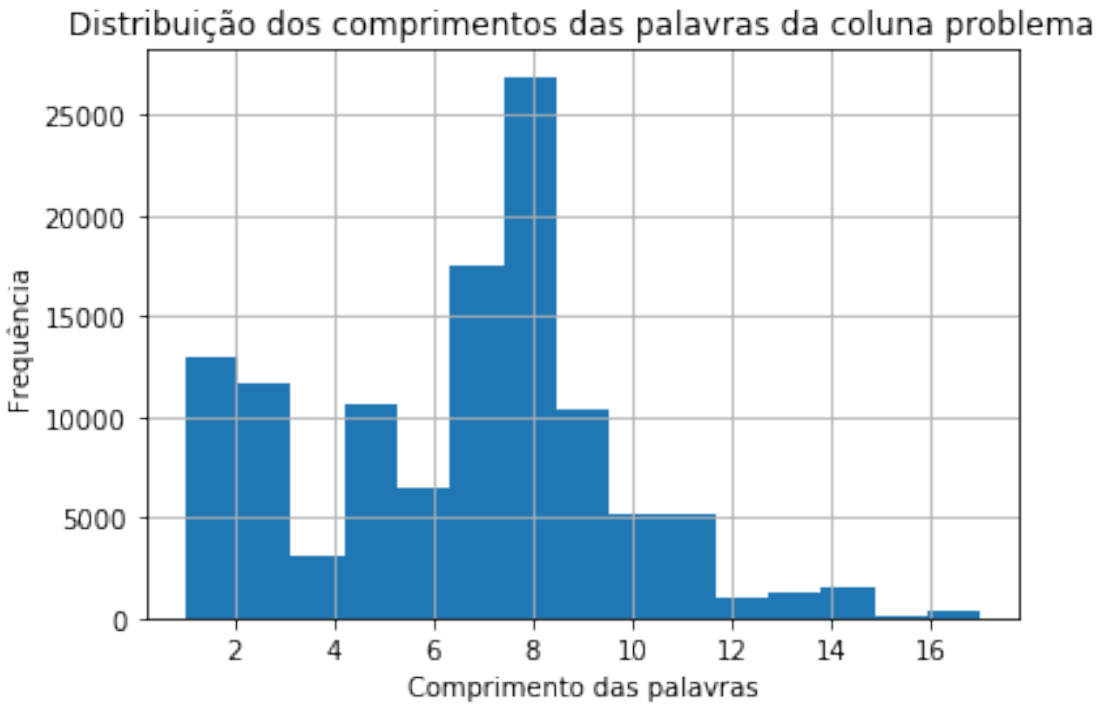
[21]:
```
from collections import Counter

#convertendo a lista em um dicionário com contagem de valores
word_counts_clean_problema = Counter(word_list_clean_problema)
a = word_counts_clean_problema
word_counts_clean_empresa = Counter(word_list_clean_empresa)
b = word_counts_clean_empresa
```

```
word_counts_clean_estado = Counter(word_list_clean_estado)
c = word_counts_clean_estado
word_counts_clean_serviço = Counter(word_list_clean_serviço)
d = word_counts_clean_serviço

#invertendo a chave / valores no dicionário para classificar
word_counts_clean_problema = list(zip(word_counts_clean_problema.values(),␣
 ↪word_counts_clean_problema.keys()))
word_counts_clean_empresa = list(zip(word_counts_clean_empresa.values(),␣
 ↪word_counts_clean_empresa.keys()))
word_counts_clean_estado = list(zip(word_counts_clean_estado.values(),␣
 ↪word_counts_clean_estado.keys()))
word_counts_clean_serviço = list(zip(word_counts_clean_serviço.values(),␣
 ↪word_counts_clean_serviço.keys()))


#classificando a lista por contagem
word_counts_clean_problema = sorted(word_counts_clean_problema, reverse=True)
word_counts_clean_empresa = sorted(word_counts_clean_empresa, reverse=True)
word_counts_clean_estado = sorted(word_counts_clean_estado, reverse=True)
word_counts_clean_serviço = sorted(word_counts_clean_serviço, reverse=True)
```

