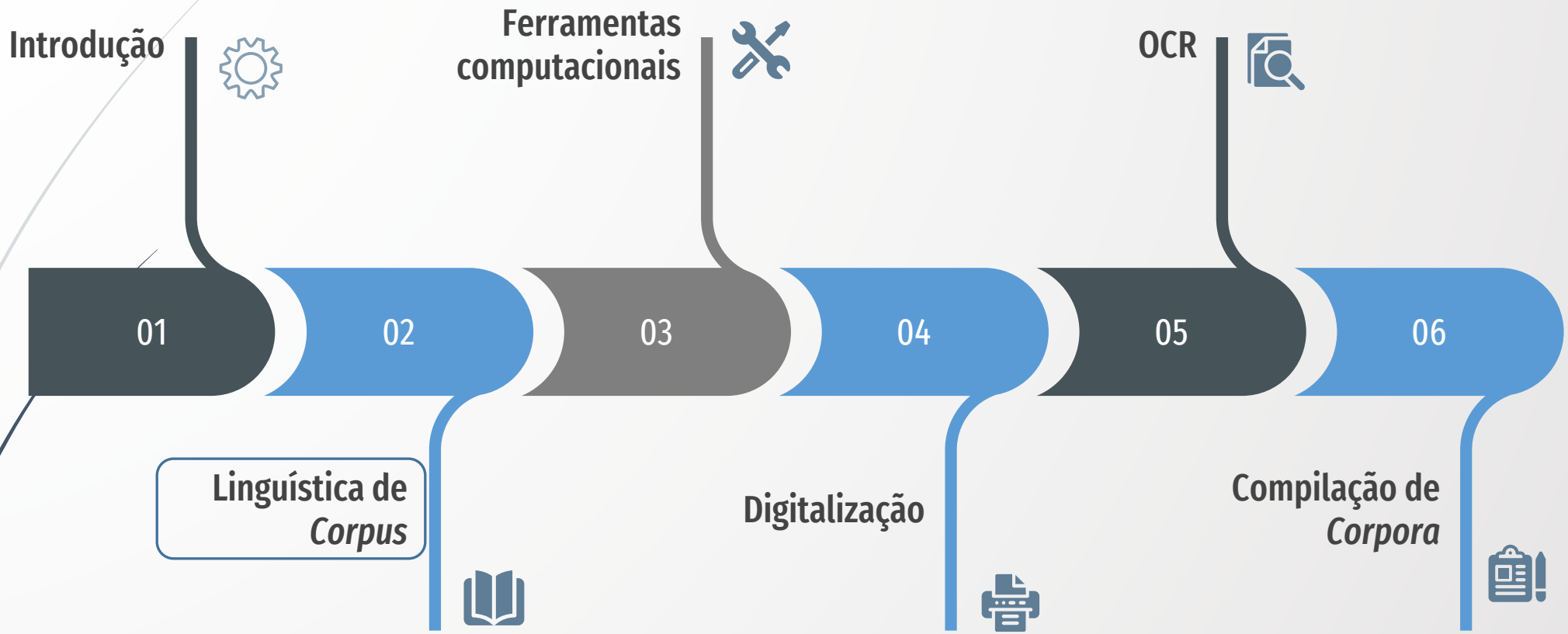
Humanities Going Digital (HUGOD)

[2020-1-CZ01-KA226-HE-094363]

Ferramentas computacionais de exploração de *Corpora*: introdução	Sessão 07	24/10/2022	Chiara Barbero Sílvia Barbosa

AULAS


Introdução
Ferramentas computacionais
OCR
01
02
03
04
05
06
Linguística de *Corpus*
Digitalização
Compilação de *Corpora*

Ferramentas computacionais para exploração de *Corpora*

O que nos permitem?

- Desconstrução do texto (ou conjunto de textos) em poucos segundos → fragmentação
- Análise quantitativa dos dados (frequências de ocorrências, distribuição etc.)
- Dados estruturados e reutilizáveis para diferentes tipos de estudo → muito usados para estudos sobre a língua, mas não só.

Importante: Permite análise crítica dos dados!

- O facto de não termos evidências relativamente a um determinado fenómeno, não implica que este não exista, pode apenas não constar no *corpus* recolhido → não podemos inferir de forma direta "evidências negativas"

Tratamento de corpora: ferramentas e procedimentos

- Separação de unidades lexicais: **tokenizador**
- Lematização verbal: **lematizador**
- Anotação morfossintática - **etiquetador/anotador**
- Análise quantitativa (frequência de ocorrência, densidade lexical, contagem de caracteres/unidades lexicais/sílabas/orações..)
- Concordanceamento - concordanciadores para seleção de contextos de ocorrências anotados
- Extração de palavras-chave
- …

Tokenizador

Para que serve?

- Indicar onde começa e acaba uma palavra/unidade de significado/unidade lexical:
  - nem sempre o que está entre dois espaços em branco é a palavra, depende da língua em uso (ex. línguas aglutinantes, línguas com ideogramas etc.)
- para identificar as palavras compostas como único itens atómicos
  - (couve-flor; rés-do-chão)

Exemplos de Tokenizadores

LX Tokenizer

Exemplo | Processar ficheiros | Notebook | Web Service | Documentação

Dentro deste parágrafo, há vários casos especiais para a separação de palavras na ortografia do português.
Tu também deste este exemplo: dar-se-lho-ia.
E prà frente é que é pelo caminho certo, etc.

