Linguateca & University of Oslo, Norway
Linguateca & 2Ai Lab, IPCA, Portugal

A Literateca

e algumas perguntas a que pode responder

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

Alberto Simões
asimoes@ipca.pt

October 23, 2019

Roteiro da apresentação

Literateca

- Introdução

Conteúdo

- Obras
- Anotação
- Tamanho

Exemplos de estudos

- Emoções ao longo do tempo
- Perfis de descrição da fala
- Referência a roupa
- Saúde, dor, e a profissão de médico

Mais pormenores

- Um ambiente para estudar textos em português para estudos linguísticos e literários

Literateca
Apresentação

- Um ambiente para estudar textos em português para estudos linguísticos e literários
- Uma infraestrutura sobre o Open CWB, um ambiente de processamento de corpos, que contém vários níveis de anotação

Literateca
Apresentação

- Um ambiente para estudar textos em português para estudos linguísticos e literários
- Uma infraestrutura sobre o Open CWB, um ambiente de processamento de corpos, que contém vários níveis de anotação
- Com uma interface embrionária para programs de computacção estatística e sua vizualização, em R

Conteúdo
Obras

A Literateca contém os seguintes tipos de obras

- Clássicas desde 1380, graças ao projeto Vercial e ao Tycho-Brahe
- Texts literários canónicos, graças aos mesmos e ao COLONIA e ao OBras
- Textos literários não canónicos, graças ao COST
- Excertos de textos literários traduzidos para outras línguas (norueguês e eventualmente inglês)

Conteúdo
Anotação

Todos os textos foram anotados pelo PALAVRAS, o mais antigo e experimentado analisador sintático para o português (Bick, 2000). Além disso:

Todos os textos foram anotados pelo PALAVRAS, o mais antigo e experimentado analisador sintático para o português (Bick, 2000). Além disso:

Anotação semântica

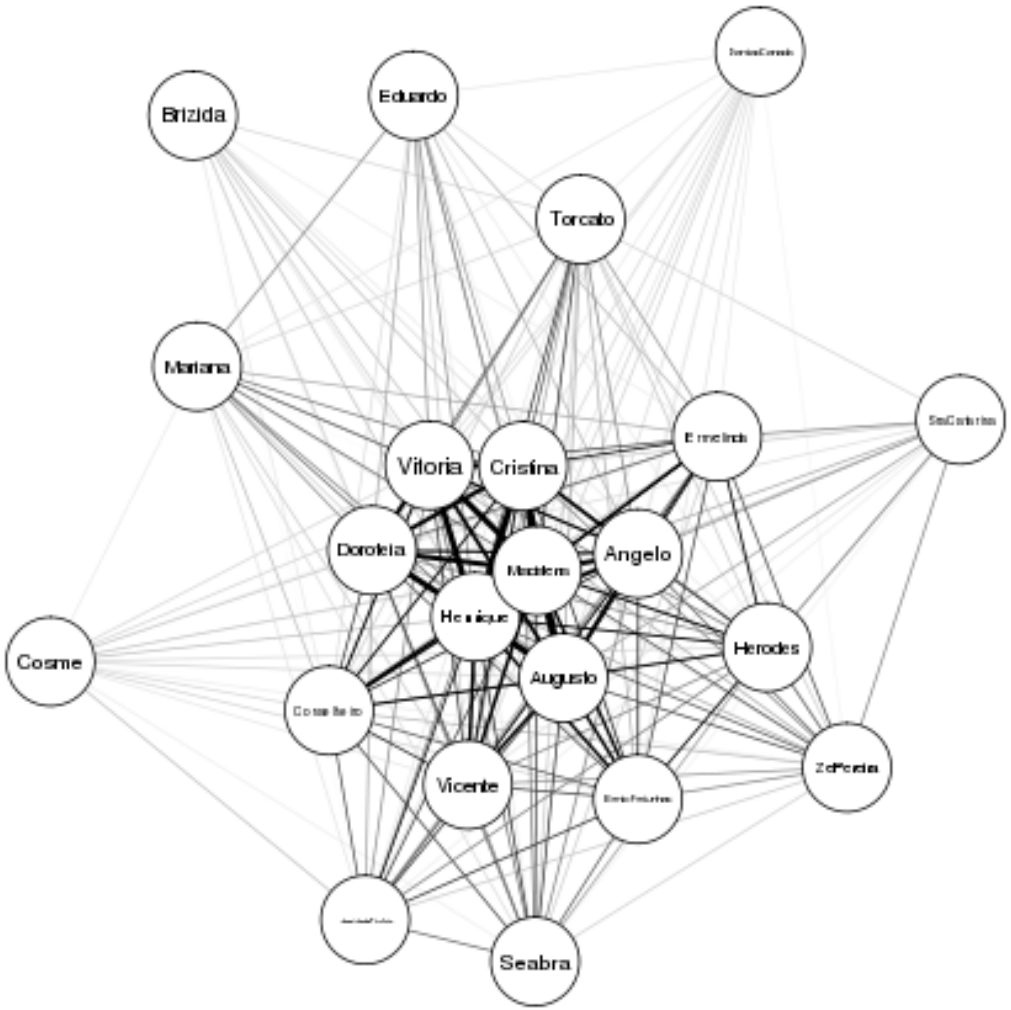
- Cores, roupa, corpo, família, emoções e saúde foram anotados
- e revistos (parcialmente)

