Grandes quantidades de informação: um olhar crítico

Humanidades Digitais HD-Rio 2020/2021

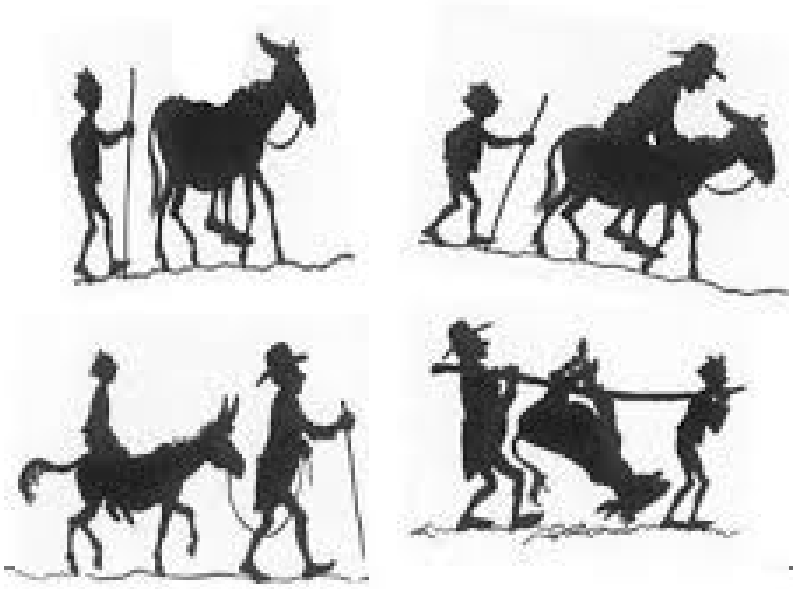
Diana Santos

d.s.m.santos@ilos.uio.no

14 de abril de 2021

Conteúdo

- Preâmbulo: o significado de *grande*
- A minha atuação como apóstola dos grandes corpos
- Um olhar crítico "de dentro"



Se olharmos para a língua...

1. A palavra *grande*, como a esmagadora maioria dos adjetivos em qualquer língua, tem um significado contextual.
   uma *formiga grande* é muito menor do que um *elefante pequeno*

Se olharmos para a língua...

1. A palavra *grande*, como a esmagadora maioria dos adjetivos em qualquer língua, tem um significado contextual.
   uma *formiga grande* é muito menor do que um *elefante pequeno*
2. *grande* pode significar tamanho (medida quantitativa), abrangência ou grandiosidade (qualitativas)
   uma *obra grande* é diferente de uma *grande obra*

Se olharmos para a língua...

1. A palavra *grande*, como a esmagadora maioria dos adjetivos em qualquer língua, tem um significado contextual.
   uma *formiga grande* é muito menor do que um *elefante pequeno*
2. *grande* pode significar tamanho (medida quantitativa), abrangência ou grandiosidade (qualitativas)
   uma *obra grande* é diferente de uma *grande obra*
3. *grande*, fazendo parte do sistema avaliativo, pode modificar coisas com conotação positiva ou negativa
   sentiu uma grande alegria ou *sentiu uma grande tristeza*
   a Grande Guerra, *Alexandre o Grande*

Se olharmos para a língua...

1. A palavra *grande*, como a esmagadora maioria dos adjetivos em qualquer língua, tem um significado contextual.
   uma *formiga grande* é muito menor do que um *elefante pequeno*
2. *grande* pode significar tamanho (medida quantitativa), abrangência ou grandiosidade (qualitativas)
   uma *obra grande* é diferente de uma *grande obra*
3. *grande*, fazendo parte do sistema avaliativo, pode modificar coisas com conotação positiva ou negativa
   sentiu uma grande alegria ou *sentiu uma grande tristeza*
   a Grande Guerra, *Alexandre o Grande*
4. mas *grande* é por omissão positivo em relação a *pequeno/a*, o seu antónimo natural
   O príncipezinho, *O pequeno príncipe*

Diana Santos (UiO) Grande 14/4/2021 3 / 23

Se olharmos para a língua...