Limpar | Separar

 <s> Dentro de_ este parágrafo ,*/ há vários casos especiais para a separação de palavras em_ a ortografia de_ o português .*/ </s> <s> Tu também deste este exemplo :*/ dar-CL-ia -se -lhe_ -o .*/ </s> <s> E para_ a frente é que é por_ o caminho certo ,*/ etc .*/ </s>

https://portulanclarin.net/workbench/lx-tokenizer/

SegmentAnt

File Help

Input Text | Clear

语料库语言学
维基百科，自由的百科全书
语料库语言学是基于语言运用的实例的语言研究。语料库语言学可以对自然语言进行语法与句法分析，还可以研究它与其他语言的关系。语料库最初由手工完成，而现在主要是由电子计算机自动完成。
语料库语言学家相信，可靠的语言分析需建立在新鲜的语料、自然的语言环境，和最小的实验干扰之上。在语料库语言学中，语料标注的意义众说纷纭，从约翰·辛克莱主张最少量的标注，并允许文本"为自己说话"，到"英语用法调查组"（设在伦敦大学学院）[2]鼓励更多的标注，并认为它是通向更完备和严谨的语言理解的道路。
历史
现代语料库语言学的个里程碑是亨利·库切拉和W.纳尔逊弗朗西斯在年出版的《当代美语的计算分析》一书。该项工作基于对布朗语料库的分析，布朗语料库是一个精心编制的美国英语语料库，规模约有一百万词次。库切拉和弗朗西斯将这些语料用于各种计算分析，获得了丰富和多样化的成果，该成果结合了语言学、语言数、心理学、统计学、和社会学元素。另一关键出版物是年伦道夫·夸克的《当代英语语法》在这本书中他介绍了"英语用法调查"项目。
此后不久，波士顿出版商霍顿米夫林邀请库切拉为其新的美国传统英语字典提供百万词次，三缕引文的来进行词典编纂。《美国传统英语字典》创新地将规定性元素（应如何使用语言）和描述性元素（语言实际上是如何被使用）结合在了一起。
其他出版社纷纷效仿。英国出版商柯林斯单语学习词典，就是为非英语母语者学习英语而出版的，它使用了"英语银行"语料库。"英语用法调查"语料库被用于由夸克等人编著的《综合英语语法》

Output Results | Clear

语料 库 语言学
维基百科 自由 的 百科全书
语料 库 语言学 是 基于 语言 运用 的 实例 的 语言 研究 语料 库 语言学 可 以对 自然语言 进行 语法 与 句法分析 还可 以 研究 它与 其他 语言 的 关系 语料 库 最初 由 手工 完成 而 现在 主要 是由 电子 计算机 自动 完成
语料 库 语言学家 相信 可靠 的 语言 分析 需 建立在 新鲜 的 语料 自然 的 语言 环境 和 最小 的 实验 干扰 之上 在 语料 库 语言学 中 语料 标注 的 意义 众说纷纭 从 约翰 辛 克莱 主张 最 少量 的 标注 并 允许 文本 为 自己 说话 到 英语 用法 调查组 设 在 伦敦大学 学院 [2] 鼓励 更多 的 标注 并 认为 它是 通向 更 完备 和 严谨 的 语言 理解 的 道路
历史
现代 语料 库 语言学 的个 里程碑 是 亨利 库切 切拉 拉和 W. 纳尔 尔逊 逊弗 弗朗 朗西 西斯 在年 出版 的 当代 美语 的 计算 分析 一书 该项 工作 基于 对 布朗 语料 库 的 分析 布朗 语料 库 是 一个 精心 编制 的 美国 英语 语料 库 规模 约有 一 百万 词次 库切 切拉 拉和 和弗 弗朗 朗西 西斯 斯将 这些 语料 用于 各种 计算 分析 获得 了 丰富 和 多样化 的 成果 该 成果 结合 了 语言学 语 言数 心理学 统计学 和 社会学 元素 另一 关键 出版物 是年 伦道夫 夸克 的 当代 英语语法 在这 本书 中他 介绍 了 英语 用法 调查 项目
此后 不久 波士顿 出版商 霍顿 米夫 林 邀请 库切 切拉 拉为 为其 其新 新的 美国 传统 英语 字典 提供 百万 词次 三缕 引文 的来 进行 词典 编纂 美国 传统 英语 字典 创新地 将 规定性 元素 应 如何 使用 语言 和 描述性 元素 语言 实际上 是 如何 被 使用 结合 在了 一起
其他 出版社 纷纷 效仿 英国 出版商 柯林 林斯 斯单 单语 学习 词典 就是 为非 英语 母语 者 学习 英语 而 出版 的 它 使用 了 英语 银行 语料库 英语 用法 调查 语料库 被 用于 由 夸克 等人 编著 的 综合 英语语法

https://www.laurenceanthony.net/software/segmentant/

Lematizador

Lema = forma canônica/neutra das palavras (entrada do dicionário)

Permite:

- agrupar diferentes ocorrências de uma mesma unidade lexical
- distinguir palavras homógrafas

Para que serve?

→ para cálculo de frequência

→ para desambiguação

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

"casa" - distribuição por formas/lemas

Exemplo de Lematizador

Stanza

Example Usage

Running the LemmaProcessor requires the TokenizeProcessor, MWTProcessor, and POSProcessor. After the pipeline is run, the `Document` will contain a list of `Sentence`s, and the `Sentence`s will contain lists of `Word`s. The lemma information can be found in the `lemma` field of each `Word`.

Accessing Lemma for Word

Here is an example of lemmatizing words in a sentence and accessing their lemmas afterwards:

```
import stanza

nlp = stanza.Pipeline(lang='en', processors='tokenize,mwt,pos,lemma')
doc = nlp('Barack Obama was born in Hawaii.')
print(*[f'word: {word.text+" "}\tlemma: {word.lemma}' for sent in doc.sentences for word in sent.words], sep='
```

As can be seen in the result, Stanza lemmatizes the word *was* as *be*.

```
word: Barack    lemma: Barack
word: Obama     lemma: Obama
word: was       lemma: be
word: born      lemma: bear
word: in        lemma: in
word: Hawaii    lemma: Hawaii
word: .         lemma: .
```

https://stanfordnlp.github.io/stanza/lemma.html#example-usage

LX Lematizador beta

Exemplo | Documentação

ter-se-lhas-ia apresentado

Clicar para caracteres especiais:

á ã ç ê é í õ ô ú ü

Limpar | apenas verbos conhecidos | Lematizar

teria apresentado (+ se + lhe + as)

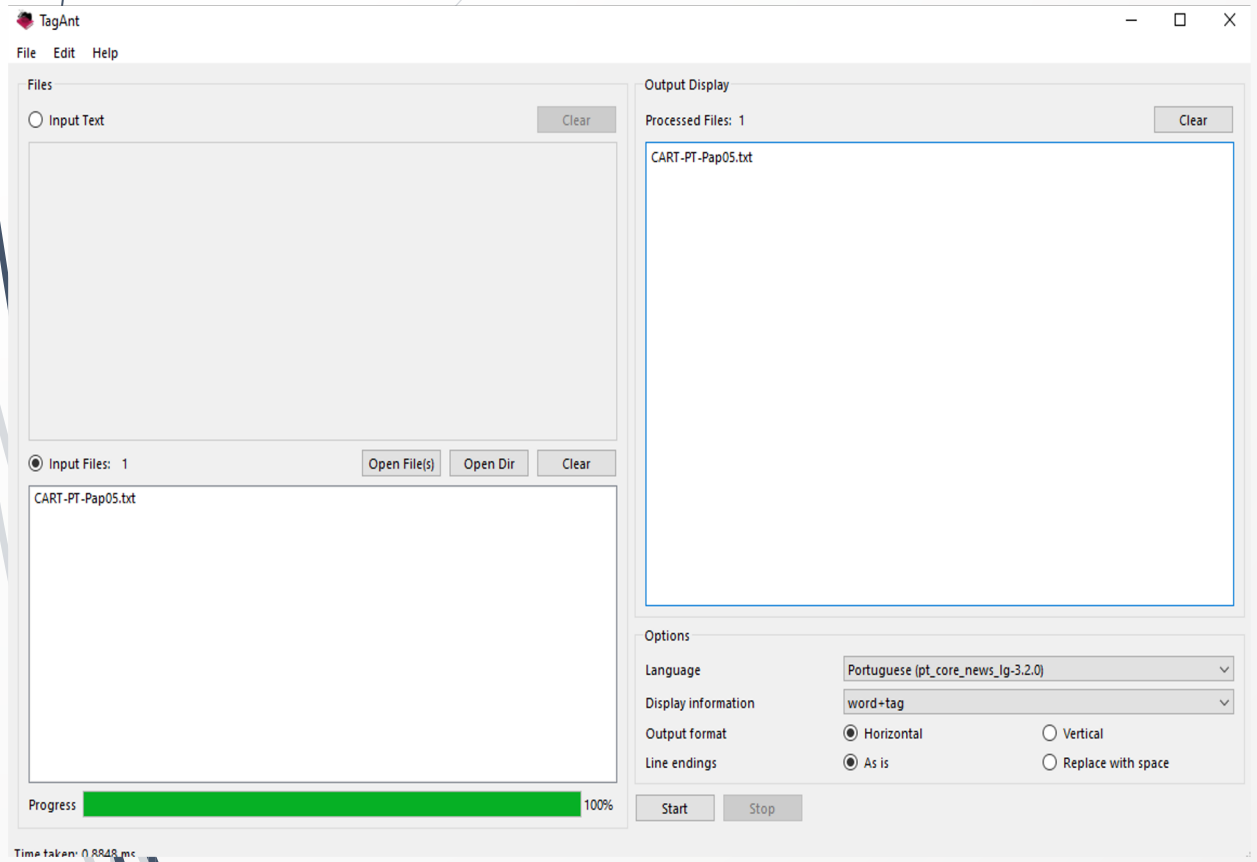
apresentar
indicativo | futuro do pretérito composto | 3ª pessoa | singular
cjg >>

https://portulanclarin.net/workbench/lx-lemmatizer/

Etiquetador

- Etiquetador morfossintático: ferramenta que atribui a cada palavra num texto a categoria morfossintática correspondente.
- POS (tagging) = Part-of-speech - categoria morfossintática
- Para que serve?
  - para efeitos de classificação das unidades lexicais que compõem o corpus
  - para desambiguação
    - https://www.linguateca.pt/acesso/corpus.php?corpus=CONDIV
    - "clara" distribuição por POS
    - "corte" distribuição por POS

