Estudo do comprimento de mensagens em fóruns online

Brian K. - 7161121

15 de dezembro de 2017

Sumário

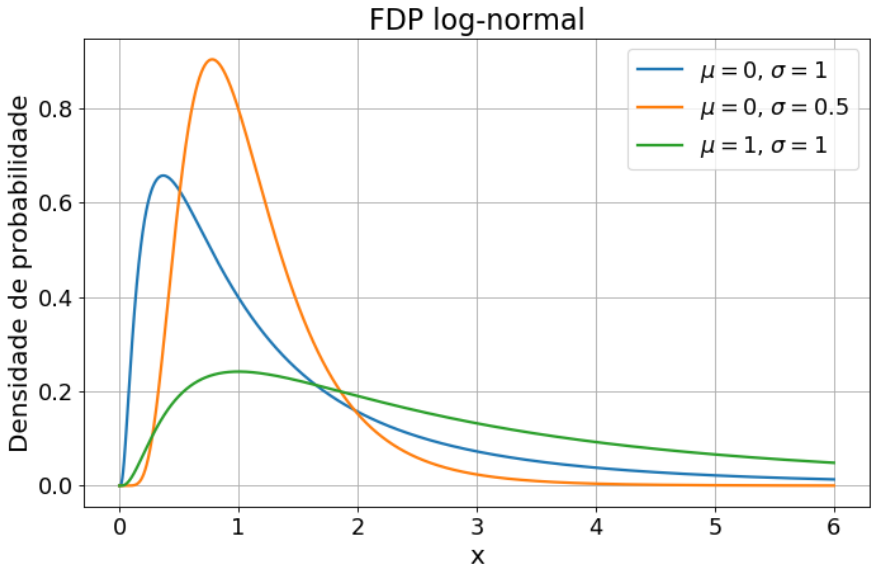
Motivação

- Trabalhar com “grande” volume de dados.
- Utilizar a internet como fonte de dados.
- Introdução à distribuição log-normal.

Sobkowicz et al. (2007)

- Lognormal distributions of user post lengths in Internet discussions - a consequence of the Weber-Fechner law?
- Trabalho com base em datasets dos fóruns da BBC, Myspace e Youtube em inglês, e fóruns poloneses.
- Comprimentos dos comentários seguem uma distribuição log-normal.
- Hipotetizam que a forma log-normal está associada com a lei de Weber-Fechner através do tempo gasto escrevendo o comentário a percepção de seu comprimento pelo próprio autor.

Distribuição log-normal



- Pode ser parametrizada com:

$$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{ln(x)-\mu}{\sigma}\right)^2} \tag{1}$$

- Se X segue uma distribuição log-normal então $ln(X) \sim \mathcal{N}(\mu, \sigma^2)$
- Obedece uma forma multiplicativa do TLC.

Distribuição log-normal(cont.)

- Mas μ e σ^2 não são a média e variância da distribuição log-normal!

Moda	Mediana	Média	Variância
$e^{\mu-\sigma^2}$	e^{μ}	$e^{\mu+\frac{\sigma^2}{2}}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$

- Utilizada freqüentemente em análise de riscos financeiros.

Procedimento

- Coleta de comentários de fóruns em diversas línguas hospedados no reddit.
- Limpeza de comentários.
- Ajuste à distribuição log-normal utilizando MMV.
- Análise resultados.

Reddit

- Agregador e plataforma de discussão.
- 8º lugar no Alexa.
- Abriga comunidades falantes de diversos idiomas ao redor do globo.
- API bem documentada com comunidade ativa de desenvolvedores.

