Professor(a):
José Wellington Franco da Silva
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**

Desafios em PLN

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

Desafios em PLN

- **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.

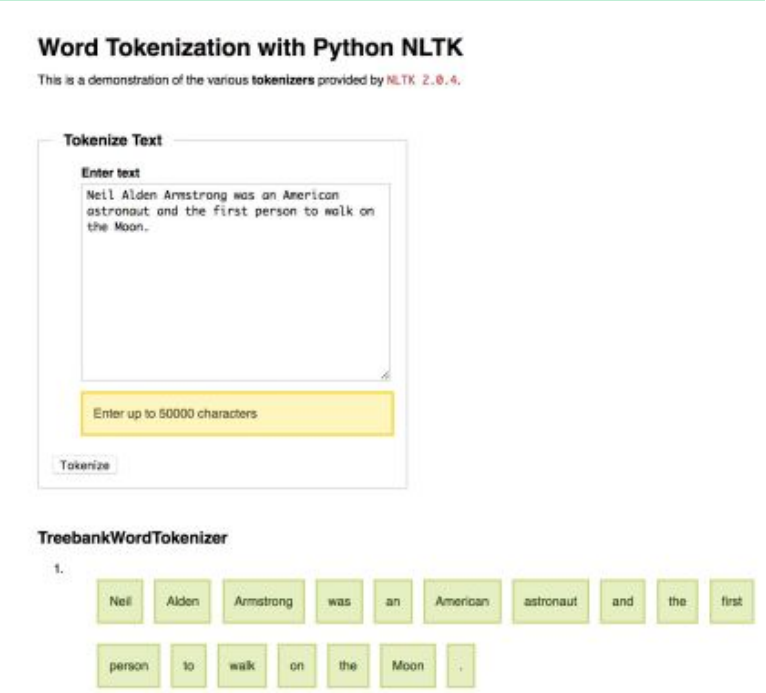
Before becoming an astronaut, Armstrong was an officer in the U.S. Navy and served in the Korean War.

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.

Desafios em PLN

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



Desafios em PLN

- **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

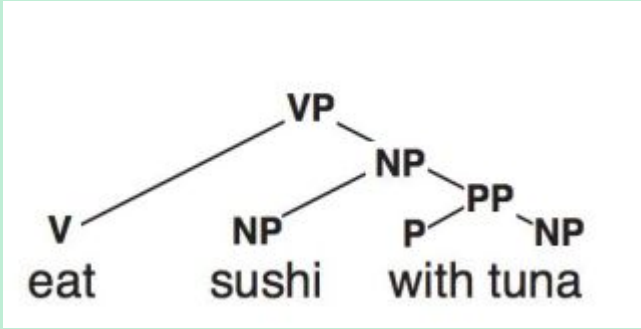
Neil|NNP Alden|NNP Armstrong|NNP (|(August|NNP 5|CD ,|, 1930|CD –|NN August|NNP 25|CD ,|, 2012|CD)|) was|VBD an|DT American|JJ astronaut|NN and|CC the|DT first|JJ person|NN to|TO walk|VB on|IN the|DT Moon|NNP

