Atividades unidade 3

TERMINOLOGIA E CONCEITOS

TC.3.1. Qual tipo de problema pode surgir na montagem de um modelo do tipo bag of words caso etapas tais como a remoção de caracteres especiais (ex. sinais de pontuação) e conversão para minúsculas/maiúsculas não sejam efetuadas? Ilustre isso para as seguintes frases que fazem parte de um mesmo corpus de texto (i.e.: são usadas para a montagem do léxico do modelo):
Frase 1: Eu quero tomar água!
Frase 2: eu, prefiro tomar café.
Na sua resposta mostre como ficaria o léxico do modelo e os bag of words correspondentes.

O problema que pode surgir na montagem de um modelo do tipo Bag of Words (BoW) é a inconsistência na representação das palavras. Problemas de sem a normalização, como se caracteres especiais não forem removidos, palavras como “água!” e “água” serão tratadas como tokens diferentes.

E se as palavras não forem convertidas para minúsculas/maiúsculas, “eu” e “Eu” serão consideradas palavras diferentes, criando duplicidade desnecessária no vocabulário.

Com o sinal de pontuação e sem a conversão de minúsculas/maiúsculas, o léxico seria: [“Eu”, “quero”, “tomar”, “água!”, “eu,”, “prefiro”, “café.”]
Bag of Words:
Frase 1: Eu quero tomar água!
[1, 1, 1, 1, 0, 0, 0]
Frase 2: eu, prefiro tomar café.
[0, 0, 1, 0, 1, 1, 1]

Para evitar esses problemas, as etapas de remoção de caracteres especiais e conversão para minúsculas devem ser realizadas.
Frase 1: Eu quero tomar água! -> eu quero tomar água
Frase 2: eu, prefiro tomar café. -> eu prefiro tomar café

O léxico seria: [“eu”, “quero”, “tomar”, “água”, “prefiro”, “café”]
Bag of Words:
Frase 1: eu quero tomar água
[1, 1, 1, 1, 0, 0]
Frase 2: eu prefiro tomar café
[1, 0, 1, 0, 1, 1]

Com a normalização, o vocabulário fica consistente e tokens representam palavras de forma uniforme, o que facilita o processamento e a análise do texto. As palavras que antes eram tratadas como diferentes, agora são reconhecidas como a mesma palavra.

PRÁTICA DE PROGRAMAÇÃO

PP.3.4. Demonstre a modelagem de tópicos, com LDA, utilizando alguns documentos representativos de revisões de produtos. Efetue todas as etapas de pré-processamento adequadas antes de efetuar a modelagem.

Foi feito as seguintes etapas de pré-processamento: conversão para minúsculo, remoção de pontuação e caracteres especiais, tokenização e remoção de stopwords para documentos de revisões de produtos.

Após, os documentos foram transformados em uma representação numérica para que o modelo LDA possa ser aplicado, usando o TF-IDF (Term Frequency-Inverse Document Frequency), onde palavras raras recebem mais peso e palavras muito frequentes em todos os documentos têm seu peso reduzido.

Cada linha da tabela indica a distribuição de probabilidade de cada documento estar associado a cada um dos tópicos. Por exemplo, o documento 0 é

fortemente representado pelo Tópico 0, com um valor de 0.793 (79,3% de associação).

A análise dos tópicos pode fornecer insights valiosos sobre as revisões de produtos, como por exemplo, identificar queixas sobre a bateria, críticas ao desempenho ou elogios à qualidade.

```
    "Este produto é maravilhoso! A bateria dura muito e o desempenho é excelente.", #0
    "Não gostei do desempenho. A bateria descarrega rápido e o suporte não responde.", #1
    "Produto ok, mas a bateria poderia ser melhor. O preço é justo pelo que oferece.", #2
    "A qualidade do material é ótima, mas a bateria não tem boa duração.", #3
    "Muito bom! Recomendo, a bateria dura o dia todo e o desempenho é ótimo!" #4
```

```
PROBLEMS  TERMINAL  DEBUG CONSOLE

PS C:\Users\Amanda\Documents\6sem\PLN\PP3_4> python app.py
Palavras do tópico #0:
['dura', 'bateria', 'desempenho', 'excelente', 'maravilhoso']

Palavras do tópico #1:
['responde', 'rápido', 'suporte', 'descarrega', 'gostei']

Palavras do tópico #2:
['preço', 'poderia', 'ok', 'justo', 'melhor']

    Topic 0   Topic 1   Topic 2
0  0.793906  0.102358  0.103736
1  0.099876  0.804268  0.095856
2  0.093310  0.091104  0.815587
3  0.797589  0.101261  0.101150
4  0.816851  0.092001  0.091147
```