Preparação para Leitura Distante em português: diálogos entre PLN e Humanidades Digitais

Luísa Rocha
PUC-Rio & Linguateca

Cláudia Freitas
PUC-Rio & Linguateca

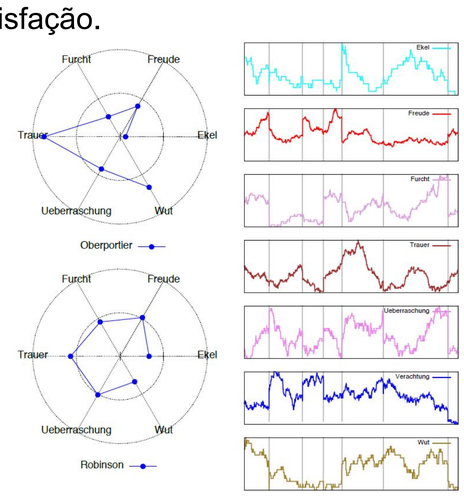
Diana Santos
Universidade de Oslo & Linguateca

INTRODUÇÃO

Humanidades Digitais são uma área que junta as humanidades e recursos digitais para criar novos pontos de vista e métodos para esses estudos. Leitura Distante, ou "ler à distância", ou “distant reading” (MORETTI, 2008), é uma das técnicas nas humanidades digitais, criada por Franco Moretti, para estudar literatura de uma forma diferente. Consiste em distanciar-se das obras literárias, não como um obstáculo, mas sim como uma forma de obter novos ângulos, vendo menos detalhes e mais relações, padrões e formas em uma obra ou entre obras.

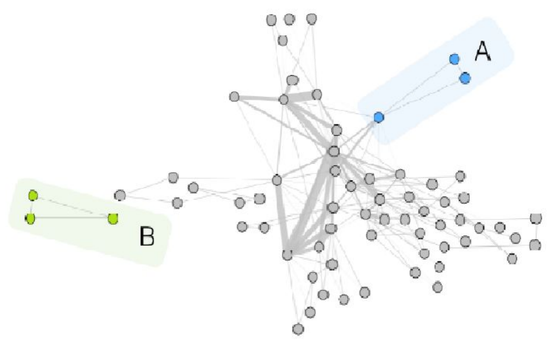
TRABALHOS RELACIONADOS:

KLINGER; SULIYA; REITER, 2016 propõem-se a detectar diferentes emoções nos textos literários, utilizando dois romances de Franz Kafka. 300 palavras foram anotadas como sentimentos pertencentes a uma das sete emoções, correspondendo aproximadamente a: raiva, desgosto, medo, felicidade, tristeza, surpresa e satisfação.



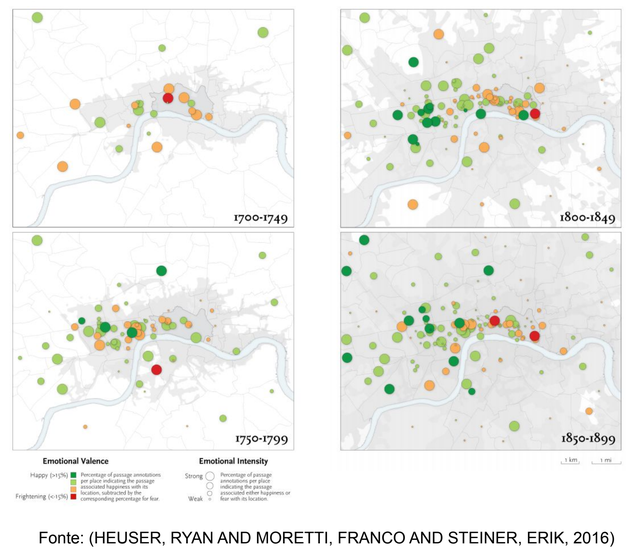
Fonte: (KLINGER; SULIYA; REITER, 2016)

GRAYSON et al., 2016 examinam as redes sociais dos personagens de Charles Dickens e Jane Austen. Foi observado que as sociedades construídas por Austen parecem ser mais compactas do que as de Dickens e que Dickens se utiliza muito de micronarrativas.



Fonte: (GRAYSON et al., 2016)

HEUSER, RYAN AND MORETTI, FRANCO AND STEINER, ERIK, 2016 utilizam-se dos romances ingleses do séc.XIX para mapear as emoções da cidade de Londres. "Frightening" (assustador) e "happy" (feliz) foram os dois polos de emoções mapeadas.



Fonte: (HEUSER, RYAN AND MORETTI, FRANCO AND STEINER, ERIK, 2016)

OBras E CORREÇÕES:

Corpus OBras: composto por 246 obras de literatura brasileira em domínio público e em permanente expansão, foi criado para ser a contraparte brasileira do Corpus Vercial. Todo material está anotado com informação morfossintática e semântica, e está disponível para consulta e para ser baixado pela página da Linguateca (SANTOS; FREITAS; BICK, 2018).

Correções: Revisão com foco nos nomes próprios:

- revisão do gênero gramatical
- revisão da segmentação
- revisão da classe semântica (atributo *sema*)
  - Pessoa (Pessoa; Pessoa:ficc; Pessoa:hist)
  - Local
  - Obra
  - Religião

Coapara
Êsse=Jaime
D.=Gaiferos
D.=Caterina
Plantei
Conversemos
Jurei
Endoenças
Pois=Paulina
Crônicas=do=Dr.=Semana

Figura 1 - Resultados da pesquisa por nomes próprios humanos com gênero indefinido.

```
a:[lema="S."] b:[lema="João"] c:[lema="do"] d:[lema="Príncipe"] >> a:[lema="S.=João=do=Príncipe" & sema="Local"] b:[lema="S.=João=do=Príncipe"
a:[lema="S."] b:[lema="Paulo"] >> a:[lema="S.=Paulo" & sema="Local"] b:[lema="S.=Paulo" & sema="Local"]
a:[lema="S."] b:[lema="Pedro"] >> a:[lema="S.=Pedro" & sema="Local"] b:[lema="S.=Pedro" & sema="Local"]
a:[lema="Rua"] b:[lema="de"] c:[lema="S."] d:[lema="Pedro"] >> a:[lema="Rua=de=S.=Pedro" & sema="Local"] b:[lema="Rua=de=S.=Pedro" & sema="Loc
a:[lema="S."] b:[lema="Francisco"] >> a:[lema="S.=Francisco" & sema="Local"] b:[lema="S.=Francisco" & sema="Local"]
a:[lema="S."] b:[lema="Francisco"] c:[lema="de"] d:[lema="Paulo"] >> a:[lema="S.=Francisco=de=paulo" & sema="Local"] b:[lema="S.=Francisco=de=
a:[lema="S."] b:[lema="Vicente"] >> a:[lema="S.=Vicente" & sema="Local"] b:[lema="S.=Vicente" & sema="Local"]
a:[lema="Capitania"] b:[lema="de"] c:[lema="S."] d:[lema="Vicente"] >> a:[lema="Capitania=de=S.=Vicente" & sema="Local"] b:[lema="Capitania=de
a:[lema="S."] b:[lema="Petersburgo"] >> a:[lema="S.=Petersburgo" & sema="Local"] b:[lema="S.=Petersburgo" & sema="Local"]
a:[lema="Rua"] b:[lema="de"] c:[lema="S."] d:[lema="José"] >> a:[lema="Rua=de=S.=José" & sema="Local"] b:[lema="Rua=de=S.=José" & sema="Local"
a:[lema="Rua"] b:[lema="de"] c:[lema="S."] d:[lema="Diogo"] >> a:[lema="Rua=de=S.=Diogo" & sema="Local"] b:[lema="Rua=de=S.=Diogo" & sema="Loc
```

Figura 2 - Correções de segmentação

```
[lema="Imaerô"] >> [sema="Pessoa" & gen="F"]
[lema="Honorina"] >> [sema="Pessoa" & gen="F"]
[lema="D.=Joana"] >> [sema="Pessoa" & gen="F"]
[lema="V.=Excia"] >> [sema="Pessoa" & gen="F"]
[lema="D.=Antônia"] >> [sema="Pessoa" & gen="F"]
[lema="D.=Carolina"] >> [sema="Pessoa" & gen="F"]
[lema="Marcelina"] >> [sema="Pessoa" & gen="F"]
[lema="D.=Benedita"] >> [sema="Pessoa" & gen="F"]
[lema="Virgília"] >> [sema="Pessoa" & gen="F"]
[lema="Domitila"] >> [sema="Pessoa" & gen="F"]
```

Figura 3 - adição da classe semântica e correção do gênero gramatical

CONCLUSÕES

Os textos estudados são fonte de várias ideias para se trabalhar com a leitura distante em português, especificamente com o corpus OBras, e para que os resultados sejam mais precisos, correções e melhorias são necessárias. Mais de mil correções envolvendo atribuições de gênero, inclusão de *sema* e correções de segmentação e pré-processamento foram efetuadas.

Agradecimentos

Luísa Rocha é bolsista de Iniciação Científica do Conselho Nacional de Desenvolvimento Científico e Tecnológico, no âmbito do projeto "Recursos para o 'distant reading' em português:diálogos entre PLN e Humanidades Digitais". Número do processo da Bolsa: 101815/2019-0

1- GRAYSON, S. et al. Discovering Structure in Social Networks of 19th Century Fiction. In:PROCEEDINGS of the 8th ACM Conference on Web Science. Hannover, Germany: ACM, 2016.(WebSci ’16), p. 325–326. ISBN 978-1-4503-4208-7. DOI:10.1145/2908131.2908196. Disponível em:<http://doi.acm.org/10.1145/2908131.2908196>.
2-HEUSER, RYAN AND MORETTI, FRANCO AND STEINER, ERIK. The Emotions of London. In:LITERARY Lab Pamphlet 13. London, England: 2016. Disponível em:<https://litlab.stanford.edu/LiteraryLabPamphlet13.pdf>.
3- KLINGER, R.; SULIYA, S. S.; REITER, N. Automatic Emotion Detection for Quantitative Literary Studies– A case study based on Franz Kafka’s “Das Schloss” and “Amerika”. In: JAGIELLONIAN UNIVERSITY e PEDAGOGICAL UNIVERSITY.DIGITAL Humanities 2016: Conference Abstracts. Kraków,Poland: [s.n.], jul. 2016. p. 826–828. Disponível em: <http://dh2016.adho.org/abstracts/318>.
4- LINGUATECA. Disponível em: <https://www.linguateca.pt/OBRAS/OBRAS.html>. (accessed:05.31.2019).
5- MORETTI, F.A Literatura Vista de Longe. Porto Alegre: Arquipélago, 2008.
6- SANTOS, D.; FREITAS, C.; BICK, E. OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in the public domain. In: OpenCor, 2018.
7- SANTOS, D.; FREITAS, C. Estudando personagens na literatura lusófona. In:ANAIS do STIL’2019.Salvador, Brasil: [s.n.], 2019.