

Tarefas de PLN

Willian Massami Watanabe

Tarefas de PLN

- Tokenização
- Parts-Of-Speech (POS) tagger
- Entidades Nomeadas
- Lematização / Stemming
- Stop words

Inicialização do Spacy



<https://spacy.io/>

```
import spacy
```

```
p1n = spacy.load("pt_core_news_sm")
```

Inicialização do Spacy



<https://spacy.io/>



```
import spacy
```

```
p1n = spacy.load("pt_core_news_sm")
```

Inicialização do Spacy



<https://spacy.io/>

```
import spacy
```



```
pln = spacy.load("pt_core_news_sm")
```

É realizado o carregamento do modelo de linguagem para o idioma Português. Existem outros modelos que podem ser carregados para diferentes idiomas:

- pt_core_news_lg
- en_core_web_sm / en_core_web_trf
- ja_core_news_trf

Inicialização do Spacy



<https://spacy.io/>

```
import spacy
```

→

```
pln = spacy.load("pt_core_news_sm")
```

Esse carregamento do modelo normalmente vai ser realizado utilizando a Internet.

Terminal \$


```
python -m spacy download pt_core_news_sm
```

Inicialização do Spacy



<https://spacy.io/>

```
import spacy
```



```
pln = spacy.load("pt_core_news_sm")
```

A chamada `spacy.load` retorna um objeto de tipo callable. Isso significa que ele pode ser executado, como se fosse uma função.

Esse objeto é a porta de entrada para a maior parte das tarefas de PLN que serão utilizadas no curso.

```
doc = pln("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```



Inicialização do Spacy



<https://spacy.io/>

```
import spacy
```

```
pln = spacy.load("pt_core_news_sm")
```



```
doc = pln("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```

Objeto de tipo Document que possui os componentes do texto, já processados por tarefas de PLN do Spacy.

Inicialização do Spacy




<https://spacy.io/>

Tokenização e Part-Of-Speech tagger



<https://spacy.io/>



```
doc = pln("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```



Objeto de tipo Document que possui os componentes do texto, já processados por tarefas de PLN do Spacy.

Um Part-Of-Speech (POS) Tagger, ou etiquetador de partes do discurso, é uma tarefa de PLN que atribui categorias gramaticais, como substantivos, verbos, adjetivos, advérbios, entre outros, a cada palavra em um texto.

Tokenização e Part-Of-Speech tagger



<https://spacy.io/>



```
doc = pln("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```

```
for token in doc:  
    print(f'{token.text:20}\t {token.tag_:4}\t {token.lemma_:20}\t {token.is_stop}')
```

Tokenização e Part-Of-Speech tagger



<https://spacy.io/>

```
doc = p1n("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```

```
for token in doc:  
    print(f'{token.text:20}\t {token.tag_:4}\t {token.lemma_:20}\t {token.is_stop}') 
```

- token.text
- token.tag_
- token.lemma_
- token.is_stop

<https://spacy.io/api/token>

Tokenização e Part-Of-Speech tagger



<https://spacy.io/>

```
doc = pln("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```

```
for token in doc:  
    print(f'{token.text:20}\t {token.tag_:4}\t {token.lemma_:20}\t {token.is_stop}')  

```

Agora	ADV	agora	True
,	PUNCT	,	False
nós	PRON	nós	True
estamos	AUX	estar	False
no	ADP	em o	True
começo	NOUN	começo	False
da	ADP	de o	True
disciplina	NOUN	disciplina	False
de	ADP	de	True
PLN	PROPN	PLN	False
e	CCONJ	e	True
Mineração	PROPN	Mineração	False
de	ADP	de	True
Texto	PROPN	Texto	False
	PUNCT		False

Tokenização e Part-Of-Speech tagger



<https://spacy.io/>

```
doc = pln("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```

```
for token in doc:  
    print(f'{token.text:20}\t {token.tag_:4}\t {token.lemma_:20}\t {token.is_stop}')  

```

Agora	ADV	agora	True
,	PUNCT	,	False
nós	PRON	nós	True
estamos	AUX	estar	False
no	ADP	em o	True
começo	NOUN	começo	False
da	ADP	de o	True
disciplina	NOUN	disciplina	False
de	ADP	de	True
PLN	PROPN	PLN	False
e	CCONJ	e	True
Mineração	PROPN	Mineração	False
de	ADP	de	True
Texto	PROPN	Texto	False
	PUNCT		False

<https://universaldependencies.org/u/pos/>

Tokenização e Part-Of-Speech tagger



<https://spacy.io/>

