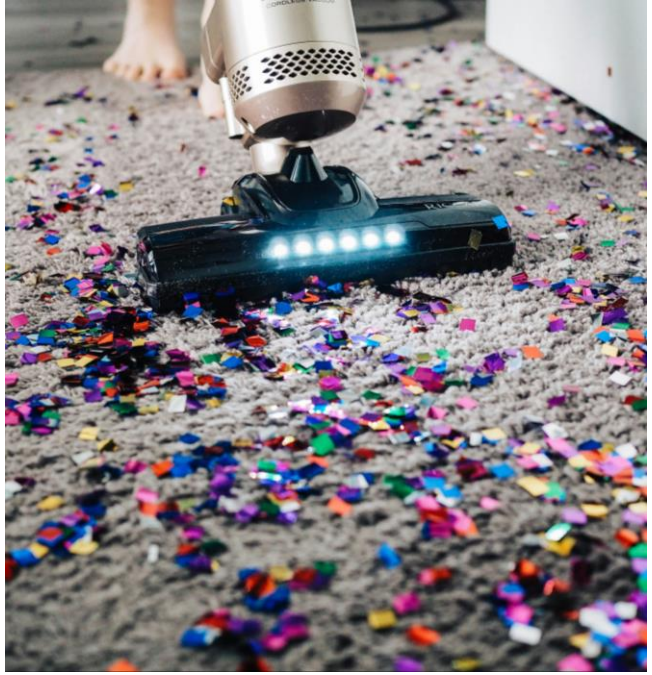
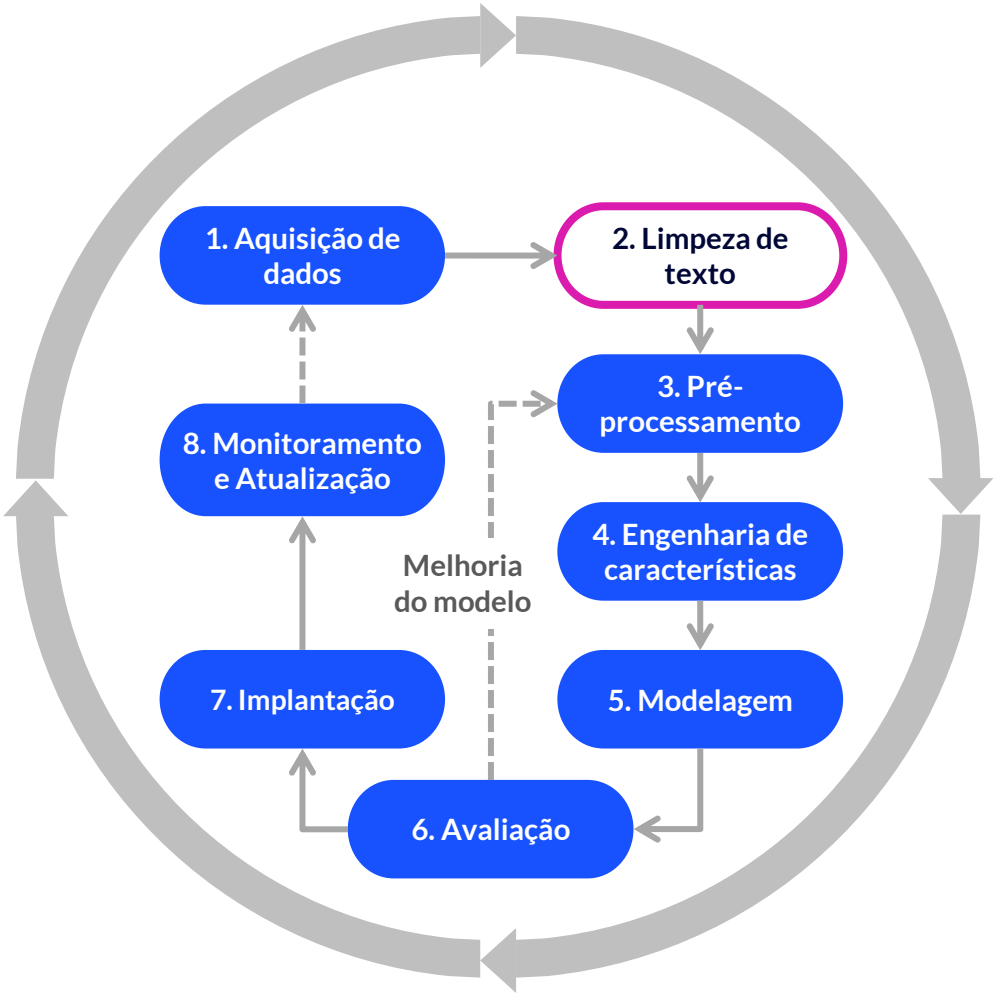
PUC Minas Virtual

