Periodização automática: Estudos linguístico-estatísticos de literatura lusófona

Diana Santos [1]
Emanoel Pires [2]
Cláudia Freitas [3]
Rebeca Schumacher Fuão [1]
João Marques Lopes [1]

[1] Linguateca & Universidade de Oslo, Noruega
[2] Universidade Estadual do Maranhão (UEMA), Brasil
[3] Linguateca & PUC-Rio, Brasil

ENAPL 2019, 9-11 de outubro de 2019

Apresentação

1 Introdução
- Motivação
- Tarefas

2 Dados
- O material
- Tamanho
- A anotação

3 Resultados
- Primeira tarefa
- Segunda tarefa

4 Observações finais

Motivação

- Aproveitar a infraestrutura da Linguateca, a Gramateca, para fazer estudos de literatura

Motivação

- Aproveitar a infraestrutura da Linguateca, a Gramateca, para fazer estudos de literatura
- Explorar as potencialidades da informação linguística para estudos literários (com a metodologida da leitura distante)

Motivação

- Aproveitar a infraestrutura da Linguateca, a Gramateca, para fazer estudos de literatura
- Explorar as potencialidades da informação linguística para estudos literários (com a metodologida da leitura distante)
- Publicitar o ambiente (a Literateca) para que outros também o possam utilizar

Tarefas

1. Replicar o trabalho de Barufaldi et al. (2009, 2010)

Objetivo

classificar 15 autores diferentes, totalizando 37 obras brasileiras, em quatro períodos literários: barroco, arcadismo, romantismo e realismo

2. Explorar o "conjunto COST"de romances e novelas portuguesas e brasileiras de 1840-1919

O material textual

`https://www.linguateca.pt/acesso/corpus.php?corpus=LITERATECA`

Tipos de textos

A Literateca contém todo o material de obras literárias da Linguateca

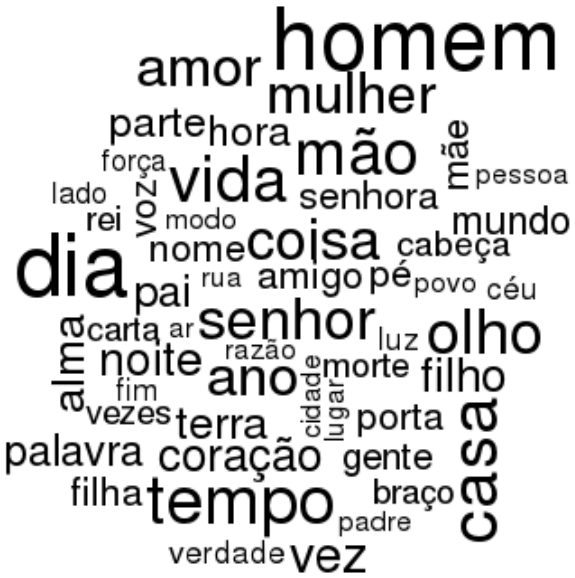
- Clássicos desde 1300
- Textos literários canónicos
- Textos literários não-canónicos
- Excertos de textos literários traduzidos de outras línguas

Estes textos provêm dos projetos Vercial, Tycho Brahe e Colonia, assim como do trabalho colaborativo na construção do OBras, incluindo especificamente a PUC-Rio e a UEMA

Dados quantitativos

Números atuais sobre a Literateca

- 28 milhões de palavras
- 784 obras de 212 autores diferentes
- 187 romances ou novelas do "período COST" (1840-1919)
- 7 romances anotados com personagens



A anotação

Todos os textos estão morfossintaticamente anotados pelo PALAVRAS (Bick, 2000).

A anotação

Todos os textos estão morfossintaticamente anotados pelo PALAVRAS (Bick, 2000). Além disso:

Domínios semânticos

- os campos das cores, roupa, corpo, família, emoções e saúde estão anotados
- e foram revistos parcialmente

A anotação

Todos os textos estão morfossintaticamente anotados pelo PALAVRAS (Bick, 2000). Além disso:

Domínios semânticos

- os campos das cores, roupa, corpo, família, emoções e saúde estão anotados
- e foram revistos parcialmente

Informação literária

- Cada obra está classificada com o sexo do autor, o género do texto, e escola literária
- Entidades mencionadas (pessoas, lugares e obras) obtidas pelo PALAVRAS foram revistas parcialmente
- Em alguns casos foi adicionada a marcação de personagens

Características linguísticas

Usámos, além das ocorrências dos campos semânticos mencionados, outras características que nos pareceram de interesse para uma possível descrição do estilo (do autor, época, escola...). Exemplos:

- o número de pronomes pessoais de primeira pessoa,
- o número de verbos de fala,
- o número de pontos de interrogação
- o discurso direto explícito

Ao todo foram criadas 128 características, ver `https://www.linguateca.pt/Gramateca/Literateca/lista_caracteristicas.txt`

Primeira tarefa
Apresentação

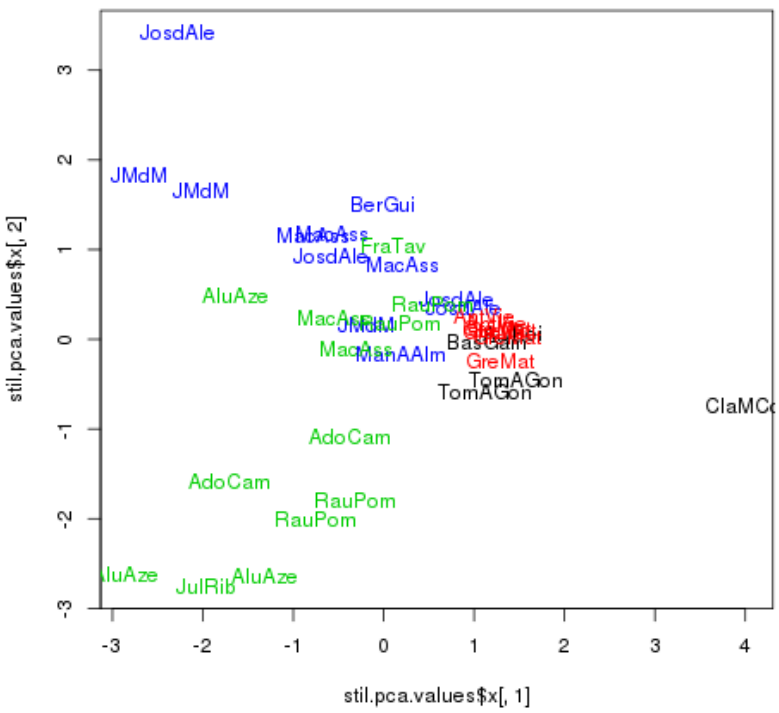
Objetivo

classificar 15 autores diferentes, totalizando 37 obras brasileiras, em quatro períodos literários: barroco, arcadismo, romantismo e realismo

Técnicas usadas

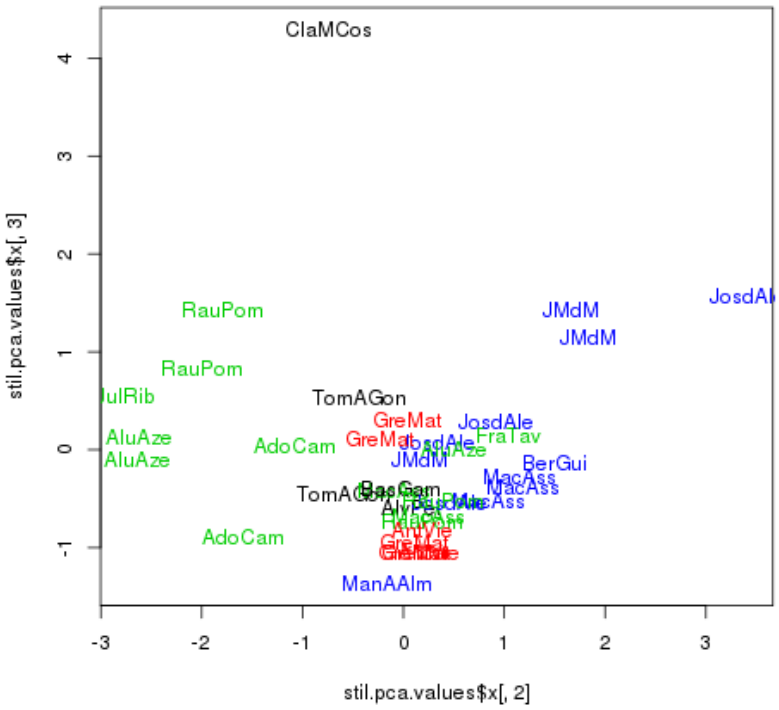
- análise de discriminantes com base em componentes principais
- máquinas de vetores de apoio (*support vector machines*)

Primeira tarefa
Resultados da análise de discriminantes: primeiro e segundo



azul - romantismo
verde - realismo
vermelho - barroco
preto - arcadismo

Primeira tarefa
Resultados da análise de discriminantes: segundo e terceiro



azul - romantismo
verde - realismo
vermelho - barroco
preto - arcadismo

Primeira tarefa
Resultados da análise de SVM

escola / previsão	arcadismo	barroco	realismo	romantismo
arcadismo	3	2	0	0
barroco	0	7	0	0
realismo	0	0	10	3
romantismo	0	1	0	11

Tabela : *Ubirajara* de José de Alencar, romântico, foi considerado barroco, assim como *O Uraguai* de Basílio da Gama e a *Coletânea de obras* de Alvarenga Peixoto, ambos arcadistas. Três romances realistas são considerados românticos.

Primeira tarefa
Comentários

- Aplicámos duas técnicas diferentes, mas ambas mostram alguma dificuldade em distinguir entre barroco e arcadismo
- Comparando com a técnica de compressão de Barufaldi et al. sobre as palavras dos textos, nós reduzimos/transformamos os mesmos em apenas um conjunto de números relativos a uma centena de características
- Vamos olhar para os casos mal classificados com mais atenção, mas podemos desde já afirmar que *Ubirajara* (com frases muito curtas) é muito diferente dos outros romances românticos aqui incluídos.

Segunda tarefa
Apresentação

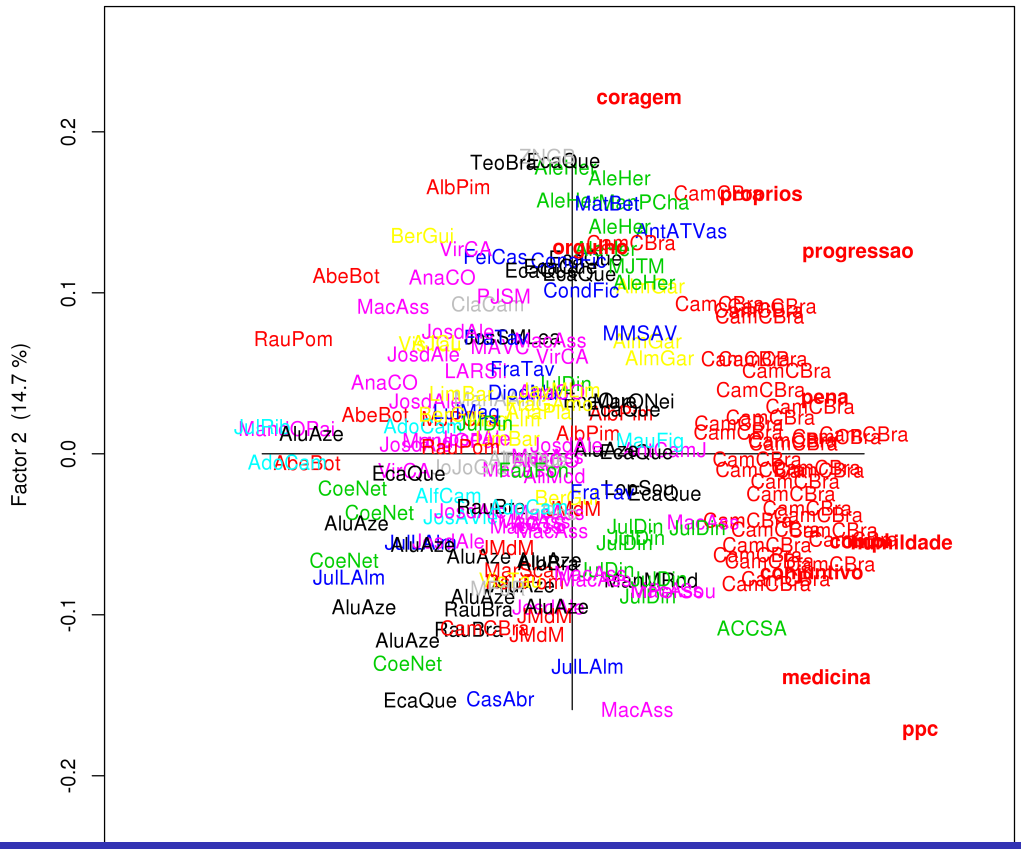
Objetivo

Organizar todos os romances e/ou novelas em português em formato eletrónico a que tínhamos acesso à data de 25 de agosto de 2019 (187 obras, das quais 120 portuguesas e 67 brasileiras)

Técnicas usadas

análise de correspondências; análise de temas (*topic modelling*)

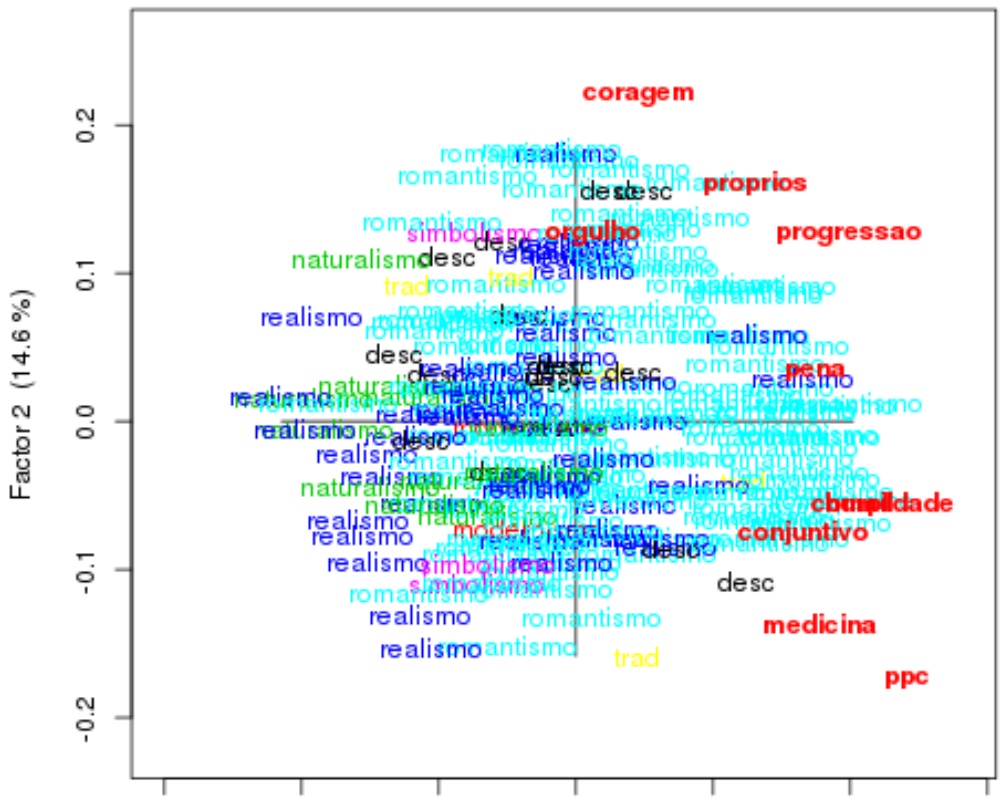
Segunda tarefa
Análise de correspondências: autores



Segunda tarefa
Análise de correspondências: autores: comentários

- A análise de correspondências parece colocar a maior parte dos autores no seu lugar
- Alguns autores, talvez mais multifacetados, encontram-se em vários quadrantes, mas pode ser que isso reflicta o seu trajeto literário. (Sabe-se que Machado de Assis teve duas fases...)
- E alguns casos são traduções! (algo que não é consensual entre os autores deste artigo)
- Um problema de que estamos conscientes é a existência de autores com livros a mais... como Camilo Castelo Branco.

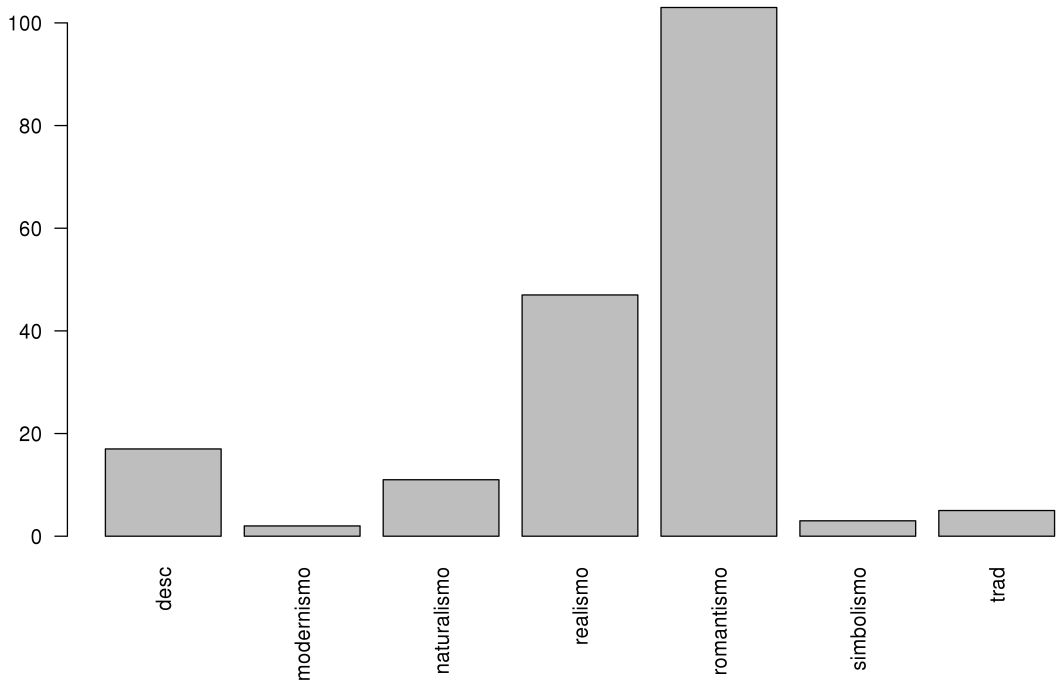
Segunda tarefa
Análise de correspondências: escola literária



Segunda tarefa
Comentários: escola literária

- A tarefa da atribuição de escola literária é complicada e em última análise arbitrária.
- Uma grande parte dos textos foram marcados como pertencendo a várias escolas, por exemplo **realismo-pósnaturalismo** ou **indianismo-regionalismo-romantismo**. Para o gráfico em questão foi feita a escolha de uma só escola: por exemplo os textos marcados como **naturalismo-realismo** foram reclassificados como simplesmente **naturalismo**.
- "desc" significa ainda não classificado

Anotação literária: distribuição de escolas



Segunda tarefa

Comentários: características discriminantes

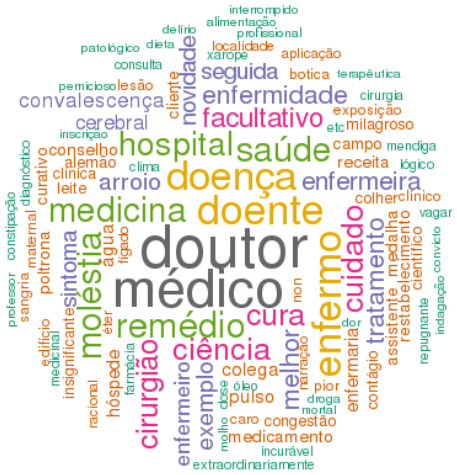
Em relação às características que discriminam mais

- algumas têm uma explicação plausível, como a CORAGEM perto dos romances históricos de Alexandre Herculano
- mas a maior parte delas exige uma análise mais detalhada

Segunda tarefa
Análise de temas (*topic modelling*): coleção toda

100 temas criados com o programa `mallet`, só nomes e adjetivos.

Tema 34



Tema 8

Segunda tarefa
Análise de temas (*topic modelling*): romantismo vs. realismo/naturalismo

Tema 21 dos temas românticos (usando 87 obras)

Tema 70 dos temas realistas (usando 51 obras)

Observações finais

- Apresentamos técnicas de leitura distante para o português, com o objetivo de dinamizar a área (*Primeiro encontro sobre leitura distante em português* em Oslo, 27/28 de outubro)
- Falta algum trabalho de classificação das obras, e muita revisão da anotação das características, que será motivada pelo seu uso nesta pesquisa
- Planeamos
  - focar os casos mais surpreendentes
  - repetir o trabalho com um máximo de 2,3 obras por autor
  - tentar outras formas de classificação das escolas (por exemplo classificadores binários, que deem percentagem de pertença a uma escola)

Bibliografia

- Baayen, Harald. *Analyzing Linguistic Data: A practical introduction to Statistics using R*. Cambridge University Press, 2008.
- Barufaldi, Bruno, Eduardo F. Santana, José Rogério B. B. Filho, Jan Kees van der Poel, Milton Marques Júnior & Leonardo Vidal Batista. "Classificação Automática de Textos por Período Literário Utilizando Compressão de Dados Através do PPM-C", *Linguamática* 2, 1, Abril 2010, pp. 35-44.
- Jockers, Matthew L. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.
- McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit". 2002. `http://mallet.cs.umass.edu`
- Moretti, Franco. "Conjectures on world literature", *New Left review* 1, Jan-Feb 2000, pp. 54-68.

Bibliografia 2

- Santos, Diana. "Literature studies in Literateca: between digital humanities and corpus linguistics". In Martin Doerr, Øyvind Eide, Oddrun Grønvik & Bjørghild Kjelsvik (eds.), *Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore*. Novus forlag, Oslo, 2019, pp. 89-109.
- Santos, Diana, Cláudia Freitas & João Marques Lopes. "Comparando a literatura lusófona com outras literaturas: recursos para leitura a distância em português". In Suemi Higuchi & Cláudio José Silva Ribeiro (eds.), *I Congresso Internacional em Humanidades Digitais no Rio de Janeiro (HdRio2018)*, CPDOC/FGV, 2018, pp. 375-383.
- R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008.

Obrigado pela atenção!