universidade de aveiro

MAPi DOCTORAL PROGRAMME IN COMPUTER SCIENCE

ASSIGNMENT #3

Advanced Information Extraction

MAPi 2018-2019

Vanessa Silva

(vanessa.silva@dcc.fc.up.pt)

TARGET APPLICATION

- Searching for social action information in the Municipality of Paredes.
  - Places,
  - Contacts,
  - Actions,
  - Events,
  - Organization,
  - ….

Programa Social
PAREDES ajuda +
PAREDES

Loja Social de Rebordosa

Projeto "Mão Solidária"

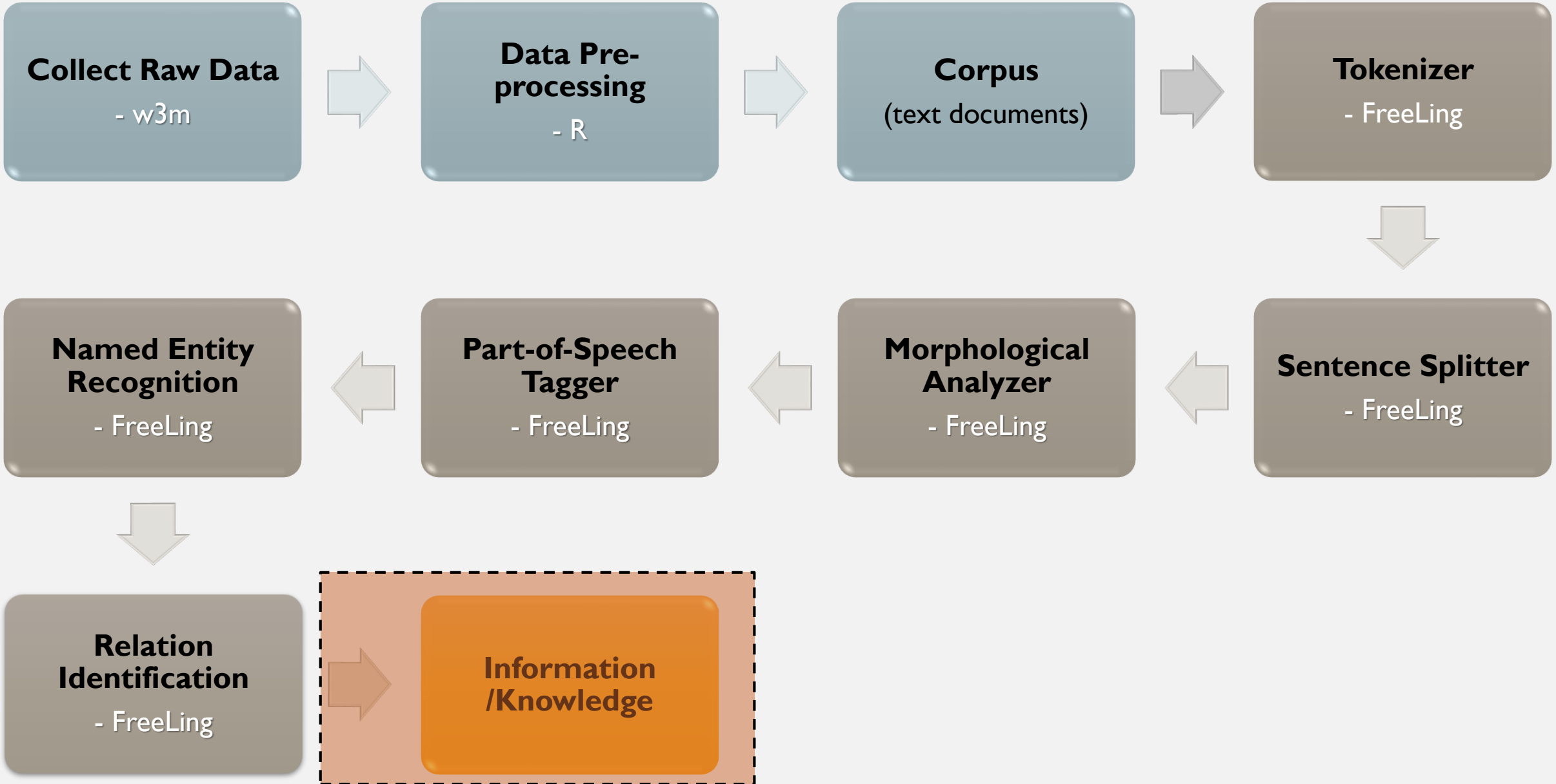
SERVIÇO DE ATENDIMENTO LOCAL SAL
› Atendimentos/Encaminhamentos
› Apoios Sociais