1. A palavra *grande*, como a esmagadora maioria dos adjetivos em qualquer língua, tem um significado contextual.
   uma *formiga grande* é muito menor do que um *elefante pequeno*
2. *grande* pode significar tamanho (medida quantitativa), abrangência ou grandiosidade (qualitativas)
   uma *obra grande* é diferente de uma *grande obra*
3. *grande*, fazendo parte do sistema avaliativo, pode modificar coisas com conotação positiva ou negativa
   sentiu uma grande alegria ou *sentiu uma grande tristeza*
   a Grande Guerra, *Alexandre o Grande*
4. mas *grande* é por omissão positivo em relação a *pequeno/a*, o seu antónimo natural
   O príncipezinho, *O pequeno príncipe*
5. em termos de frequência (medida em 1.500.000.000 palavras em português)
   2,3 milhões de *grande*, 900 mil de *pequeno*

Grandes quantidades de informação

O que significa?

- O valor de *grande* varia com o tipo de informação

Grandes quantidades de informação

O que significa?

- O valor de *grande* varia com o tipo de informação
- O conceito varia com o tempo – cada vez maior! Crescimento exponencial?

Grandes quantidades de informação

O que significa?

- O valor de *grande* varia com o tipo de informação
- O conceito varia com o tempo – cada vez maior! Crescimento exponencial?
- Existe um limite a partir do qual mais já não é preciso? Ou a partir do qual não é possível aprender?

Grandes quantidades de informação

O que significa?

- O valor de *grande* varia com o tipo de informação
- O conceito varia com o tempo – cada vez maior! Crescimento exponencial?
- Existe um limite a partir do qual mais já não é preciso? Ou a partir do qual não é possível aprender?

Resultados da procura

14 de abril de 2021

Procura: **[lema="grande|pequeno"] [lema="quantidade"]**
Distribuição de **lema**
Corpo: os corpos todos v. 7.2

29534 casos.

Distribuição

Houve **2** valores diferentes de **lema**.

grande	22559
pequeno	6975

Para informação sobre os códigos internos do atributo **lema**, consulte a página de anotação.

[voltar]
[nova pesquisa]

Acesso a grandes quantidades de...

Procura: **[lema="grande|pequeno"] [lema="quantidade"] "de" @[pos="N.*"]**
Distribuição de **lema**
Corpo: os corpos todos v. 7.2

19609 casos.

Distribuição

Houve **2880** valores diferentes de **lema**. Apresentam-se apenas 999, por ordem decrescente de frequência.

água	637
informação	547
material	392
dado	384
energia	328
alimento	232
matéria	232
gordura	206
resíduo	192
produto	186

recurso	174

Se olharmos para os corpos linguísticos

- grande em 1970:
  - Português fundamental: 700 mil palavras, entrevistas orais
  - Brown corpus: 1 milhão de palavras, escritas
- grande em 1990: WSJ: 25 milhões de palavras
  - BNC: 100 milhões de palavras
  - Hansards: 2,1 milhões de pares de frases
  - Susanne: floresta de 192 mil palavras
- grande em 2000
  - Bosque: 9.368 frases (correspondendo a 212 mil palavras)
  - COMPARA: corpo paralelo com 1,5 milhões de palavras em cada língua
  - WPT-2005: 7 milhões de documentos
  - Corpus brasileiro: 1 bilhão de palavras
- grande em 2010
  - Common crawl: 38,5 milhões de páginas do domínio `.br`
  - Google books dataset: 3 milhões de livros
- grande em 2020 ?

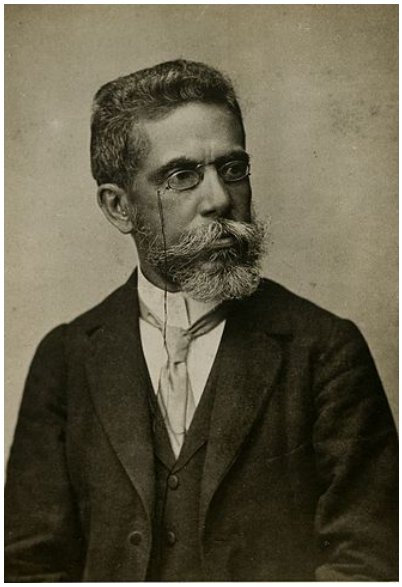
Outra visão de *grande*, agora nos estudos literários