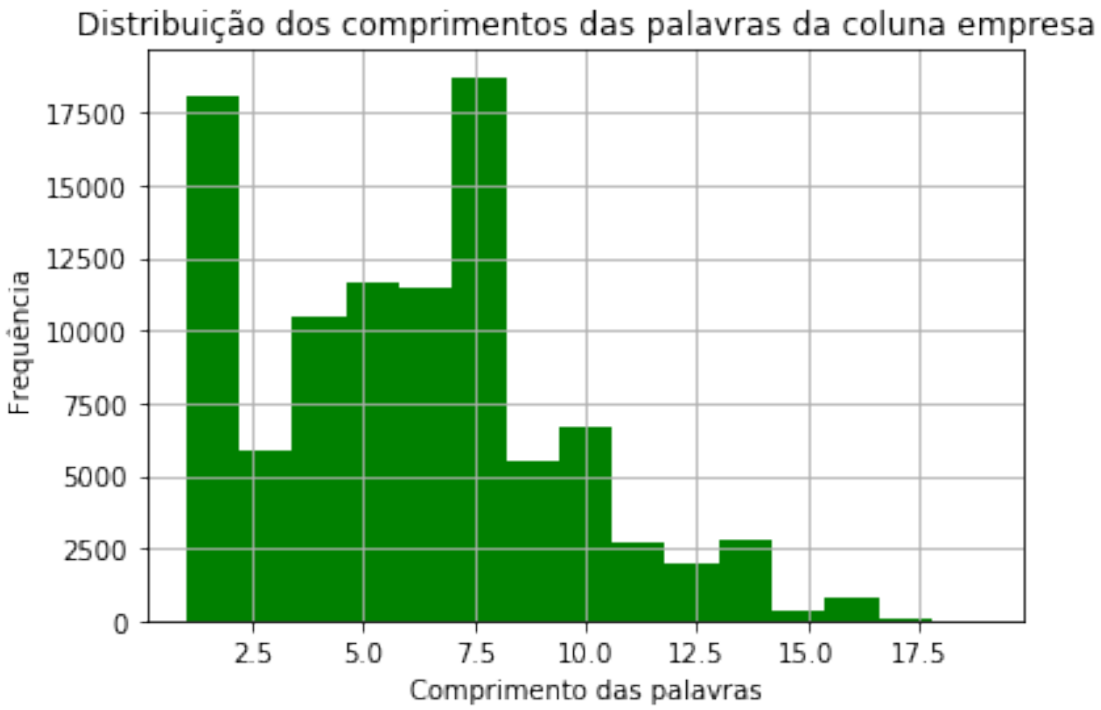
[22]:
```
#imprimindo as 10 palavras mais comuns
#word_counts_clean_problema[:10]
word_counts_clean_empresa[:10]
#word_counts_clean_estado[:10]
#word_counts_clean_serviço[:10]
```

[22]:
```
[(6315, 'ltda'),
 (4177, 'brasil'),
 (2666, 'banco'),
 (1375, 'comercio'),
 (1126, 'telefonica'),
 (1126, 'oi'),
 (1103, 'sa'),
 (922, 'companhia'),
 (794, 'claro'),
 (698, 'celular')]
```