LOJA SOCIAL DE REBORDOSA:
Horário de Atendimento
› Segundas-feiras: 09h30 | 12h00
› Quartas-feiras: 09h30 | 12h00

PAREDES

CPCJ PAREDES
COMISSÃO DE PROTECÇÃO DE CRIANÇAS E JOVENS

APPLICATION PIPELINE



Collect Raw Data

- w3m

- **Source:** the web page of the Municipality of Paredes: https://www.cm-paredes.pt/pages/382
  - Initially it was also defined some online PDF files, however, so as not to make the corpus very heterogeneous.
- **Language:**
  - portuguese
- **To obtain the documents:**
  - To use **w3m** software which, compared to the *wget* software initially chosen, is more efficient for the desired return content.
    - The **w3m -dump** command automatically returns the information "clean" from the meta-information of the web document structure .

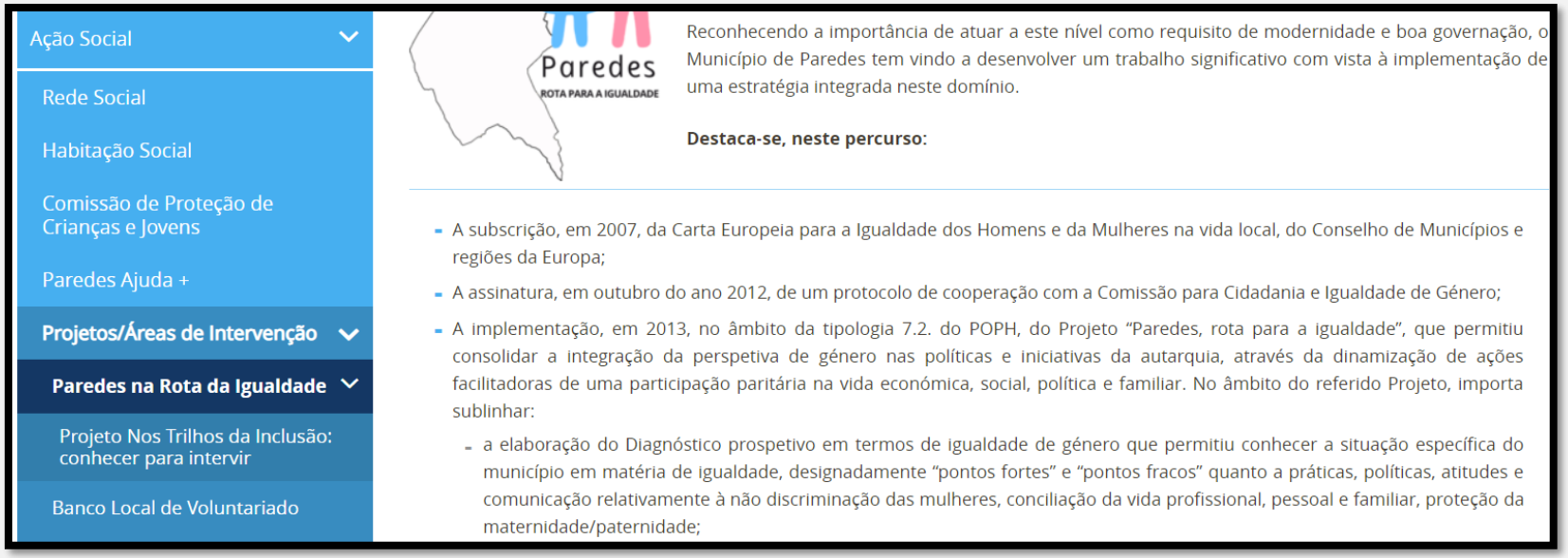
Data Pre-processing

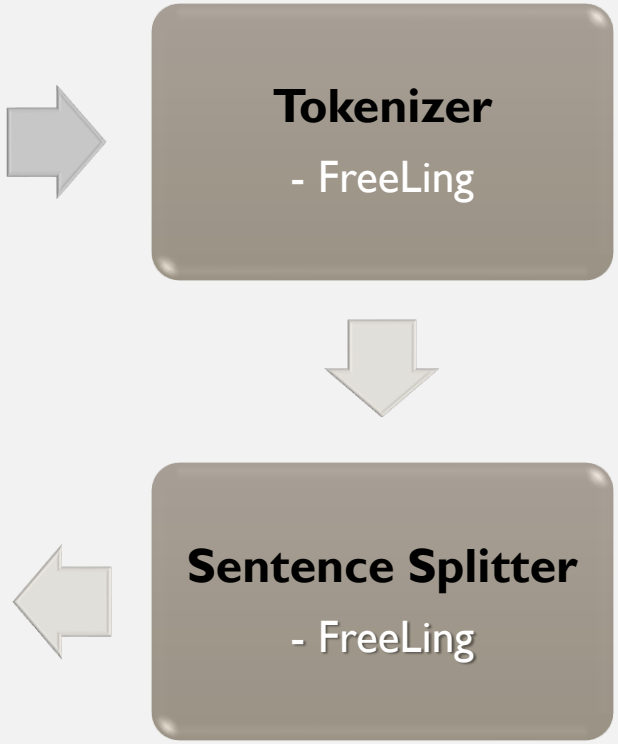
- R

Corpus

(text documents)

- To use the software R to handle the raw data.
  - Several lines of text before and after the lines with the desired text have been deleted: headers and footers of web pages.
  - Page titles and empty lines have been deleted.
  - Text with items has been converted to plain text.





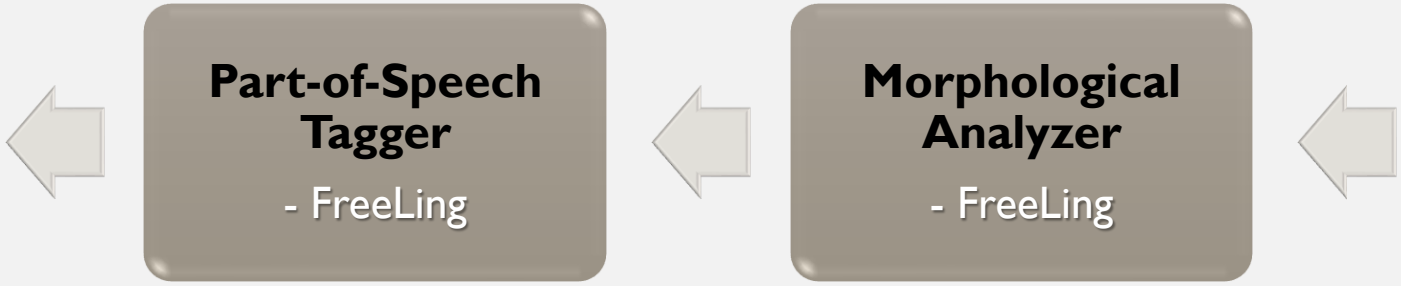
- FreeLing tokenizer converts plain text into a list of word objects corresponding to the tokens created.