Questões "literárias"

- Metadados como género do autor, género literário e escola têm sido adicionados
- Categorias de entidades mencionadas como Pessoa, Lugar e Obra têm sido revistas
- Identificação de personagens foi feita (até agora) para 7 obras

Contents

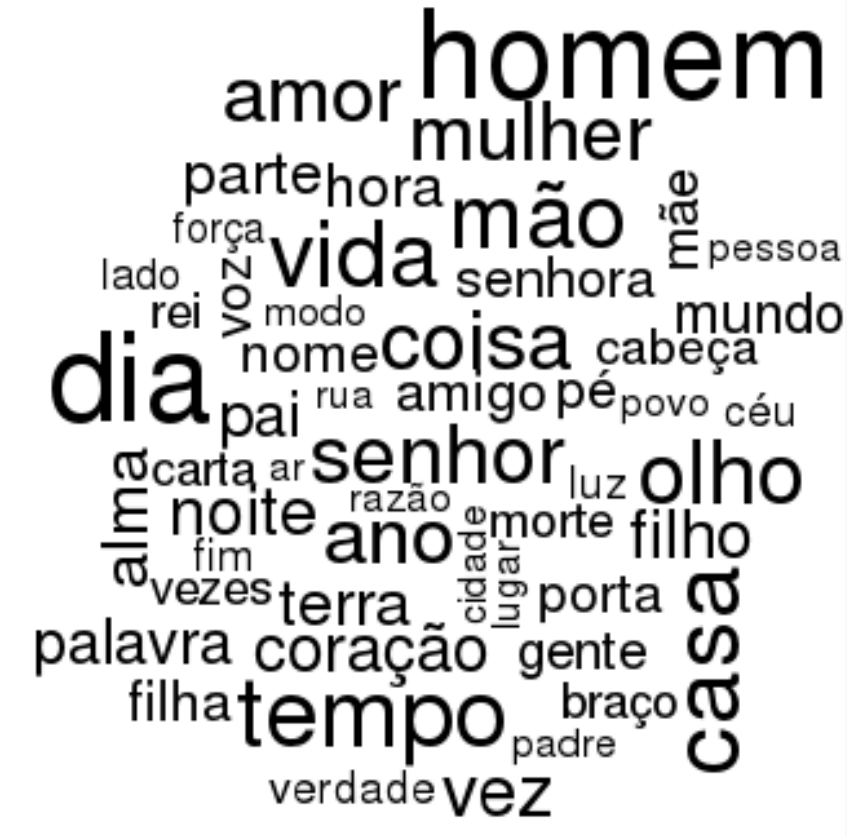
Exemplo de rede de personagens



Conteúdo
Tamanho

Números correspondendo à versão 2.7 (20 de outubro de 2019)

- 28,6 milhões de palavras
- 785 obras de 197 autores
- 189 romances ou novelas do "período COST"
- 7 romances com personagens anotadas



