Comparação entre a Distribuição de Etiquetas PoS
– Estatísticas –

1 Sobre os Coporas

1.1 DanteStocks

Descrição: corpus com tweets do mercado financeiro que foram anotados manualmente com informações morfológicas e morfosintaxicas

Onde encontrar o corpus: DanteStocks

1.2 DanteShots

Descrição: corpus com tweets sobre a vacinação contra a COVID-19 que foram anotados automaticamente com informações morfológicas e morfosintaxicas

Onde encontrar o corpus: DanteShots

1.3 Porttinari-base

Descrição: contém sentenças da Folha de S.Paulo que foram anotadas manualmente com informações morfológicas e morfosintaxicas

Onde encontrar o corpus: Porttinari-base

1.4 PetroGold

Descrição: contém sentenças de artigos acadêmicos do domínio do petróleo

Onde encontrar o corpus: PetroGold

2 Dados Gerais dos Corpus

2.1 Quantidade de Tweet ou Sentença

DanteStocks	DanteShots	Porttinari-base	PetroGold
4048	7056	8420	9127

2.2 Total de Etiquetas nos Corpus

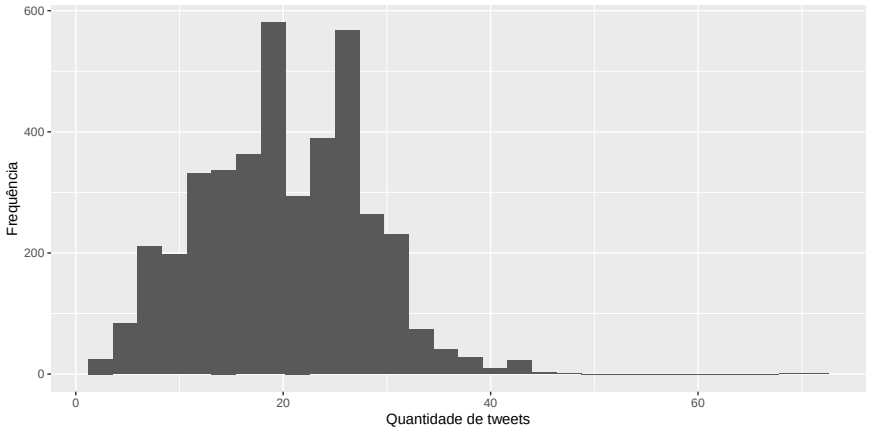
DanteStocks	DanteShots	Porttinari-base	PetroGold
81050	296849	168400	253640

2.3 Etiquetas em um tweet ou numa sentença

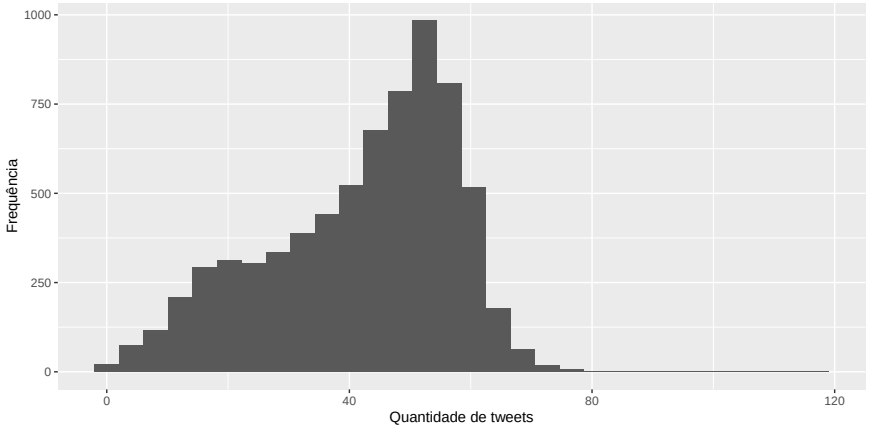
Medida	DanteStocks	DanteShots	Porttinari-base	PetroGold
Mínimo	2	2	4	1
Máximo	71	119	79	239
Média	20.0222332	42.0704365	20	27.7900734
Mediana	20	46	18	26

2.3.1 Gráficos de Dispersão

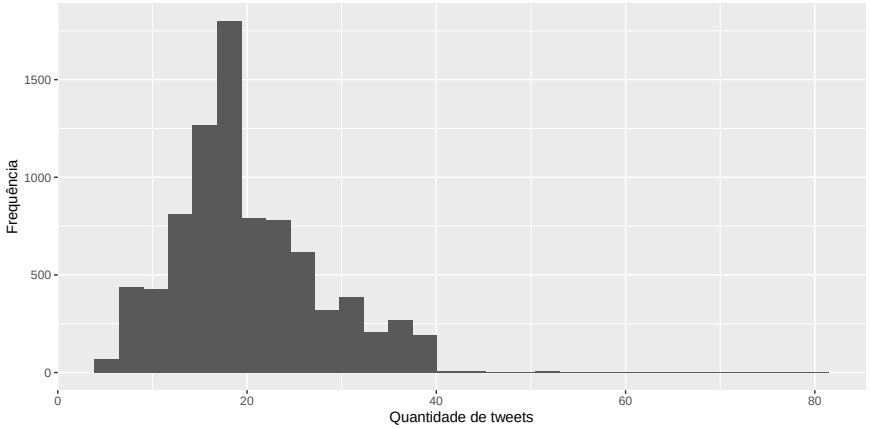
DanteStocks:



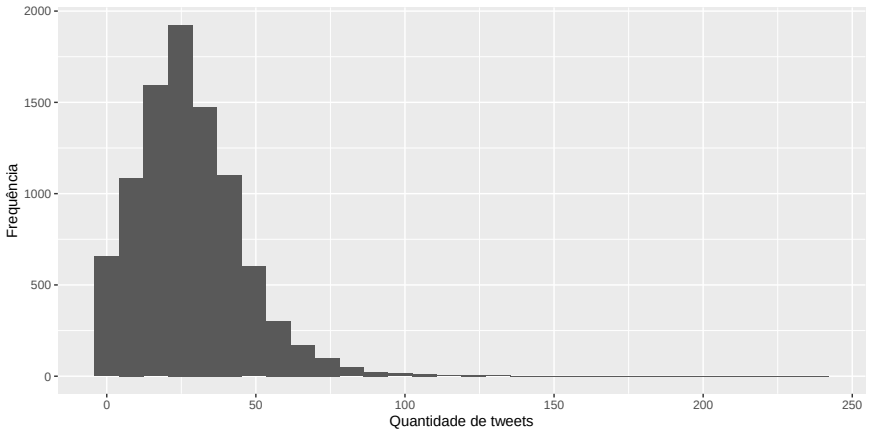
DanteShots:



Porttinari-base:



PetroGold:

