Humanidades digitais para estudantes de História: algumas observações

Diana Santos

Linguateca & Universidade de Oslo d.s.m.santos@ilos.uio.no

Picture from https://lithub.com/the-places-we-read/

ISCTE, 16 de dezembro de 2021

Plano da apresentação

- Minha apresentação
- Grandes quantidades de dados
  - Questões políticas: a quem aproveita a "Big data"?
  - Recolha de informação, extração de informação, prospeção de dados
- Técnicas/áreas usadas nas HD
  - Leitura distante
  - Métodos estatísticos
  - Aprendizagem automática
  - Redes de personagens
  - Bases de dados
  - Sistemas de informação geográficos
  - Modelos de tópicos
- A luta entre a anotação/interpretação e a virgindade dos dados

Minha apresentação

- Engenheira eletrotécnica, Instituto Superior Técnico, 1985
- Mestrado em tradução automática (1988), doutoramento em tradução humana (1996) (área: processamento de linguagem natural)
- Desde 1998 lidero a Linguateca, para o desenvolvimento e avaliação de sistemas e recursos para o processamento computacional da língua portuguesa (financiada até 2011 pelo MCT português)
- Desde 2011 sou professora na Universidade de Oslo (de Linguística Portuguesa), e ensinei as cadeiras "Estatística para linguística e literatura" (2014-2018) e "Humanidades digitais" (2022-)
- Participo na ação COST "Distant reading for European literary history"
- Tenho dado algumas palestras sobre temas associados

Subjetivamente: sempre adorei História.

Humanidades digitais

- Uma definição por investigador...
- A minha: grandes quantidades de dados exigem uma abordagem informática, no sentido de usar o computador como ajudante
- Política científica: quem ganha? quem manda?
- não precisamos de interdisciplinaridade, mas sim de multidisciplinaridade (para todos os cursos, é preciso aprender informática, "pensamento digital")
- mas essa informática é diferente para cada curso... e deverá ser ensinada pelas pessoas que são desse curso ou que têm suficientes conhecimentos da área
- e do digital também faz parte o estatístico, idem idem aspas aspas

Alves, Daniel. "Humanidades Digitais e História: da evolução geral ao caso português". In I. Galina Russell, M. Peña Pimentel, E. Priani Saisó, J. F. Barrón Tovar, D. Domínguez Herbón, & A. Álvarez Sánchez (Eds.), *Humanidades digitales: recepción, institucionalización y crítica*, pp. 163-194. Bonilla Artigas Editores.

Não sou apóstola do digital

- A tecnologia é invasora, e não necessariamente inovadora ou libertadora.
- O problema é que muitas das decisões sobre como queremos fazer a nossa investigação, ou a nossa vida, não dependem (só) de nós
- A educação, neste caso num doutoramento, deve apresentar um leque vasto de possibilidades, de forma a não comprometer os futuros dos estudantes – que cada vez são menos previsíveis
- A postura com que se recebe uma tecnologia deve ser crítica, e pessoal.

Mas sou apologista do conhecimento

E portanto acho que tenho o dever de ensinar, ou pelo menos mencionar, as seguintes técnicas/áreas do saber:

- leitura distante
- técnicas estatísticas
- aprendizagem automática
- modelos de tópicos
- bases de dados
- redes de personagens
- sistemas de informação geográfica

assim como a importância da Wikipédia.

Grandes quantidades de informação : um olhar crítico

Um olhar crítico "por dentro":
`https://www.youtube.com/watch?v=Qi-3QzP0NxM` (3:45-42:10)

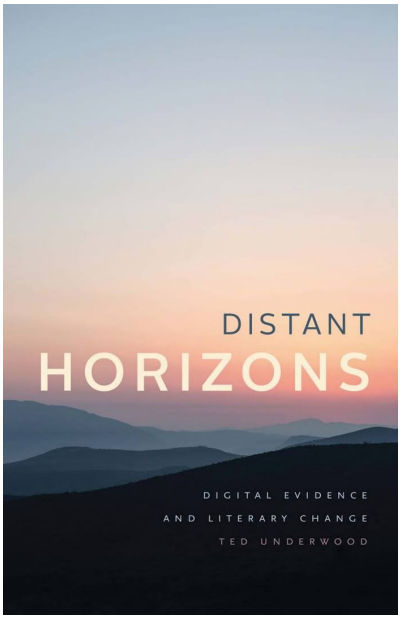
- O significado de *grande* na língua
- O significado de *grande* associado a informação
- O significado de *grande* associado a corpos
- O significado de *grande* nos estudos literários
- O que é que se conta?
- Diferença entre a linguística com corpos, a extração de informação, e o garimpo de dados.
- Quantos trabalhos seriam melhores se usassem MENOS dados?
- Consequências políticas de algumas opções tecnológicas