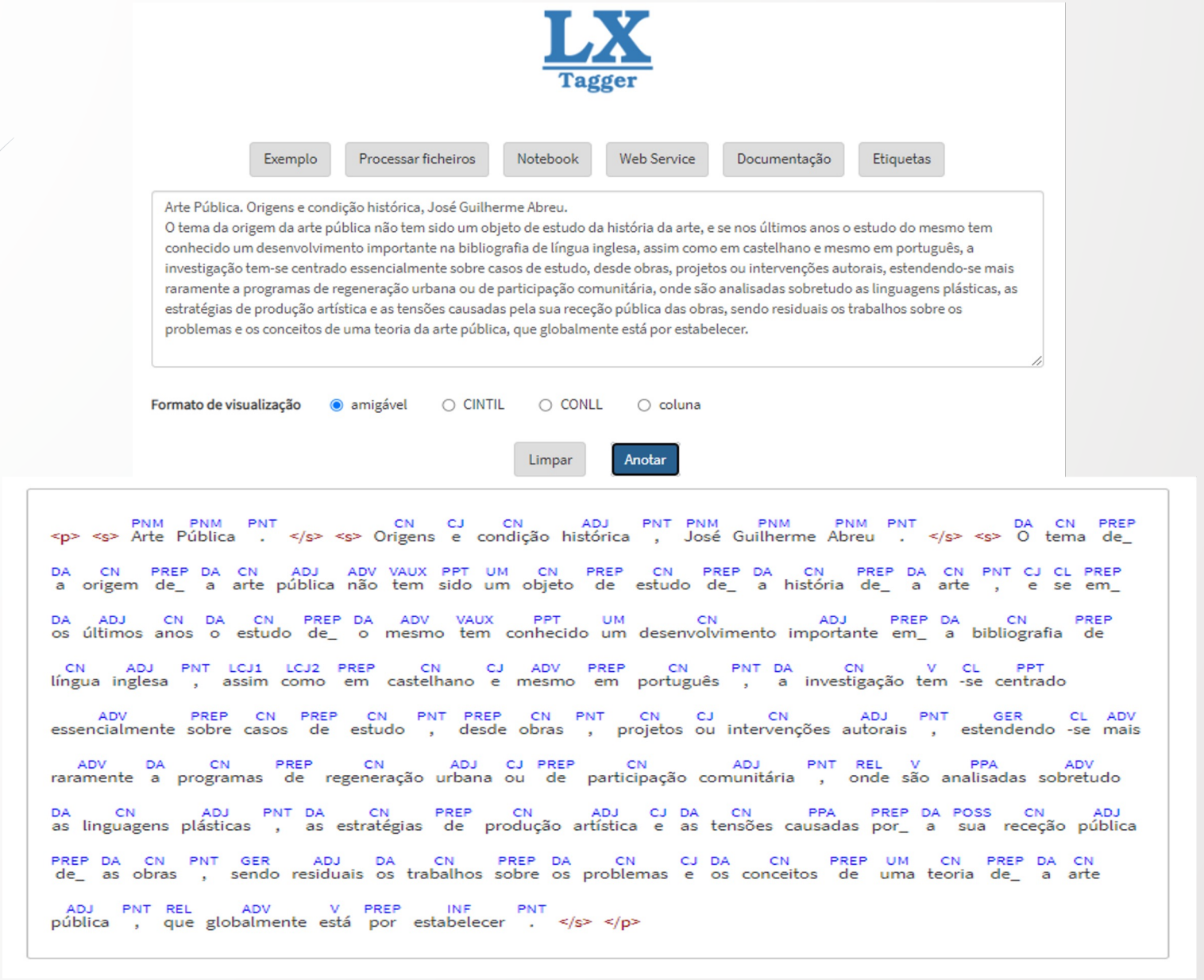
Exemplo de Etiquetador




```
Arte_PROPN Pública_PROPN ._PUNCT Origens_NOUN e_CCONJ condição_NOUN histórica_ADJ ,_PUNCT José_PROPN Guilherme_PROPN Abreu_PROPN ._PUNCT
1_NUM ._PUNCT Introdução_NOUN
O_DET tema_NOUN da_ADP origem_NOUN da_ADP arte_NOUN pública_ADJ não_ADV tem_AUX sido_AUX um_DET objeto_NOUN de_ADP estudo_NOUN da_ADP histó
Assim_ADV sendo_AUX ,_PUNCT importa_VERB desde_ADP já_ADV referir_VERB que_SCONJ apesar_ADV do_DET seu_DET título_NOUN algo_ADV ambicioso_A
2_NUM ._PUNCT Excerto_NOUN etimológico_ADJ ._PUNCT
Antes_ADV de_ADP entrar_VERB no_ADP estudo_NOUN histórico_ADJ propriamente_ADV dito_VERB ,_PUNCT importa_VERB interrogar_VERB a_DET origem_
Etimologicamente_ADV ,_PUNCT público_NOUN é_AUX uma_DET palavra_NOUN que_PRON deriva_NOUN do_ADP latim_NOUN clássico_ADJ "_PUNCT publicus_N
O_DET processo_NOUN de_ADP formação_NOUN do_ADP vocábulo_NOUN publicus_NOUN decorre_VERB ,_PUNCT por_NOUN sua_ADP vez_NOUN ,_PUNCT da_ADP a
•_PUNCT Enquanto_ADV coisa_NOUN da_ADP população_NOUN designa_VERB o_PRON que_PRON é_AUX comum_ADJ e_CCONJ partilhável_ADJ por_ADP todos_PR
•_PUNCT Enquanto_ADP coisa_NOUN do_ADP povo_NOUN designa_VERB o_PRON que_PRON é_AUX comum_ADJ e_CCONJ partilhado_VERB coletivamente_ADV ._P
•_PUNCT Enquanto_ADV coisa_NOUN dirigida_VERB à_ADP população_NOUN designa_VERB uma_DET comunicação_NOUN destinada_VERB a_ADP todos_PRON ._
•_PUNCT Enquanto_ADP coisa_NOUN oposta_ADJ ao_ADP privado_ADJ designa_VERB oposição_NOUN à_ADP exclusão_NOUN da_ADP partilha_NOUN do_PRON q

Função_NOUN mediadora_ADJ da_ADP Esfera_PROPN Cívica_PROPN


Sentido_NOUN de_ADP identidade_NOUN ._PUNCT
O_DET sentido_NOUN polissémico_ADJ da_ADP noção_NOUN de_ADP público_NOUN reflete-se_VERB na_ADP linguística_NOUN de_ADP forma_NOUN muito_AD
A-_ADP Como_ADP substantivo_ADJ ,_PUNCT público_NOUN designa_VERB :_PUNCT
•_PUNCT 1_NUM ._PUNCT A_DET comunidade_NOUN ou_CCONJ povo_NOUN como_ADP uma_DET totalidade_NOUN ._PUNCT Ex_NOUN ._PUNCT :_PUNCT os_DET cida
•_PUNCT 2_NUM ._PUNCT Um_DET setor_NOUN da_ADP população_NOUN que_PRON compartilha_VERB um_DET interesse_NOUN comum_ADJ ._PUNCT Ex_NOUN ._P
•_PUNCT 3_NUM ._PUNCT Os_DET admiradores_NOUN ou_CCONJ seguidores_NOUN de_ADP uma_DET pessoa_NOUN famosa_ADJ ._PUNCT Ex_NOUN ._PUNCT :_PUNC
B-_PROPN Como_ADP adjetivo_NOUN ,_PUNCT público_NOUN discrimina_VERB :_PUNCT
•_PUNCT 1_NUM ._PUNCT  _SPACE O_PRON que_PRON pertence_VERB ,_PUNCT afeta_VERB ou_CCONJ se_PRON relaciona_VERB com_ADP a_DET comunidade_NO
•_PUNCT 2_NUM ._PUNCT  _SPACE O_PRON que_PRON é_AUX mantido_VERB e/_NOUN ou_CCONJ usado_VERB pelo_ADP povo_NOUN ou_CCONJ pela_ADP comunida
•_PUNCT 3_NUM ._PUNCT  _SPACE O_PRON que_PRON éparticipado_VERB ou_CCONJ dirigido_VERB ao_ADP povo_NOUN ou_CCONJ à_ADP comunidade_NOUN ._P
•_PUNCT 4_NUM ._PUNCT  _SPACE O_PRON que_PRON se_PRON fag_VERB em_ADP benefício_NOUN do_ADP povo_NOUN ,_PUNCT da_ADP comunidade_NOUN ou_CC
•_PUNCT 5_NUM ._PUNCT  _SPACE O_PRON que_PRON se_PRON integra_VERB ou_CCONJ funciona_VERB numa_ADP escola_NOUN pública_ADJ ._PUNCT Ex_NOUN
•_PUNCT 6_NUM ._PUNCT  _SPACE O_PRON que_PRON é_AUX do_ADP conhecimento_NOUN e_CCONJ juígo_PROPN de_ADP todos_PRON ._PUNCT Ex_INTJ ._PUNCT
C-_PROPN Como_ADP verbo_NOUN ,_PUNCT publicar_VERB significa_VERB :_PUNCT
•_PUNCT 1_NUM ._PUNCT Tornarpatente_ADV e_CCONJ manifestar_VERB algo_PRON ao_ADP público_NOUN ._PUNCT Ex_INTJ ._PUNCT :_PUNCT a_DET publici
•_PUNCT 2_NUM ._PUNCT Revelar_VERB ou_CCONJ diger_VERB o_PRON que_PRON estava_AUX secreto_ADJ ou_CCONJ oculto_ADJ ._PUNCT Ex_INTJ ._PUNCT :
•_PUNCT 3_NUM ._PUNCT Difundirpor_PROPN medio_PROPN da_ADP imprensa_NOUN ou_CCONJ de_ADP qualquer_DET outro_DET procedimento_NOUN técnico_A

3_NUM ._PUNCT O_DET Conceito_PROPN de_ADP Público_PROPN em_ADP Hannah_PROPN Arendt_PROPN
Sobre_ADP conceito_NOUN de_ADP público_NOUN ,_PUNCT a_DET definição_NOUN de_ADP Hannah_PROPN Arendt_PROPN expressa_VERB com_ADP clareza_NOU
O_DET termo_NOUN público_ADJ denota_VERB dois_NUM fenómenos_NOUN intimamente_ADV relacionados_VERB ,_PUNCT mas_CCONJ não_ADV completamente_
E_CCONJ mais_ADV adiante_ADV ,_PUNCT acrescenta_VERB :_PUNCT
Em_ADP segundo_ADJ lugar_NOUN ,_PUNCT o_DET termo_NOUN público_ADJ significa_VERB o_DET próprio_DET mundo_NOUN ,_PUNCT na_ADP medida_NOUN e
Em_ADP Arendt_PROPN ,_PUNCT a_DET dialética_NOUN da_ADP noção_NOUN de_ADP público_NOUN opõe_VERB a_DET delimitação_NOUN passiva_ADJ do_ADP
Ontologia_NOUN e_CCONJ Genealogia_PROPN de_ADP Público_PROPN
Sentido_NOUN abstrato_ADJ -_PUNCT Integração_NOUN -_PUNCT O_DET substantivo_NOUN
Público_PROPN é_AUX o_PRON que_PRON é_AUX comum_ADJ (_PUNCT o_DET ser_NOUN :_PUNCT o_DET conceito_NOUN não_ADV inscrito_VERB )_PUNCT
Sentido_NOUN de_ADP atividade_NOUN -_PUNCT síntese_NOUN operativa_ADJ ._PUNCT o_DET verbo_NOUN
Publicar_VERB é_AUX comunicar_VERB o_PRON que_PRON é_PUNCT comum_ADJ (_PUNCT o_DET modo_NOUN de_ADP ser_VERB :_PUNCT o_DET inscrever_VERB )
Sentido_NOUN concreto_ADJ -_PUNCT diferenciação_NOUN -_PUNCT o_DET adjetivo_NOUN
Público_PROPN é_AUX o_PRON que_PRON é_AUX partilhado_ADJ (_PUNCT o_DET ente_NOUN :_PUNCT o_DET conceito_NOUN inscrito_VERB )_PUNCT
4_NUM ._PUNCT Complexo_NOUN conceptual_ADJ de_ADP Arte_PROPN Pública_PROPN
```