Desafios em PLN

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

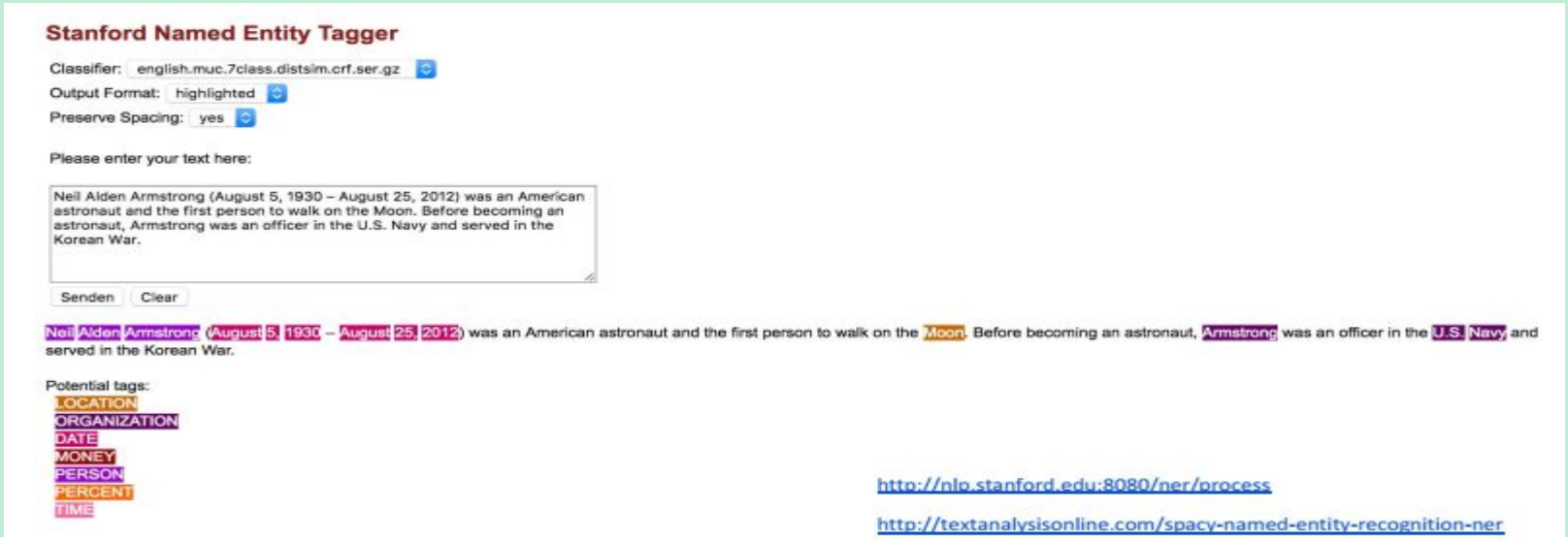
Parse

```
(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)))))))))
    (. .)))
```