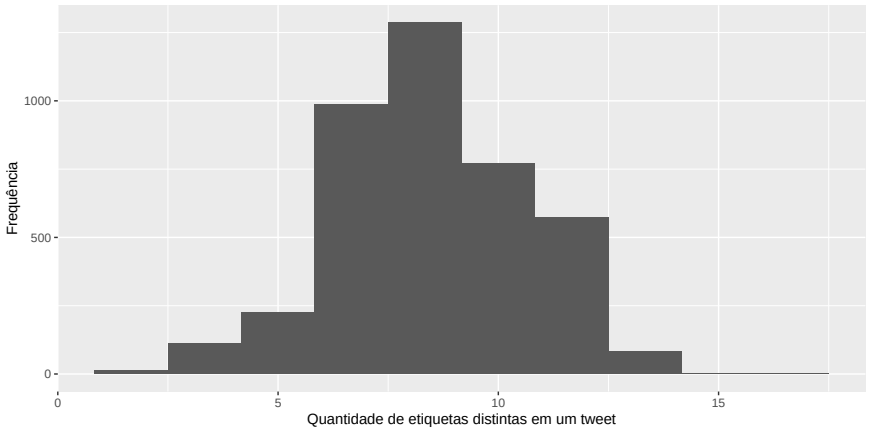

2000
1500
1000
500
0
Frequência
0
50
100
150
200
250
Quantidade de tweets

2.4 Etiquetas Distintas em um Tweet ou Sentença

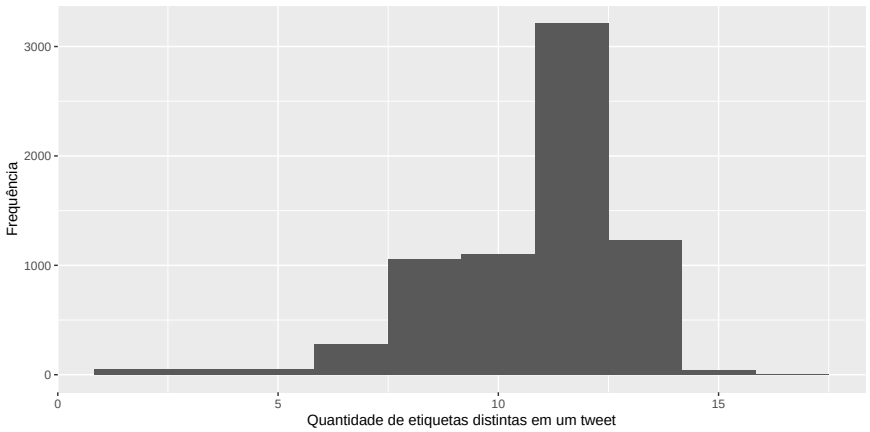
Medida	DanteStocks	DanteStocks	Porttinari-base	PetroGold
Mínimo	1	1	2	1
Máximo	16	16	14	15
Média	8.4807312	10.7993197	8.328266	8.3635368
Mediana	9	11	8	9

2.4.1 Gráficos de Dispersão

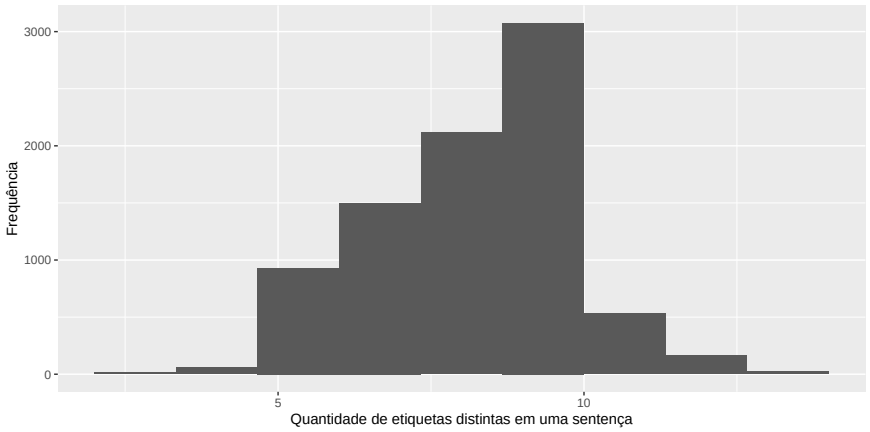
DanteStocks:



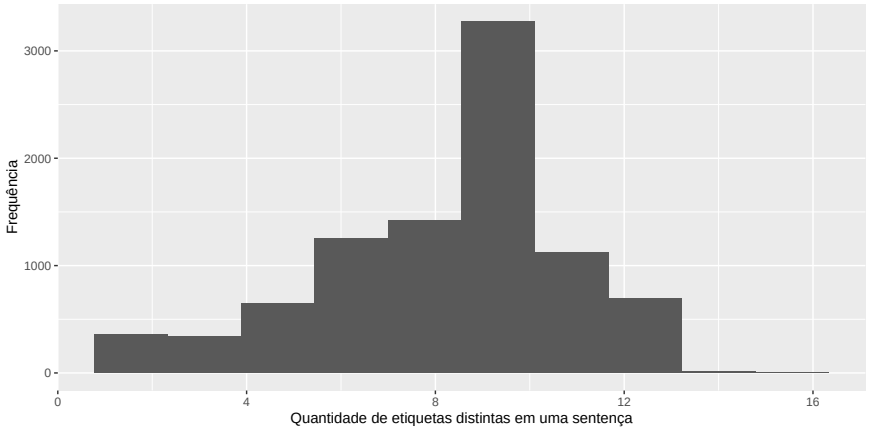
DanteShots:



Porttinari-base:


3000
2000
1000
0
Frequência
5
10
Quantidade de etiquetas distintas em uma sentença

PetroGold:


3000
2000
1000
0
Frequência
0
4
8
12
16
Quantidade de etiquetas distintas em uma sentença

2.5 Qual a etiqueta mais frequente?

No DanteStocks, A etiqueta mais frequente é PUNCT (com 13056 tags no corpus), seguida de NOUN (11936 tags no corpus) e PROPN (11440 tags no corpus).

No DanteShots, A etiqueta mais frequente é NOUN (com 57020 tags no corpus), seguida de ADP (44867 tags no corpus) e DET (42137 tags no corpus).

No Portinari, a etiqueta mais frequente é NOUN (com 31462 tags no corpus), seguida de ADP (25159 tags no corpus) e DET (24273 tags no corpus).

No PetroGold, a etiqueta mais frequente é NOUN (com 57106 tags no corpus), seguida de ADP (42305 tags no corpus) e DET (36710 tags no corpus).

2.6 Qual a etiqueta menos frequente?

No outro extremo, em DanteStocks, temos PART como a etiqueta menos frequente (com 3 tags no corpus), seguida de INTJ (142 tags no corpus) e SCONJ (732 tags no corpus).

No DanteShots, temos PART como a etiqueta menos frequente (com 0 tags no corpus), seguida de INTJ (439 tags no corpus) e X (2961 tags no corpus).

No Portinari, temos PART como a etiqueta menos frequente (com 0 tags no corpus), seguida de INTJ (35 tags no corpus) e X (275 tags no corpus).

No PetroGold, temos PART como a etiqueta menos frequente (com 3 tags no corpus), seguida de INTJ (11 tags no corpus) e X (229 tags no corpus).