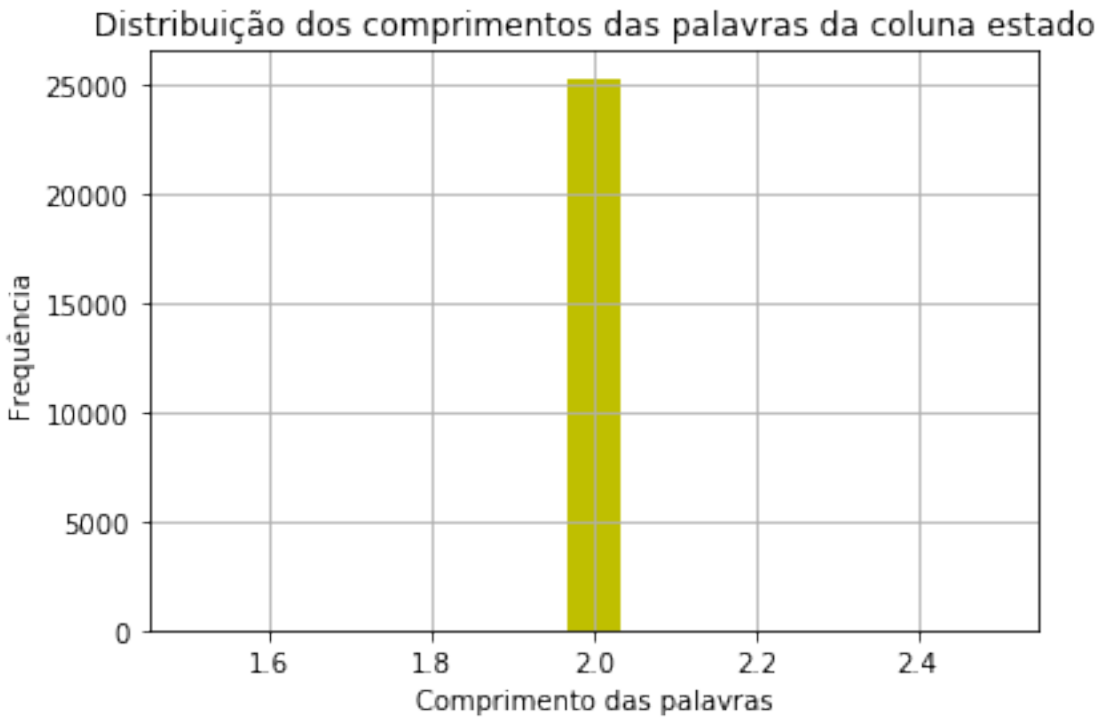
[23]:
```
#distribuição dos comprimentos das palavras

word_lengths_problema = pd.Series([len(x) for x in word_list_problema])

ax = word_lengths_problema.hist(bins=15)
ax.set(xlabel='Comprimento das palavras', ylabel='Frequência',␣
 ↪title='Distribuição dos comprimentos das palavras da coluna problema');
```



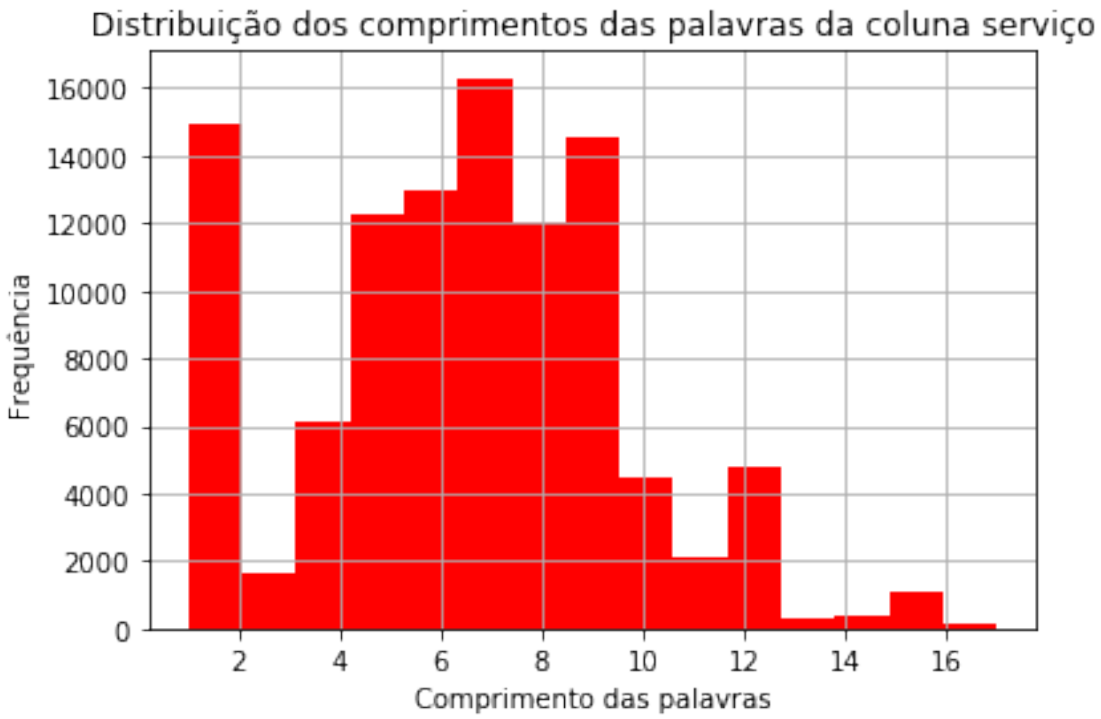
[24]:
```
word_lengths_empresa = pd.Series([len(x) for x in word_list_empresa])

bx = word_lengths_empresa.hist(bins=15, color='g')
bx.set(xlabel='Comprimento das palavras', ylabel='Frequência',␣
 ↪title='Distribuição dos comprimentos das palavras da coluna empresa');
```