```
doc = pln("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```

```
for token in doc:  
    print(f'{token.text:20}\t {token.tag_:4}\t {token.lemma_:20}\t {token.is_stop}') 
```

Agora	ADV	agora	True
,	PUNCT	,	False
nós	PRON	nós	True
estamos	AUX	estar	False
no	ADP	em o	True
começo	NOUN	começo	False
da	ADP	de o	True
disciplina	NOUN	disciplina	False
de	ADP	de	True
PLN	PROPN	PLN	False
e	CCONJ	e	True
Mineração	PROPN	Mineração	False
de	ADP	de	True
Texto	PROPN	Texto	False
	PUNCT		False

Tokenização e Part-Of-Speech tagger



<https://spacy.io/>

```
doc = pln("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```

```
for token in doc:  
    print(f'{token.text:20}\t {token.tag_:4}\t {token.lemma_:20}\t {token.is_stop}')  

```

Agora	ADV	agora	True
,	PUNCT	,	False
nós	PRON	nós	True
estamos	AUX	estar	False
no	ADP	em o	True
começo	NOUN	começo	False
da	ADP	de o	True
disciplina	NOUN	disciplina	False
de	ADP	de	True
PLN	PROPN	PLN	False
e	CCONJ	e	True
Mineração	PROPN	Mineração	False
de	ADP	de	True
Texto	PROPN	Texto	False
	PUNCT		False

Tokenização e Part-Of-Speech tagger

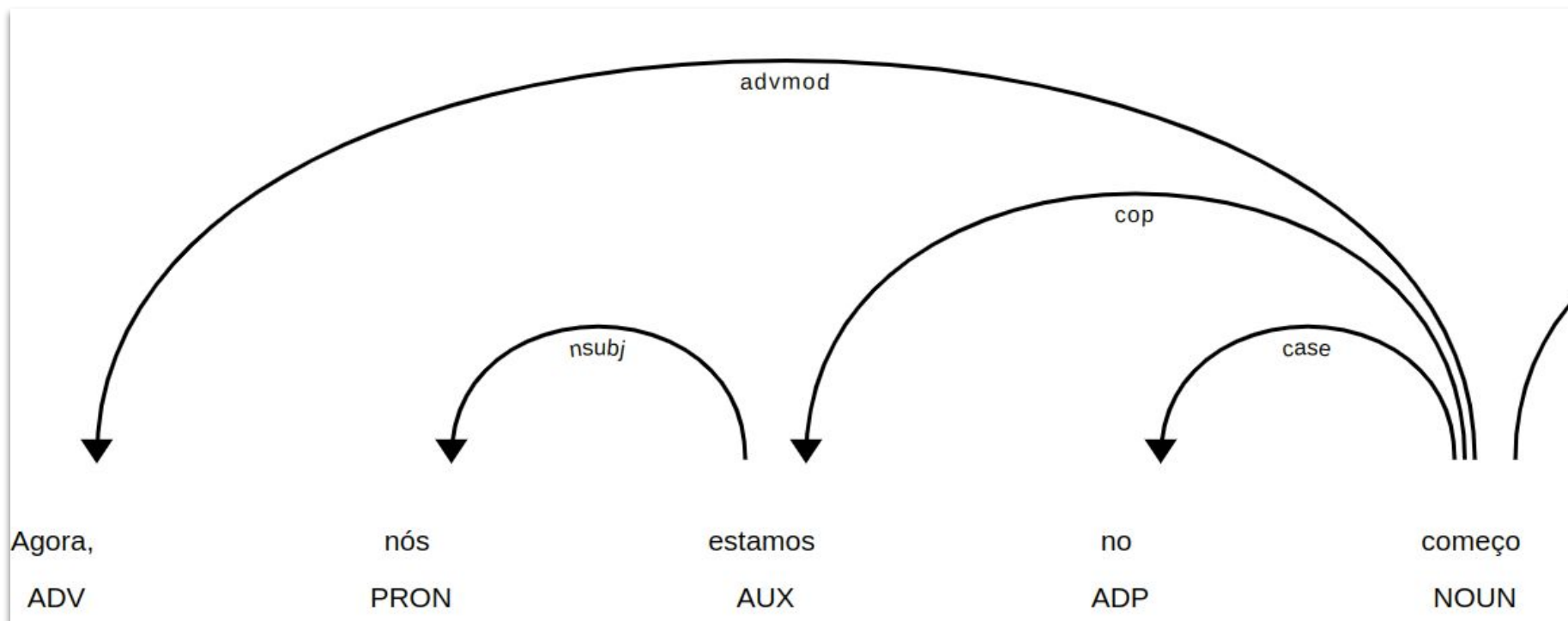


<https://spacy.io/>