2.7 Quantas etiquetas diferentes há no corpus?

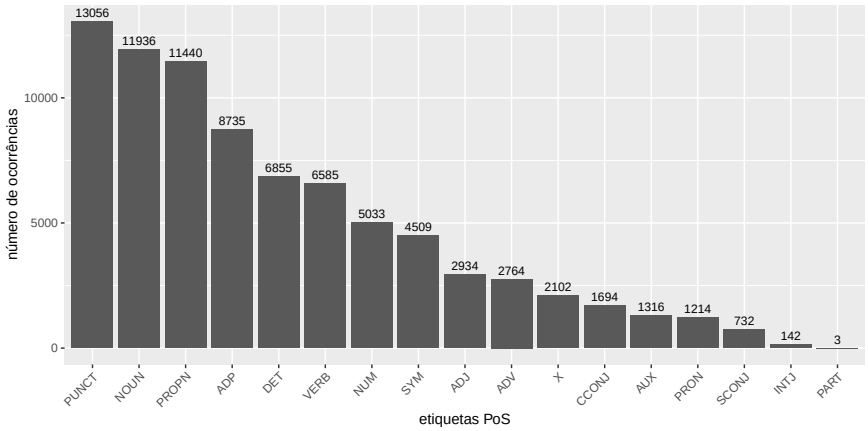
De acordo com UD v2 temos o total de 17 etiquetas na UD. Dessa forma, temos que:

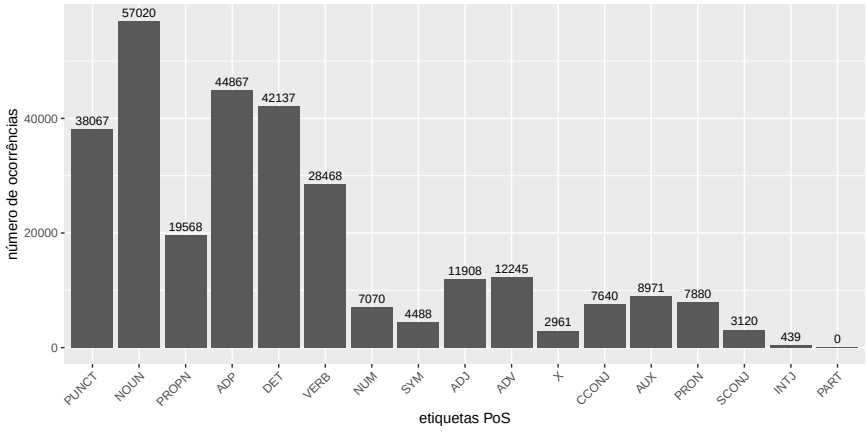
- No DanteStocks, há um total de 17 aparecem no corpus. Estas são (em ordem decrescente de frequência no corpus): PUNCT, NOUN, PROPN, ADP, DET, VERB, NUM, SYM, ADJ, ADV, X, CCONJ, AUX, PRON, SCONJ, INTJ, PART. As etiquetas faltantes são:
- No DanteShots, há um total de 16 aparecem no corpus. Estas são (em ordem decrescente de frequência no corpus): NOUN, ADP, DET, PUNCT, VERB, PROPN, ADV, ADJ, AUX, PRON, CCONJ, NUM, SYM, SCONJ, X, INTJ. As etiquetas faltantes são: PART
- No Portinari, há um total de 16 aparecem no corpus. Estas são (em ordem decrescente de frequência no corpus): NOUN, ADP, DET, PUNCT, VERB, PROPN, ADJ, ADV, PRON, AUX, CCONJ, NUM, SCONJ, SYM, X, INTJ. As etiquetas faltantes são: PART
- No PetroGold, há um total de 17 aparecem no corpus. Estas são (em ordem decrescente de frequência no corpus): NOUN, ADP, DET, PUNCT, VERB, ADJ, PROPN, NUM, AUX, CCONJ, ADV, PRON, SCONJ, SYM, X, INTJ, PART. As etiquetas faltantes são:

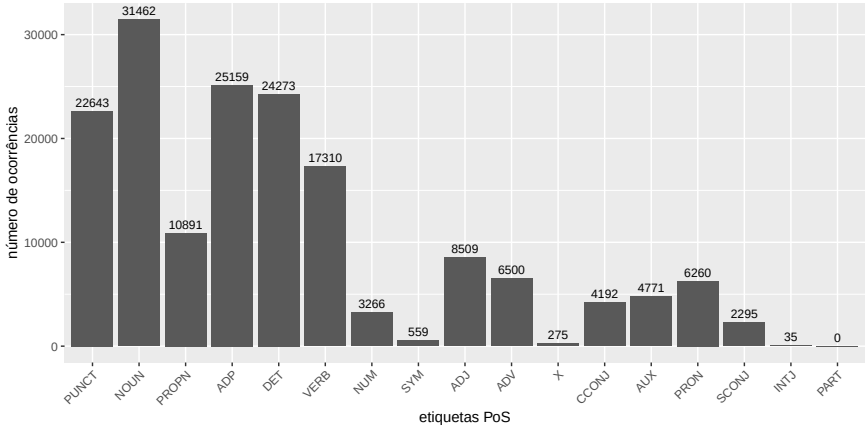
3 Distribuição Geral das Etiquetas PoS

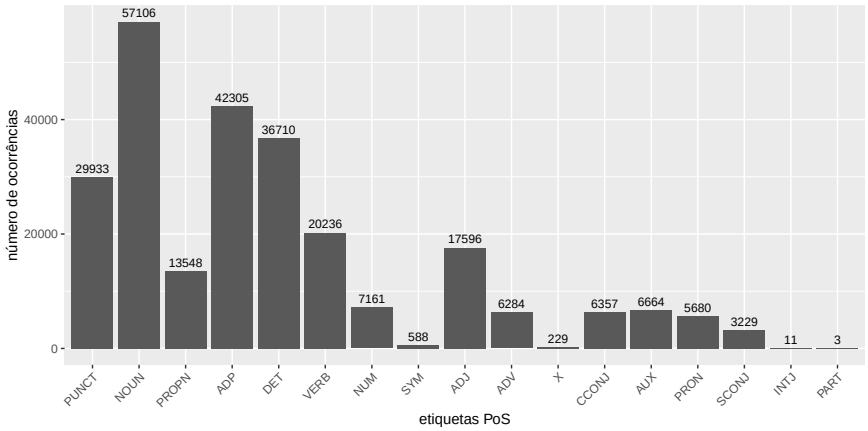
3.1 Frequencia Absoluta

Frequência absoluta de etiquetas em gráficos separados, com as etiquetas seguindo a ordem decrescente de frequência do DanteStocks:

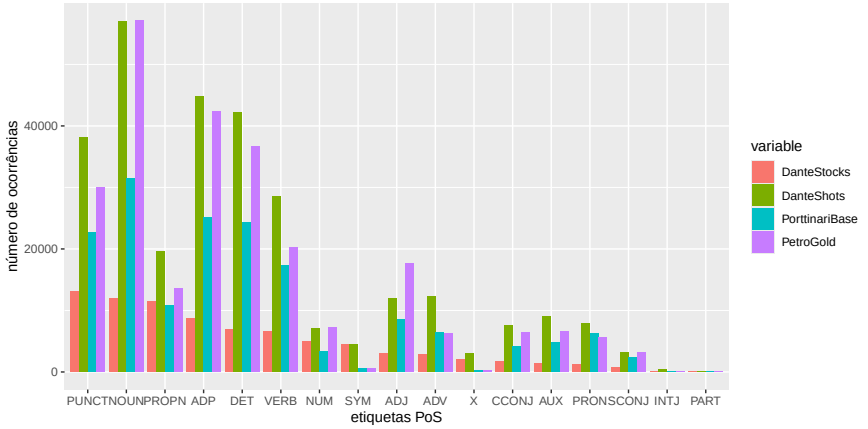








3.1.1 Grafico com todas as distribuições:

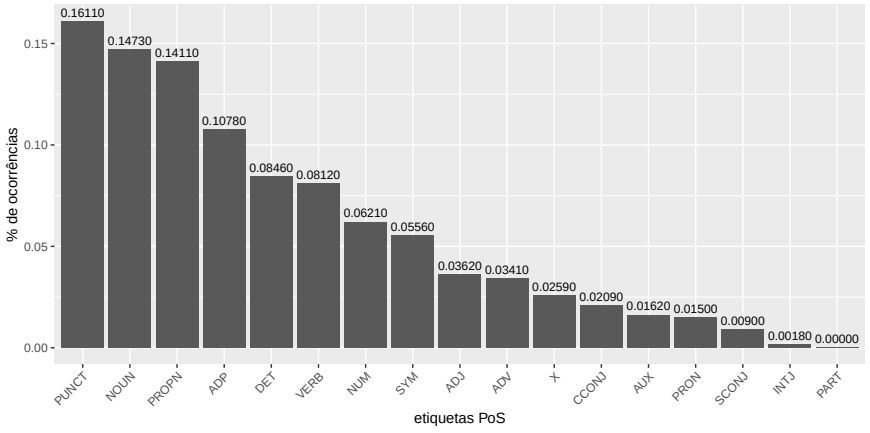


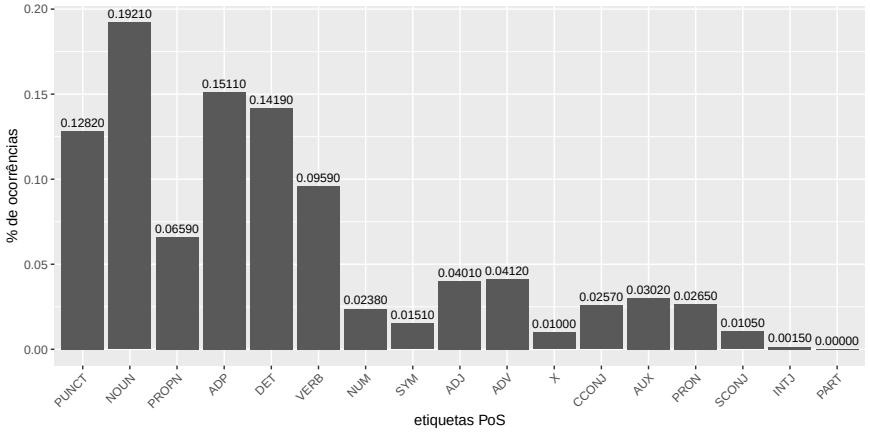
3.2 Frequencia Relativa

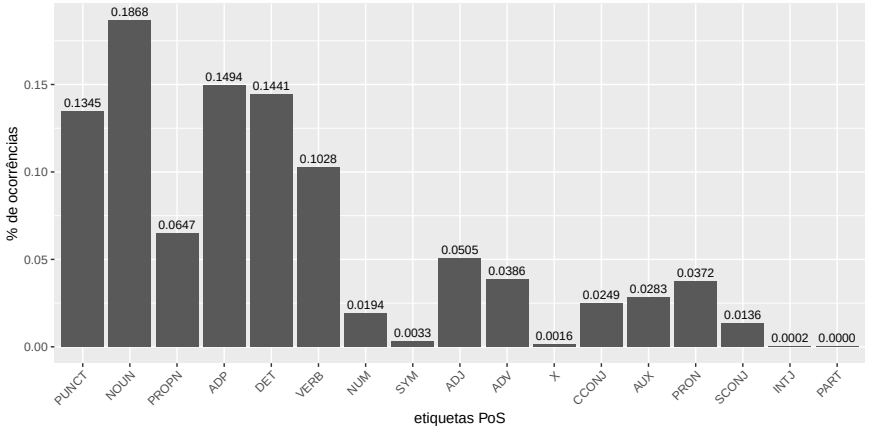
Como os corpora tem tamanhos diferentes, a imagem acima fica um pouco dificil de visualizar. Então vamos plotar em porcentagem:

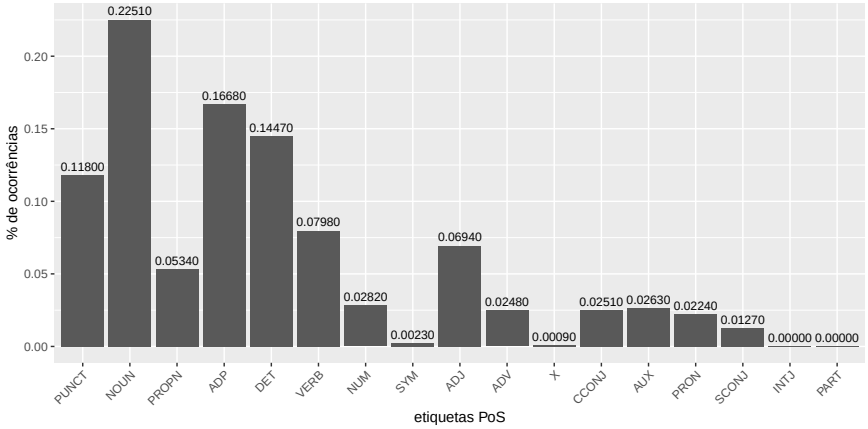
```
##      PoS   DanteStocks  DanteShots PorttinariBase     PetroGold
## 1  PUNCT 0.16108574954 0.128236915   0.1344596200 0.11801372023
## 2   NOUN 0.14726711906 0.192084191   0.1868289786 0.22514587604
## 3  PROPN 0.14114743985 0.065919036   0.0646733967 0.05341428797
## 4    ADP 0.10777297964 0.151144184   0.1494002375 0.16679151553
## 5    DET 0.08457742134 0.141947590   0.1441389549 0.14473269200
## 6   VERB 0.08124614436 0.095900609   0.1027909739 0.07978236871
## 7    NUM 0.06209747070 0.023816823   0.0193942993 0.02823292856
## 8    SYM 0.05563232572 0.015118798   0.0033194774 0.00231824633
## 9    ADJ 0.03619987662 0.040114671   0.0505285036 0.06937391579
## 10   ADV 0.03410240592 0.041249928   0.0385985748 0.02477527204
## 11     X 0.02593460827 0.009974768   0.0016330166 0.00090285444
## 12 CCONJ 0.02090067859 0.025736991   0.0248931116 0.02506308153
## 13   AUX 0.01623689081 0.030220752   0.0283313539 0.02627345845
## 14  PRON 0.01497840839 0.026545483   0.0371733967 0.02239394417
## 15 SCONJ 0.00903146206 0.010510394   0.0136282660 0.01273064185
## 16  INTJ 0.00175200494 0.001478866   0.0002078385 0.00004336855
## 17  PART 0.00003701419 0.000000000   0.0000000000 0.00001182779
```