[A]; [subscrição]; [,]; [em]; [2007]; [,]; [da]; [Carta]; [Europeia]; [para]; [a]; [Igualdade]; [dos]; [Homens]; [e]; [da]; [Mulheres]; [na]; [vida]; [local]; [,]; [do]; [Conselho]; [de]; [Municípios]; [e]; [regiões]; [da]; [Europa]; [;]; [A]; [assinatura];

- FreeLing's sentence splitter receives lists of word objects (produced by the tokenizer) and buffers them until a sentence limit is detected, and returns a list of sentence objects. The sentence limits are: **.**, **?** and **!**.

[Destaca-se , neste percurso : A subscrição , em 2007 , da Carta Europeia para a Igualdade dos Homens e da Mulheres na vida local , do Conselho de Municípios e regiões da Europa ; (…)]

[Dia Internacional pela Eliminação da Violência Contra as Mulheres : A ONU (Organização das Nações Unidas) fixou o 25 de Novembro como o Dia internacional pela eliminação da violência contra as mulheres .]



- Morphological analyzer in FreeLing is a meta-module that does not perform any processing itself, it calls submodules necessary for morphological analysis, such as: punctuation detection, number detection, dates detection, dictionary search, multiword recognition, quantity recognition and affix checking.
  - Each submodule receives which configure files should be used to create them.
  - To some of these files, information directed to the application domain has been added.
- FreeLing has a hmm_tagger class that implements a classical trigram Markovian tagger.

