

Professor(a):

José Wellington Franco da Silva

[wellington@crateus.ufc.br](mailto:wellington@crateus.ufc.br)



**Análise de Dados Textuais**

ANO:2019

# O que é Processamento de Linguagem Natural?

- É um campo de pesquisa da ciência da computação, inteligência artificial e linguística computacional;
- Que se preocupa com as interações entre computadores e linguagens humanas (naturais) e, em particular;
- Usa a programação de computadores para processar frutuosamente grandes corpora de linguagem natural;
- Especificamente, a tarefa de extrair informações significativas da entrada em linguagem natural ou de produzir saída em linguagem natural.

# O que é a Linguística?

- **É o estudo científico da linguagem e sua estrutura;**
- **e envolve uma análise de:**
  - **Forma (fonética, sintaxe e gramática)**
  - **Significado**
  - **Contexto**



- Segmentação de Texto
  - Dividir um texto em unidades significativas, como frases, seções ou parágrafos;



- **Segmentação de Sentenças**
  - **Dividindo um texto em frases;**

## Original Text

Neil Alden Armstrong was an American astronaut and the first person to walk on the Moon. He was also an aerospace engineer, naval aviator, test pilot, and university professor. Before becoming an astronaut, Armstrong was an officer in the U.S. Navy and served in the Korean War. After the war, he earned his bachelor's degree at Purdue University and served as a test pilot at the National Advisory Committee for Aeronautics (NACA) High-Speed Flight Station, where he logged over 900 flights. He later completed graduate studies at the University of Southern California.

## Analysis Result

Neil Alden Armstrong was an American astronaut and the first person to walk on the Moon.

He was also an aerospace engineer, naval aviator, test pilot, and university professor.

Before becoming an astronaut, Armstrong was an officer in the U.S. Navy and served in the Korean War.

After the war, he earned his bachelor's degree at Purdue University and served as a test pilot at the National Advisory Committee for Aeronautics (NACA) High-Speed Flight Station, where he logged over 900 flights.

He later completed graduate studies at the University of Southern California.

- **Tokenização**

- **Dividir um fluxo de texto em palavras, frases, símbolos ou outros elementos significativos chamados tokens;**

## Word Tokenization with Python NLTK

This is a demonstration of the various tokenizers provided by [NLTK 2.0.4](#).

### Tokenize Text

Enter text

Neil Alden Armstrong was an American astronaut and the first person to walk on the Moon.

Enter up to 50000 characters

Tokenize

### TreebankWordTokenizer

1.

Neil Alden Armstrong was an American astronaut and the first  
person to walk on the Moon .

- **Pos Tagger (*Part-of-Speech Tagging*)**
  - **Marcar uma palavra em um texto como correspondendo a uma parte específica do discurso, com base em sua definição e em seu contexto;**

## Text Analysis Result -- NLTK POS Tagging

### Original Text

Neil Alden Armstrong (August 5, 1930 – August 25, 2012) was an American astronaut and the first person to walk on the Moon

<http://textanalysisonline.com/nltk-pos-tagging>

### Analysis Result

Neil|NNP Alden|NNP Armstrong|NNP (( August|NNP 5|CD ,|, 1930|CD –|NN August|NNP 25|CD ,|, 2012|CD )) was|VBD an|DT American|JJ astronaut|NN and|CC the|DT first|JJ person|NN to|TO walk|VB on|IN the|DT Moon|NNP

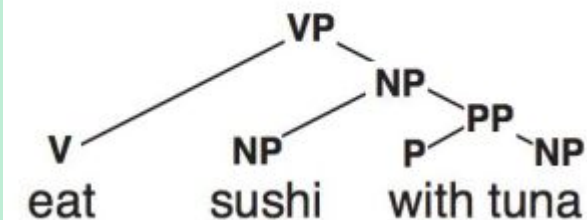


- **Parsing**

- **Analizando uma sequência de símbolos em conformidade com as regras da gramática formal e construindo uma árvore sintática de uma frase;**

**Parse**

```
(ROOT
 (S
  (NP (NNP Neil) (NNP Alden) (NNP Armstrong))
  (VP (VBD was)
    (NP
     (NP (DT an) (JJ American) (NN astronaut))
     (CC and)
     (NP (DT the) (JJ first) (NN person)
       (S
        (VP (TO to)
          (VP (VB walk)
            (PP (IN on)
              (NP (DT the) (NNP Moon))))))))
      (. .)))
```





- **Reconhecedor de Entidades Nomeadas**
  - **Localização e classificação de elementos atômicos em categorias predefinidas, como nomes, pessoas, organizações, locais, expressões de tempo, quantidades, valores monetários etc.**

**Stanford Named Entity Tagger**

Classifier:

Output Format:

Preserve Spacing:

Please enter your text here:

Neil Alden Armstrong (August 5, 1930 – August 25, 2012) was an American astronaut and the first person to walk on the Moon. Before becoming an astronaut, Armstrong was an officer in the U.S. Navy and served in the Korean War.

Neil Alden Armstrong (August 5, 1930 – August 25, 2012) was an American astronaut and the first person to walk on the Moon. Before becoming an astronaut, Armstrong was an officer in the U.S. Navy and served in the Korean War.

