

Professor(a):

José Wellington Franco da Silva

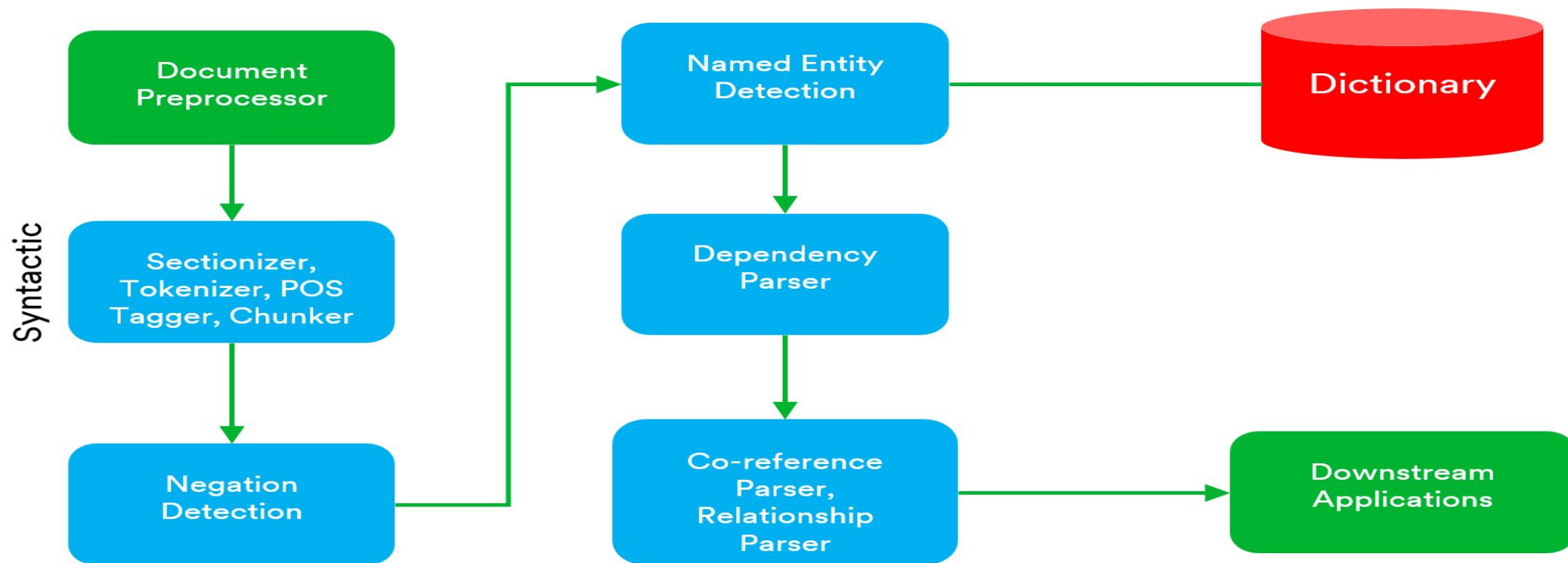
wellington@crateus.ufc.br



Análise de Dados Textuais

ANO:2019

Typical Components of an NLP Application



Nem todas os documentos são iguais...

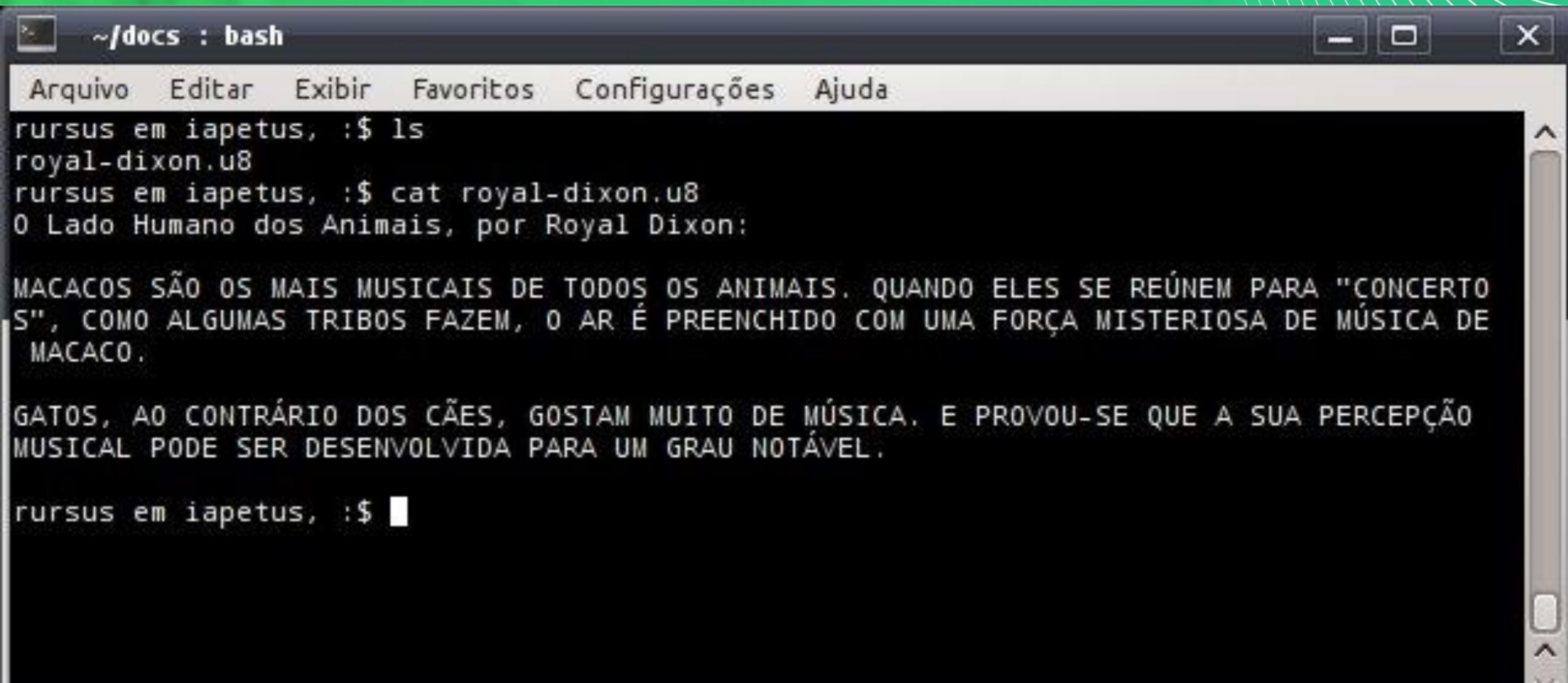
- **Existem uma diversidade de formatos e documentos textuais;**
- **Precisamos saber tratar cada documento de forma diferente para poder extrair o conhecimento necessário;**