- a obra toda de X
- OBras: romances brasileiros no domínio público: 85
- Literateca: romances em português: 339
- ELTeC: 1.300 romances em 18 línguas
- 3.821 livros em inglês (literatura e biografia) de 1700 a 2000 (Underwood, 2019)
- 93.960 livros da HathiTrust Digital Library (Underwood, 2019)
- Google books dataset: 3 milhões de livros

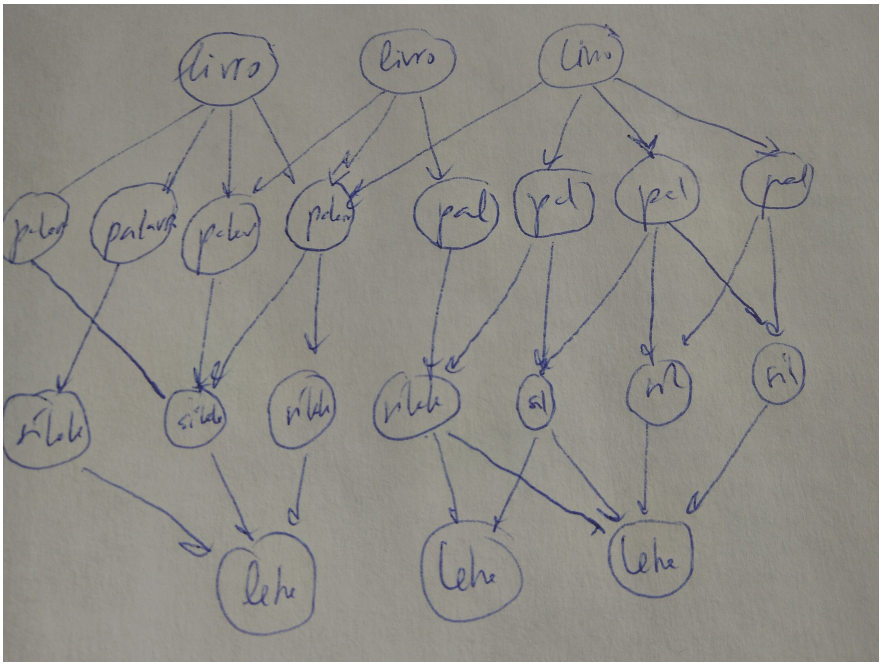
Importante perceber o que se conta

Um problema no OBras:

- quantas obras?
- depende do que se conta: volumes, ou obras literárias?
- particularmente dramático para livros de contos
- Machado de Assis tem neste momento 162 "obras" no OBras, embora tenha apenas cerca de 3 vezes mais palavras do que Coelho Neto (15) e 4 vezes mais palavras do que Aluísio Azevedo (7 obras)
- Machado de Assis tem 129 contos, 1 novela, 12 romances (2 traduções), 17 crónicas e 3 peças de teatro



O que contamos?



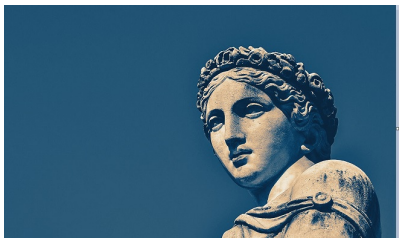
Autopromoção

Responsável pela criação e desenvolvimento da Linguateca (1998-), cujo objetivo é criar recursos públicos para o processamento da língua portuguesa. Muitos dos maiores corpos e recursos públicos para o português foram criados em colaboração com a Linguateca, ou são disponibilizados por ela. `http://www.linguateca.pt`

Mas: e parecendo premonitório, em Santos (1999) escrevi que em geral há três comunidades associadas aos corpos: os compiladores, os desenvolvedores de ferramentas, e os usuários. É muito raro que sejam o mesmo grupo...

Apóstola? Confesso

1. O uso de corpos para compreender a língua – colocações, correlações, exemplificações: por exemplo Santos (2007)
2. O uso de corpos para compreender o mundo – tomando os documentos como informação sobre o mundo: por exemplo Santos (2014)



Duas tarefas estanques?

Não são duas tarefas estanques, naturalmente, mas podemos arranjar facilmente perguntas que apenas se referem a uma das vertentes:

1. existe na língua portuguesa uma preferência por orações relativas em que o pronome tem a função de sujeito, ou de objeto?
2. quantos/que autarcas portugueses foram acusados de corrupção no Minho na década de 90?
3. a partir de quando se começou a falar dos direitos dos homossexuais nos jornais de grande tiragem?

