Universidade Federal do Marannao - Departamento de Informática
Processamento de Linguagem Natural com Deep Learning
Prof. Anselmo Paiva 2025.1
Primeira Avaliação

1) Escreva uma expressão regular para os seguintes padrões descritos abaixo:
a) Número do RG. O formato completo de um RG é XX.XXX.XXX-X, em que X é um dígito (0 a 9); A expressão deve aceitar que os caracteres . e - estejam ausentes, mas não deve aceitar esses caracteres presentes em posições inesperadas.
b) Validação de senhas fortes
Crie uma expressão regular para validar senhas que atendam aos seguintes critérios:

- Comprimento mínimo de 8 caracteres
- Pelo menos uma letra maiúscula
- Pelo menos uma letra minúscula
- Pelo menos um número
- Pelo menos um caractere especial entre: !@#$%^&*()-_+=
- Não deve conter espaços em branco

c) Palavras quick money e as seguintes ofuscações que os criadores de spam usam, como:

- qu!ck m0ney
- qu!ck m@ney
- qu1ck m0n€y

2) Calcule a distância mínima de edição (Levenshtein). Mostre as operações realizadas.
a) ABACAXI -> ABACATE
b) LIVRO -> BIBLIOTECA

3) Considere a tabela com a frequência dos termos nos documentos Doc1, Doc2 e Doc3 e o valor df para cada termo t. Para uma coleção de documentos de tamanho N = 2000, calcule os pesos tf-idf de cada termo para cada documento.

Termo	Doc1	Doc2	Doc3	df
coffee	20	5	15	400
tea	5	25	0	250
breakfast	0	30	35	150
morning	18	0	22	600

Doc1 ⇒ coffee = $20 \cdot \log\left(\frac{400}{2000}\right)$
tea = $5 \log\left(\frac{250}{2000}\right)$
break = $0 \log\left(\frac{150}{2000}\right)$
morning = $18 \cdot \log\left(\frac{600}{2000}\right)$

4) Usando as frases a seguir como um pequeno corpus de treinamento:

```
<start> ele bebe café <end>
<start> ele bebe chá <end>
<start> ela adora um bom café <end>
<start> ela toca boa música <end>
<start> ele pede para descansar <end>
<start> por favor, traga outra xícara <end>
<start> ele encanta os outros clientes bebendo um bom café <end>
```

a) Calcule o número de tokens, tipos, bigrams e trigrams.
b) Estime as probabilidades P(t) que um modelo de bigrama com suavização de Laplace retornaria, para cada uma das seguintes sentenças t:

$$\frac{t+1}{+(t_{tipos})}$$