Potential tags:

- LOCATION
- ORGANIZATION
- DATE
- MONEY
- PERSON
- PERCENT
- TIME

<http://nlp.stanford.edu:8080/ner/process>

<http://textanalysisonline.com/spacy-named-entity-recognition-ner>

- **Desambiguação do sentido da palavra**
  - **Descobrir o significado exato de uma entidade, ou seja, vincular um segmento de texto à sua entidade correspondente em uma base de conhecimento;**



Neil Armstrong (1930 - 2012)

<b>Neil Armstrong</b>	
Neil Alden Armstrong (August 5, 1930 – August 25, 2012) was an American astronaut and the first person to walk on the Moon. He was born, the first person to walk on the Moon. He was also a test pilot, and university professor. Armstrong was the first person to walk on the Moon. He was born, the first person to walk on the Moon. He was also a test pilot, and university professor. Armstrong was the first person to walk on the Moon.	
<b>birth year</b>	1930
<b>death year</b>	2012
<b>death place</b>	Cincinnati
<b>type</b>	NASA
<b>occupation</b>	Naval aviation
<b>death place</b>	Ohio

Armstrong's Youth and

Neil Armstrong was born in an auditor for the Ohio. Armstrong experienced his Trimotor. Armstrong attended Wapakoneta airfield. In 19

to Stephen Koenig Armstrong, Engel. Already at the age of five, his father took a ride in a Ford flying lessons at the grassy flying aeronautical engineering at

Purdue University, funded the Holloway Plan, where he was committed to two years of study, followed by three years of service in the U.S. Navy, and then completion of the final two years of the degree. Armstrong's call-up from the Navy arrived in 1949, requiring him to report to Naval

- **Semantic Role Labeling**

- **Extração de triplas sujeito-predicado-objeto de uma sentença;**

	Neil	Alden	Armstrong	was	an	American	astronaut	and	the	first	person	to	walk	on	the	Moon	.
SRL	extent [AM-EXT]			V: be.01	topic [A1]												
SRL													walker [A0]	V: walk.01		direction [AM-DIR]	
Nom																	
Prep													Governor	Location (on)		Object	

## Key

### Verb

**V** verb

### Arguments

**A0** subject

**A1** object

**A2** indirect object

### Other

**C-arg** continuity of an argument/adjunct of type arg

**R-arg** reference to an actual argument/adjunct of type arg

### Adjunct

**AM-ADV** adverbial modification

**AM-DIR** direction

**AM-DIS** discourse marker

**AM-EXT** extent

**AM-LOC** location

**AM-MNR** manner

**AM-MOD** general modification

**AM-NEG** negation

**AM-PNC** proper noun component

**AM-PRD** secondary predicate

**AM-PRP** purpose

- **Paráfrase**

- **Uma paráfrase é uma reafirmação do significado de um texto ou passagem usando outras palavras;**



- **Palavra**

- **É a menor unidade de uma língua**
- **Não depende de outras palavras**
- **Pode ser separado de outras unidades**
- **Pode mudar de posição**

# Conceitos Linguísticos Básicos

- **Vocabulário**
  - **É um conjunto de palavras;**



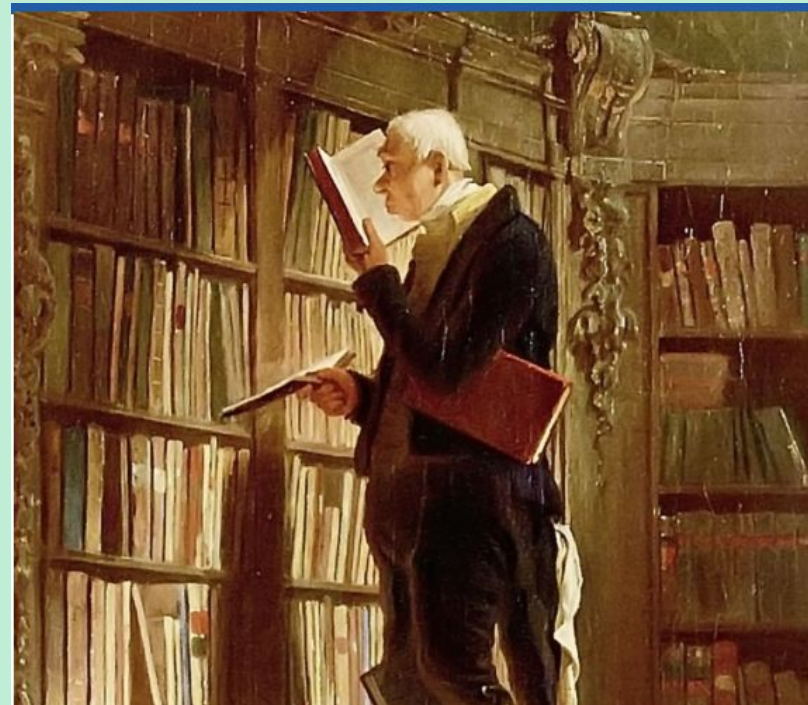
- **Texto**

- **Um texto é composto de uma sequência de palavras de um vocabulário.**

Shall I compare thee to a summer's day?  
Thou art more lovely and more temperate:  
Rough winds do shake the darling buds of May,  
And summer's lease hath all too short a date:  
Sometime too hot the eye of heaven shines,  
And often is his gold complexion dimm'd,  
And every fair from fair sometime declines,  
By chance, or nature's changing course untrimm'd:  
But thy eternal summer shall not fade,  
Nor lose possession of that fair thou ow'st,  
Nor shall death brag thou wander'st in his shade,  
When in eternal lines to time thou grow'st,  
    So long as men can breathe, or eyes can see,  
    So long lives this, and this gives life to thee.



- **Linguagem**
  - **É construída de conjunto de textos possíveis.**





- **Stemming**

- É o processo de reduzir palavras flexionadas (ou às vezes derivadas) ao seu tronco (stem), base ou raiz, geralmente uma forma da palavra escrita;

*Example: **cats** → **cat***

- **Lematização**

- **A lematização reduz a palavra ao seu lema, que é a forma no masculino e singular. No caso de verbos, o lema é o infinitivo.**

*Example: better → good*

# Dúvidas são bem vindas!!

- [wellington@crateus.ufc.br](mailto:wellington@crateus.ufc.br)