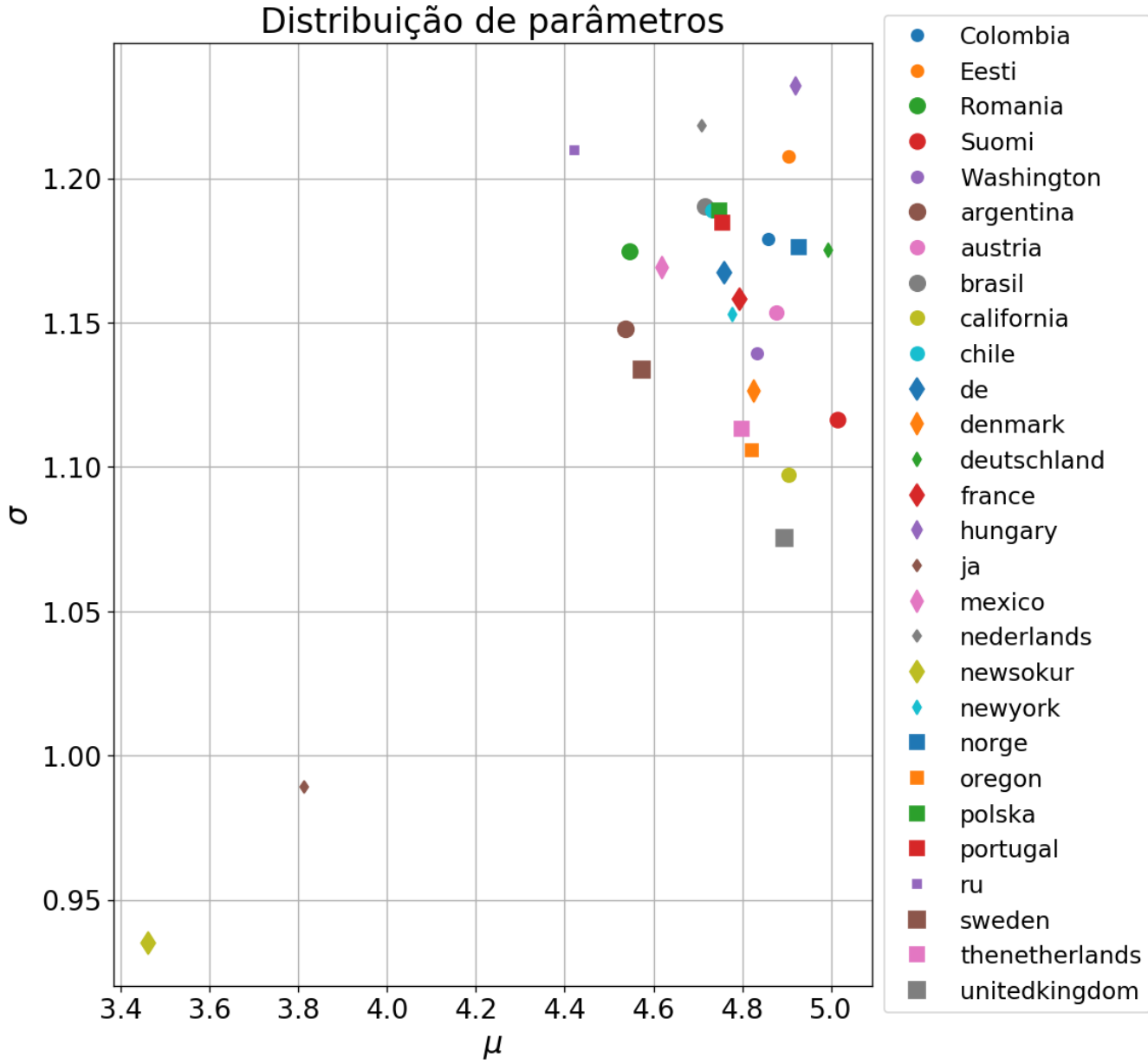
Reddit - Problemas

- Comunidade anglófona enorme e difusa.
- Limites da API/servidores.
- Bots.
- Formatação.

Dados coletados

- Pouco mais de 12 milhões de comentários processados em aproximadamente duas semanas.
- 28 subreddits em 16 idiomas completamente arquivados.

