

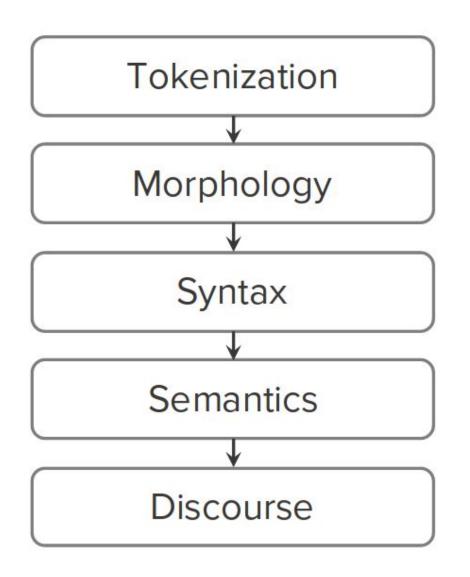
# Análise de Mídias Sociais e Mineração de Texto

# Aquisição e Modelagem de Dados

Laura de Oliveira F. Moraes

## **NLP Pipeline**





Break text into sentences and words, lemmatize

Part of speech (POS) tagging, stemming, NER

Constituency/dependency parsing

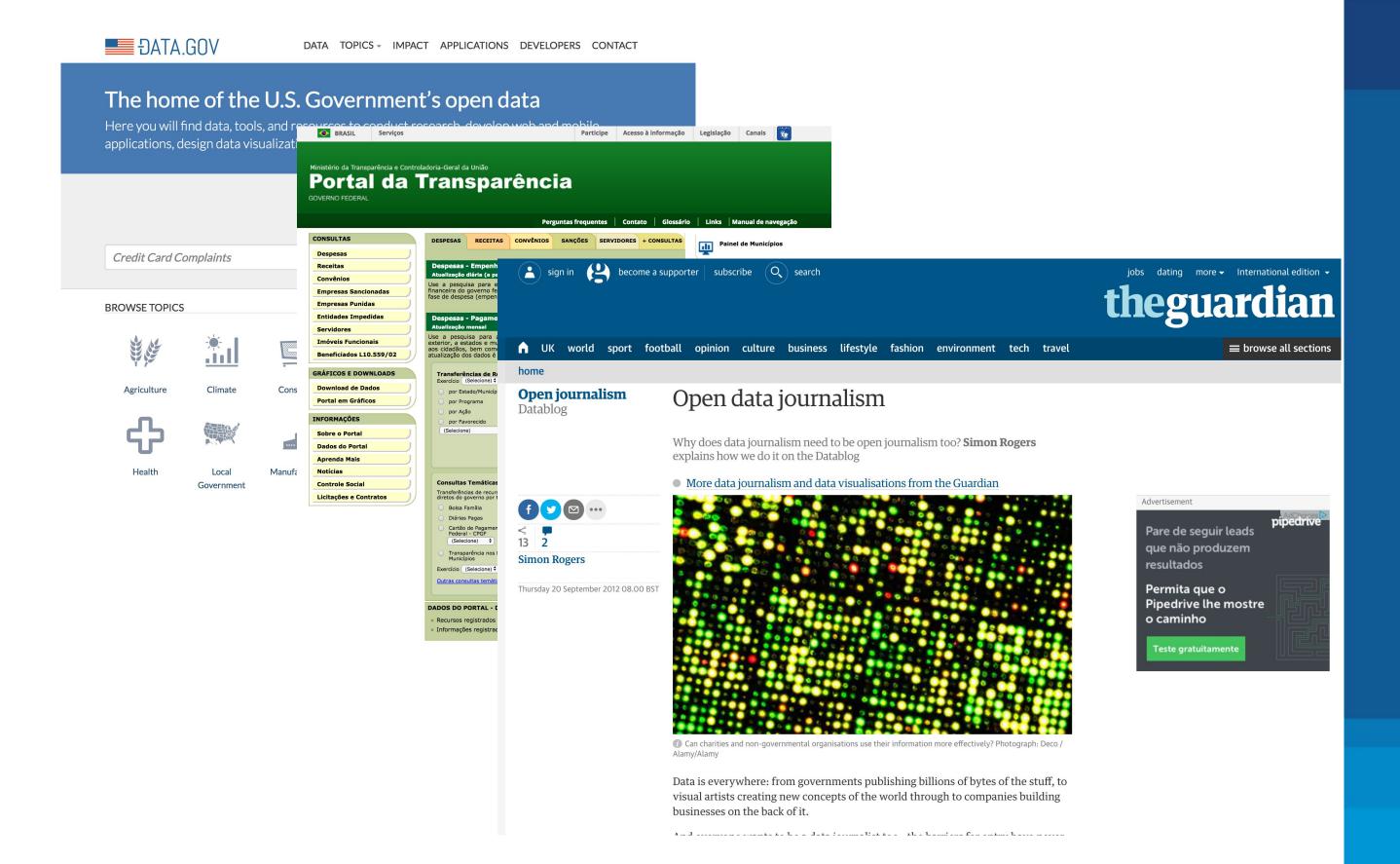
Coreference resolution, wordsense disambiguation

Task-dependent (sentiment, ...)