https://www.laurenceanthony.net/software/tagant/

Exemplo de Etiquetador



https://portulanclarin.net/workbench/lx-tagger/

Análise quantitativa

- Análise objetiva de dados (totais ou parciais) ou parâmetros específicos

Ex.

- lista de ocorrências (=*wordlist*)
- nº de frases, parágrafos, lemas, formas etc..
- tokens por texto
- distribuição do lema ou da expressão pelo corpus (com diferentes visualizações: diagramas, plots etc.)

Exemplo de Análise Quantitativa

LX Quantitative beta

Exemplo | Processar ficheiros | Notebook | Web Service | Documentação

A final do Campeonato Europeu de Futebol de 2016 realizou-se em 10 de julho de 2016 no Stade de France em Saint-Denis, França. Foi disputada entre Portugal e a França, que era a equipa anfitriã. Os portugueses ganharam a partida e sagraram-se campeões europeus de futebol. Esta foi a segunda participação numa final deste campeonato para Portugal e a terceira para a França. Os portugueses haviam participado anteriormente nas edições de 1984 e em todas as edições desde 1996. O seu melhor resultado anterior foi em 2004, com o título de vice-campeão. Já os franceses participaram em 1960, 1984 e em todas as edições desde 1992, tendo-se sagrado campeões nas edições de 1984 e de 2000.