Leitura distante - no campo da história da literatura

- Argumento surgido inicialmente na literatura, por Franco Moretti, na chamada "Guerra dos cânones"
- Grandes quantidades ("grande" sendo subjetivo) – donde a extração de características sobre cada obra
- Essencialmente estudos comparativos (em vez de substituir um cânone por outro, comparar vários conjuntos). A cada estudo, a sua amostra!
- Todas as coleções têm faltas (segundo Underwood, mesmo os 16 milhões de livros do HathiTrust não incluem metade dos livros de outra coleção)
  - metadados extremamente deficientes
  - quase impossível de separar automaticamente prefácios etc

Research on this scale is about measuring error, not eliminating it. (p. 182)

Leitura distante segundo Underwood



- investigação em grande escala, quantitativa
- não necessariamente para salvar os esquecidos (moralismo de Moretti)
- outra forma de olhar: lentes que abarcam maiores períodos

  Can distant readers write quantitative literary history that is nevertheless detailed enough, streamlined enough, and lively enough to interest a wide range of readers? If we can't, then no argument will save us: what we are doing may be important, but it will belong in the social sciences. (Underwood, 2019, p. xxii)

Leitura distante segundo Underwood, cont.

- leitura distante traz novos objetos de estudo: padrões ainda sem nome
- exemplo: aumento da referência a cores durante o século XIX –> passagem de contar para mostrar ("telling to showing")? E então?
- partamos antes de temas literários, obtenhamos uma hipótese, e tentemos testá-la. Qual a diferença entre uma obra literária e uma biografia? Um contínuo entre obras imaginárias e reais. Fig. p. 23, 3832 obras

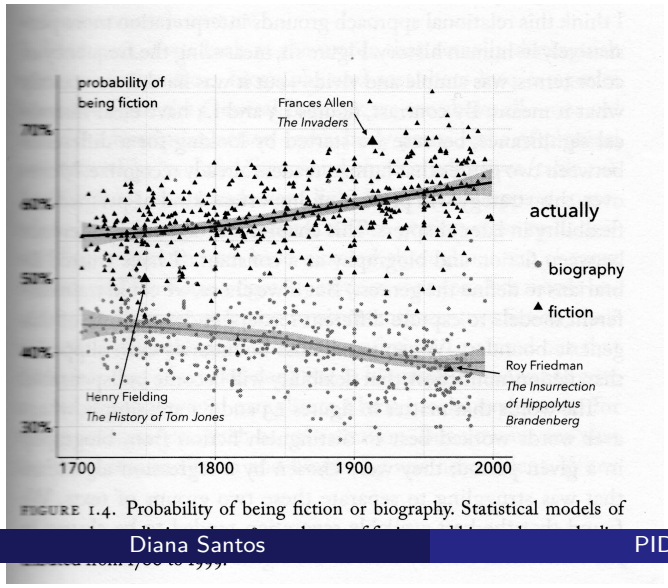
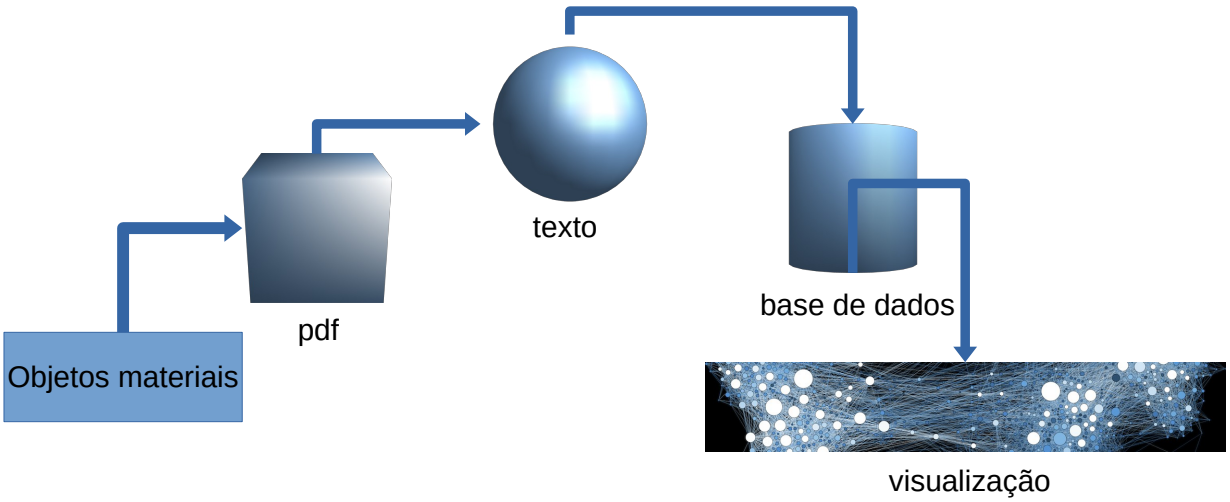


