

Professor(a):

José Wellington Franco da Silva

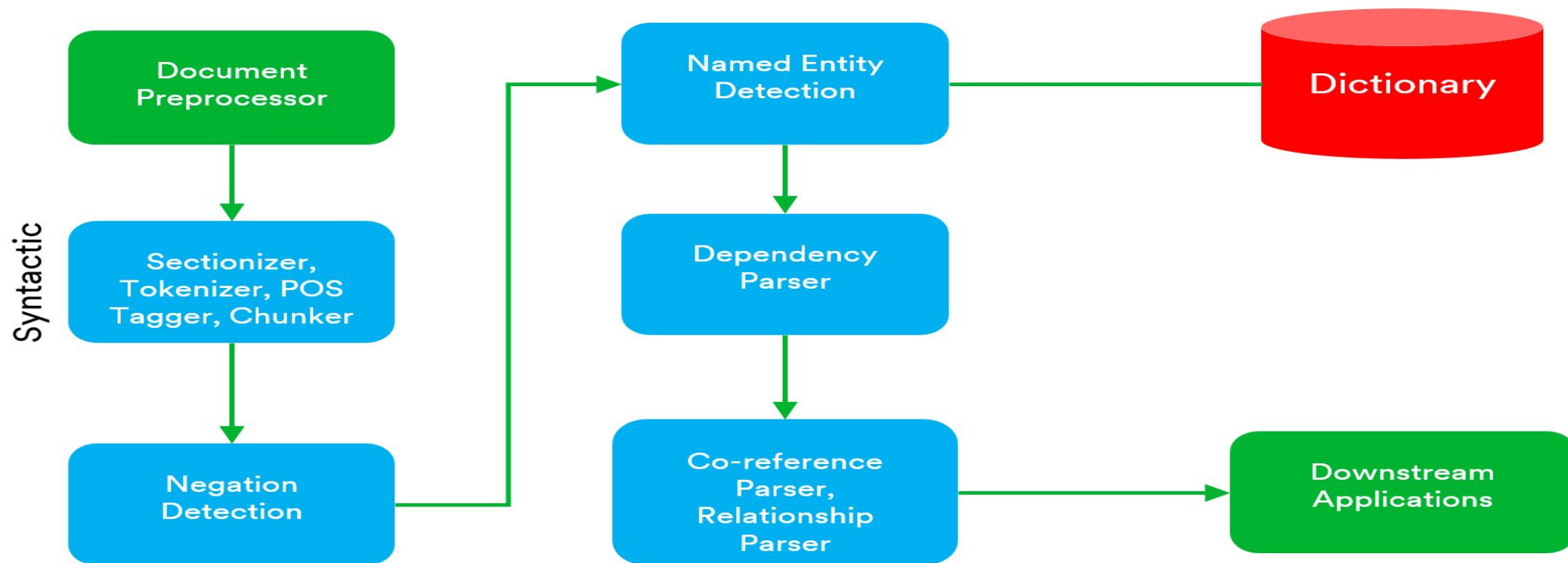
[wellington@crateus.ufc.br](mailto:wellington@crateus.ufc.br)



**Análise de Dados Textuais**

ANO:2019

## Typical Components of an NLP Application



# Term Frequency, Inverse Document Frequency

- **tf-idf** (abreviação do inglês *term frequency-inverse document frequency*, que significa frequência do termo–inverso da frequência nos documentos), é uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos ou em um corpus linguístico;
- **Tf-idf** de uma palavra aumenta proporcionalmente à medida que aumenta o número de ocorrências dela em um documento, no entanto, esse valor é equilibrado pela frequência da palavra no corpus



# Term Frequency, Inverse Document Frequency

## TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

Term frequency

Number of times term  $t$  appears in a doc,  $d$

Inverse document frequency

$$\log \frac{1 + n}{1 + \text{df}(d, t)} + 1$$

# of documents  $n$

Document frequency of the term  $t$   $\text{df}(d, t)$

# Dúvidas são bem vindas!!

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