Emoções ao longo do tempo

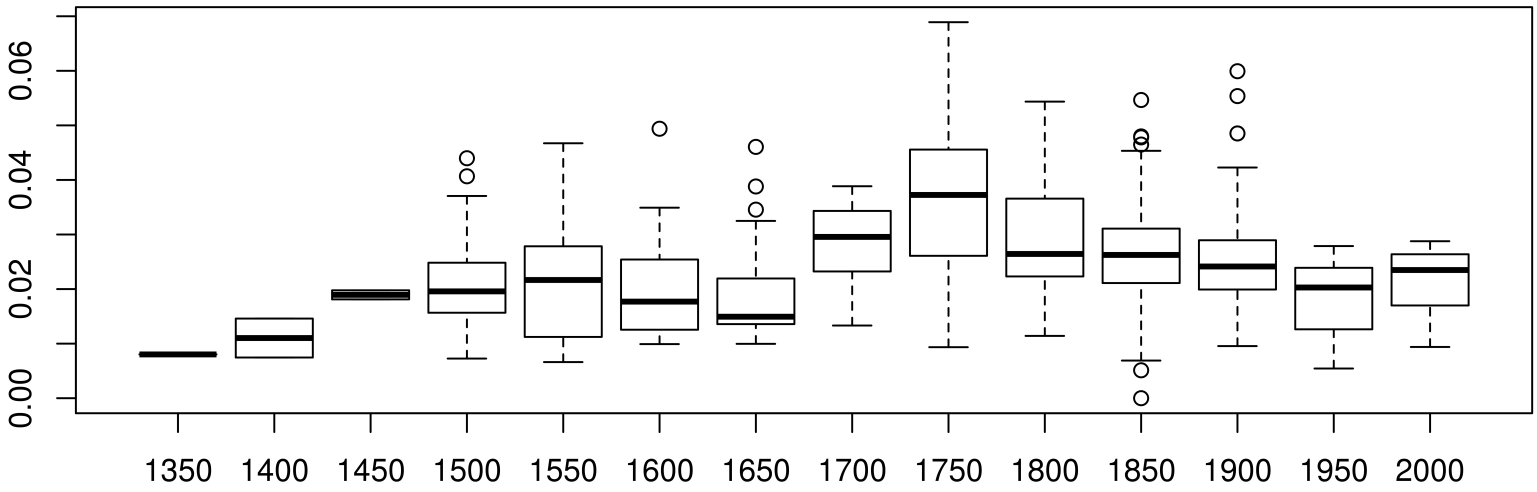
Um léxico com palavras de emoção (e expresões)

Mais de 4.000 lemas categorizados em 24 grupos. Ainda não revistos.

Presença nos textos (dados de 1 de setembro

amor	107,203	desejo	72,242
feliz	61,541	infeliz	61,488
medo	38,473	gen	30,237
vergonha	28,135	orgulho	25,439
feliz & satisfeito	22,889	coragem	21,117
surpresa	20,637	humildade	20,282
odio	19,502	esperanca	19,259
furia	14,943	satisfeito	14,651
desespero	14,601	saudade	13,037