```
doc = pln("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```

```
from spacy import displacy
```

```
displacy.render(doc, jupyter=True)
```



Tokenização e Part-Of-Speech tagger

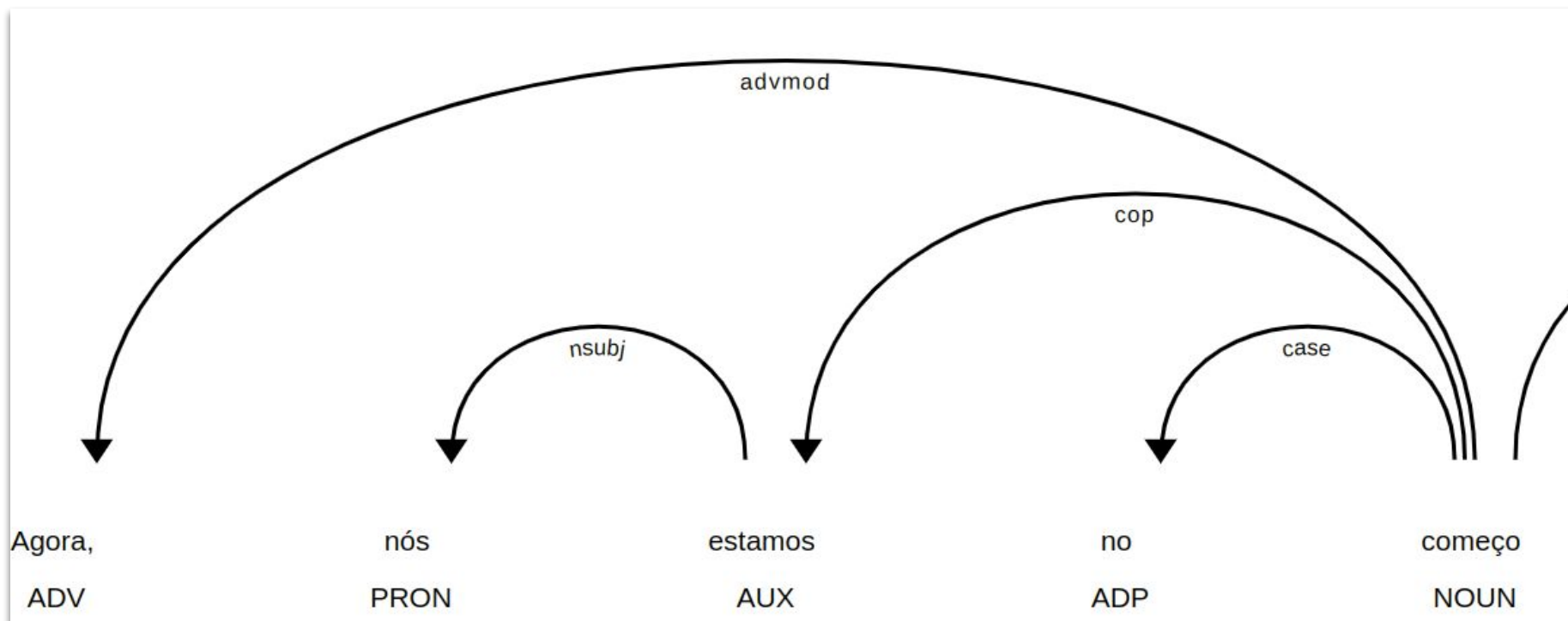


<https://spacy.io/>

```
doc = pln("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```