Limpar | Analisar

Ocorrências de letras:	559
Média de letras por palavra:	4,47
Número de sílabas:	243
Média de sílabas por palavra:	1,94
Ocorrências de palavras:	125
Média de palavras por frase:	17,86
Proporção de palavras únicas:	52,00%
Frases:	7
Orações simples:	11
Orações passivas:	0 (0,00%)
Orações subordinadas:	1 (9,09%)
Orações coordenadas:	6
Índice de Flesch:	24,25

Densidade lexical:

Verbos	12	8,76%
Nomes	20	14,60%
Adjetivos	4	2,92%
Advérbios	2	1,46%
Preposições	28	20,44%

Frequências de palavras:

de	12
em	10
a	7
e	6
as	4
edições	4
o	4
1984	3
foi	3
frança	3
os	3
se	3
2016	2
campeonato	2
campeões	2
desde	2
final	2
futebol	2
para	2
portugal	2
portugueses	2
todas	2
10	1
1960	1
1992	1
1996	1
2000	1
2004	1
anfitriã	1
anterior	1
anteriormente	1
com	1
disputada	1

https://portulanclarin.net/workbench/lx-quantitative/

Concordanceador

- Permite dentro do *corpus* pesquisar, identificar e recuperar uma sequência específica de caracteres - uma palavra, uma parte de palavra, um sintagma etc.

- Que tipo de resultado fornece?
  - A sequência em causa está sempre evidenciada (**negrito**, cor diferente), normalmente é apresentada no meio da linha, mas não necessariamente. O contexto **antes** e **depois** da forma em questão, pode corresponder com uma frase completa ou com parte desta.

- Para que serve?
  - para analisar o contexto de ocorrência de formas ou sequências (é possível aplicar filtros para isolar coocorrentes mais ou menos distantes antes e depois da forma em causa)
  - para identificar padrões linguísticos recorrentes
  - para inferir informações - confirmar ou negar hipóteses prévias

Exemplo de Concordanceador

Linguateca

Concordância: linha de texto em que ocorre a sequência de caracteres em causa. Tipicamente para cada ocorrência corresponde uma linha.

Procura: **[lema="jogador"]**
Pedido de uma concordância em contexto
Corpo: CONDIVport 11.2

8090 ocorrências.

Número de ocorrências excessivo! Tente restringir a sua procura a menos de 4000 casos.

Concordância

Procura: **[lema="jogador"]**.

Apresenta-se uma amostra aleatória de 4000 das **8090** ocorrências encontradas.

par=fut-PT-Bola-50-5912: Dos três defesas, Alfredo continua a ser o mais **jogador** -- um jogador autêntico, aliás .

par=fut-PT-Bola-50-6215: Os recortes de alguns jornais que damos, a seguir devem habilitar os nossos leitores, a ajuizar melhor do reflexo causado pela decisão tomada por Cândido de Oliveira, e da forma como foram recebidas pelos **jogadores**, críticos e dirigentes as suas primeiras intervenções como orientador técnico do importante clube carioca .

par=fut-BR-GazetaEsp-50-60003: Até mesmo antes da leitura dos jornais, fervilhavam os dos «fans» em torno da atuação dos **jogadores**, cada qual, como é natural, procurando colocar em o seu idolo .

par=fut-BR-GazetaEsp-70-63841: — Ainda não sei se poderei contar com o **jogador** .

par=fut-BR-GazetaEsp-70-62295: **Jogadores** -- Armando, Osvaldo Cunha, Mendes, Vagner, Pedro Rodrigues, Ademir, Adãozinho, Peri, Servilio, Rocha, Lima, Mauri, Leonete, Ico, Nelson Lopes, José Eduardo, Neilo e Carlos Alberto .

par=fut-PT-Bola-50-2563: -- Em quatro meses de permanente labor, não registámos a mais pequena deselegância da parte dos **jogadores** .

par=fut-PT-Bola-50-9653: **Jogadores** correctos e público correctíssimo valorizaram o espectáculo o mais possível .

par=fut-PT-Bola-50-8988: Aproveitámos a ocasião para ouvir alguns **jogadores** que se dirigiam para o balneário .

par=fut-PT-Bola-70-17741: Coexistirão na equipa jogadores não amadores e **jogadores** profissionais ?

par=fut-BR-JSports-50-37601: Quando joga o scratch inglês os **jogadores** do scratch abandonam os seus clubes até em dias de matches decisivos do campeonato inglês .

par=fut-BR-JSports-70-43169: Mas, ainda que já tenha o seu time na cabeça, Zagalo enfrenta dois problemas sérios, resultantes de um mesmo **jogador**: Tostão .

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

Exemplo de Concordanceador

16

BNCweb (CQP-Edition)

Your query "time" returned 152502 hits in 3860 different texts (98,313,429 words [4,048 texts]; frequency: 1551.18 instances per million words) [0.289 seconds]

|< << >> >| Show Page: 1 | Show KWIC View | Show in random order | Show extended audio data controls | New Query | Go!