FIGURE 1.4. Probability of being fiction or biography. Statistical models of

Importante: Diferentes objetos digitais

Fases e objetos diferentes



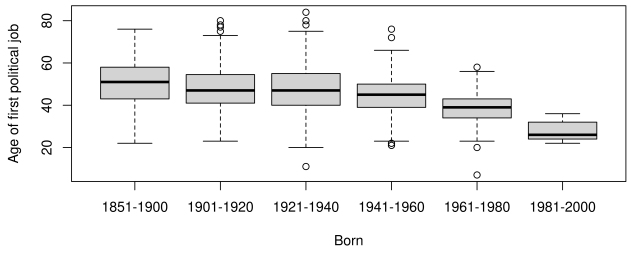
Leitura distante segundo Higuchi et al. (2019, 2022)

- E se olhássemos agora para materiais em português, que têm a ver com História, e usássemos, como unidade, não a obra mas o verbete (de uma enciclopédia)?
- Conversão da enciclopédia num corpo linguístico
- Anotação e definição de novos campos, e sua revisão
  - unificação de diversos nomes de políticos
  - adição do campo semântico da família
- Extração de informação
- Agregação e visualização

Leitura distante segundo Higuchi et al. (2019, 2022)

Exemplos de visualizações:

- um mapa do Brasil com a origem dos políticos
- uma rede de contactos entre políticos (e uma rede de famílias)
- um gráfico com as profissões ao longo do tempo
- número de filhos de políticos homens e mulheres
- idade de entrada na política



Local	*Nasc.*	*Morte*
Rio de Janeiro	1039	1506
São Paulo	126	257
Recife	211	81
Salvador	201	81
Porto Alegre	163	86
Belo Horizonte	95	104
Fortaleza	131	43
Niterói	102	68
Curitiba	86	82
Campos	137	14

Informação geográfica na literatura

- Relações entre literatura e História... podem usar-se romances para estudar história? ou história das ideias?
- O exemplo seguinte é apenas para problematizar algumas questões
  - Nem todas as "localizações" são lugares
  - A anotação de lugares depende do objetivo
  - A anotação exige sempre interpretação
  - as palavras em linguagem natural são vagas (Lisboa é uma cidade, é Grande Lisboa, é o círculo eleitoral de Lisboa, é a região demarcada de Lisboa, é a dicoese de Lisboa, é o distrito de Lisboa, ...?)

Vamos usar os 100 romances portugueses do período 1840-1919 que representam o português na coleção ELTeC.

Aljubarrota: in which works does it occur? (186 cases)

Distribuição

Houve **17** valores diferentes de **obra**.

A_Ala_dos_Namorados	110	271393
O_soldado_de_Aljubarrota	42	42498
O_Monge_de_Cister_I	6	69542
O_Monge_de_Cister_II	6	81895
A_Ilustre_Casa_de_Ramires	4	130722
Os_quatro_reis_impostores	4	156385
Febo_Moniz	3	81649
Arzila:_Romance_do_Século_XV	2	50783
A_Casa_dos_Fantasmas	1	108552
A_ermida_de_Castromino	1	95729
A_morte_vence	1	74311
O_Anel_Misterioso:_Cenas_da_Guerra_Peninsular	1	66351
O_Conde_de_Castel_Melhor	1	235046
O_Manuelinho_de_Évora	1	55402
Os_Maias	1	265179
Os_tripeiros:_Crónica_do_século_XIV	1	44785
Providência	1	85356

Aljubarrota: which authors mention it by name?