PROCESSAMENTO DE LINGUAGEM NATURAL

LIMPEZA DE DADOS

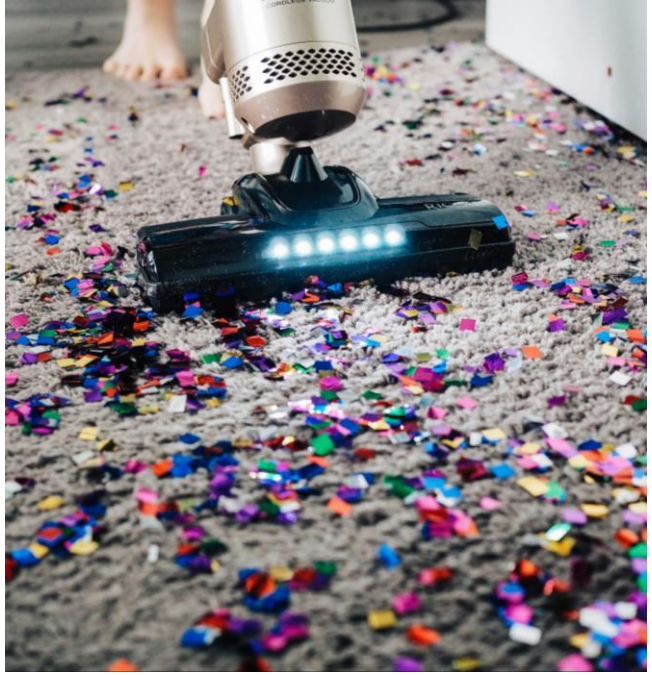


Limpeza de Dados



Limpeza de Dados

Extração do texto bruto dos dados de entrada, removendo informações não textuais



Limpeza de Dados

01 Remoção de dados não textuais

02 Correção ortográfica

03 Normalização Unicode

Remoção de dados não textuais

Tipos de Documentos

Portable Document Format

HyperText Markup Language

Texto Embutidos em Imagens

BeautifulSoup

BeautifulSoup

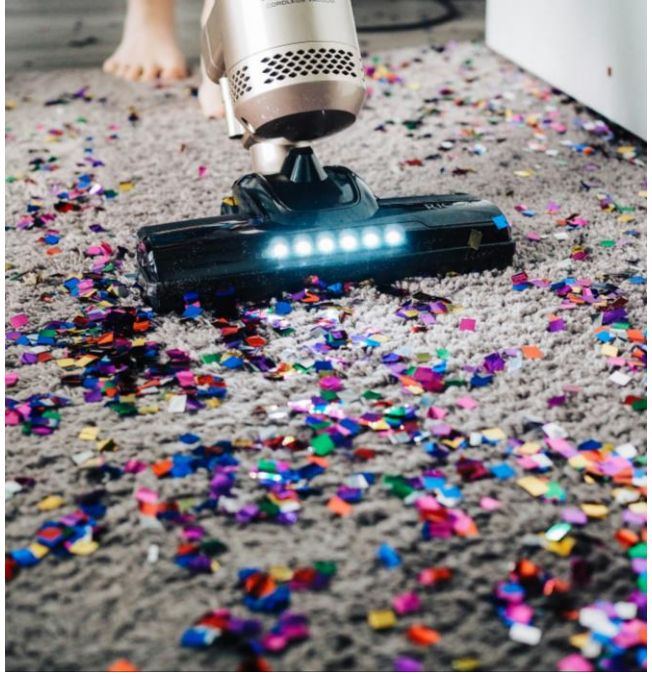
```
pip install beautifulsoup4
```

01
```
from bs4 import BeautifulSoup
```

02
```
page = requests.get(url, params)
```

03
```
soup = BeautifulSoup(p.content,"html.parser")
```

05
```
print(soup.title.get_text().strip())
```



Limpeza de Dados

01 Remoção de dados não textuais

02 Correção ortográfica

03 Normalização Unicode

Spellchecker

SpellChecker

pip install pyspellchecker

```
01  from spellchecker import SpellChecker

02  sp = SpellChecker(language='pt')

03  erros =  sp.unknow(['bola', 'futbol'])

04  for palavra in erros:
        print(sp.correction(palavra))
```

SpellChecker

pip install pyspellchecker

```
01  from spellchecker import SpellChecker