[25]:
```
word_lengths_estado = pd.Series([len(x) for x in word_list_estado])

cx = word_lengths_estado.hist(bins=15, color='y')
cx.set(xlabel='Comprimento das palavras', ylabel='Frequência',␣
 ↪title='Distribuição dos comprimentos das palavras da coluna estado');
```

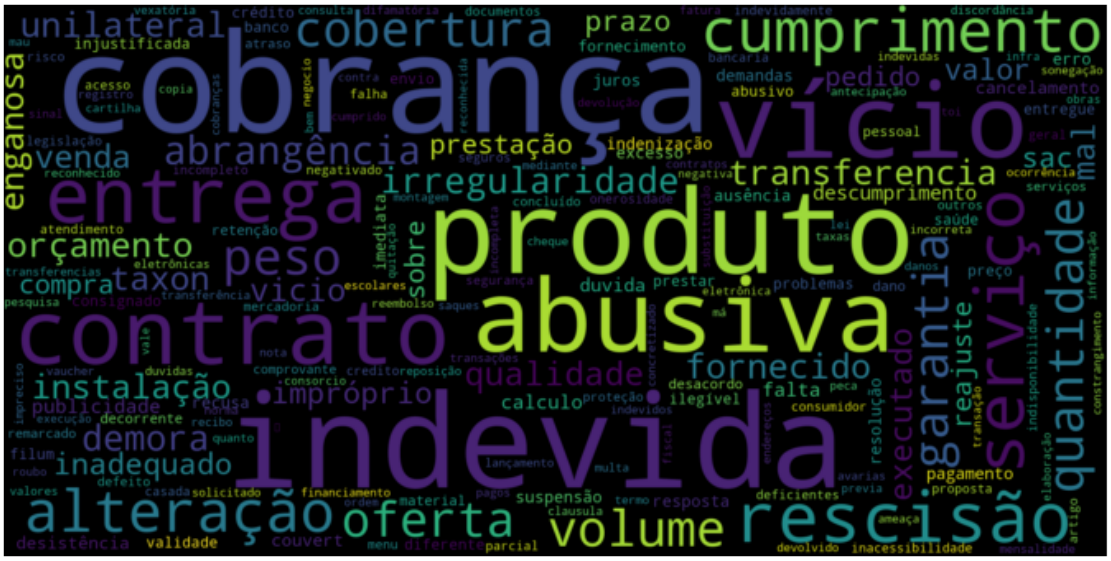


[26]:
```
word_lengths_serviço = pd.Series([len(x) for x in word_list_serviço])

dx = word_lengths_serviço.hist(bins=15, color='r')
dx.set(xlabel='Comprimento das palavras', ylabel='Frequência',␣
 ↪title='Distribuição dos comprimentos das palavras da coluna serviço');
```