Distribuição

Houve **15** valores diferentes de **autor**.

AntCamJ	110	271393
MatBet	42	42498
AleHer	12	1484356
EcaQue	5	2773552
MarMes	4	181339
OliMar	3	81649
BerPPin	2	50783
AJCL	1	44785
AlbPim	1	241132
AntATVas	1	95729
AntFBar	1	198230
AugSar	1	85356
JoJoGra	1	74311
JoadCam	1	235046
LARSil	1	108552

Aljubarrota: selected concordances

- (AleHer1) *Há três anos, não longe da morada de meu velho pai, em Aljubarrota, pelejava eu na Ala dos Namorados* Three years ago was I fighting not far away from my old father's abode, in Aljubarrota,
- (AleHer1) *Sim, depois de Aljubarrota, quando no seu castelo de Sintra já não podia ter voz* Yes, after Aljubarrota, when in his Sintra castle he was no longer heard
- (AleHer1) *o bom cavaleiro da ala de Mem Rodrigues nos campos de Aljubarrota* The good gentleman from M.R. battalion in the Aljubarrota battlefield
- (AleHer1) *Estando eu na tenda d' el-rei, naquela noite depois da de Aljubarrota* I was in the king's tent, that night after Aljubarrota's night

Aljubarrota: selected concordances

- (EcaQue1) *E os outros Ramires, o de Silves, o de Aljubarrota, os de Arzila, os da índia !* And the other Ramires: the one from Silves, the one from Aljubarrota, ...
- (EcaQue2) *sua nora não tivera avós mortos em Aljubarrota !* His daughter-in-law could not boast of ancestors dead in Aljubarrota!
- (AntCJ1) *para os lados de Aljubarrota se estreitava consideravelmente* toward vicinity of Aljubarrota it would become narrower
- (AntCJ1) *irão primeiro jantar a Aljubarrota* they will first have dinner in Aljubarrota
- (AntCJ1) *nunca mais depois daquele crepúsculo épico de Aljubarrota* never again since that epic sunset of Aljubarrota

Aljubarrota: selected concordances

- (AntCJ1) *tangera as Ave-marias na igreja matriz de Aljubarrota* he had tolled Ave-maria at Aljubarrota's main church
- (AntCJ1) *resto de algum cerrado feito pelos pastores de Aljubarrota para guarida do gado* the remains of some fence built by Aljubarrota's shepherds...
- (AntCJ1) *o rei, sempre de luto por causa de Aljubarrota,* the king, still mourning because of Aljubarrota
- (AntCJ1) *Valverde foi o corolário de Aljubarrota* Valverde (battle) was the corollary of Aljubarrota (battle)
- (MatBet1) *perto do rio Lena que atravessa a planicie de Aljubarrota* near the river that crosses Aljubarrota's plain

Aljubarrota: selected concordances

- (MatBet1) *estabelecer uma padaria nas visinhanças de Aljubarrota* open a bakery near Aljubarrota
- (MatBet1) *A pobre vida de Aljubarrota não offerecia um abrigo de caridade* The poor life in Aljubarrota did not offer a shelter
- (MatBet1) *destacamento que saía de Aljubarrota* batallion that was leaving Aljubarrota
- (MatBet1) *onde jazia o corpo da sobrinha da padeira de Aljubarrota* where was buried the body of the Aljubarrota baker(woman)
- (OliMar1) *e lhes daremos outra Aljubarrota !* and we will give them another Aljubarrota

Further examples, now from *Lamego*

Let me also bring other cases, now with the place name *Lamego*.

- *cónego de Lamego* - Lamego vicar
- *cortes de Lamego* - Lamego assembly
- *um presunto de Lamego* - a ham produced in Lamego, in the way they are produced there
- *uma menina de Lamego* - a young woman from Lamego
- *uma patrícia dos presuntos de Lamego* - someone from the same place as Lamego ham
- *estrada de Lamego* - road to Lamego
- *descendo de Lamego* - coming down from Lamego (southwards)

Quantitative description

Currently (25 November 2021), the ELTeC-por collection in Literateca (v. 7.5) has

Total tokens	8,162,799
Distinct tokens	213,618
Distinct lemmas	111,539
Total place tokens (revised)	27,698
Total places (revised)	24,510
Distinct places (revised)	1,066
Total place tokens (unrevised)	7,660
Total places (unrevised)	6,206
Distinct places (unrevised)	1,903

Difference between place tokens and places: *Campo Grande* has **two** place tokens, but counts as **one** place.