A; o; DA0FS0

subscrição; subscrição; NCFS000

,; ,; Fc

em; em; SP

2007; 2007; Z

,; ,; Fc

de; de; SP

a; o; DA0FS0

Carta_Europeia_para_a_Igualdade_de_os_Homens; carta_europeia_para_a_igualdade_de_os_homens; NP00000

e; e; CC

de; de; SP

a; o; DA0FS0

Mulheres; mulheres; NP00000

em; em; SP

a; o; DA0FS0

vida; vida; NCFS000

- local; local; AQ0CS00
- ,; ,; Fc
- de; de; SP
- o; o; DA0MS0
- Conselho_de_Municípios; conselho_de_municípios; NP00000
- e; e; CC
- regiões; região; NCFP000
- de; de; SP
- a; o; DA0FS0
- Europa; europa; NP00000
- ;; ;; Fx
- A; o; DA0FS0
- assinatura; assinatura; NCFS000
- ,; ,; Fc
- em; em; SP
- outubro_do_ano_2012; [??:??/10/2012:??.??:??]; W

Named Entity Recognition

- FreeLing

Relation Identification

- FreeLing

- Freeling offers two different modules capable of performing the recognition of NE:
  - **Basic NER:** detects word sequences in capital letters, taking into account some functional words (for example, *Município de Paredes*) and capitalization at the beginning of the sentence.
  - **BIO NER:** is based on machine learning, uses a classification algorithm (*AdaBoost* classifier) to decide whether each word is in a **begin** (B), **inside** (I) or **outside** (O).

Municípios; municípios; NP00G00

Município_de_Paredes; município_de_paredes; NP00G00

Carta_Europeia_para_a_Igualdade_de_os_Homens; carta_europeia_para_a_igualdade_de_os_homens; NP00V00

Conselho_de_Municípios; conselho_de_municípios; NP00V00

Europa; europa; NP00G00

Comissão_para_Cidadania_e_Igualdade_de_Género; comissão_para_cidadania_e_igualdade_de_género; NP00O00

