UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO

Processamento de Linguagem Natural
Professor: Thales Vieira

1a lista de exercícios **12 de julho de 2023**

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: 26/07/2023.

1. Escreva uma função que recebe uma senha como entrada e verifica se ela atende aos seguintes critérios de uma senha forte:

- Pelo menos 8 caracteres
- Pelo menos uma letra maiúscula e uma letra minúscula
- Pelo menos um dígito
- Pelo menos um caractere especial (por exemplo, !@#$%&*)

A função deve retornar True se a senha for forte e False caso contrário. Mostre exemplos.

2. Escreva uma função que recebe um endereço de e-mail como entrada e verifica se ele é válido de acordo com as regras básicas de formação de um e-mail. Utilize expressões regulares para realizar a validação. A função deve retornar True se o endereço de e-mail for válido e False caso contrário. Mostre exemplos.

3. 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.

4. Usando sua base de textos, determine a distribuição de comprimentos dos textos (em quantidade de caracteres), listando estas quantidades e plotando um histograma.

5. 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.

6. 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.

7. Escolha 5 documentos da sua base pré-processada, e determine seu documento mais parecido, dentre todos os documentos da base (não pode ser ele mesmo), usando:

a) Representação vetorial CountVectorizer com similaridade do cosseno;
b) Representação vetorial TF-IDF com similaridade do cosseno.

Justifique, para cada par de documentos mais parecidos, porque cada representação considerou o par semelhante.