Desafios em PLN

- **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.**

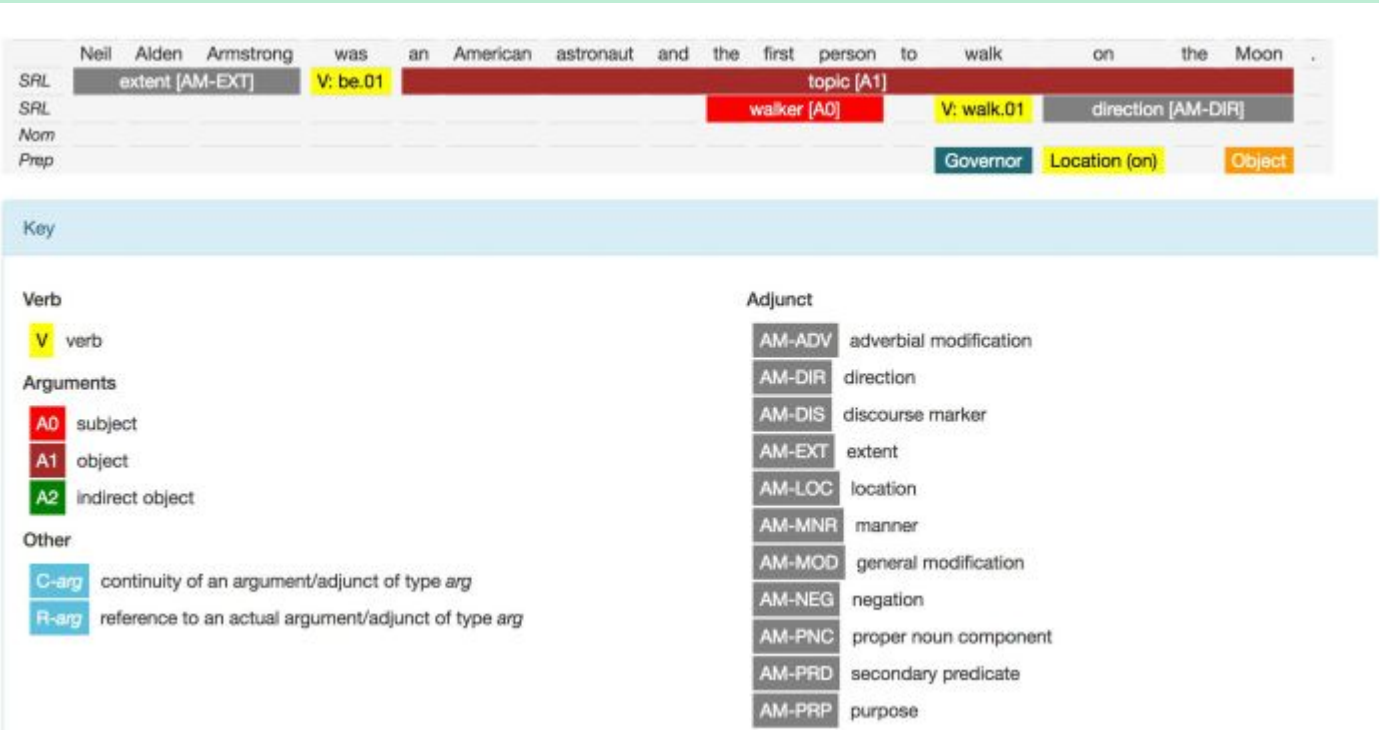


Desafios em PLN

- **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;**

Desafios em PLN

- **Semantic Role Labeling**
  - **Extração de triplas sujeito-predicado-objeto de uma sentença;**



Desafios em PLN

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

Conceitos Linguísticos Básicos

- **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;**

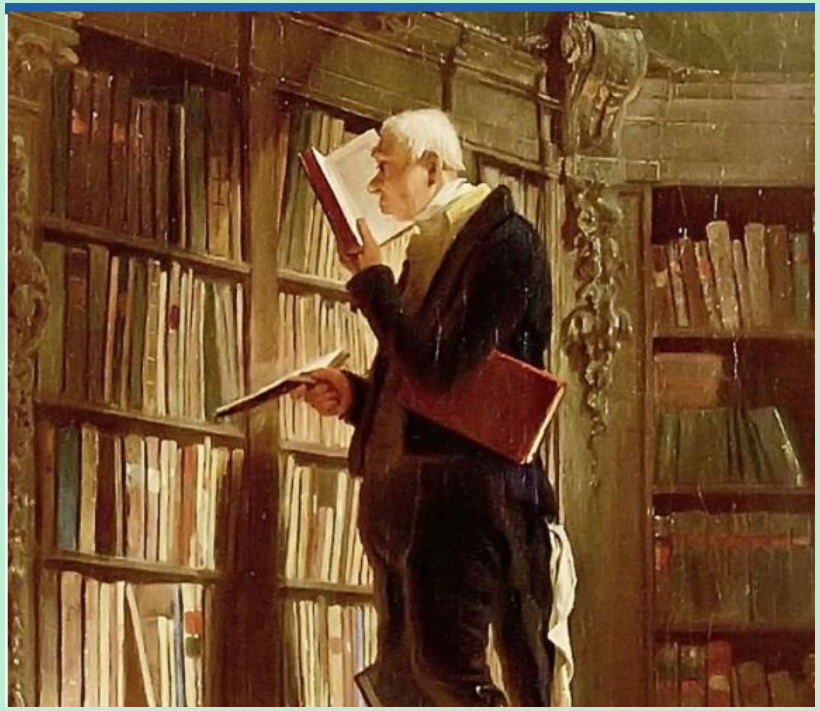
Conceitos Linguísticos Básicos

- **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.

Conceitos Linguísticos Básicos

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



Conceitos Linguísticos Básicos

- **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

Conceitos Linguísticos Básicos

- **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