3.2.1 Separado

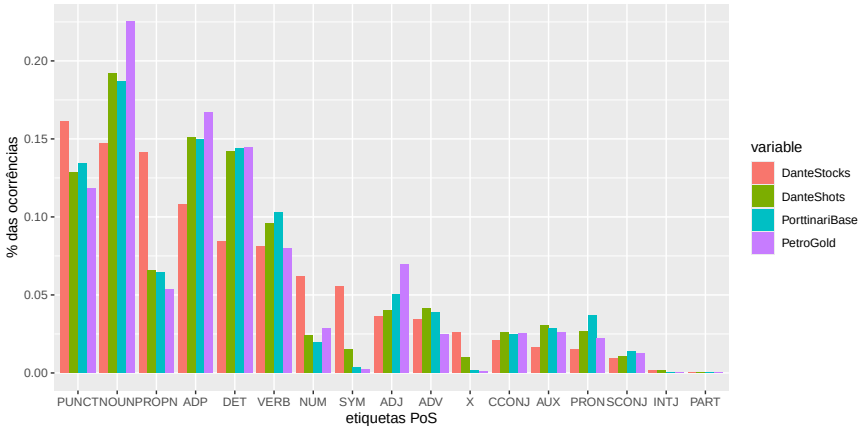






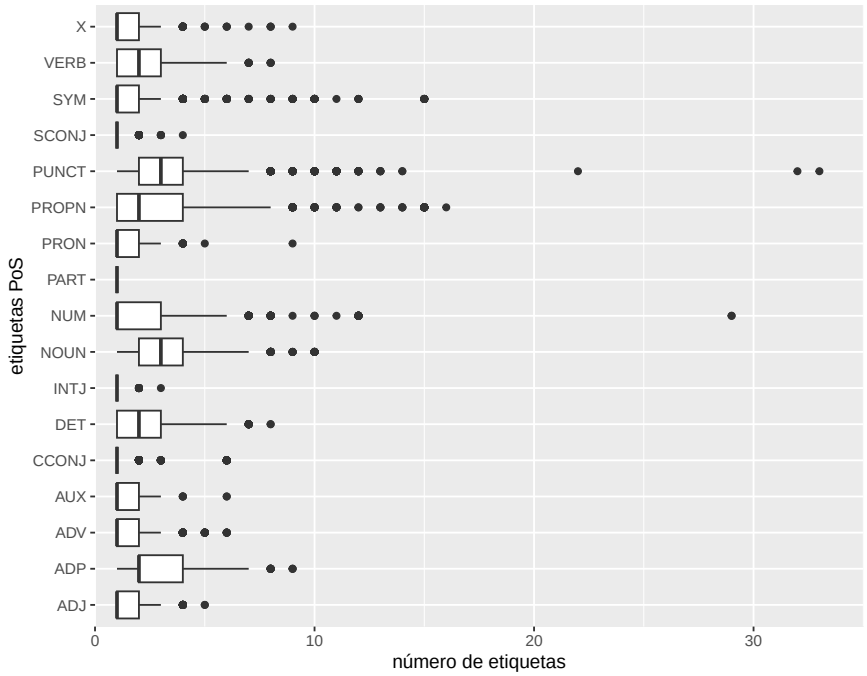

0.22510
0.16680
0.14470
0.11800
0.07980
0.06940
0.05340
0.02820
0.02480
0.02510
0.02630
0.02240
0.01270
0.00230
0.00090
0.00000
0.00000
0.20
0.15
0.10
0.05
0.00
% de ocorrências
PUNCT
NOUN
PROPN
ADP
DET
VERB
NUM
SYM
ADJ
ADV
X
CCONJ
AUX
PRON
SCONJ
INTJ
PART
etiquetas PoS

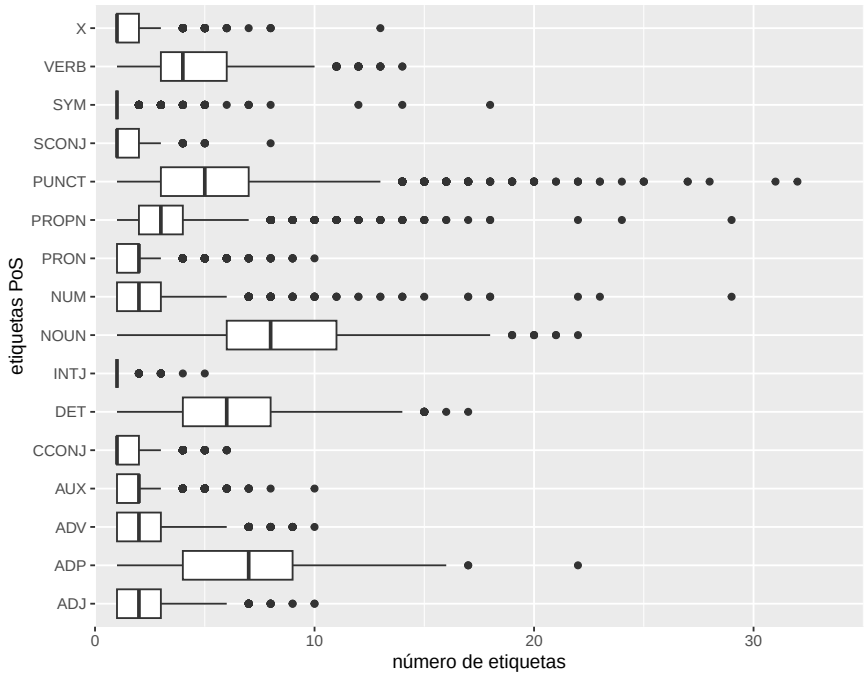
Junto:

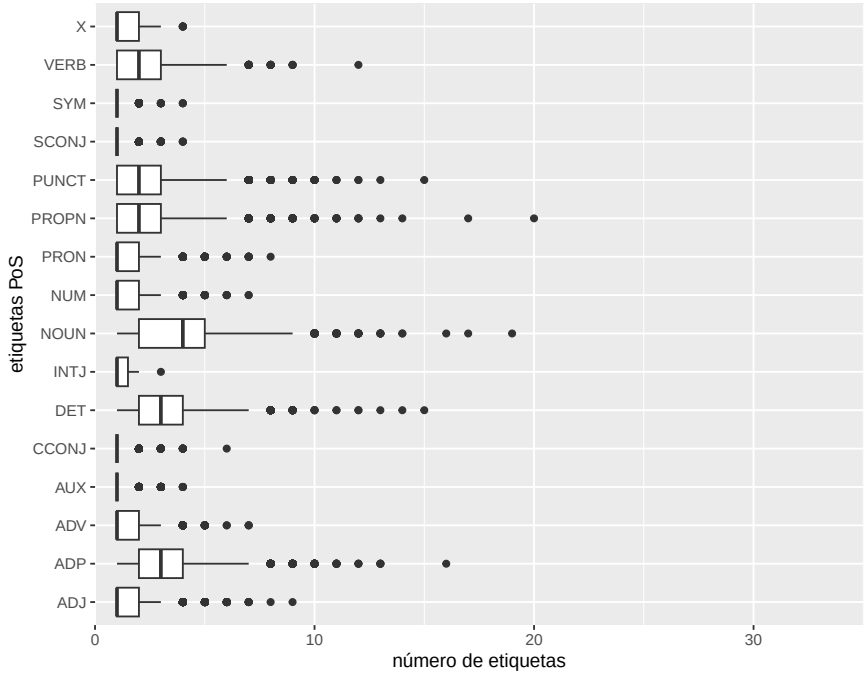

0.20
0.15
0.10
0.05
0.00
% das ocorrências
PUNCT
NOUN
PROPN
ADP
DET
VERB
NUM
SYM
ADJ
ADV
X
CCONJ
AUX
PRON
SCONJ
INTJ
PART
etiquetas PoS
variable
DanteStocks
DanteShots
PorttinariBase
PetroGold