De facto, existem vários domínios científicos que se preocupam com estas questões de maneira diferente. Em particular, a **linguística com corpos**, **a extração de informação**, e o **garimpo/mineração de textos**.

Uma reflexão sobre as definições de Hearst (2003)

Linguística com corpos Estudo da língua usando corpos

Extração de informação Obtenção de informação estruturada a partir de texto

Garimpo de texto Descoberta de informação implícita em grandes quantidades de texto

No primeiro caso, o corpo é uma ferramenta fundamental para a análise da língua. Nos dois outros casos, não precisamos de corpos em sentido estrito. Mas: podemos fazer extração de informação linguística... e garimpo linguístico de texto...
E podemos enriquecer os corpos com a informação que deles extraímos, de forma a obter um recurso mais rico.

O que é um corpo?

- Uma colecção **classificada** de **objectos linguísticos** para **uso** em PLN/LC/L
- Uso: estudo, medição, teste, avaliação
- Objectos linguísticos: textos, frases, palavras, entrevistas, erros ortográficos, entradas de dicionário, citações, pareceres jurídicos, filmes, imagens com legendas, traduções, correcções, telefonemas, WOZ, programas ...

Santos (2006), Santos (2008)

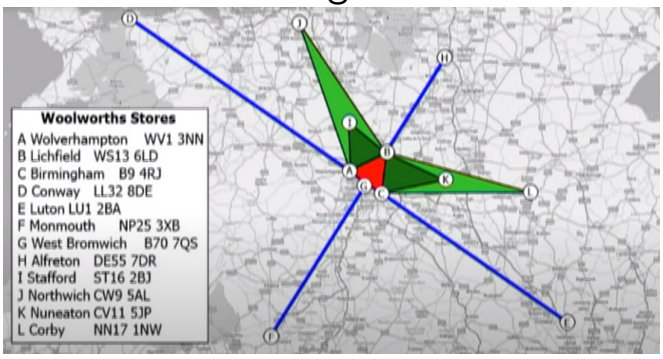
Antídoto: uma posição crítica

Há duas questões para que quero chamar a atenção:

1. é sempre necessário mais material para ter melhores resultados?
2. com um grande número de pontos/dados, é sempre possível encontrar padrões/regularidades!
   Ver Matt Parker:
   `https://www.youtube.com/watch?v=sf5OrthVRPA`

O problema de demasiados dados

É sempre possível encontrar qualquer padrão se tivermos dados suficientes.
Matt Parker: triângulos exatos...

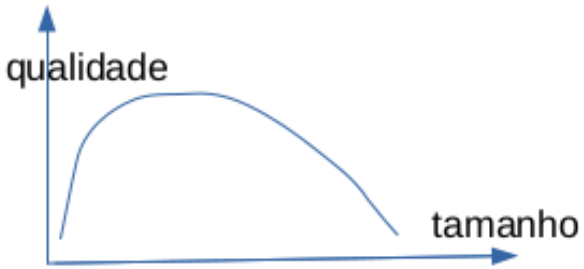


O pesadelo dos estatísticos, e o companheiro das teorias da conspiração.

Quantos trabalhos seriam melhores se usassem mais dados?

- Não é muito comum fazer esta pergunta. Pelo contrário, grande parte dos trabalhos publicados diz que pretende confirmar os resultados em maiores quantidades.
- Mas é raro ver trabalhos de confirmação – replicação, que juntem muito ao trabalho inicial.
- Muitas vezes, olhar para "poucos" dados com atenção dá mais resultados do que olhar para muito mais dados com menos atenção.