Resultados dos ajustes



Errata

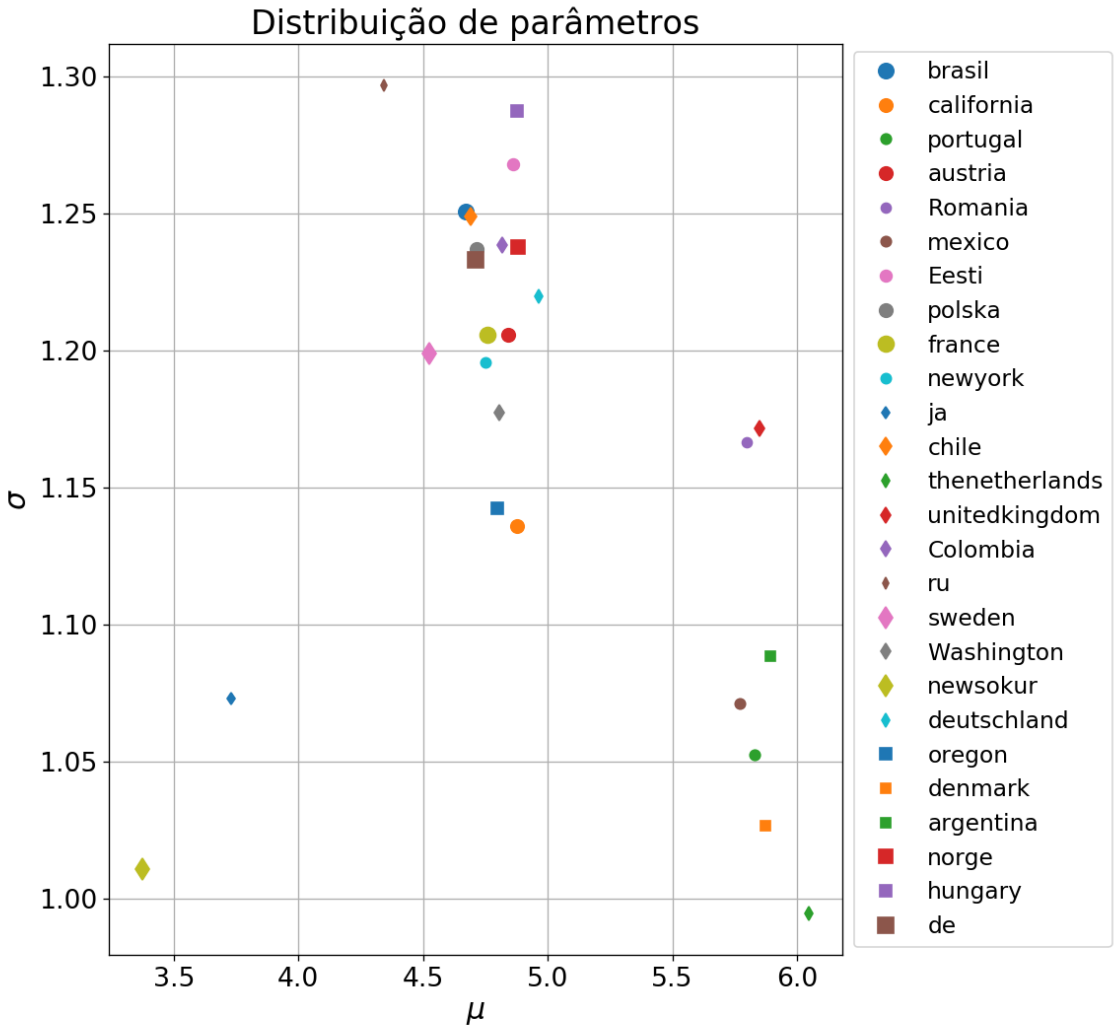