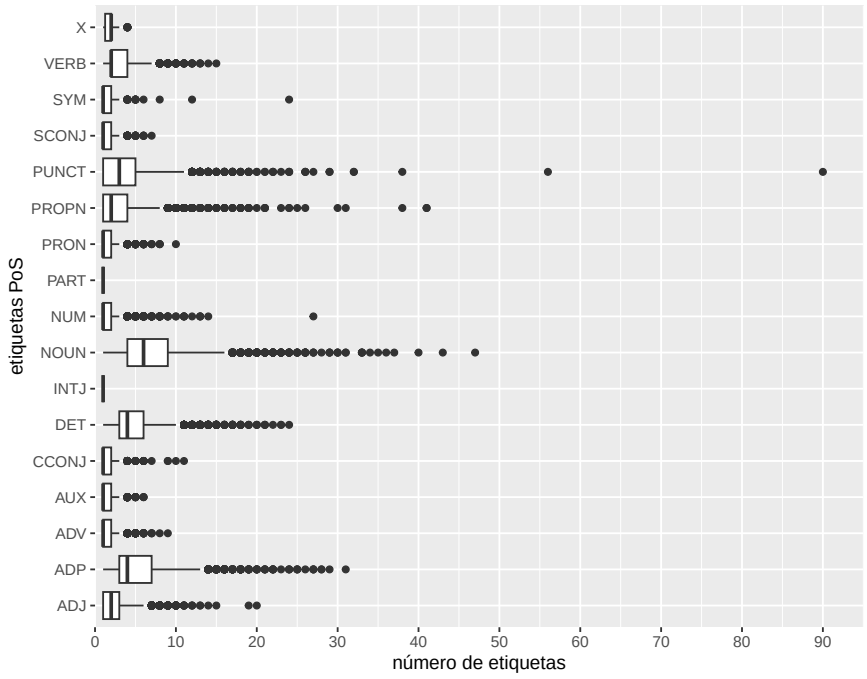
4 Distribuição de Etiquetas Conforme sua Frequência em Tweets/Sentences em que Ocorrem

O gráfico abaixo ilustra o número mínimo, máximo, bem como quartis, de vezes que cada etiqueta ocorre por tweet, considerando apenas os tweets em que ela ocorre:






X
VERB
SYM
SCONJ
PUNCT
PROPN
PRON
NUM
NOUN
INTJ
DET
CCONJ
AUX
ADV
ADP
ADJ
etiquetas PoS
0
10
20
30
número de etiquetas


X
VERB
SYM
SCONJ
PUNCT
PROPN
PRON
PART
NUM
NOUN
INTJ
DET
CCONJ
AUX
ADV
ADP
ADJ
etiquetas PoS
0
10
20
30
40
50
60
70
80
90
número de etiquetas

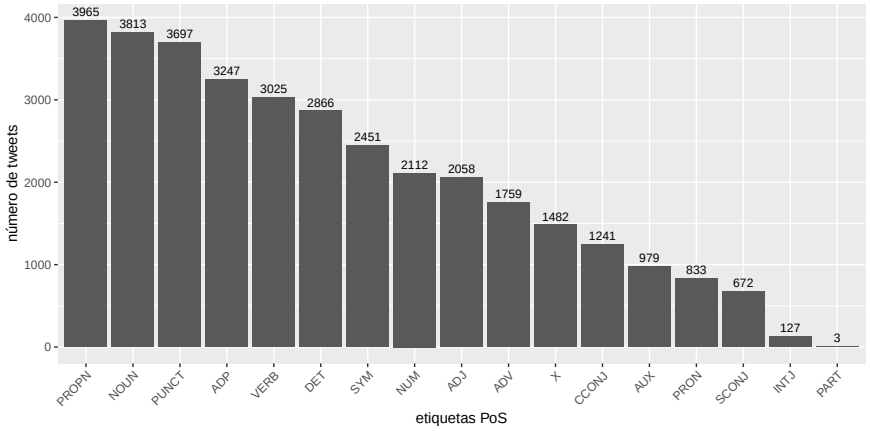
5 Abrangência de cada etiqueta

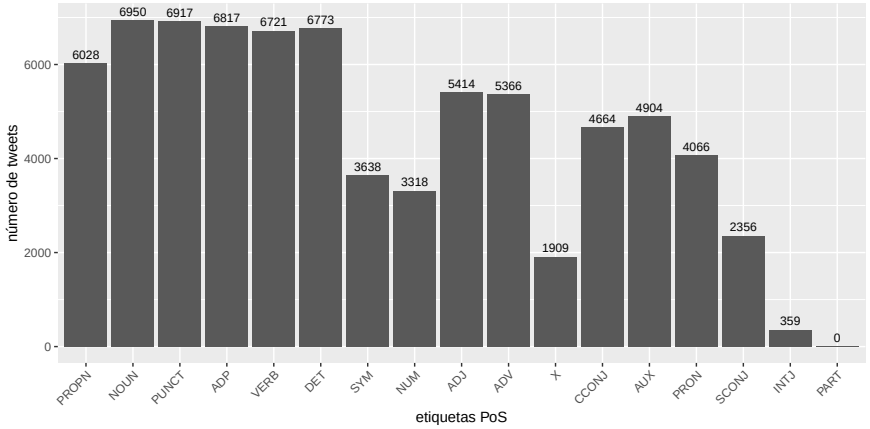
5.1 Frequência Absoluta

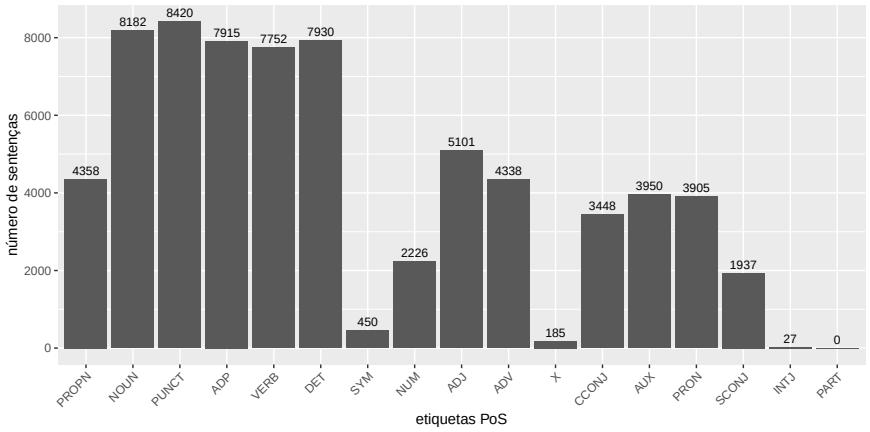
Quantos tweets ou sentenças possuem certa tag:

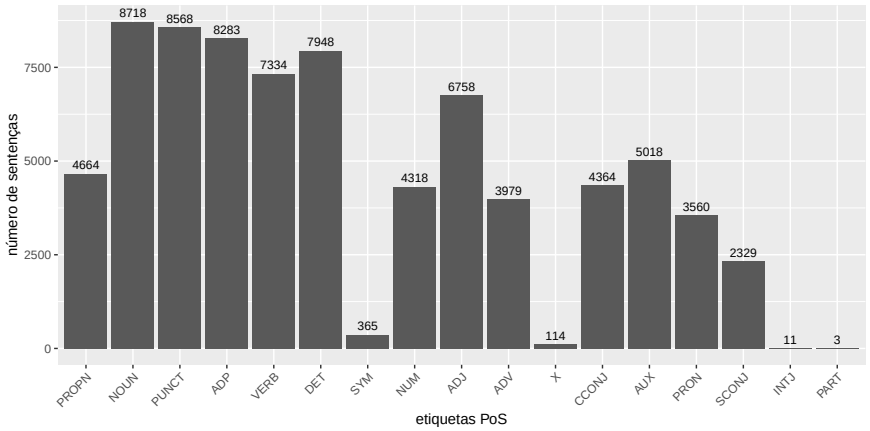
```
##      PoS DanteStocks DanteShots PorttinariBase PetroGold
## 1  PROPN        3965       6028           4358      4664
## 2   NOUN        3813       6950           8182      8718
## 3  PUNCT        3697       6917           8420      8568
## 4    ADP        3247       6817           7915      8283
## 5   VERB        3025       6721           7752      7334
## 6    DET        2866       6773           7930      7948
## 7    SYM        2451       3638            450       365
## 8    NUM        2112       3318           2226      4318
## 9    ADJ        2058       5414           5101      6758
## 10   ADV        1759       5366           4338      3979
## 11     X        1482       1909            185       114
## 12 CCONJ        1241       4664           3448      4364
## 13   AUX         979       4904           3950      5018
## 14  PRON         833       4066           3905      3560
## 15 SCONJ         672       2356           1937      2329
## 16  INTJ         127        359             27        11
## 17  PART           3          0              0         3
```

Separado:

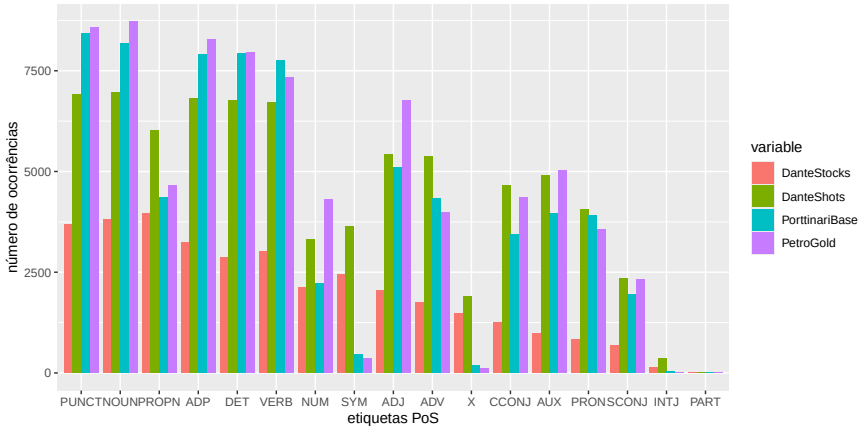

número de tweets
4000
3000
2000
1000
0
3965
3813
3697
3247
3025
2866
2451
2112
2058
1759
1482
1241
979
833
672
127
3
PROPN
NOUN
PUNCT
ADP
VERB
DET
SYM
NUM
ADJ
ADV
X
CCONJ
AUX
PRON
SCONJ
INTJ
PART
etiquetas PoS


número de tweets
6000
4000
2000
0
6028
6950
6917
6817
6721
6773
3638
3318
5414
5366
1909
4664
4904
4066
2356
359
0
PROPN
NOUN
PUNCT
ADP
VERB
DET
SYM
NUM
ADJ
ADV
X
CCONJ
AUX
PRON
SCONJ
INTJ
PART
etiquetas PoS


número de sentenças
8000
6000
4000
2000
0
4358
8182
8420
7915
7752
7930
450
2226
5101
4338
185
3448
3950
3905
1937
27
0
PROPN
NOUN
PUNCT
ADP
VERB
DET
SYM
NUM
ADJ
ADV
X
CCONJ
AUX
PRON
SCONJ
INTJ
PART
etiquetas PoS


número de sentenças
7500
5000
2500
0
4664
8718
8568
8283
7334
7948
365
4318
6758
3979
114
4364
5018
3560
2329
11
3
PROPN
NOUN
PUNCT
ADP
VERB
DET
SYM
NUM
ADJ
ADV
X
CCONJ
AUX
PRON
SCONJ
INTJ
PART
etiquetas PoS

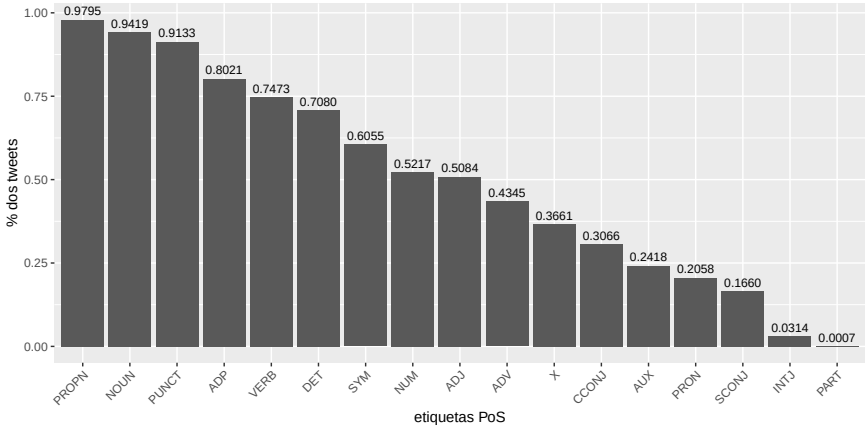
Junto:

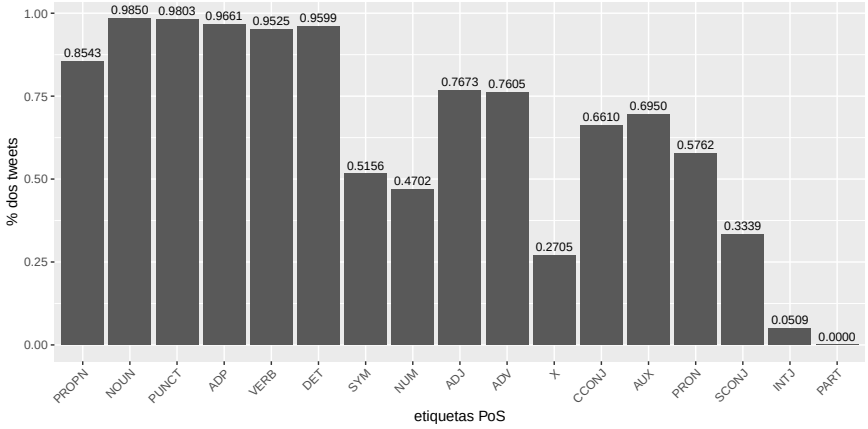

número de ocorrências
7500
5000
2500
0
PUNCT
NOUN
PROPN
ADP
DET
VERB
NUM
SYM
ADJ
ADV
X
CCONJ
AUX
PRON
SCONJ
INTJ
PART
etiquetas PoS
variable
DanteStocks
DanteShots
PorttinariBase
PetroGold