```
from spacy import displacy
```

```
displacy.render(doc, jupyter=True)
```



Tokenização e Part-Of-Speech tagger

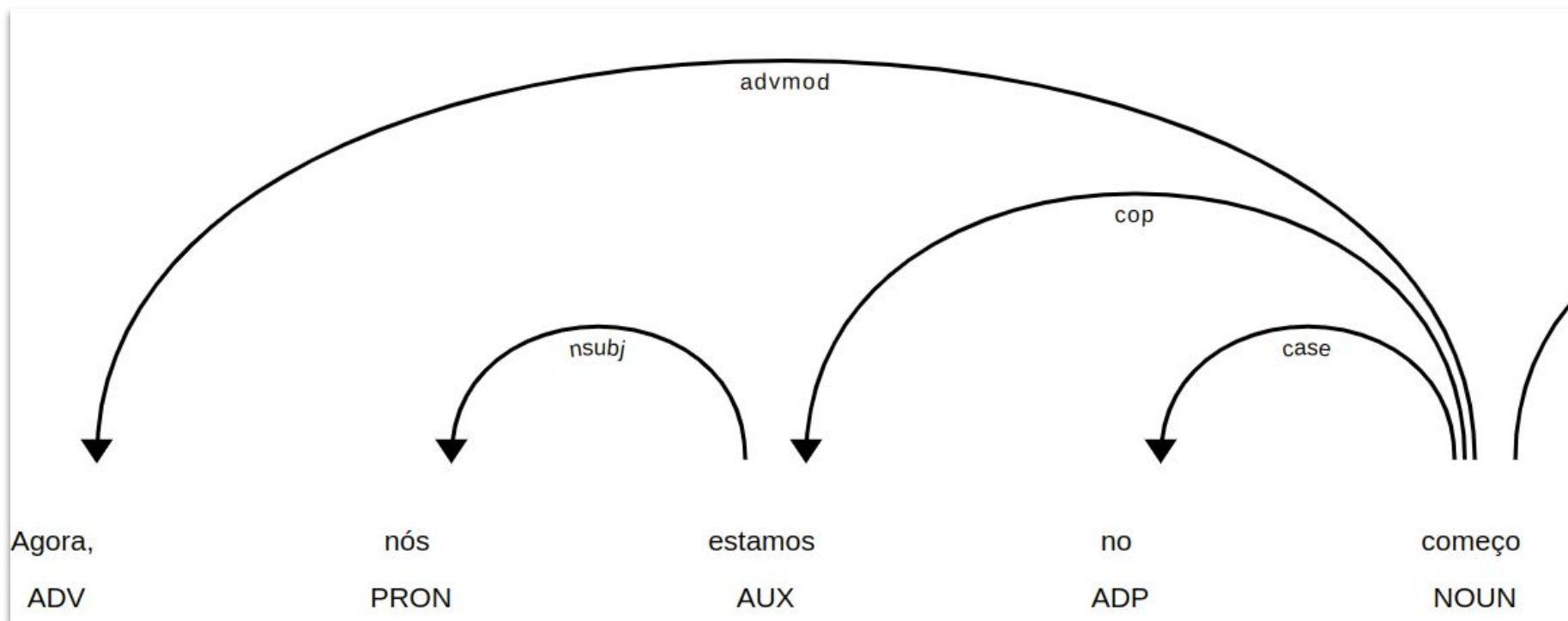


<https://spacy.io/>

```
doc = pln("Agora, nós estamos no começo da disciplina de PLN e Mineração de Texto.")
```

```
from spacy import displacy
```

```
displacy.render(doc, jupyter=True)
```



Entidades Nomeadas



<https://spacy.io/>

Entidades Nomeadas



<https://spacy.io/>

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Núvem.  
No Brasil, essa área de atuação exige conhecimentos de Programação Web  
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )  
  
for ent in doc.ents:  
    print(f'{ent.text:20}\t{ent.label_:20}')
```

Entidades nomeadas são referências a pessoas, lugares, organizações, eventos e outros conceitos específicos que podem ser identificados em um texto.