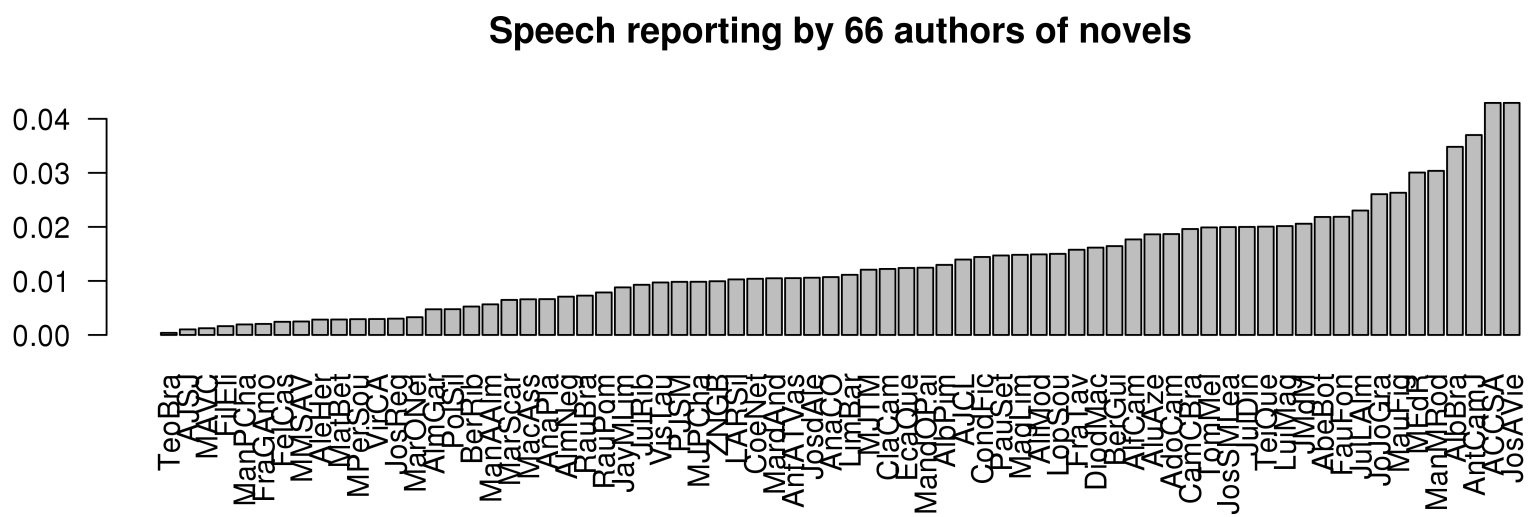
Emoções ao longo do tempo



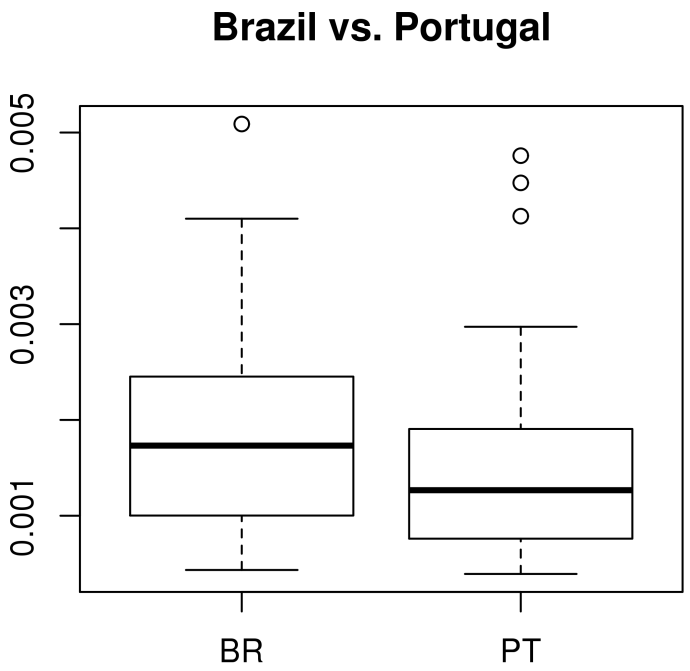
Perfis de descrição da fala

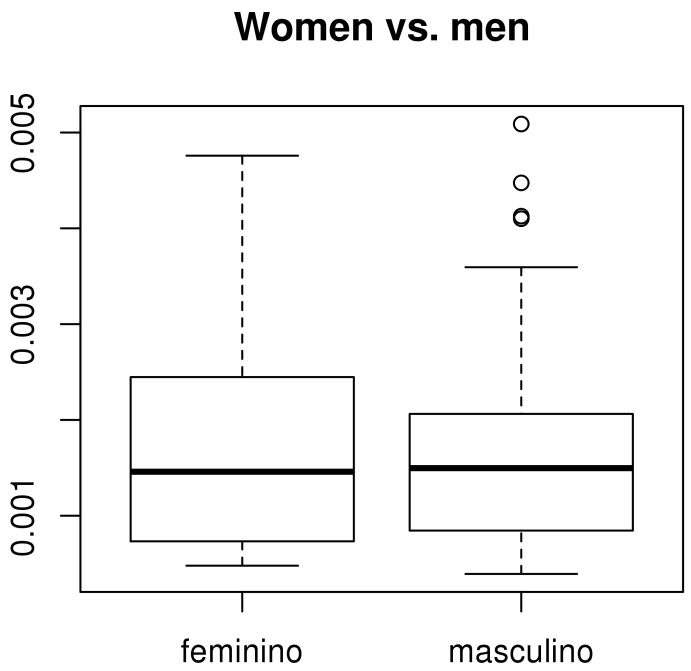
Verbos de fala

- três tipos de relato (discurso direto, indireto e misto) e simples menção