POPH; poph; NP00O00

Diagnóstico; diagnóstico; NP00V00

Rede_de_Municípios_Solidários; rede_de_municípios_solidários; NP00O00

Protoloco; protoloco; NP00V00

Governo; governo; NP00O00

Secretária_de_Estado; secretária_de_estado; NP00O00

Assuntos_Parlamentares; assuntos_parlamentares; NP00V00

Teresa_Morais; teresa_morais; NP00SP0

Reforma_Administrativa; reforma_administrativa; NP00O00

Paulo_Simões_Júlio; paulo_simões_júlio; NP00SP0

Named Entity Recognition

- FreeLing

Relation Identification

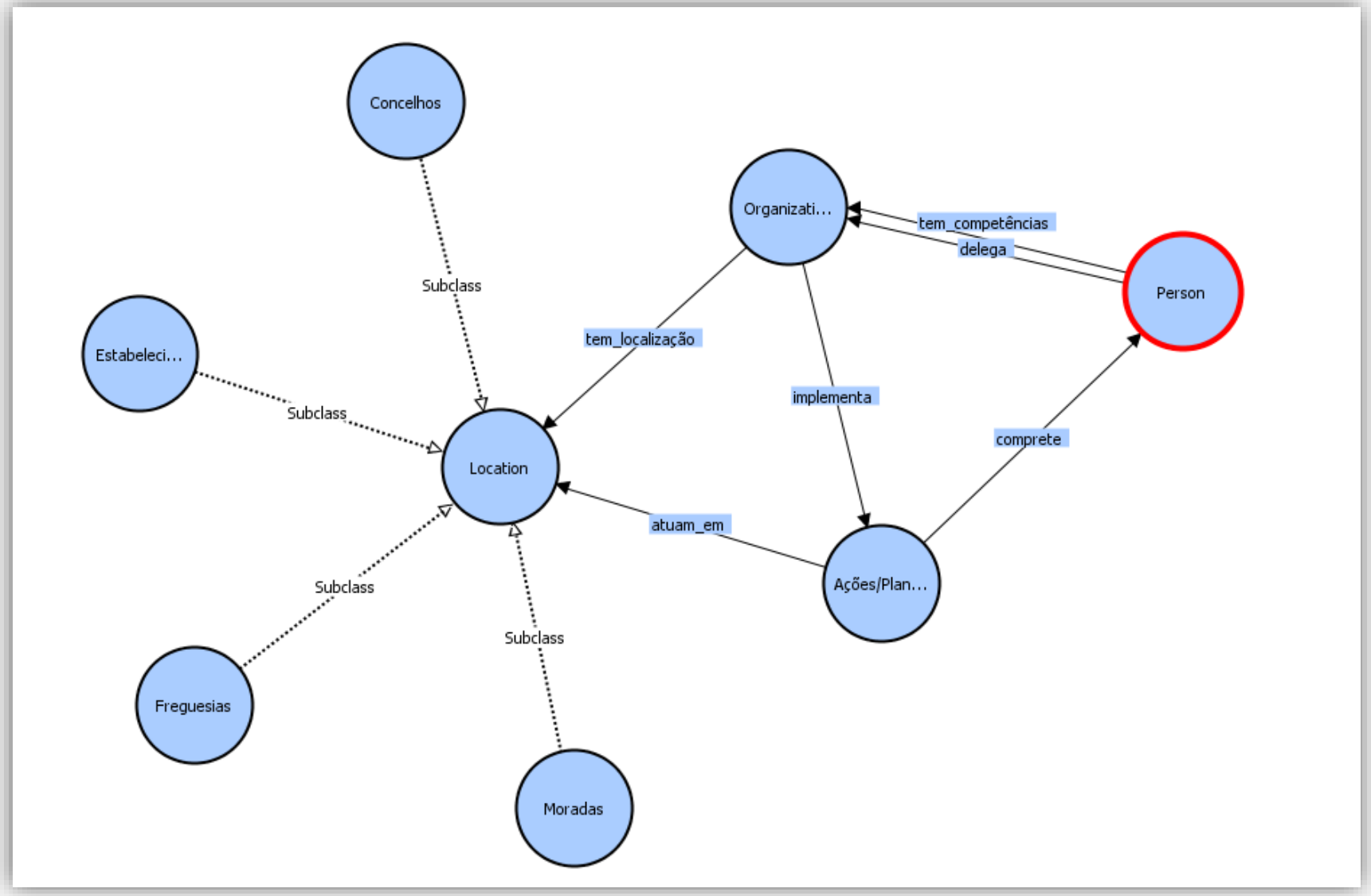
- FreeLing

- FreeLing has a statistical dependency analysis module, this is based on Treeler machine learning library.

1 Reconhecendo	reconhecendo	NP00V00	- - B-MISC	- - 19 nsubj - -
2 a	o	DA0FS0	- - -	- - 3 det - -
3 importância	importância	NCFS000	- - -	- - 19 nsubj - -
4 de	de	SP	- - -	- - 5 mark - -
5 atuar	atuar	VMN0000	- - -	- - 3 acl - -
6 a	a	SP	- - -	- - 8 case - -
7 este	este	DD0MS0	- - -	- - 8 nmod - -
8 nível	nível	NCMS000	- - -	- - 5 nmod - -
9 como	como	CS	- - -	- - 19 mark - -
10 requisito	requisito	NCMS000	- - -	- - 19 nsubj - -
11 de	de	SP	- - -	- -
12 modernidade	modernidade	NCFS000	- - -	- - 10 nmod - -
13 e	e	CC	- - -	- - 10 cc - -
14 boa	bom	AQ0FS00	- - -	- - 10 conj - -
15 governação	governação	NCFS000	- - -	- - 10 conj - -
16 ,	,	Fc	- - -	- - 10 conj - -
17 o	o	DA0MS0	- - -	- - 18 det - -
18 Município_de_Paredes	município_de_paredes	NP00G00	- - B-LOC	- - 19 nsubj - -
19 tem	ter	VMIP3S0	- - -	- - 0 root - -
20 vindo	vir	VMP00SM	- - -	- - 19 xcomp - -

(…)

ONTOLOGY


Concelhos
Organizati...
tem_competências
delega
Person
Subclass
Estabeleci...
tem_localização
implementa
Subclass
Location
comprete
atuam_em
Ações/Plan...
Subclass
Subclass
Freguesias
Moradas

- I could not finish.