MATE34 – PLN

Atividade 01 – REN/Expressões Regulares

Setembro, 2022

Aluno: Anderson Boa Morte

PGCOMP
Universidade Federal da Bahia

Escopo

- No contexto de PLN (Processamento de Linguagem Natural), realizar um experimento de REN (Reconhecimento de Entidades Nomeadas).
- O experimento consiste em reconhecer a entidade **Pessoa** em textos de contexto geral escritos na Língua Portuguesa.

Descrição do experimento

- Utilizamos a biblioteca python re (*regular expression*) para descrever o padrão de nome de pessoas na língua portuguesa.

- O objetivo é descrever um padrão genérico de nomes de pessoas englobando contextos médico, acadêmico e jurídico.

- Descreve-se adiante o padrão construído e trecho de código de exemplo.

Padrão para nome de pessoas

- Regras para nomes completos de pessoas:
  1. Nome completo é composto por **prenome** e **sobrenome**, podendo haver a ocorrência de prefixo, conjunção e sufixo.
  2. As partes do nome completo (palavras) são compostas por letras, podendo haver a ocorrência de hifens e apóstrofos.
  3. A primeira letra de cada palavra é maiúscula.
  4. Palavras (exceto conjunções) devem ter entre 2 e 30 letras.
  - Obs. Possíveis domínios: Medico, Acadêmico, Jurídico

Trecho de código-exemplo

```
PREFIXO = '(?:(?:'
PREFIXO += 'Doutor|Dr|Dr.|Doutora|Dra|Dra.|Dr.ª|'
PREFIXO += 'Professor|Prof|Prof.|Professora|Profa|Prof.ª|'
PREFIXO += 'Advogado|Adv.|Adva.|'
PREFIXO += 'd''|D''|O''|Mc|Mac|al\-|'
PREFIXO += '))?'

PRENOME = '[A-Z]{1}[a-z]{1,30}'
ESPACO =  '\s'
ESPACO_OPT = '\s*'
CONJUNCAO = '(?:(?:de|do|dos|da|das|del|van|von|bin|le)\s)?'
SOBRENOME = '[A-Z]{1}[a-z]{1,30}'

NOME_COMPLETO = '^' + PREFIXO + ESPACO_OPT + PRENOME + ESPACO + CONJUNCAO + SOBRENOME + '$'
```

Colab: https://colab.research.google.com/drive/1knV4MVWNmZWFPx7ApPvcVDgGEKUuX2zO#scrollTo=doTeCqodms9E

Perguntas?

Contatos:
Anderson Boa Morte
andersonmorte@ufba.br