Entidades Nomeadas



<https://spacy.io/>

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Núvem.  
No Brasil, essa área de atuação exige conhecimentos de Programação Web  
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )  
  
for ent in doc.ents:  
    print(f'{ent.text:20}\t{ent.label_:20}')
```

Entidades Nomeadas

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Nuvem.
No Brasil, essa área de atuação exige conhecimentos de Programação Web
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )

for ent in doc.ents:
    print(f'{ent.text:20}\t{ent.label_:20}')
```

Microsoft	ORG
Computação	LOC
Nuvem	LOC
Brasil	LOC
Programação Web	MISC
Front-End	LOC
Back-End	LOC
Banco de Dados	ORG
JavaScript	MISC

Entidades Nomeadas

```
doc = pln('''A Microsoft tem atuado forteme  
No Brasil, essa área de atuação exige conhe  
(Front-End e Back-End), Banco de Dados, Jav  
  
for ent in doc.ents:  
    print(f'{ent.text:20}\t{ent.label_:20}')
```

Microsoft	ORG
Computação	LOC
Núvem	LOC
Brasil	LOC
Programação Web	MISC
Front-End	LOC
Back-End	LOC
Banco de Dados	ORG
JavaScript	MISC

PERSON - People, including fictional.

NORP - Nationalities or religious or political groups.

FAC - Buildings, airports, highways, bridges, etc.

ORG - Companies, agencies, institutions, etc.

GPE - Countries, cities, states.

LOC - Non-GPE locations, mountain ranges, bodies of water.

PRODUCT - Objects, vehicles, foods, etc. (Not services.)

EVENT - Named hurricanes, battles, wars, sports events, etc.

WORK_OF_ART - Titles of books, songs, etc.

LAW - Named documents made into laws.

LANGUAGE - Any named language.

DATE - Absolute or relative dates or periods.

TIME - Times smaller than a day.

PERCENT - Percentage, including "%".

MONEY - Monetary values, including unit.

QUANTITY - Measurements, as of weight or distance.

Entidades Nomeadas

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Núvem.
No Brasil, essa área de atuação exige conhecimentos de Programação Web
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )
```

```
for ent in doc.ents:
    print(f'{ent.text:20}\t{ent.label_:20}')
```

```
displacy.render(doc, style='ent', jupyter=True)
```

A **Microsoft** **ORG** tem atuado fortemente na área de **Computação** **LOC** em **Núvem** **LOC** .

No **Brasil** **LOC** , essa área de atuação exige conhecimentos de **Programação Web** **MISC**

(**Front-End** **LOC** e **Back-End** **LOC**), **Banco de Dados** **ORG** , **JavaScript** **MISC** , entre outras tecnologias.

Entidades Nomeadas



<https://spacy.io/>

```
doc2 = pln('''Vinícius de Moraes é um cantor/compositor de renome de músicas  
no estilo de bossa nova. Parte de suas músicas referenciavam a cidade do Rio de  
Janeiro. Ele não tinha iPhone e nem gostava de carros no modelo Tesla.''' )  
displacy.render(doc2, style='ent', jupyter=True)
```

Vinícius de Moraes **PER** é um cantor/compositor de renome de músicas
no estilo de bossa nova. Parte de suas músicas referenciavam a cidade do **Rio de Janeiro LOC** . Ele não tinha **iPhone MISC** e nem
gostava de carros no modelo **Tesla PER** .

Lemas e Stop-words

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Nuvem.
No Brasil, essa área de atuação exige conhecimentos de Programação Web
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )
```

```
for token in doc:
    print(f'{token.text:30} -> {token.lemma_}')
```

A tarefa de lematização é uma tarefa de PLN que consiste em reduzir palavras flexionadas a sua forma base ou lema, ou seja, a forma como elas aparecem em um dicionário.

