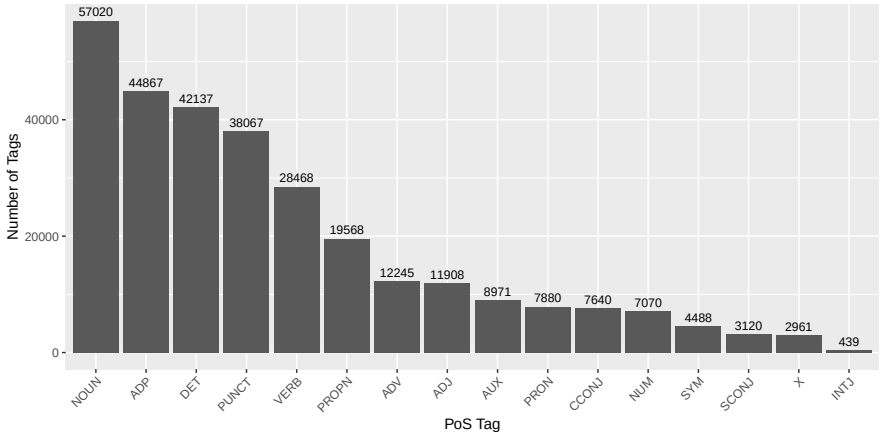
Distribuição de Etiquetas PoS no DANTEShots
– Estatísticas –

1 Dados Gerais do Corpus

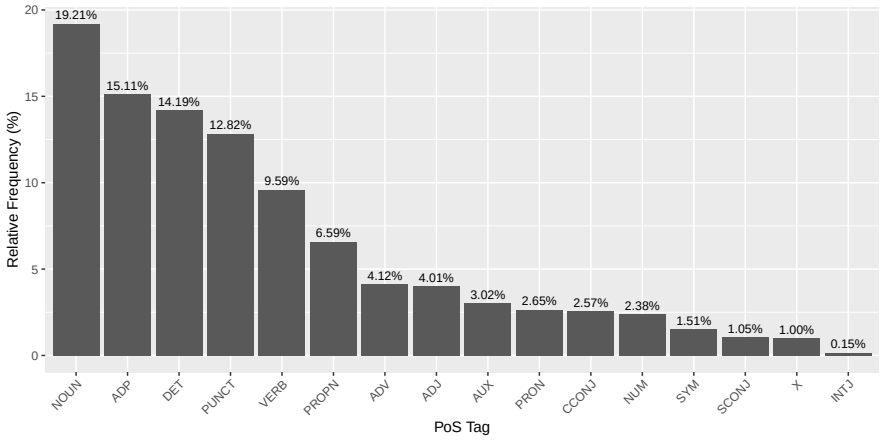
- Quantidade de tweets: 7056
- Total de etiquetas no corpus: 296849
- Número mínimo de etiquetas em um tweet: 2
- Número máximo de etiquetas em um tweet: 119
- Número máximo de etiquetas distintas em um tweet: 16

2 Distribuição Geral das Etiquetas PoS

Frequência absoluta de etiquetas, e sua distribuição, no corpus como um todo:



Frequência relativa de etiquetas, e sua distribuição, no corpus como um todo:



2.1 Qual a etiqueta mais frequente?

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

2.2 Qual a etiqueta menos frequente?

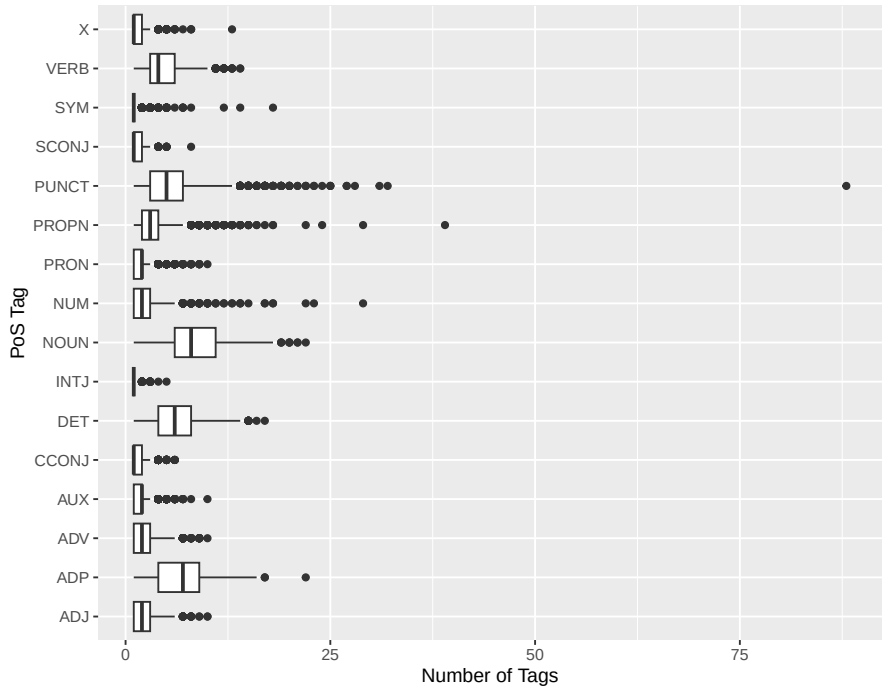
No outro extremo, temos INTJ como a etiqueta menos frequente (com 439 tags no corpus), seguida de X (2961 tags no corpus) e SCONJ (3120 tags no corpus).

2.3 Quantas etiquetas diferentes há no corpus?

Das 17 etiquetas possíveis na UD v2 (https://universaldependencies.org/v2/postags.html), 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.

3 Distribuição de Etiquetas Conforme sua Frequência em Tweets 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:



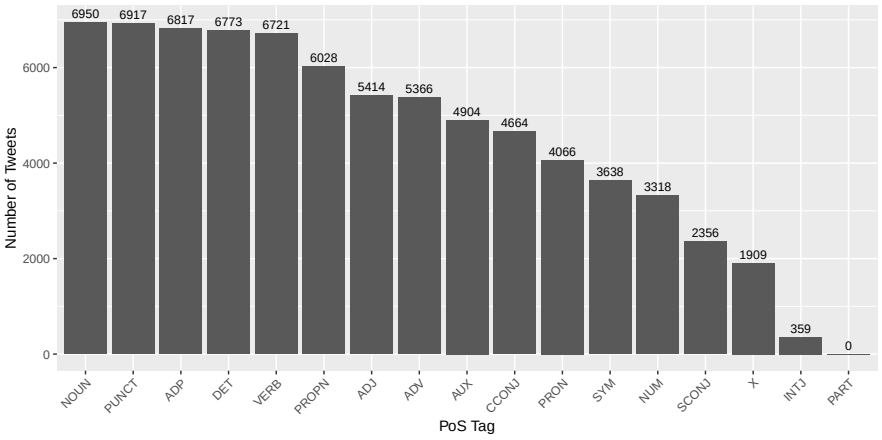
Note que, em 5 etiquetas (das 16), a mediana iguala o valor mínimo, indicando que 50% dos dados correspondem a esse valor, e que em 11 a mediana se afasta da base do corpo da caixa. Também é notória a presença de outliers apenas na parte superior do intervalo (acima de 1.5 $\times$ IQR – inter-quartile range – da borda), tudo isso indicando uma concentração das frequências dos tags nos valores mais baixos.

Podemos observar que existe um tweet cuja quantidade de pontuações está muito além da quantidade encontrada nos demais, o tweet em questão possui o índice dante_02_1440451736863318016. E o seu conteúdo é mostrado abaixo:

"Vacinam adolescentes:\n\nEUA????Canadá????Chile????Uruguai????Áustria????Itália????Suiça????Alem????Israe
França????Espanha????Inglaterra??????????????Irlanda????Escócia??????????????Lituânia????China????Singapura
Filipina????Romênia????Portugal????Israel????Catar????Japão????Polônia????Hungria????Noruega????Bélgica???
Argentina????Índia????"

4 Abrangência de cada etiqueta

A distribuição das etiquetas, conforme o número de tweets em que ocorrem, é:



Note que:

- Das 17 tags, 12 ocorrem em mais de 50% dos tweets (**i.e.** mais de 3528 tweets)
- As 4 tags mais abrangentes, em número de tweets em que ocorrem, também são as mais comuns, em número de vezes em que ocorrem no corpus, embora a ordem mude.