No	Filename	Hits 1 to 50 Page 1 / 3051
1	A00 81	Many people with AIDS have to spend long periods of **time** in hospital unless there is someone at home who can help and look after them.
2	A00 82	ACET volunteers work as part of a team and provide help in many different ways to ensure that people don't spend **time** in hospital unnecessarily.
3	A00 88	How much **time** to I need to give?
4	A00 134	Dr Dixon said, 'With up to 20 years from infection to illness, we just have to ask how many of our congregation have been added during that **time**?
5	A00 158	It was the first **time** our national and international network had gathered together in one place and made us all realise just how much the work has grown.'
6	A00 206	By working co-operatively, long-term, with the people around me, I hope to continue for some **time** yet.
7	A00 223	After spending some **time**, often years, with us, people move on to other things and we need to fill the gaps.'
8	A00 226	In London, volunteer training programmes will now take place every September and February and will require an individual's **time** for one evening a week over a six-week programme.
9	A00 252	The biggest changes are in the length of **time** people ill with the disease are now surviving and in the nature of the illnesses themselves.
10	A00 275	These factors help explain some of the reasons why the total number of ACET clients covered at any one **time** by our on call service in London has more than doubled from 70 in April 1990 to over 150 by March 1991; and why the nature of the services required has become so much more sophisticated.
11	A00 340	By working co-operatively, longterm, with the people around me, I hope to continue for some **time** yet.
12	A00 363	Within a short space of **time** referrals were regularly coming in.
13	A00 392	He is not at all well, very breathless, and by the **time** we are in the care he is gasping for air.
14	A00 401	4.00pm — We set off again; this **time** via Tony's home to collect a variety of possessions, finally arriving at hospital no.3.
15	A01 61	Over a period of **time** they will all be ill.
16	A01 146	Strong friendship takes **time** to build.
17	A01 148	Friendship takes **time**.
18	A01 149	In most happy marriages, husband and wife continue to make **time** to be with each other, and to understand each other .
19	A01 188	Without ACET's practical support at home they could spend long periods of **time** in hospital unnecessarily.
20	A01 189	The reality of AIDS is that the person can die at any **time**.
21	A01 247	At the **time** you enter a Deed of Covenant, the covenant should be capable of lasting for more than 3 years , and there should be the intention by you that it does so.
22	A01 256	If you use the covenant form attached to this leaflet you can make each annual payment at any **time**, or by any instalments, you wish as long as you make the full annual payment by the end of each 12-month period.
23	A01 340	The intention of a Gift Aid scheme is to encourage giving without the donor being tied to a particular charity for this length of **time**.
24	A01 342	Under Gift Aid there are no formalities at the **time** of the gift, just a cheque or cash gift to the charity.
25	A01 343	The certificate on form R190(SD) must be completed but this can be done at any **time** and amounts to little more than a claim procedure.
26	A01 373	The donor must be resident in the United Kingdom at the **time** that the gift is made.
27	A01 381	If you wish to give small amounts regularly, e.g. monthly, you could accumulate the money in a separate account and then convert this to a gift to ACET every **time** the amount reaches £600.
28	A01 410	Moreover, death is a **time** of great stress to those you love most.
29	A01 470	This means you can get professional help any **time** of the day or night, and at weekends.
30	A01 529	A recent survey of church youth groups shows that 1 in 4 have had sex by the **time** they are eighteen years old, 1 in 10 under the age of sixteen.

http://corpora.lancs.ac.uk/BNCweb (http://corpora.lancs.ac.uk/clmtp/2-conc.php)

Extração de palavras-chave

- Processo automático utilizado para a extração de informação relevante a partir de um ou mais textos
- Baseada em cálculos estatísticos para identificar as palavras mais relevantes dentro do *input* textual
- Existem 2 abordagens para extrair palavras-chaves e para treinar as ferramentas automáticas:
  - por comparação com um standard (*reference corpus* ou *golden standard*), aplicando ou não algoritmos de *machine-learning* para melhorar constantemente o desempenho da ferramenta -> *language-dependent,* parcialmente *domain-dependent* e precisa do auxílio de ferramentas complementares (vocabulários controlados, recursos lexicais etc.)
    - (https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf)
  - baseada em medidas estatísticas (de coocorrência dos elementos do próprio texto) e nas características textuais do texto ou do conjunto de textos em análise
    - (https://www.sciencedirect.com/science/article/pii/S0020025519308588)

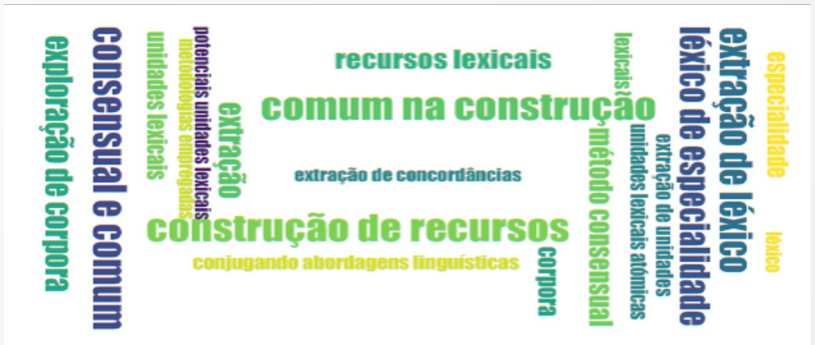
Exemplo de Palavra-chave

YAKE! LIAAD INESCTEC LABORATÓRIO ASSOCIADO U.PORTO beta

Texto anotado

A **exploração de corpora** para a **extração de léxico** de **especialidade** é um método **consensual e comum** na **construção de recursos** **lexicais** . No entanto, as **metodologias empregadas** não são explicitamente discutidas, dificultando a comparação e a determinação de abordagens robustas. Para preencher esta lacuna, neste artigo apresentamos e discutimos uma metodologia detalhada para **extração de léxico** de **especialidade** a partir de **corpora** , **conjugando abordagens linguísticas** e estatísticas. O método proposto prevê tanto o uso de **corpora** de **especialidade** como de **corpora** monitores e inclui: i) análise de dados de frequência; ii) **extração** de concordâncias e colocações; iii) **extração** de informação de ordem textual, permitindo a **extração** de **unidades lexicais** atómicas e multipalavra e de relações semânticas relevantes. Como princípio base de garantia de qualidade, a proposta inclui validação dos dados finais por parte de especialistas. Deste modo, o objetivo da metodologia é a determinação de listas de potenciais **unidades lexicais** de **especialidade** e de informações relevantes para a sua descrição que permitam uma validação final rápida e eficiente, maximizando o valor informacional da interação com os especialistas.