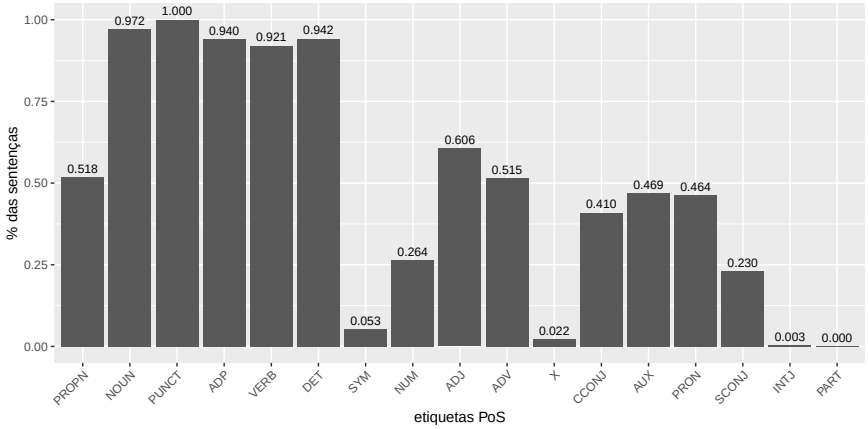
5.2 Frequência Relativa

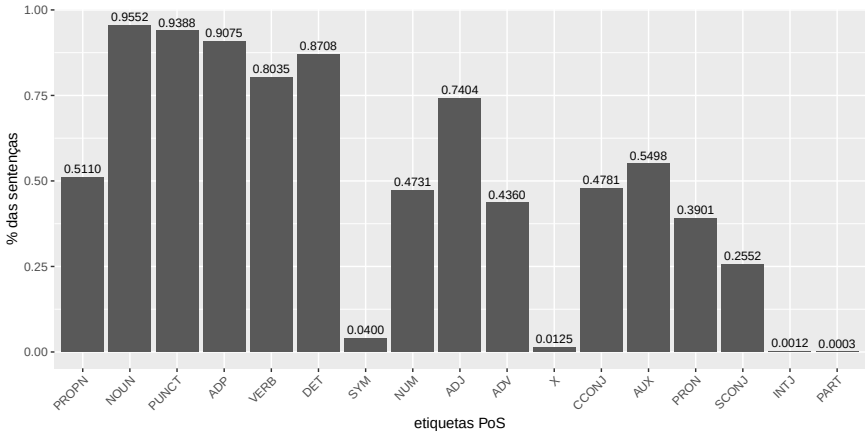
```
##       PoS  DanteStocks DanteShots PorttinariBase    PetroGold
## 1   PROPN 0.9794960474 0.85430839    0.517577197 0.5110112852
## 2    NOUN 0.9419466403 0.98497732    0.971733967 0.9551879040
## 3   PUNCT 0.9132905138 0.98030045    1.000000000 0.9387531500
## 4     ADP 0.8021245059 0.96612812    0.940023753 0.9075271173
## 5    VERB 0.7472826087 0.95252268    0.920665083 0.8035499069
## 6     DET 0.7080039526 0.95989229    0.941805226 0.8708228334
## 7     SYM 0.6054841897 0.51558957    0.053444181 0.0399912348
## 8     NUM 0.5217391304 0.47023810    0.264370546 0.4731017859
## 9     ADJ 0.5083992095 0.76729025    0.605819477 0.7404404514
## 10    ADV 0.4345355731 0.76048753    0.515201900 0.4359592418
## 11      X 0.3661067194 0.27054989    0.021971496 0.0124904131
## 12 CCONJ 0.3065711462 0.66099773    0.409501188 0.4781417771
## 13    AUX 0.2418478261 0.69501134    0.469121140 0.5497973047
## 14   PRON 0.2057806324 0.57624717    0.463776722 0.3900514956
## 15 SCONJ 0.1660079051 0.33390023    0.230047506 0.2551769475
## 16   INTJ 0.0313735178 0.05087868    0.003206651 0.0012052153
## 17   PART 0.0007411067 0.00000000    0.000000000 0.0003286951
```

Separado:

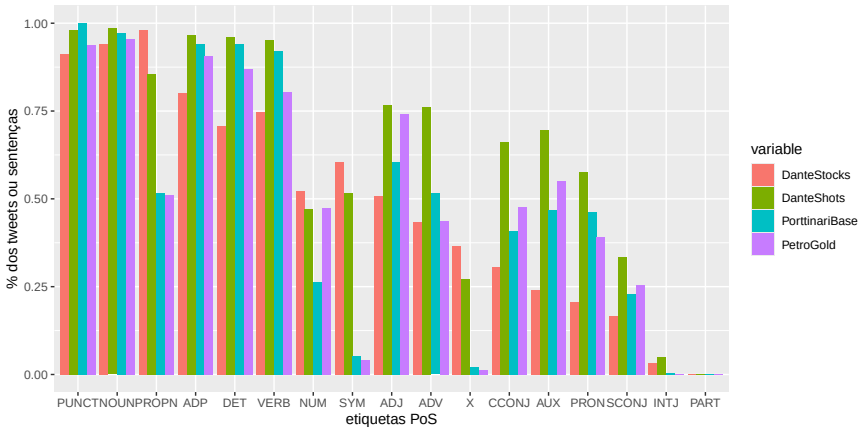





% das sentenças
1.00
0.75
0.50
0.25
0.00
0.518
0.972
1.000
0.940
0.921
0.942
0.053
0.264
0.606
0.515
0.022
0.410
0.469
0.464
0.230
0.003
0.000
PROPN
NOUN
PUNCT
ADP
VERB
DET
SYM
NUM
ADJ
ADV
X
CCONJ
AUX
PRON
SCONJ
INTJ
PART
etiquetas PoS


% das sentenças
1.00
0.75
0.50
0.25
0.00
0.5110
0.9552
0.9388
0.9075
0.8035
0.8708
0.0400
0.4731
0.7404
0.4360
0.0125
0.4781
0.5498
0.3901
0.2552
0.0012
0.0003
PROPN
NOUN
PUNCT
ADP
VERB
DET
SYM
NUM
ADJ
ADV
X
CCONJ
AUX
PRON
SCONJ
INTJ
PART
etiquetas PoS

Junto:


% dos tweets ou sentenças
1.00
0.75
0.50
0.25
0.00
PUNCT
NOUN
PROPN
ADP
DET
VERB
NUM
SYM
ADJ
ADV
X
CCONJ
AUX
PRON
SCONJ
INTJ
PART
etiquetas PoS
variable
DanteStocks
DanteShots
PorttinariBase
PetroGold