02  sp = SpellChecker(language='pt')
03  erros =  sp.unknow(['bola', 'futbol'])
04  for palavra in erros:
        print(sp.correction(palavra))
```

SpellChecker

pip install pyspellchecker

```
01 from spellchecker import SpellChecker
02 sp = SpellChecker(language='pt')
03 erros =  sp.unknow(['bola', 'futbol'])
04 for palavra in erros:
       print(sp.correction(palavra))
```

SpellChecker

pip install pyspellchecker

```
01 from spellchecker import SpellChecker
02 sp = SpellChecker(language='pt')
03 erros =  sp.unknow(['bola', 'futbol'])
04 for palavra in erros:
       print(sp.correction(palavra))
```

SpellChecker

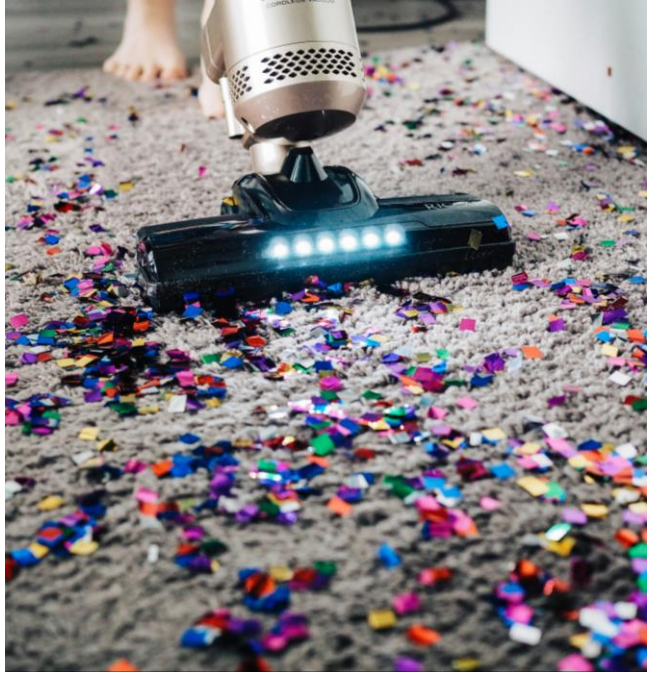
pip install pyspellchecker

01 from spellchecker import SpellChecker

02 sp = SpellChecker(language='pt')

03 erros = sp.unknow(['bola', 'futbol'])

04
```
for palavra in erros:
    print(sp.correction(palavra))
```



Limpeza de Dados

01 Remoção de dados não textuais

02 Correção ortográfica

03 Normalização Unicode

encode()

encode()

01 texto = 'Eu amo pizza 🍕 !

02 print(texto.encode(“utf-8”))

Eu amo pizza \xf0\x9f\x8d\x95!

encode()

01 texto = 'Eu amo pizza 🍕 !

02 print(texto.encode(“utf-8”))

Eu amo pizza \xf0\x9f\x8d\x95!

PUC Minas
Virtual