Fonte: https://www.slideshare.net/YuriyGuts/natural-language-processing-nlp

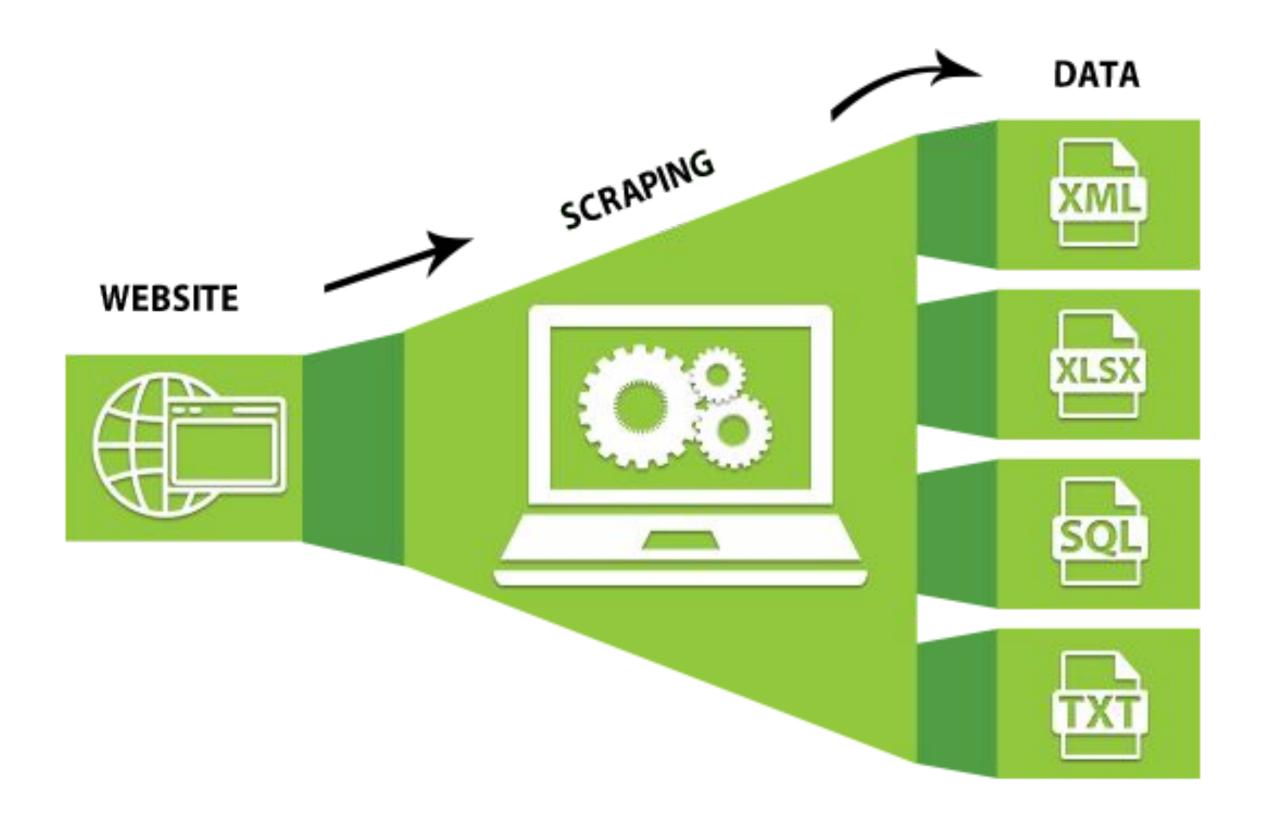
#### **Open Data**





## **Web Scraping**

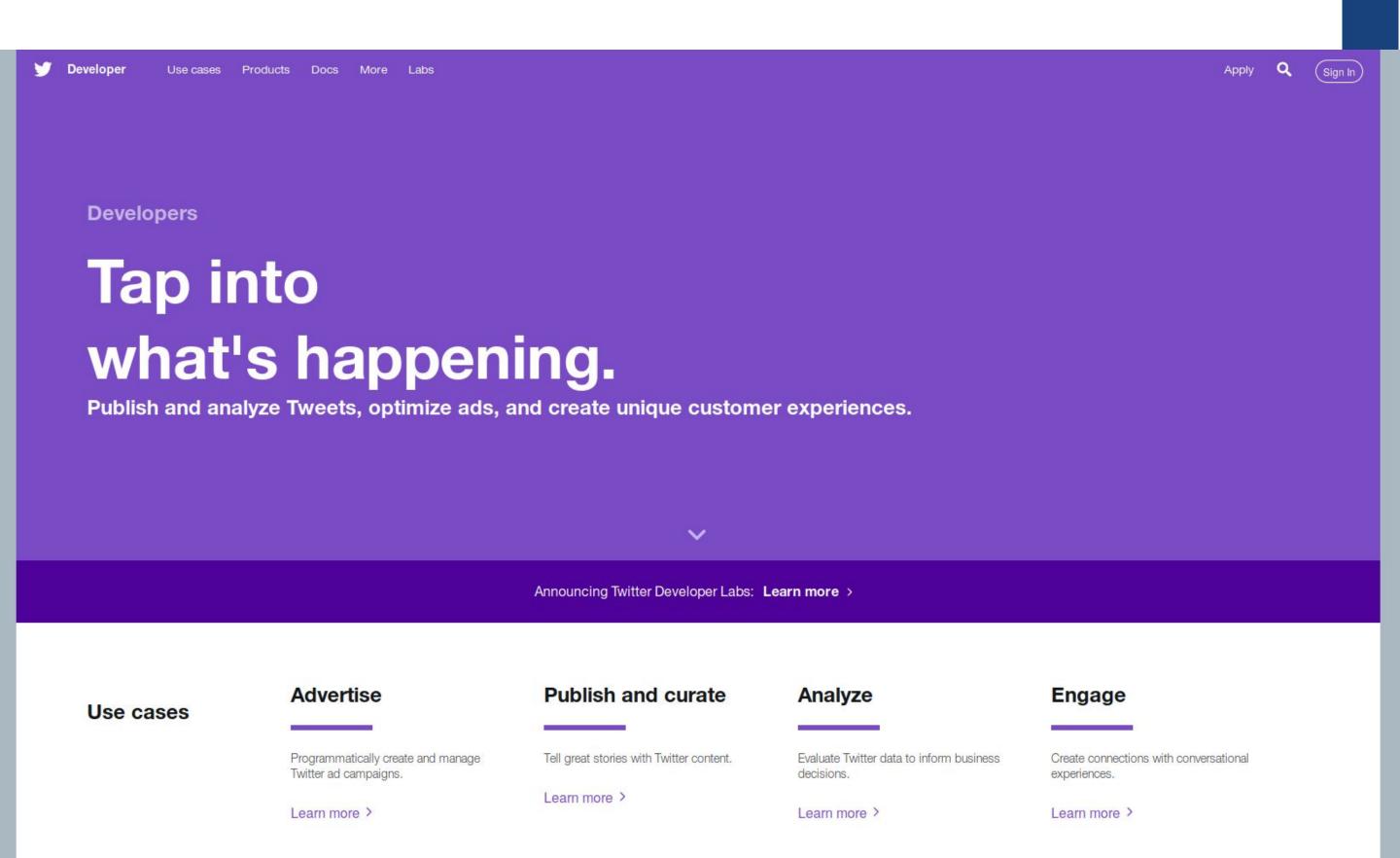




Fonte: http://prowebscraping.com/web-scraping/



## https://developer.twitter.com/



## Bag of words e N-grams



## the dog is on the table

```
N = 1 : This is a sentence unigrams: this, is, a, sentence

N = 2 : This is a sentence bigrams: this is, is a, is a, a sentence

N = 3 : This is a sentence trigrams: this is a, is a sentence
```

Fonte: https://www.slideshare.net/YuriyGuts/natural-language-processing-nlp

```
def lessthan5(num_list):
def light(switchA, switchB):
    if switchA == 1 and switchB == 1:
                                            new list = []
                                            for item in num list:
        return True
                                                if item < 5:
   else:
                                                    new list.append(item)
        return False
                                            return new list
```

```
def light switchA switchB is block
    is_indent if switchA is_op_logic is_number and switchB is_op_logic
is_number is_block
      is_indent return True
      is_dedent else is_block
      is indent return False
      is dedent
is_dedent
```

```
def lessthan5 num_list is_block
      is indent new list is attribution is list
      for item in num list is block
      is_indent if item is_op_logic is_number is_block
            is indent new list is class append item
      is dedent
is_dedent return new list is dedent
```

## N-grams → Bag of words



term\document	doc	doc
	#1	#2
is_number	2	1
is_op_logic	2	1
if	1	1
append	0	1
is_attribution	0	1
for	0	1
is_list	0	1
is_op_logic is_number	2	1
is_attribution is_list	0	1
is_attribution is_list for	0	1