Teste - Versão 1

15/11/2024

Tecnologias de Análise de Comportamento

90 minutos

Nome: _____

ID: _____

Problema	Valores	Classificação
1	2.5	
2	2.5	
3	2.5	
4	1.5	
5	4	
6	2.5	
7	2	
8	3	
Total	20.5	

Problema 1 (Escolha Múltipla Geral, 5 valores)

Circula a opção correta.

1.1 Qual das seguintes representações de texto tem em consideração a frequência de palavras num documento em relação ao seu uso numa coleção de documentos?

a) Bag of Words (BoW)

b) Embeddings

c) N-grams

d) TF-IDF

Resposta correta: d) TF-IDF

1.2 Em extração de informação, qual das seguintes técnicas é usada para identificar nomes de pessoas, lugares, ou organizações num texto?

a) Stemming

b) Análise de sentimentos

c) Named Entity Recognition (NER)

d) Classificação de documentos

Resposta correta: c) Named Entity Recognition (NER)

1.3 Qual das alternativas seguintes descreve corretamente a tokenização?

a) Converter o texto num formato numérico

b) Remover palavras irrelevantes

c) Separar o texto em palavras (tokens)

d) Converter todas as palavras para a sua forma raiz

Resposta correta: c) Separar o texto em palavras (tokens)

1.4 No pré-processamento de texto, por que removemos stopwords?

a) Estas contêm informação importante para a análise de sentimentos

b) Estas ajudam a identificar entidades nomeadas

c) Estas geralmente são palavras comuns e não agregam valor semântico

d) Estas são necessárias para identificar a estrutura gramatical

Resposta correta: c) Estas geralmente são palavras comuns e não agregam valor semântico

1.5 Qual das opções seguintes representa um desafio comum na análise de sentimentos?

a) Dificuldade de identificar entidades nomeadas

b) Ironia e sarcasmo podem dificultar uma análise mais precisa

c) Frequência das palavras

d) Indexação dos documentos

Resposta correta: b) Ironia e sarcasmo podem dificultar uma análise mais precisa

Problema 2 (Representação de Texto, 2.5 valores)

Relaciona cada Tipo de Representação de Texto com a sua Descrição
Abaixo estão alguns tipos de representações de texto e a suas descrições. Associa a descrição correta a cada tipo de representação de texto.

Representação de Texto	Descrição
Bag of Words	
N-grams	
TF-IDF	
Embeddings	
One-hot Encoding	

A. Representa palavras como vetores de alta dimensionalidade, com um valor “1” para a posição da palavra no vocabulário e “0” nas demais posições.

B. Agrupa palavras em sequências de tamanho definido, como pares de palavras, para capturar a coocorrência entre elas.
C. Considera a frequência das palavras num documento em relação à sua frequência numa coleção de documentos, ajudando a dar mais peso a palavras mais importantes.
D. Representa palavras como vetores densos com dimensionalidade baixa, onde palavras com significados semelhantes têm vetores próximos.
E. Cria um vetor onde cada palavra é tratada como uma unidade independente, sem considerar a ordem das palavras.

Solução:

Representação de Texto	Descrição
Bag of Words	E
N-grams	B
TF-IDF	C
Embeddings	D
One-hot Encoding	A

Problema 3 (TF-IDF, 2.5 valores)

Dado o seguinte conjunto de documentos e vocabulário, calcule a matriz TF-IDF.

$$tf_{t,d} = \frac{\text{number of } t \text{ in } d}{\text{total number of terms in } d}$$

$$tfidf_{t,d} = tf_{t,d} \cdot idf_t$$

$$idf_t = log \frac{\text{total number of documents}}{\text{number of documents with } t}$$

Documentos:

1. "gato branco e preto"
2. "cão branco"
3. "gato preto"

Vocabulário:
{gato, branco, preto, cão, e}

	Termos				
Documentos	gato	branco	preto	cão	e
1					
2					
3					

Solução:

	Termos				
Documentos	gato	branco	preto	cão	e
1	0.25*log(3/2)	0.25*log(3/2)	0.25*log(3/2)	0	0.25*log(3/1)
2	0	0.5*log(3/2)	0	0.5*log(3/1)	0
3	0.5*log(3/2)	0	0.5*log(3/2)	0	0

Problema 4 (Similaridade de Vetores, 1.5 valores)

Dado os seguintes vetores binários:

Vetor A: [1, 0, 1, 1, 0, 0, 1]
Vetor B: [1, 1, 1, 0, 0, 1, 0]

$$sim_{\text{Jaccard}}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

4.1 Qual é o valor da interseção para os vetores A e B?

Solução: 2

4.2 Qual é o valor da união para os vetores A e B?

Solução: 6

4.3 Qual é o valor da similaridade de Jaccard entre o Vetor A e o Vetor B?

Solução: 2/6

Problema 5 (Classificação de Texto, 4 valores)

A classificação de texto é uma tarefa essencial em processamento de linguagem natural, onde um modelo é treinado para categorizar textos em classes com base no seu conteúdo. Considera que estás a desenvolver um sistema de análise de

comentários de clientes de uma empresa, classificando-os em duas categorias: "Satisfação" e "Insatisfação".

5.1 Qual é a diferença entre classificação binária e classificação multiclasse? Indique em qual dessas categorias se encaixa o exemplo de análise de comentários de clientes referido acima.

Solução: Classificação binária envolve duas classes (ex: satisfação e insatisfação), enquanto classificação multiclasse possui mais de duas classes. A análise de sentimentos de comentários de clientes é uma classificação binária.

5.2 Indica três passos de pré-processamento que poderiam ser aplicados aos comentários antes de treinar um classificador.

Solução: passar para letra minúscula, remover stopwords, stemming/lemmatization.

5.3 Explica o papel da vetorização de texto na classificação e mencione dois métodos comuns de representação de texto.

Solução: Converter o texto em números para poder ser usado pelos modelos. TF-IDF, Bag-of-Words.

5.4 Supõe que treinaste um classificador de sentimentos e obtiveste os seguintes resultados na matriz de confusão:

Verdadeiro Positivo (TP): 70
Falso Positivo (FP): 15
Verdadeiro Negativo (TN): 100
Falso Negativo (FN): 20

Calcula a accuracy, precision e recall do modelo.

Solução:

Accuracy = 70 + 100 / 70 +15 + 100 +20 = 170 / 205 = 0.83
Precision = 70 / 70 + 15 = 70 / 85 = 0.82
Recall = 70 / 70 + 20 = 70 / 90 = 0.78

Problema 6 (WebScraping e APIs – Verdadeiro e Falso, 2.5 valores)

Classifica as seguintes afirmações como verdadeiras (V) ou falsas (F).

6.1 Numa API RESTful, o método GET é usado para enviar dados para o servidor, enquanto o método POST é usado para solicitar dados. _____
Solução: Falso (O método GET é usado para solicitar dados, enquanto o método POST é usado para enviar dados para o servidor.)

6.2 Web scraping é o processo de extrair informações diretamente de páginas web, sem a necessidade de uma API específica para esse propósito. _____
Solução: Verdadeiro

6.3 Uma das vantagens de usar uma API em vez de web scraping é que as APIs são projetadas para fornecer dados estruturados e, geralmente, são menos sujeitas a mudanças frequentes em comparação com o HTML de uma página. _____
Solução: Verdadeiro

6.4 Usar técnicas de web scraping para recolher dados de qualquer site é sempre legal, independentemente dos termos de serviço do site. _____
Solução:Falso (As leis sobre web scraping variam, e é importante respeitar os termos de serviço e os direitos autorais de cada site.)

6.5 As APIs RESTful exigem que os dados sejam sempre transferidos em formato JSON, pois é o único formato que elas suportam. _____
Falso (APIs RESTful podem transferir dados em vários formatos, como JSON, XML, ou até mesmo HTML, dependendo do que é suportado e especificado.)

Problema 7 (Huffman Trees, 2 valores)

Para a seguinte tabela de contagem de palavras constrói a respetive Huffman Tree.

Palavra	Frequência
análise	6
dados	3
ciência	6
aplicada	2
texto	3
gráfico	7
python	12

Solução: Desenho agenda.

Problema 8 (Word Embeddings, 3 valores)

8.1 O que são word embeddings e quais as suas vantagens em relação ao uso de representações como Bag of Words?
Solução: Word embeddings são representações vetoriais de palavras num espaço contínuo de menor dimensão. A principal vantagem é que eles capturam a semântica das palavras, permitindo que palavras com significados semelhantes tenham vetores semelhantes, o que não acontece em representações como Bag of Words.

8.2 Qual a diferença entre as abordagens CBOW (Continuous Bag of Words) e Skip-gram no Word2Vec?
Solução: CBOW prevê a palavra central com base nas palavras de contexto ao redor, enquanto Skip-gram prevê as palavras de contexto usando a palavra central. CBOW é mais rápido e preferido para grandes conjuntos de dados, enquanto Skip-gram funciona bem em pequenos conjuntos e captura melhor as relações de palavras raras.

8.3 Em word embeddings, o que significa dizer que o modelo pode capturar relações entre palavras, como "rei - homem + mulher = rainha"?
Solução: Isso significa que os embeddings conseguem capturar a estrutura semântica e relações de analogia entre palavras. Num espaço de embeddings, operações vetoriais podem representar relações semânticas, como género ou hierarquia, permitindo que o modelo infira novas relações de forma significativa.