Quantitative description: kinds of places

Distribuição

Houve **83** valores diferentes de **sema**.

Local:cidade	9998
Local:pais	5671
Local:vila	1276
Local:continente	892
Local:territorio	804
Local:rua	734
Local:relig	567
Local:freguesia	509
Local:regiao	460
Local:municipio	458
Local:bairro	344
Local:provincia	287
Local:organizado	283
Local:aldeia	209
Local:cidade_Local:vila	194
Local:ludico	188
Local:ilha	186

Quantitative description: which places

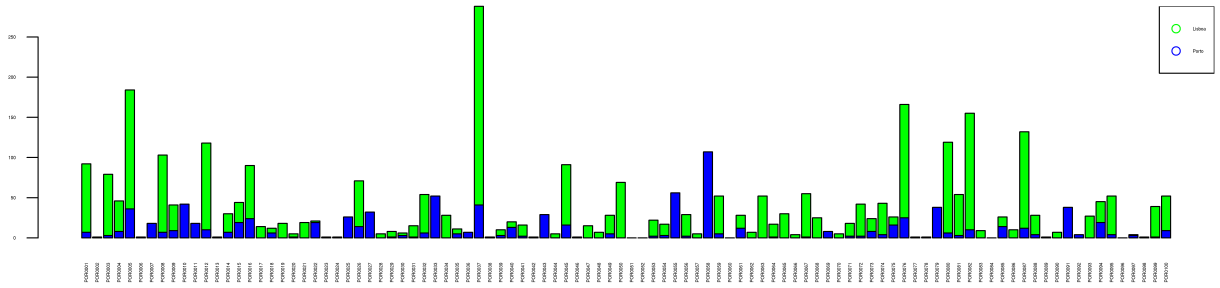
Distribuição

Houve **1049** valores diferentes de **lema**. Apres

Lisboa	3375
Portugal	2429
Coimbra	959
Porto	909
França	714
Paris	699
Castela	476
Inglaterra	451
Europa	415
Roma	386
Hespanha	328
Santarém	315
Evora	303
Guimarães	303
Índia	289
Espanha	277

Presence of Lisbon vs. Porto (absolute)

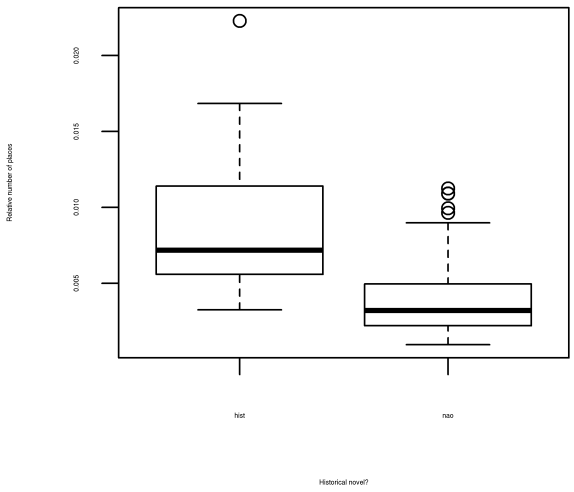
There is often some competition between the capital and the second city, although novelists seem to be well distributed between the two.

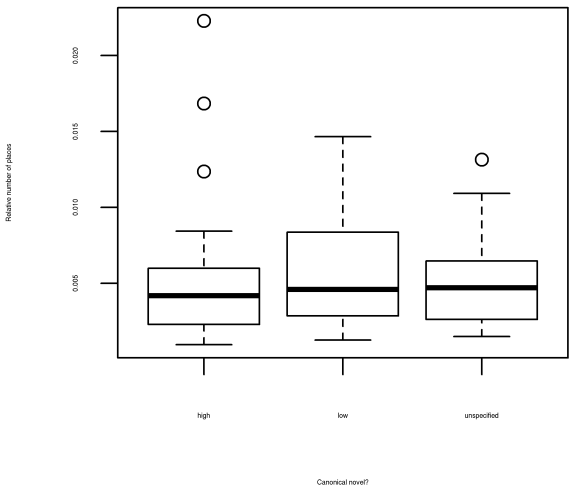


But Lisbon clearly wins in number of mentions.