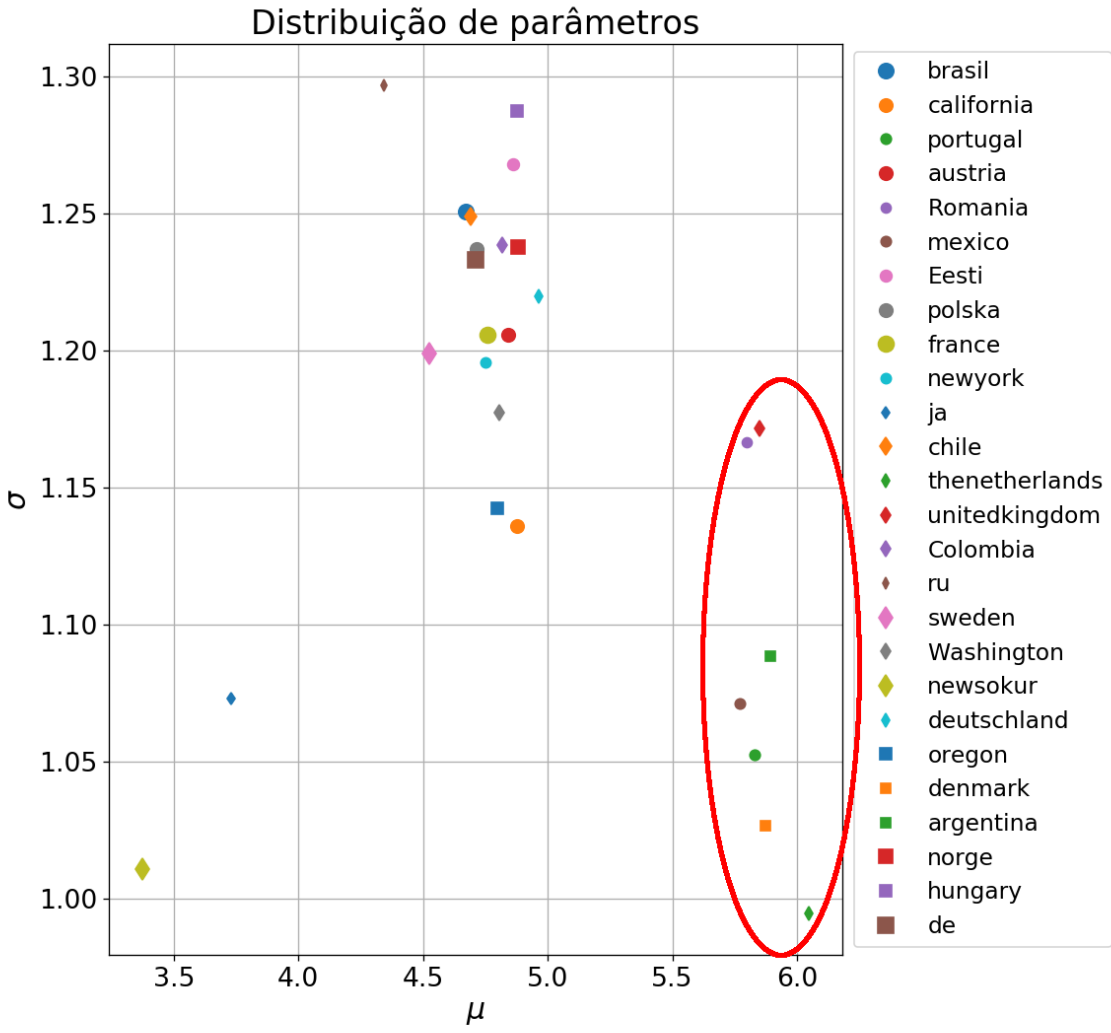