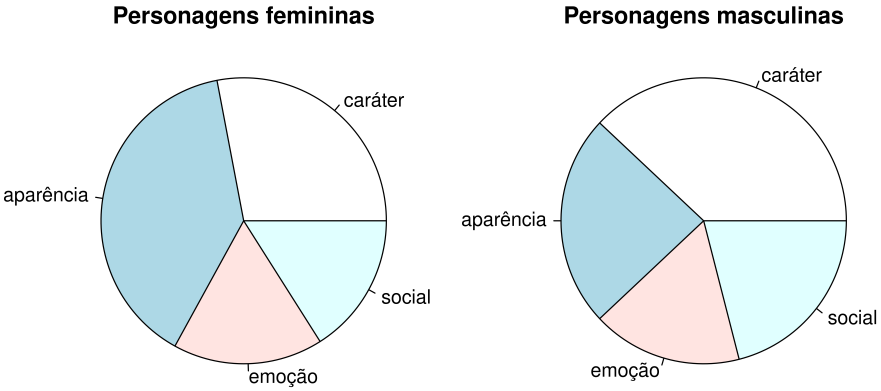
A "mesma" pergunta, ou a oposta, é: de quantos dados eu preciso para propor uma dada hipótese com alguma confiança?



Um caso como exemplo

Silva (2021, p.78) usa o OBras de 2019 (com 248 obras) para

- estudar os predicadores adjetivos em relação a personagens femininas e masculinas:



- detetar as palavras mais frequentes atribuídas a homens e a mulheres.

só mulheres $<$ mais a mulheres que a homens $<$ mais a homens que a mulheres $<$ só homens

formosa $<$ ... $<$ *bela* $<$ ... $<$ *mau* $<$ ... $<$ *honrado*

cogitar $<$ ... $<$ *exercer* (atração) $<$ ... $<$ *ganhar* $<$ ... $<$ *montar*

Não estaremos a tentar matar moscas com um canhão?

- Agora a moda, em PLN, é usar palavras pulverizadas ("word embeddings") – que são modelos estatísticos da coocorrência de palavras, treinados em grandes quantidades de texto – para qualquer tarefa.
- Mesmo para tarefas que os seres humanos fazem com muito menos esforço, possivelmente usando um conjunto muitíssimo mais reduzido de características, que correspondem a conceitos e generalizações a um nível muito mais elevado.

Além da falta de eficiência, e do gasto de recursos (um mata-moscas é incomensuravelmente mais barato do que um míssil), estamos a dar o monopólio à Google e à Amazon – só elas têm dados de dimensão suficiente.

Concluindo, ou desconcluindo

- É importante uma reflexão sobre o que fazemos, e quais as suas consequências: científicas, e políticas.
- É importante equacionarmos o equilíbrio entre quantidade e qualidade, para cada tarefa ou problema.

Referências

- Hearst, Marti. What is Text Mining? Berkeley, 2003.
- Santos, Diana. Disponibilização de corpora através da WWW. In Palmira Marrafa & Maria Antónia Mota (eds.), *Linguística Computacional: Investigação Fundamental e Aplicações*, Colibri, 1999, pp. 323-346.
- Santos, Diana. Desenho, construção e aplicação de corpora. *Primeira Escola de Verão da Linguateca*, Universidade do Porto, Julho de 2006.
- Santos, Diana. Breves explorações num mar de língua. *Ilha do Desterro* 52, 1, Jan/Jun 2007, pp. 127-150.
- Santos, Diana. Corporizando algumas questões. In Stella E. O. Tagnin Oto Araújo Vale (orgs.), *Avanços da Lingüística de Corpus no Brasil*, Editora Humanitas/FFLCH/USP, São Paulo, 2008, pp. 41-66
- Santos, Diana. Procura em grandes quantidades de texto ... e suas consequências. Apresentação na UERJ, 9 de outubro de 2014.
- Silva, Flávia Martins da. Caracterização e personagens na literatura brasileira quanto ao gênero: uma proposta metodológica. Tese de mestrado, PUC-Rio. Abril de 2021.

Imagens

https://upload.wikimedia.org/wikipedia/commons/4/40/Machado_de_Assis_aos_57_anos.jpg
http://guiaturismobrasil.eco.br/wp-content/uploads/2020/03/15192790_604176959767387_6779047347616211245_n5.jpg
http://rollene.no/wp-content/uploads/2020/08/junia-en-kvinnelig-apostel.jpg http://www.martharobles.com/blog/day/month/year/burocracia-cultural
https://fotos.web.sapo.io/i/oea1289ac/19959114_juAis.jpeg
https://www.spondylitten.no/wp-content/uploads/2015/11/Sp%C3%B8rsm%C3%A5l-tastatut-300x219.jpg

Perguntas? Opiniões?