- uso de – (travessão) para discurso direto
- frequentemente exprimindo atitude ou sentimento

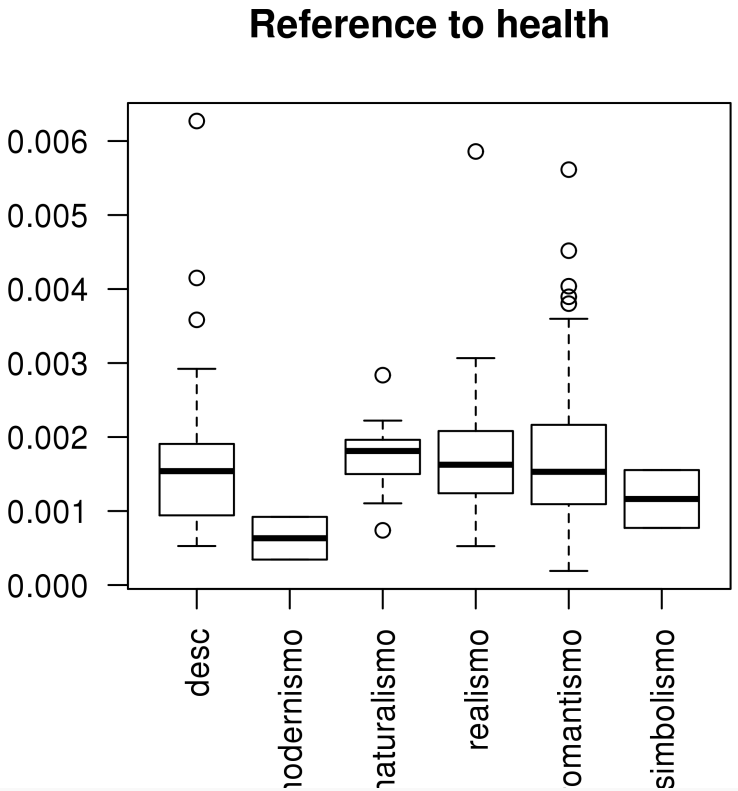


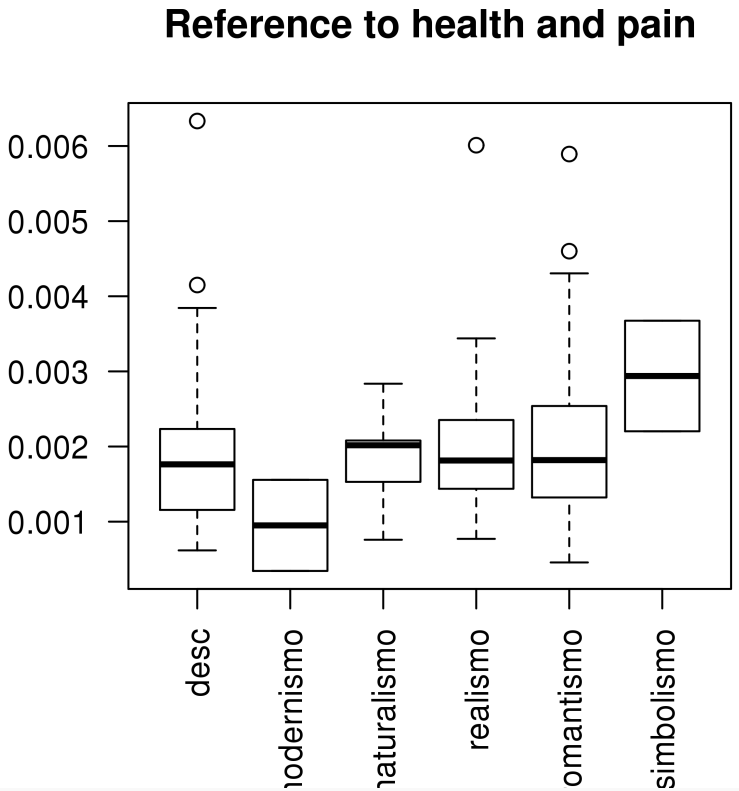
Referência a roupa em romances e novelas