Imagem da apresentação original

Errata(cont.)



Discrepantes das demais comunidades ocidentais.

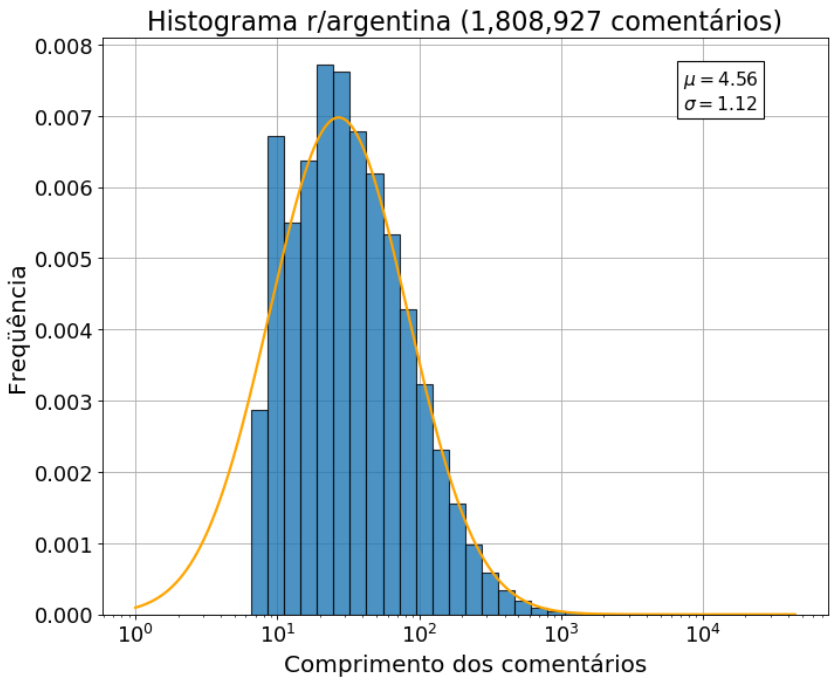
Errata(cont.)

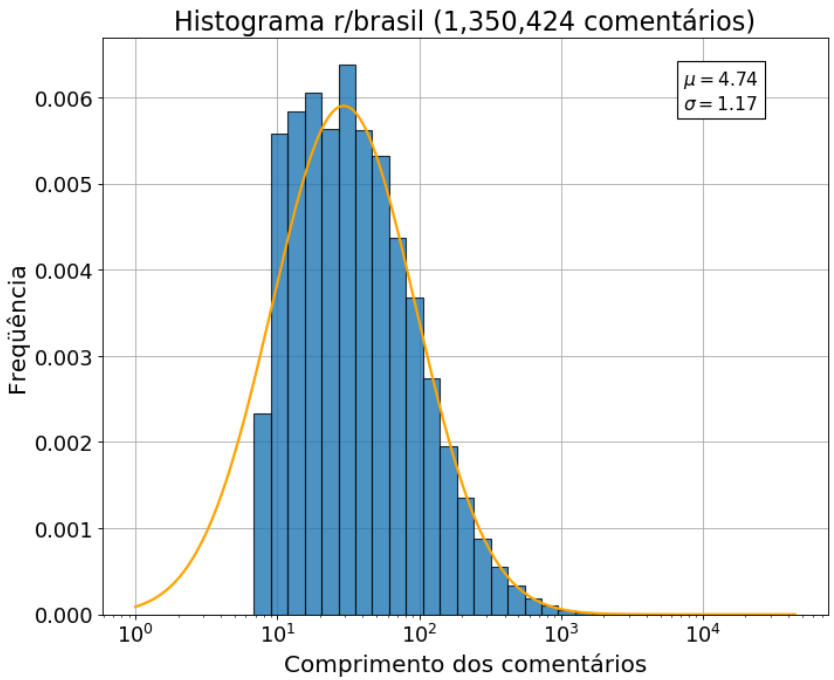
- Alguns subreddits apenas parcialmente arquivados foram acidentalmente selecionados para análise.
- Para estes subs apenas a primeira postagem de cada discussão havia sido arquivada.
- Sobkowicz já havia mostrado que as primeiras postagens tendem a ser mais longas que comentários subseqüentes.
- Exatamente o que aconteceu com os subs destacados.

Resultados - Gráficos

Histogramas de alguns ajustes.