Nuvem de palavras



Palavras-chave

Pontuação	n-grama
0,021910596993882824	consensual e comum
0,021910596993882824	comum na construção
0,021910596993882824	construção de recursos
0,024277956086531063	extração de léxico
0,028746159502555813	léxico de especialidade
0,03643768539079015	exploração de corpora
0,04035126286249145	extração
0,04601595761636467	método consensual
0,046663892103767375	recursos lexicais
0,05921738501646989	especialidade
0,06002427740848347	corpora
0,06835989446700896	unidades lexicais
0,0866008319943846	extração de unidades
0,09639310023377093	unidades lexicais atómicas
0,09671816241649626	conjugando abordagens linguísticas
0,10149199943030174	lexicais
0,11141158535628319	potenciais unidades lexicais
0,11476075390596947	léxico
0,1149336208382416	metodologias empregadas
0,1160654588669079	extração de concordâncias

https://portulanclarin.net/workbench/liaad-yake/

Exemplo de Palavra-chave

YAKE! LIAAD INESCTEC LABORATÓRIO ASSOCIADO U.PORTO beta

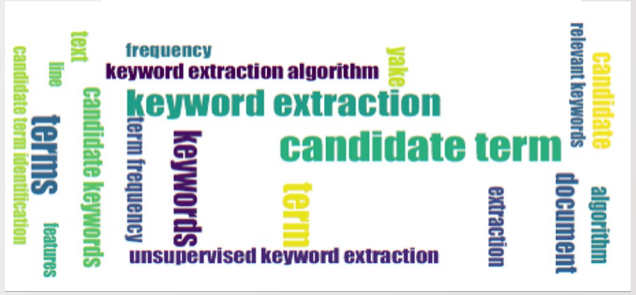
Texto anotado

future work Keyword extraction algorithms have become the key components in many computer science applications. In this article, we have presented YAKE !, a novel feature-based online system for multi-lingual keyword extraction from individual documents that can be used both as a stand-alone tool and as a support for several applications, including summarization, clustering, indexing, and information visualization, to name just a few. YAKE ! follows an unsupervised approach. It does not require a corpus or a dictionary (corpus independent), which means it can be quickly adapted across different languages or domains (domain and language independent). Our solution scales to any document length in a linear manner in the number of candidate terms identified and is term frequency-free, meaning that no conditions are set with respect to the minimum frequency or sentence frequency that a candidate keyword must have. Our experiments confirm that YAKE ! can lead to superior keyword extraction results compared with ten popular state-of-the-art unsupervised keyword extraction algorithms, under a large number of text documents belonging to twenty different datasets. It also offers similar or superior effectiveness to state-of-the-art supervised methods, without the need for an extensive training set for multiple languages and domains. As a further contribution, we provide an app on Google Play, a demo, an API, and a python package so that YAKE ! can be tested and used by the research community. YAKE ! has already been used for temporal clustering and ranking, and won first prize at the Arquivo.pt 2018 Portuguese internet archive contest with "Tell me Stories" [archive.tellmestories.pt], a news summarization platform which generates a timeline summary for a given news topic in a storytelling-style fashion. In the future, we plan to develop a supervised solution, which will be built using the statistical features introduced in this work. An analysis of the importance of each feature has been already conducted in this research. Furthermore, we are currently working on an extended version of YAKE ! to tackle the problem of absent keywords , for instance when a large number of manually-assigned keywords cannot be found in the text itself. We plan to do this based on word embeddings.

Palavras-chave

Pontuação	n-grama
0,0026072961246104177	candidate term
0,003392069824422952	term
0,003422799305155864	keyword extraction
0,0036689734835595195	terms
0,004687604124718088	keywords
0,006948990799399872	document
0,007403780495156102	candidate
0,008610707179194595	candidate keywords
0,00974898610265248	yake
0,010598117626789114	text
0,010984486480505979	term frequency
0,0120421914184087	algorithm
0,012438828392054232	keyword extraction algorithm
0,012615110881503395	extraction
0,012760619978787826	unsupervised keyword extraction
0,015886390222924134	frequency
0,017680157572175777	candidate term identification
0,01787379293168973	relevant keywords
0,020651804058790486	line
0,02120125058770971	features

Nuvem de palavras



https://portulanclarin.net/workbench/liaad-yake/

Corpus manager, corpus management systems ou corpus query systems

Existem ferramentas/interfaces que permitem explorar as potencialidades dos *corpora*, enquanto coletâneas de pistas linguísticas para análises/inferências linguísticas

Exemplos de programas a explorar:

- Sketch Engine (https://www.sketchengine.eu/)
- AntConc (https://www.laurenceanthony.net/software/antconc/)

- Linguateca (https://www.linguateca.pt/)

Instalar e Explorar alguns destes softwares

- Pensar num tópico de trabalho ex. *vinho*
- Pesquisar repositórios onde haja os textos relativos ao tópico *revistas, sites, …*
- Fazer um resumo sobre o trabalho a apresentar *Observando o panorama português, pouco tem sido feito no sentido de compreender e desenvolver estudos no domínio da Análise Sensorial Enológica ou prova de vinho ou prova organoléptica. Neste sentido, o objetivo deste trabalho é compreender quais os descritores utilizados nas notas de prova provenientes das provas de vinho, que constituem um tipo de discurso especializado em Enologia, com características próprias e realizadas por especialistas.*
- Inserir a informação relativa ao trabalho no formulário (para avaliação)

Avaliação

Trabalho de casa

- Escrever o resumo para submeter até ao dia 31/10/2022

Para a próxima aula

- Pensar no Enquadramento Teórico do trabalho
- Ter uma amostra do corpus para treinar em aula

Erasmus+

NOVAFCSH

EN	The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.
PT	O apoio da Comissão Europeia à produção desta publicação não constitui um aval do seu conteúdo, que reflete unicamente o ponto de vista dos autores, e a Comissão não pode ser considerada responsável por eventuais utilizações que possam ser feitas com as informações nela contidas.