Do historical novels differ from the others?

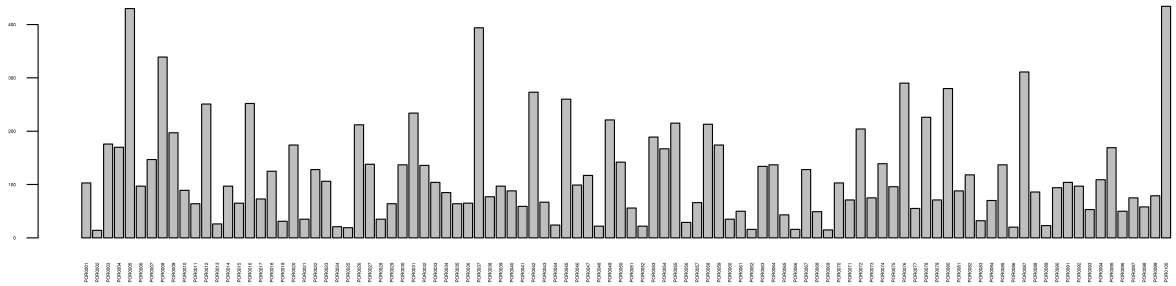
There are 32 historical novels in the ELTeC collection. And there are 26 canonical novels (operationalised by having more than one reprint in the period 1980-2010).

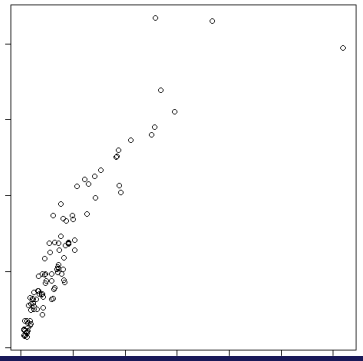




And what about place diversity?

Are there works which are more diverse placewise than others? *A senhora duquesa* (most) and *Sacrificada* (least)





Does the number of distinct places correlate with the number of places? Yes, 0.90

Técnicas estatísticas

A estatística tem três grandes áreas

1. estatística descritiva: para caracterizar (grandes) conjuntos de dados
2. estatística inferencial: para criar modelos de previsão, para testes de hipóteses, para medir correlações
3. estatística exploratória: para descobrir padrões e regularidades

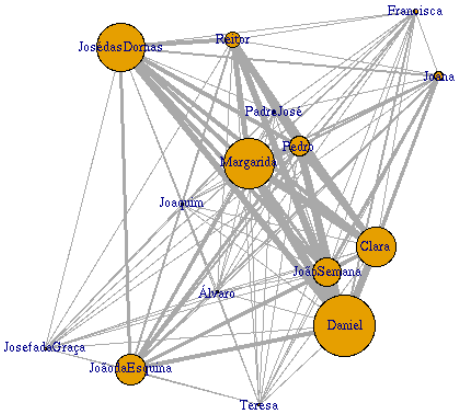
Nas humanidades digitais é geralmente a terceira vertente que é usada.

Bases de dados

- É uma tecnologia para dar conta de informação estruturada sobre muitos objetos.
- Quando cada objeto no "mundo" que queremos modelar tem mais informação associada
- Tecnologia eficiente para lidar com grandes quantidades de dados, sem repetir informação
- Com um conjunto de formas de agregar a informação, também muito eficientes
- E um conjunto de linguagens de interrogação
- Existe também a possibilidade de criar um modelo de base de dados a partir de sistemas de modelação

Muito importante: uma folha de cálculo não é uma base de dados!

Redes de personagens



- Uma outra técnica muito usada nos estudos literários e na História é a das redes (representadas matematicamente por um grafo)
- Embora exista grande conhecimento informático e matemático sobre grafos, as redes nas humanidades digitais são em geral usadas sobretudo para visualização

Modelos de tópicos

- Uma técnica estatística muito utilizada, tanto nos estudos literários como na História, para identificar "temas" recorrentes numa coleção, temas esses representados por um grupo de palavras.
- Um tópico ou tema é "a recurring pattern of co-occurring words."
- Todos os documentos partilham os mesmos tópicos, mas em proporções diferentes. O objetivo é, a partir da coleção de documentos, inferir os tópicos.
- Por um lado, é preciso conhecer um pouco o teor da coleção, para poder fazer um pré-processamento fiável (retirar nomes próprios? retirar siglas?)
- E depois, é preciso interpretar os temas (que são dados por um conjunto de palavras)