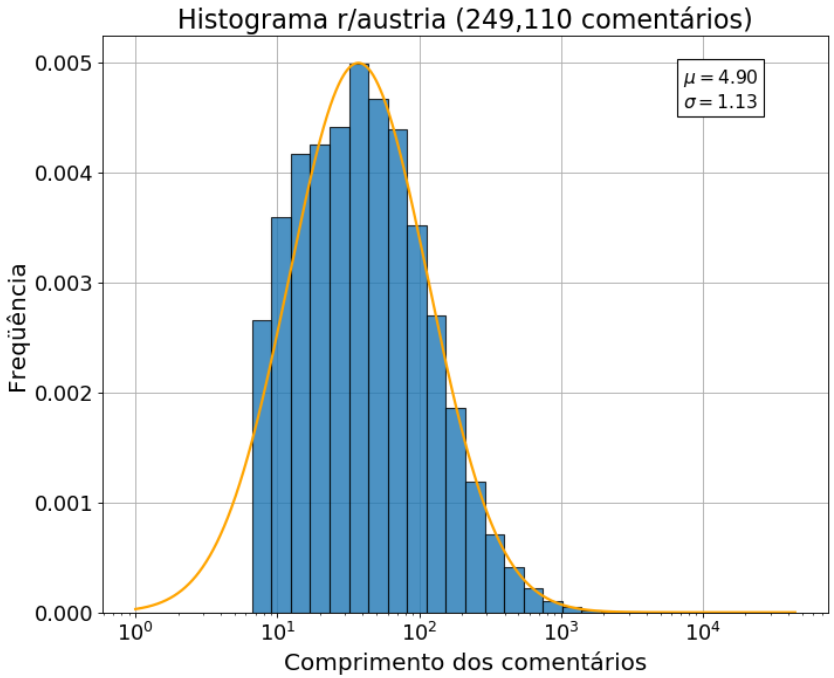
Principais subs Argentina e Brasil

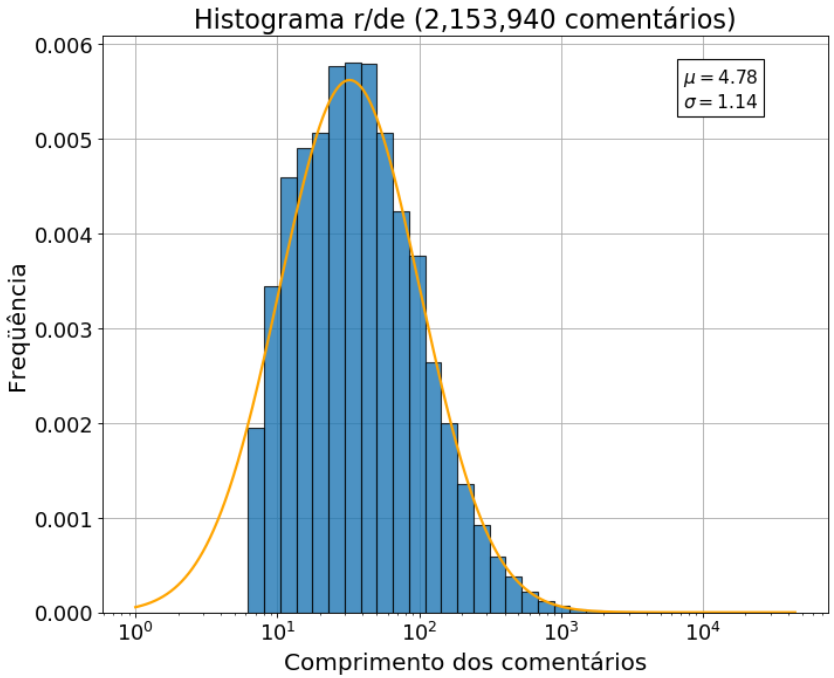




Resultados - Gráficos

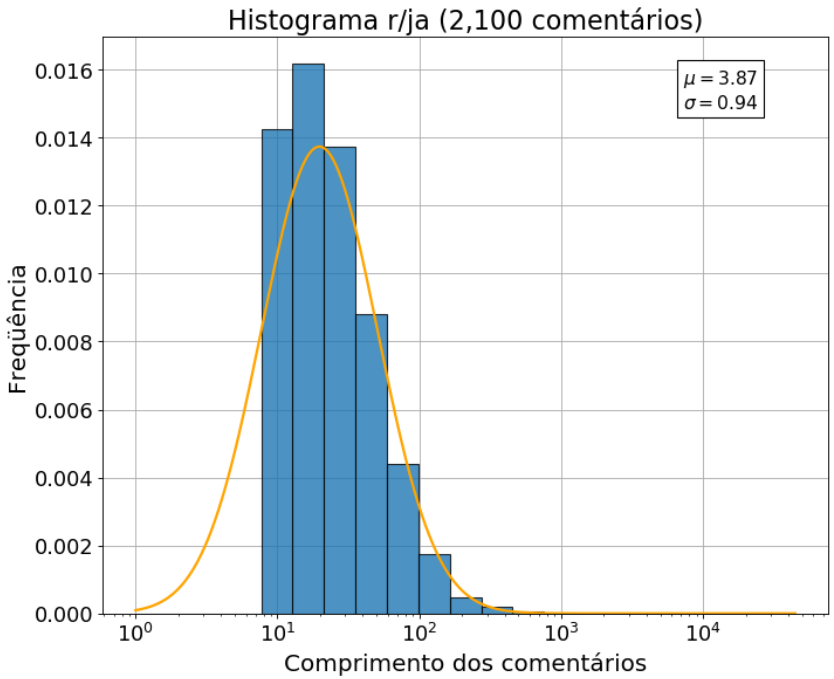
Principais subs Áustria e Alemanha

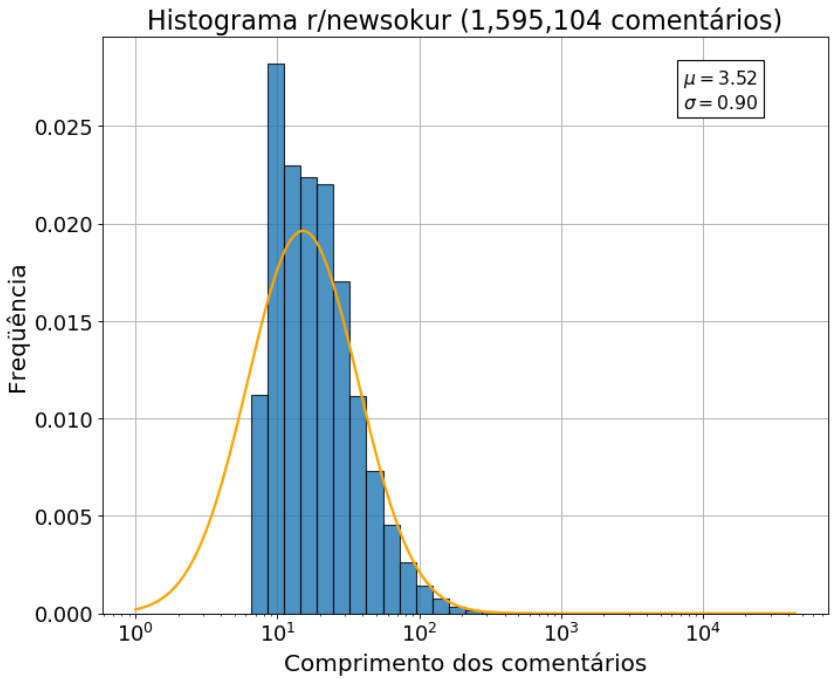




Resultados - Gráficos

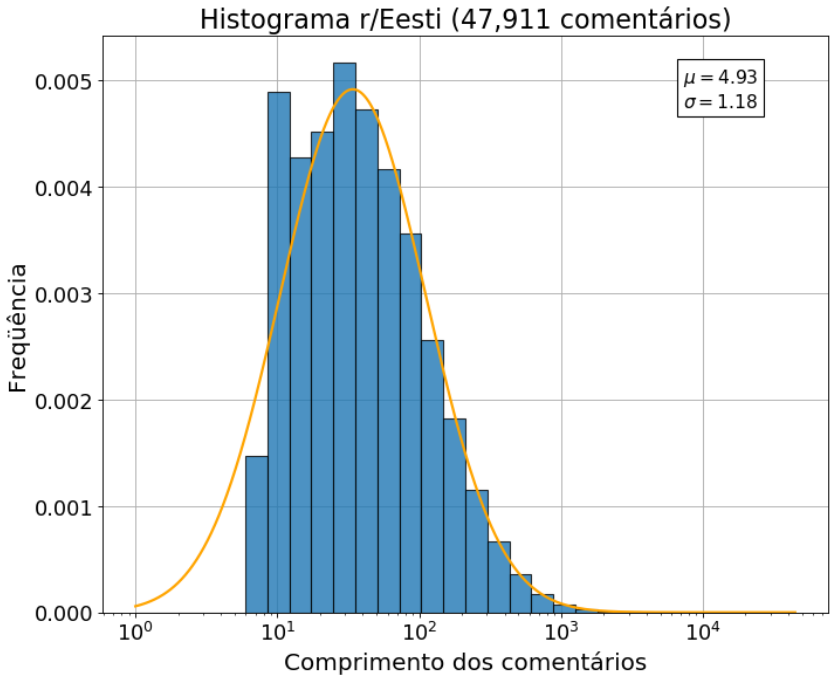
Principais subs japoneses, geral e notícias