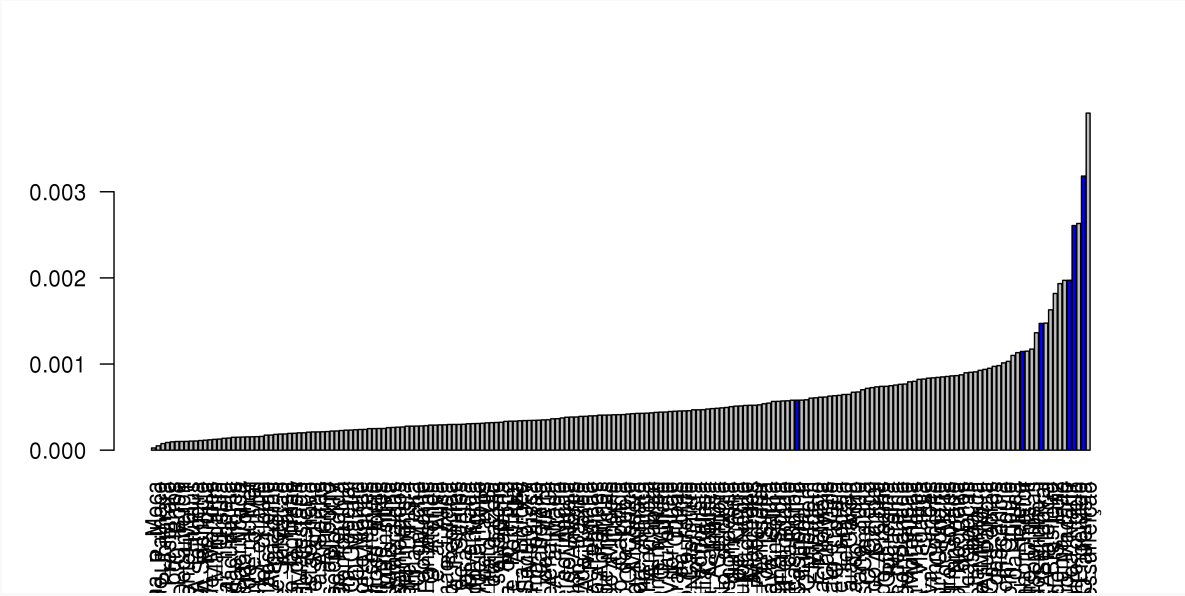


Saúde e dor


Reference to health
0.006
0.005
0.004
0.003
0.002
0.001
0.000
desc
modernismo
naturalismo
realismo
romantismo
simbolismo


Reference to health and pain
0.006
0.005
0.004
0.003
0.002
0.001
desc
modernismo
naturalismo
realismo
romantismo
simbolismo

As profissões hospitalares: Médicos e enfermeiros



História da Literateca

- Um subconjunto da Gramateca (ver o artigo inicial)
- A necessidade de retirar repetidos
- A necessidade de uniformizar a informação associada
- A possibilidade de obter os dados através da internete
- A possibilidade de pôr os dados acessíveis, e verificáveis, ao exterior

Nos bastidores da Literateca

Existem vários programas acessórios:

- Existem dois programas que calculam todas as características invocando o CWB, e mais um que calcula o tamanho em frases, e os resultados são tornados acessíveis da página do projeto Literateca
- existe outro programa que faz as contagens para a modelação de temas (topic modelling)
- e outro que calcula os valores para as redes de personagens

Toda a vizualização e cálculos estatísticos são feitos em R

Obrigada!