Principais formatos

- **Texto Plano**
- **HTML**
- **CSV**
- **PDF**

- Em computação, texto simples ou texto puro é o conteúdo de um arquivo sequencial ordinário legível como material textual sem muito processamento.
- Texto puro é diferente do texto formatado, onde informações de estilo são incluídas e "arquivos binários" nos quais algumas porções devem ser interpretadas como objetos binários (inteiros codificados, números reais, imagens, etc.).

Texto Plano



```
~/docs : bash
Arquivo  Editar  Exibir  Favoritos  Configurações  Ajuda
rursus em iapetus, :$ ls
royal-dixon.u8
rursus em iapetus, :$ cat royal-dixon.u8
0 Lado Humano dos Animais, por Royal Dixon:

MACACOS SÃO OS MAIS MUSICAIS DE TODOS OS ANIMAIS. QUANDO ELES SE REÚNEM PARA "CONCERTO S", COMO ALGUMAS TRIBOS FAZEM, O AR É PREENCHIDO COM UMA FORÇA MISTERIOSA DE MÚSICA DE MACACO.

GATOS, AO CONTRÁRIO DOS CÃES, GOSTAM MUITO DE MÚSICA. E PROVOU-SE QUE A SUA PERCEPÇÃO MUSICAL PODE SER DESENVOLVIDA PARA UM GRAU NOTÁVEL.

rursus em iapetus, :$
```

- HTML (abreviação para a expressão inglesa *HyperText Markup Language*, que significa Linguagem de Marcação de Hipertexto) é uma linguagem de marcação utilizada na construção de páginas na Web.
- Documentos HTML podem ser interpretados por navegadores.
- Todo documento HTML possui marcadores (do inglês: tags), palavras entre parênteses angulares (chevron) (< e >); esses marcadores são os comandos de formatação da linguagem.

```
<!DOCTYPE html>
<html lang=en>
<meta charset=UTF-8>
<title>Introduction to
  The Mating Rituals of
  Bees</title>
<h1>Introduction</h1>
<p>This companion guide to
  the highly successful
  <cite>Introduction to
  Medieval Beekeeping</cite>...
```


- **Comma-separated values**
- **São arquivos de texto de formato regulamentado pelo RFC 4180, que faz uma ordenação de bytes ou um formato de terminador de linha, separando valores com vírgulas;**
- **Ele comumente é usado em softwares offices, tais como o Microsoft Excel e o LibreOffice Calc.**

File Form Template Config AutoSend Help

CSV spreadsheet file: ARES_Members.c2s

Import CSV

Export CSV

View CSV

Edit CSV

CALL, FIRSTNAME, LASTNAME, Skywn, IS-100, IS-700, PUBLIC
AC5NT, LARRY, SWANSON, , FALSE, FALSE, TRUE
AC5T, LYNN, JACKSON, , FALSE, FALSE, TRUE
AE5BK, BILL, KRUEGER, , FALSE, FALSE, TRUE
AE5CF, BRIAN, WEIDENMAIER, , FALSE, FALSE, TRUE
AE5FF, DALE, LEWIS, 3/26/2011, TRUE, TRUE, TRUE
AE5IT, WALTER, LEMONS, , TRUE, TRUE, TRUE
AE5IV, MIKE, CHITTENDEN, , FALSE, FALSE, TRUE
AE5NS, BRIAN, BAUGH, , FALSE, FALSE, TRUE
K1WL, DONALD, PATTERSON, 2/26/2011, TRUE, TRUE, TRUE
K3DHB, DON, BARBER, 5/19/2011, TRUE, TRUE, TRUE
K5AGO, JACK, O'TOOLE, , TRUE, FALSE, TRUE
K5AJP, ANDY, PONDER, 2/26/2011, TRUE, TRUE, TRUE
K5BP, BERNIE, PARKER, , TRUE, TRUE, TRUE
K5BYX, CRAIG, WAGGONER, , FALSE, FALSE, TRUE
K5EMI, WILLIAM, STEWART, , FALSE, FALSE, FALSE
K5HUD, SETH, HUDSON, , TRUE, TRUE, TRUE
K5IW, BILL, GRUBBS, , FALSE, FALSE, TRUE
K5JFD, JOHN, DINNEEN, 3/4/2010, TRUE, TRUE, TRUE

Dúvidas são bem vindas!!

- wellington@crateus.ufc.br