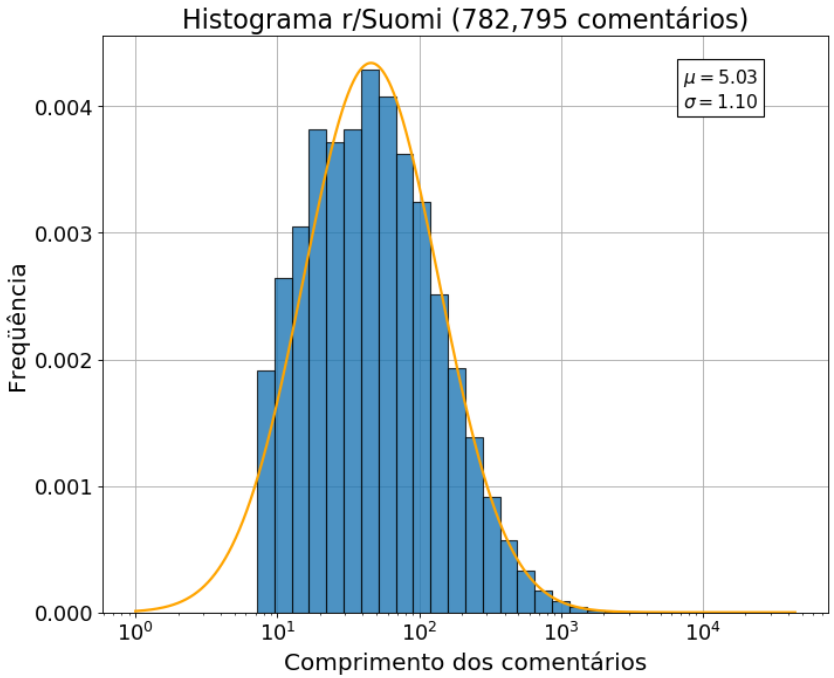




Resultados - Gráficos

Principais subs, Estônia e Finlândia





Resultados(discussão)

- Os comentários seguem uma log-normal de forma bastante clara.
- Os parâmetros μ e σ das distribuições não parecem refletir similaridades entre idiomas.
- De fato comunidades falantes do mesmo idioma podem ter distribuições com parâmetros extremamente discrepantes(eg Argentina e Colombia)
- Há a possibilidade de os parâmetros representarem bem diferenças em sistemas de escrita(cf comunidades japonesas).

Agradecimentos

- À comunidade de desenvolvedores do Reddit.
- Em particular ao /u/GoldenSights pelo auxílio e por disponibilizar o timesearch.