**UNIVERSIDADE FEDERAL DE ALAGOAS**
**INSTITUTO DE COMPUTAÇÃO**

**Processamento de Linguagem Natural**
**Professor: Thales Vieira**

**1a lista de exercícios** **30 de agosto de 2022**

**Instruções:**
A lista deve ser respondida por grupos de até 2 pessoas.
Resoluções idênticas de grupos distintos serão desconsideradas.
O código e demais dados devem ser anexados a cada questão.
Data limite para entrega: 13/09/2022.

**1.** Implemente uma função usando expressões regulares que testa se um string representa um CPF no formato correto. Mostre exemplos.

**2.** Implemente uma função usando expressões regulares que seja capaz, com uma única expressão regular, de extrair rua, número, apartamento (opcionalmente), bairro (opcionalmente), CEP, cidade e estado. Esta função deve ser capaz de funcionar em todos os exemplos abaixo:

- Rua José da Silva, 346, Farol, CEP 57002-220. Maceió, AL.
- Rua da Consolação, 9999, apt 302, Consolação, CEP 11022-202. São Paulo, SP.
- Avenida Atlântica, 420, ap 1001, Copacabana, cep 22011-010. Rio de Janeiro, RJ.

**3.** Implemente uma função usando expressões regulares que encontre todas as URLs em um string. Mostre exemplos.

**4.** Considere o seguinte exemplo de referência de livro em formato de citação APA:

Manning, C. D., Manning, C. D., & Schutze, H. (1999). Foundations of statistical natural language processing. MIT press.

Implemente uma função usando expressões regulares que extraia cada autor, ano de publicação, título e editora do livro, para qualquer referência de livro neste formato.

**5.** Implemente uma função usando expressões regulares que encontre e substitua todas as datas em formato *dd/mm/yyyy* por *mm-dd-yy*. Mostre exemplos.

Usando sua base de textos, resolva as seguintes questões:

**6.** Determine a distribuição de comprimentos dos textos (em quantidade de caracteres), listando estas quantidades e plotando um histograma.

**7.** Aplique os seguintes passos de pré-processamento aos textos:

- Remova todas as palavras que contêm números;
- Converta as palavras para minúsculas;
- Remova pontuação;
- Tokenize os textos em palavras, gerando um dicionário único com $n$ tokens e convertendo cada texto em um vetor de dimensão $n$ com a respectiva contagem de palavras.

Em seguida, encontre as 10 palavras mais frequentes da base de textos.

**8.** Aplique os seguintes passos de pré-processamento aos textos processados na questão anterior:

- Remova *stopwords*;
- Realize rotulação de POS;
- Realize stemização;

a) Exiba os resultados em alguns textos.
b) Verifique quais são as 10 palavras mais frequentes e compare com as 10 palavras mais frequentes da questão anterior.
c) Repita a letra b) usando os tokens stemizados.
d) Verifique quais são as classes gramaticais mais frequentes.