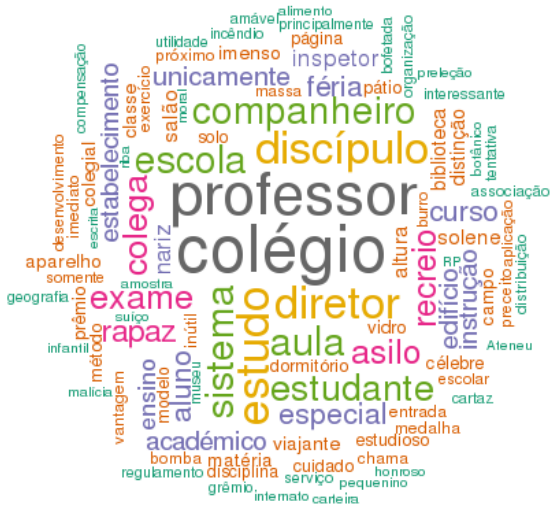
Modelos de tópicos em História

(do Programming historian, `https://programminghistorian.org/en/lessons/topic-modeling-and-mallet`)

- O jornal *Daily Dispatch* de Richmond (EUA) de 1860 a 1865 - 112.000 artigos, 24 milhões de palavras, traduzidos em 40 tópicos
- O diário de Martha Ballard (de 1785 a 1812), com 10 mil entradas. Uma parteira de Maine (EUA)
- O jornal americano Pennsylvania Gazette (1728-1800), 100 mil documentos, 40 tópicos

Exemplos de tópicos em literatura lusófona

Santos et al. 2020 (trabalho muito preliminar) 192 romances, 100 tópicos (só substantivos e adjetivos).
Os bons:



Exemplos de tópicos em literatura lusófona

Santos et al. 2020 (trabalho muito preliminar) 192 romances, 100 tópicos (só substantivos e adjetivos).
Os maus

Outras técnicas em humanidades digitais

De que eu não vou falar, mas que é provavelmente importante mencionar

- realidade virtual
- jogos
- literatura digital
- sonificação
- grafos de referências
- semântica distribucional e palavras pulverizadas ("word embeddings")

Anotação ou dados virgens

Está tudo latente, ou podemos e devemos marcar a nossa interpretação?

- PLN ou RI? Em inúmeras subáreas
- N-grams ou pares com relações sintáticas
- em leitura distante: adicionar conhecimento, ou deixar o modelo descobrir?
- depende do tamanho dos dados?

Qual é o problema de fazer anotação?

- É que introduz erros
- Pode introduzir vieses
- É mais difícil de avaliar (qual a contribuição do método, e dos erros de anotação?)
- Pode levar a que se descubra o que se quer descobrir

Para finalizar

Mensagens principais

- O digital é um "método" que cada vez mais é usado em História (e nas "ciências humanas" em geral)
- Há diferentes métodos digitais para diferentes disciplinas: história digital é diferente de linguística digital, de estudos literários digitais, de física digital ou de sociologia digital
- É preciso muito trabalho para fazer qualquer projeto, seja ele digital ou não (o que se poupa num lado, gasta-se noutro)
- O sentido crítico é sempre essencial na investigação
- Na minha opinião, a maior vantagem das humanidades digitais é a possibilidade de partilha, e portanto de interação com outros investigadores, que podem por sua vez enriquecer os nossos dados e desafiar as nossas hipóteses

Se quiserem ter impacto, escrevam na Wikipédia!

Referências

- Blei, David M., Andrew Y. Ng & Michael I. Jordan. "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3 (2003), pp. 993-1022.
- Graham, Shawn, Ian Milligan & Scott Weingart. *Exploring Big Historical Data: The Historian's Macroscope*. Imperial College Press. 2015
- Higuchi, Suemi, Diana Santos, Cláudia Freitas & Alexandre Rademaker. "Distant reading Brazilian history". *Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries (Copenhagen, March 6-8 2019)*, 2019, pp. 190-200.
- Santos, Diana, Emanoel Pires, João Marques Lopes, Rebeca Schumacher Fuão & Cláudia Freitas. "Periodização automática: Estudos linguístico-estatísticos de literatura lusófona". *Linguamática* 12 (1), 2020, pp. 80-95.
- Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press. 2019

Obrigada!