Lemas e Stop-words

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Núvem.
No Brasil, essa área de atuação exige conhecimentos de Programação Web
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )
```

```
for token in doc:
    print(f'{token.text:30} -> {token.lemma_}')
```

A	-> o
Microsoft	-> Microsoft
tem	-> ter
atuado	-> atuar
fortemente	-> fortemente
na	-> em o
área	-> área
de	-> de
Computação	-> Computação
em	-> em
Núvem	-> Núvem
.	-> .
No	-> em o

Lemas e Stop-words

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Núvem.
No Brasil, essa área de atuação exige conhecimentos de Programação Web
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )
```

```
for token in doc:
    print(f'{token.text:30} -> {token.lemma_}')
```

A	-> o
Microsoft	-> Microsoft
tem	-> ter
atuado	-> atuar
fortemente	-> fortemente
na	-> em o
área	-> área
de	-> de
Computação	-> Computação
em	-> em
Núvem	-> Núvem
.	-> .
No	-> em o

Lemas e Stop-words

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Núvem.
No Brasil, essa área de atuação exige conhecimentos de Programação Web
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )
```

```
for token in doc:
    print(f'{token.text:30} -> {token.lemma_}')
```

A	-> o
Microsoft	-> Microsoft
tem	-> ter
atuado	-> atuar
fortemente	-> fortemente
na	-> em o
área	-> área
de	-> de
Computação	-> Computação
em	-> em
Núvem	-> Núvem
	->
No	-> em o

Lemas e Stop-words



<https://spacy.io/>

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Nuvem.  
No Brasil, essa área de atuação exige conhecimentos de Programação Web  
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )
```

```
from spacy.lang.pt import stop_words  
  
print(stop_words.STOP_WORDS)
```

```
{'como', 'estas', 'qual', 'máximo', 'iniciar', 'aí', 'nas', 'exemplo', 'diante',
```

Stop words em português são palavras comuns que são frequentemente removidas de um texto em aplicações de mineração de texto porque elas não contribuem significativamente para o significado do texto.

Lemas e Stop-words

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Núvem.  
No Brasil, essa área de atuação exige conhecimentos de Programação Web  
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )
```

```
for token in doc:  
    is_stop = 'é stop word' if token.is_stop else 'não'  
    print(f'{token.text:20} -> {is_stop}')
```


Lemas e Stop-words



<https://spacy.io/>

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Nuvem.  
No Brasil, essa área de atuação exige conhecimentos de Programação Web  
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )
```

```
for token in doc:  
    is_stop = 'é stop word' if token.is_stop else 'não'  
    print(f'{token.text:20} -> {is_stop}')
```

Lemas e Stop-words

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Nuvem.
No Brasil, essa área de atuação exige conhecimentos de Programação Web
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )
```

```
for token in doc:
    is_stop = 'é stop word' if token.is_stop else 'não'
    print(f'{token.text:20} -> {is_stop}')
```

Lemas e Stop-words

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Nuvem.
No Brasil, essa área de atuação exige conhecimentos de Programação Web
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )
```

```
for token in doc:
    is_stop = 'é stop word' if token.is_stop else 'não'
    print(f'{token.text:20} -> {is_stop}')
```

A -> é stop word

Microsoft -> não

tem -> é stop word

atuado -> não

fortemente -> não

na -> é stop word

área -> é stop word

de -> é stop word

Computacao -> não

em -> é stop word

Nuvem -> não

. -> não

No -> é stop word

Lemas e Stop-words

```
doc = pln('''A Microsoft tem atuado fortemente na área de Computação em Nuvem.
No Brasil, essa área de atuação exige conhecimentos de Programação Web
(Front-End e Back-End), Banco de Dados, JavaScript, entre outras tecnologias.''' )
```

```
for token in doc:
    is_stop = 'é stop word' if token.is_stop else 'não'
    print(f'{token.text:20} -> {is_stop}')
```

A	-> é stop word
Microsoft	-> não
tem	-> é stop word
atuado	-> não
fortemente	-> não
na	-> é stop word
área	-> é stop word
de	-> é stop word
Computação	-> não
em	-> é stop word
Nuvem	-> não
.	-> não
No	-> é stop word
Brasil	-> não

Aplicações de Tarefas de PLN

- Parts-Of-Speech (POS) tagger
 - Corretores ortográficos
 - Identificação de relação entre palavras
- Entidades Nomeadas
 - Identificação de Entidades em textos
 - Extração de preços/cotações em documentos
- Lematização / Stemming e Stop words
 - Redução de dimensionalidade