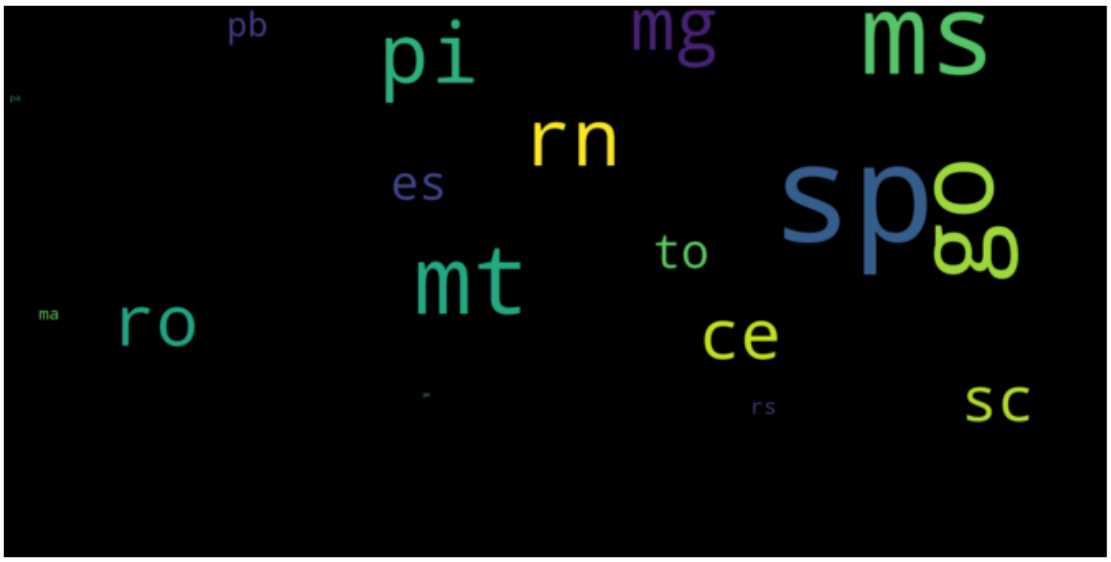
7 Wordclouds

[27]:
```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

text_problema = word_list_clean_problema
wordcloud_problema = WordCloud(width=1600, height=800, max_font_size=200).
 ↪fit_words(a)

plt.figure(figsize=(12,10))
plt.imshow(wordcloud_problema, interpolation='bilinear')
plt.axis("off")
plt.show()
```



[28]:
```
text_empresa = word_list_clean_empresa
wordcloud_empresa = WordCloud(width=1600, height=800, max_font_size=200).
 ↪fit_words(b)

plt.figure(figsize=(12,10))
plt.imshow(wordcloud_empresa, interpolation='bilinear')
plt.axis("off")
plt.show()
```

[29]:
```
text_estado = word_list_clean_estado
wordcloud_estado = WordCloud(width=1600, height=800, max_font_size=200).
 ↪fit_words(c)

plt.figure(figsize=(12,10))
plt.imshow(wordcloud_estado, interpolation='bilinear')
plt.axis("off")
plt.show()
```



[30]:
```
text_serviço = word_list_clean_serviço
wordcloud_serviço = WordCloud(width=1600, height=800, max_font_size=200).
 ↪fit_words(d)

plt.figure(figsize=(12,10))
plt.imshow(wordcloud_serviço, interpolation='bilinear')
plt.axis("off")
plt.show()
```


celular
telefone
interfone
convencional
telefonia
compra
plano
locação
venda
cartão
expansão
fixa
outros
banco
elétrica
comercial
crédito
energia
contratos
informática
móveis
laser
televisão
água
loja
cabo
seguro
agências
viagens
livres
cassete
vídeo
video
produtos
satélite
filmadora
financeira
escola
microcomputador
assinatura
roupa
artigo
esgoto
máquina
operadoras
exaustor
aquecedor
fogão
acessório
cursos
superior
freezer
assistência
pacotes
internet
vestuário
secadora
automóvel
construtoras
turísticos
loteamento

[ ]: