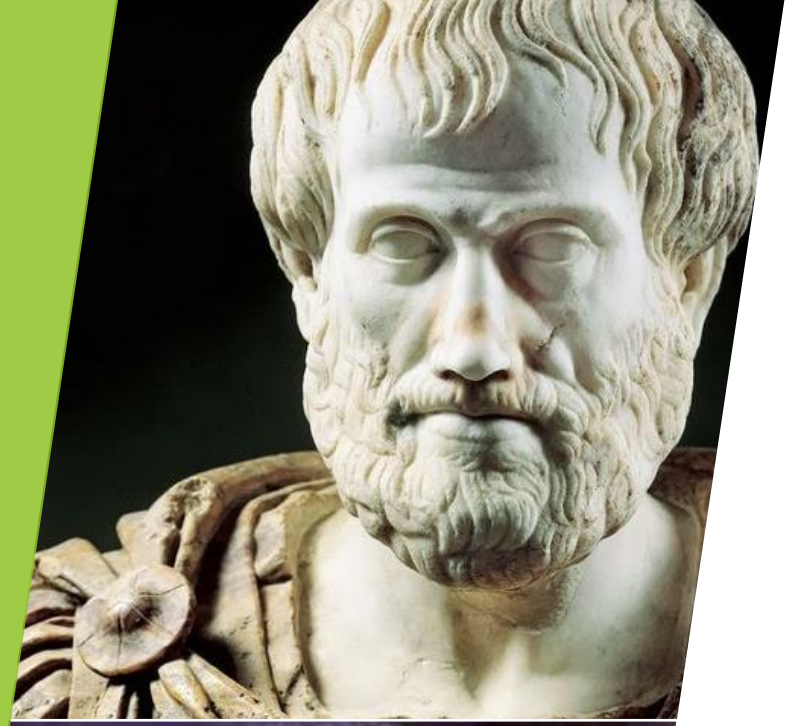
Part-of-Speech Tagging

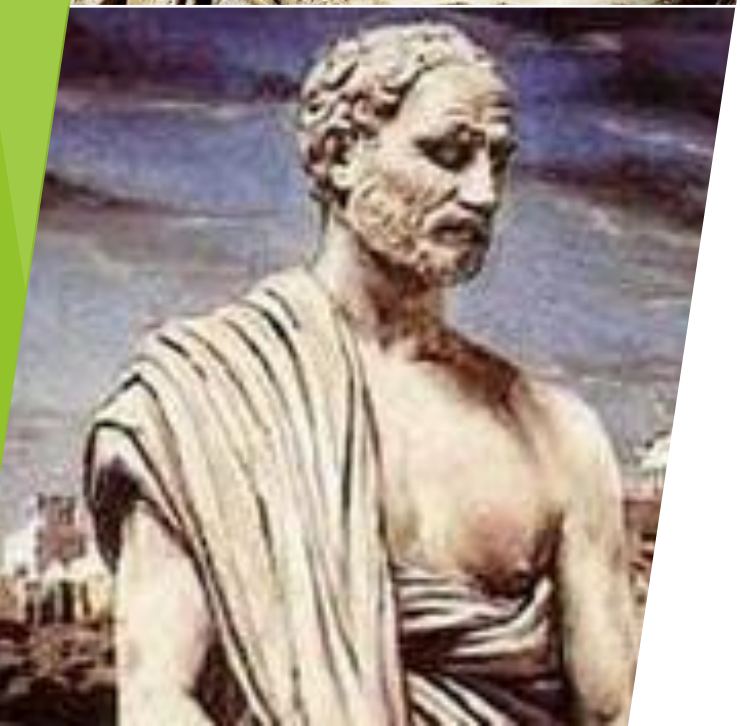
Processamento de Linguagem Natural

Prof. Leandro Alvim, D. Sc.

Agenda

- História
- Classes de POS
- Por que precisamos resolver ?
- Aplicações
- Qualidade dos sistemas de identificação
- Alguns *corpora*
- Extração da informação em Corpora via ACDC
- Extração da informação em Corpora via Python

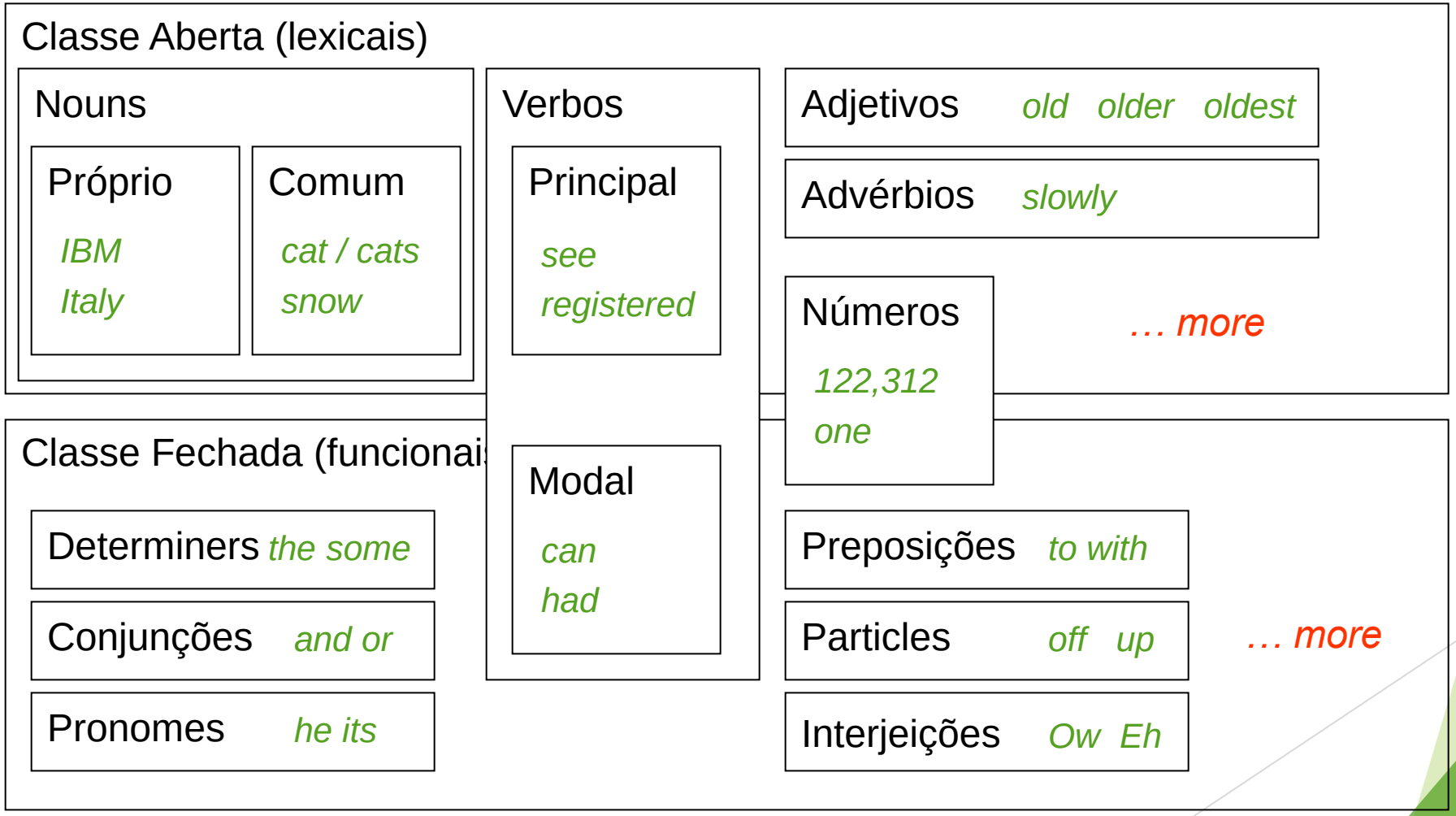




Part of Speech

- Aristóteles (384-322 A.C.)
  - Categorias léxicas, classes gramaticais, tags, ...
- Dionysius Thrax de Alexandria (100 A.C.)
  - Ideia de que ainda persiste de que existem apenas oito POS (porém não as mesmas)
    - Thrax: substantivo, verbo, artigo, advérbio, preposição, conjunção, particípio, pronome
    - Gramática atual: substantivo, verbo, adjetivo, advérbio, preposição, conjunção, pronome, interjeição

Classes


Classe Aberta (lexicais)
Nouns
Próprio
IBM
Italy
Comum
cat / cats
snow
Verbos
Principal
see
registered
Modal
can
had
Adjetivos old older oldest
Advérbios slowly
Números
122,312
one
… more
Classe Fechada (funcionais)
Determiners the some
Conjunções and or
Pronomes he its
Preposições to with
Particles off up
Interjeições Ow Eh
… more

Aberto vs Fechado

Fechado	Artigos: a, an, the
	Pronomes: she, he, I
	Preposições: on, under, over, near, by
Aberto	Substantivos
	Verbos
	Adjetivos
	Advérbios

Por que precisamos de um sistema para resolver ?

- As palavras, comumente, têm mais de um POS
  - The *back* door = JJ
  - On my *back* = NN
  - Win the voters *back* = RB
  - Promised to *back* the bill = VB
- O problema de POS tagging consiste em determinar o POS tag para uma particular instância da palavra em seu contexto.

Por que precisamos de um sistema para resolver ?

Entrada	Plays	well	with	others
Ambiguidade	NNS/VBZ	UH/JJ/NN/RB	IN	NNS
Saída	Plays/VBZ	well/RB	with/IN	others/NNS

Aplicações

- Texto para Fala
  - Como pronunciar a palavra *lead* ?
    - Verbo (“lid”) ou Substantivo (“led”)
- Recuperação da Informação
  - **“[lema=“mulher"] @[pos="ADJ" & func="N<"]”**
- A base para tarefas mais complexas de classificação
  - Reconhecimento de Entidades
  - Identificação de Anáforas e Catáforas
  - Question Answering

Qualidade dos Sistemas de POS Tagging

- Quantas tags estão corretas ? (Tag accuracy)
  - Em torno de 97%
  - Entretanto, baseline já atinge 90%
    - Baseline é o método mais estúpido possível
      - Marque cada palavra com o seu associado POS mais frequente
      - Marque palavras desconhecidas como substantivo
  - Parcialmente fácil porque
    - Muitas palavras não são ambíguas
    - Alguns tokens ajudam a subir muito o resultado (*the*, *a*, etc.) e também as marcações de pontuação

A decisão do POS correto pode ser difícil até para humanos

- Mrs/NNP Shaefer/NNP never/RB got/VBD **around/RP** to/TO joining/VBG
- All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB **around/IN** the/DT corner/NN
- Chateau/NNP Petrus/NNP costs/VBZ **around/RB** 250/CD

Evolução dos Sistemas de POS Tagging

- Acurácia por Token
  - Most freq tag: ~90% / ~50%
  - Trigram HMM: ~95% / ~55%
  - Maxent P(t|w): 93.7% / 82.6%
  - TnT (HMM++): 96.2% / 86.0%
  - MEMM tagger: 96.9% / 86.9%
  - Bidirectional dependencies: 97.2% / 90.0%
  - Limite Superior: ~98% (human agreement)
- Acurácia por frase
  - ~56%

Corpus

- **Brown corpus: Corpus of American English**
  - The **Brown corpus** (full name **Brown University Standard Corpus of Present-Day American English**) was the first text corpus of American English. The original corpus was published in 1963-1964 by W. Nelson Francis and Henry Kučera at Department of Linguistics, Brown University Providence, Rhode Island, USA.
  - The corpus consists of 1 million words (500 samples of 2000+ words each) of running text of edited English prose printed in the United States during the year 1961 and it was revised and amplified in 1979.
- **Part-of-speech tagset**
  - The Brown corpus is PoS tagged with the Penn TreeBank tagset. The Brown family corpus has POS tags from the CLAWS tagset version 7.
- **Tools to work with the Brown corpus**
  - A complete set of tools is available to work with the Brown corpus online (without registration) to generate:
  - word sketch- English collocations categorized by grammatical relations
  - thesaurus- synonyms and similar words for every word
  - keywords- terminology extraction of one-word and multi-word units
  - word lists – lists of English nouns, verbs, adjectives etc. organized by frequency
  - n-grams- frequency list of multi-word units
  - concordance – examples in context
- **Availability**
- The access to the corpus is freely available for research.

Corpus

- Mac-Morpho
  - Mac-Morpho is a corpus of Brazilian Portuguese texts annotated with part-of-speech tags. Its first version was released in 2003 [1], and since then, two revisions have been made in order to improve the quality of the resource [2, 3].
  - The corpus is available for download split into train, development and test sections. These are 76%, 4% and 20% of the corpus total, respectively (the reason for the unusual numbers is that the corpus was first split into 80%/20% train/test, and then 5% of the train section was set aside for development). This split was used in [3], and new POS tagging research with Mac-Morpho is encouraged to follow it in order to make consistent comparisons possible.

- Referências

  - Aluísio, S., Pelizzoni, J., Marchi, A.R., de Oliveira, L., Manenti, R., Marquiafável, V. 2003. **An account of the challenge of tagging a reference corpus for brazilian portuguese.** In: Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language. PROPOR 2003 [link]
  - Fonseca, E.R., Rosa, J.L.G. 2013. **Mac-morpho revisited: Towards robust part-of-speech.** In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology – STIL [link]
  - Fonseca, E.R., Aluísio, Sandra Maria, Rosa, J.L.G. 2015. **Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese.** Journal of the Brazilian Computer Society. [link]

Corpus

- Bosque
  - O **Bosque** é composto por 9.368 frases, retiradas os primeiros 1000 extractos (aprox.) dos corpora CETENFolha e CETEMPúblico. Desde 2007, o Bosque vem passando por um novo processo de revisão, em que foram corrigidas algumas pequenas inconsistências e acrescentadas novas etiquetas. A versão final, disponível para consulta e download, é o Bosque 8.0.
  - Este é o corpus mais correto da Floresta, e por isso o mais aconselhado para pesquisas em que não se prioriza tanto a quantidade, mas sim a precisão dos resultados.
  - Uma quantificação das etiquetas usadas no Bosque pode ser encontrada no anexo 4 da Bíblia Florestal, uma extensa documentação das opções linguísticas tomadas durante o projecto.

- Referências
  - https://www.researchgate.net/publication/327224726_Tagsets_and_Datasets_Some_Experiments_Based_on_Portuguese_Language_13th_International_Conference_PROPOR_2018_Canela_Brazil_September_24-26_2018_Proceedings

Repositório Floresta Sintática

- Chamamos de "Floresta Sintática" um conjunto de frases (corpus) analisadas (morfo) sintaticamente. Como, além da indicação das funções sintáticas, a análise também explicita hierarquicamente informação relativa à estrutura de constituintes, dizemos que uma frase sintaticamente analisada se parece com uma árvore, donde um conjunto de árvores constitui uma floresta sintática (em inglês, *treebank*).
- O ***projecto Floresta Sintá(c)tica*** é uma colaboração entre a Linguateca e o projecto VISL. Contém textos em português (do Brasil e de Portugal) anotados (analisados) automaticamente pelo analisador sintáctico PALAVRAS (Bick 2000) e revistos por linguistas.
- Atualmente, a Floresta Sintá(c)tica tem quatro partes, que diferem quanto ao gênero textual, quanto ao modo (escrito *vs* falado) e quanto ao grau de revisão linguística: o **Bosque**, totalmente revisto por linguistas; a **Selva**, parcialmente revista, a **Floresta Virgem** e a **Amazônia**, não revistos. Junto, todo esse material soma cerca de 261 mil frases (6,7 milhões de palavras) sintaticamente analisadas

Repositório Floresta Sintática

corpus	Floresta Virgem	Amazônia	Bosque	Selva Lit.	Selva Fal.	Selva Cie.
palavras	c. 1.640.000	c. 4.580.000	c. 186.000	c. 105.000	c. 170.000	c. 125.000
frases	c. 96.000	c. 275.000	9.368	c. 7.900	c. 14.000	c. 6.200
revisão	não	não	integral	parcial	parcial	parcial
variantes	PT BR	BR	PT BR	PT BR	PT BR	PT BR
gênero	jornalístico	opinião	jornalístico	literário	entrevistas / debates	académico / informativo
domínio	genérico	cultura brasileira	genérico	genérico	biografia / política	educação / psicolinguística / computação / economia /ciências *
registro	formal	formal e informal	formal	formal	formal e informal	formal
modo	escrito	escrito	escrito	escrito	falado	escrito
origem	jornais Folha de São Paulo e Público	blog Overmundo	jornais Folha de São Paulo e Público	livros *	Museu da Pessoa (PT, BR) debates parlamentares	bibliotecas universitárias banco centrais Wikipedia

Penn TreeBank Tagset

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	there is
FW	foreign word	les
IN	preposition, subordinating conjunction	in, of, like
IN/that	that as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables

Penn TreeBank Tagset

NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
PDT	predeterminer	both the boys
POS	possessive ending	friend's
PP	personal pronoun	I, he, it
PP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give up
SENT	Sentence-break punctuation	. ! ?
SYM	Symbol	/ [= *
TO	infinitive 'to'	togo
UH	interjection	uhhuhhuhh
VB	verb be, base form	be

Penn TreeBank Tagset

VBD	verb be, past tense	was, were
VBG	verb be, gerund/present participle	being
VBN	verb be, past participle	been
VBP	verb be, sing. present, non-3d	am, are
VBZ	verb be, 3rd person sing. present	is
VH	verb have, base form	have
VHD	verb have, past tense	had
VHG	verb have, gerund/present participle	having
VHN	verb have, past participle	had
VHP	verb have, sing. present, non-3d	have
VHZ	verb have, 3rd person sing. present	has
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, sing. present, non-3d	take

Penn TreeBank Tagset

VVZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when
#	#	#
$	$	$
"	Quotation marks	' "
``	Opening quotation marks	' "
(	Opening brackets	({
)	Closing brackets	) }
,	Comma	,
:	Punctuation	- ; : — ...

História dos corpora do Linguateca

- Tornar acessível e agrupar em um único lugar da web o que já existia (AC/DC)
- Criar ou melhorar analisando os corpora (AC/DC)
- Adicionar a dimensão da tradução (COMPARA, CorTrad)
- Adicionar a dimensão da revisão humana (Floresta Sintática, COMPARA)
- Adicionar a possibilidade de criar corpora novos (Corpógrafo)
- Adicionar a dimensão de corpora comparáveis (Corpógrafo)

Anotação dos corpos

Adaptado de Rocha (2007)

Cada corpo é anotado sintacticamente.

PALAVRAS

```
$START
Cada      [cada] <quant> DET M S @>N
corpo     [corpo] N M S @SUBJ>
é         [ser] <fmc> V PR 3S IND VFIN @FAUX
anotado   [anotar] V PCP M S @IMV @#ICL-AUX<
sintacti  ente ALT sintaticamente       [sintático] ADV @<ADVL
$.
```

Formato AC/DC

forma | lema | classe gramatical (PoS) | tempo | num/pes | gênero | função

Cada	cada	DET_quant	0	S	M	>N
corpo	corpo	N	0	S	M	SUBJ>
é	ser	V_fmc	PR_IND	3S	0	FAUX
anotado	anotar	V	PCP	S	M	IMV_#ICL-AUX<
sintaticamente	sintático	ADV	0	0	0	<ADVL

Projeto AC/DC: corpo Colonia

AC/DC : Linguateca

O **Colonia** é um corpo eletrônico anotado compilado para pesquisa sobre a história da língua portuguesa, com textos escritos entre 1500 e 1936, desenvolvido pela Universidade de Colónia (Köln). A sua página principal é http://corporavm.uni-koeln.de/colonia/ onde todas as informações estão disponíveis. Veja também Zampieri & Becker (2013). (Nota: À versão do AC/DC faltam ainda cinco textos.)

Procurar: OK

Resultado:

- (•) Concordância
- ( ) Distribuição das formas (`word`)
- ( ) Distribuição dos lemas (`lema`)
- ( ) Distribuição da categoria gramatical (PoS) (`pos`)
- ( ) Distribuição do tempo verbal e/ou do caso pronominal (`temcagr`)
- ( ) Distribuição de pessoa e/ou número (`pessnum`)
- ( ) Distribuição do género morfológico (`gen`)
- ( ) Distribuição da função sintáctica (`func`)
- ( ) Distribuição por variante do português (`variante`)
- ( ) Distribuição pelas obras (`obra`)
- ( ) Distribuição por autores (`autor`)
- ( ) Distribuição por século (`seculo`)
- ( ) Distribuição pela corrente literária (`escola`)
- ( ) Distribuição pelo sexo do entrevistado, do biografado ou do autor (`sexo`)
- ( ) Distribuição por campo semântico (`sema`)
- ( ) Distribuição por grupo (de cor, roupa, etc.) (`grupo`)
- ( ) Distribuição das dependências (`dependencias`)

Opções

- [ ] Resultados por ordem alfabética (só distribuições)
- [ ] Ignorar maiúsculas/minúsculas (não admite parâmetros)

Fazer nuvem com limite de [ ]

Amostra aleatória de [ ] linhas.

Tipo	Textos vários
Variante(s)	PT BR
Tamanho (unidades)	6.1 milhões
Tamanho (palavras)	5.0 milhões

Carateres úteis: | { } [ ]

Página principal

Procure noutros corpos:

AmostRA-NILC ANCIB Avante! Corpus Brasileiro CD HAREM CETEMPúblico CHAVE Ciência Viva **Colonia** CONDIVport CONDIVport2 CoNE C-Oral-Brasil DHBB DiaCLAV Diáspora TL-PT ECI-EBR ECI-EE ENPCPUB (parte em português) Floresta FrasesPB FrasesPP Mariano Gago Literateca Marielle, presente! Moçambula Museu da Pessoa Natura/Minho NOBRE OBras P'lo Norte Português Falado - Documentos Autênticos ReLi NILC/São Carlos todos juntos Tycho Brahe Vercial

Informações sobre expressões de busca

- http://www.linguateca.pt/acesso/PJR.html
- http://www.linguateca.pt/acesso/exemplos.html
- http://www.linguateca.pt/Diana/download/instrACDC.pdf
- http://www.linguateca.pt/acesso/anotacao.html

Exemplos de consulta

- Quais as palavras terminadas em *mente* que não são advérbios?
  - Procura: [lema=".*mente" & pos!="ADV.*"]
  - Resultado: distribuição de lemas

Exemplos de consulta

- Qual classe gramatical delas?
  - Procura: [lema=".*mente" & pos!="ADV.*"]
  - Resultado: distribuição de pos

Exemplos de consulta

- Quais palavras terminam em *udo* ou *ento* e que não são ADJ e N?
  - Procura: [lema=".*udo|.*ento" & pos!="ADJ|N"]
  - Resultado: distribuição de lemas

Exemplos de consulta

- Quais palavras que termina em *ando* ou *endo* ou *indo* e que *não e*stão no gerúndio e apenas no português do Brasil ?
  - Procura: [lema=".*[aei]ndo" & temcagr!="GER" & variante="BR"]
  - Resultado: distribuição de lemas

Exercícios

- Como era adjetivado o personagem feminino nos romances literários ? Era diferente do personagem masculino ?
  - Encontre adjetivos associados aos substantivos
    - Mulher e Homem (incluir moça, rapariga, esposa(o), companheira(o), ...)
- Encontre textos relativos ao racismo
- Encontre em textos do que é que se tem *medo*
- Encontre textos sobe a **emoção** ódio
- Encontre o padrão X é um Y que (ex. ...sonambulismo é um fenômeno que...)

Extração em Python + Spacy

- Google Colab
  - https://colab.research.google.com/drive/1l6oCWa7uvQx9eiSYtw__PboBv_aiB0wD?usp=sharing