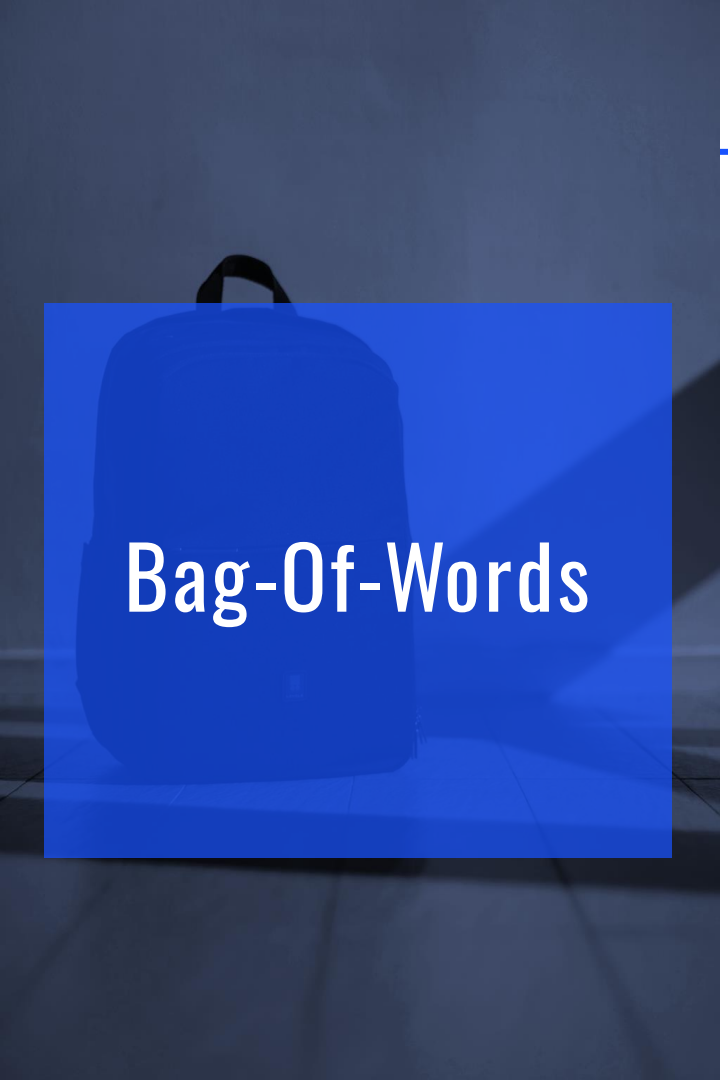
PUC Minas Virtual

PROCESSAMENTO DE LINGUAGEM NATURAL

Bag-Of-Words



Bag-Of-Words

Técnica clássica de representação textual que transforma textos em vetores de frequência de palavras

Exemplo de Bag-of-Words

C1 Este filme é muito assustador e longo

C2 Este filme não é assustador e é lento

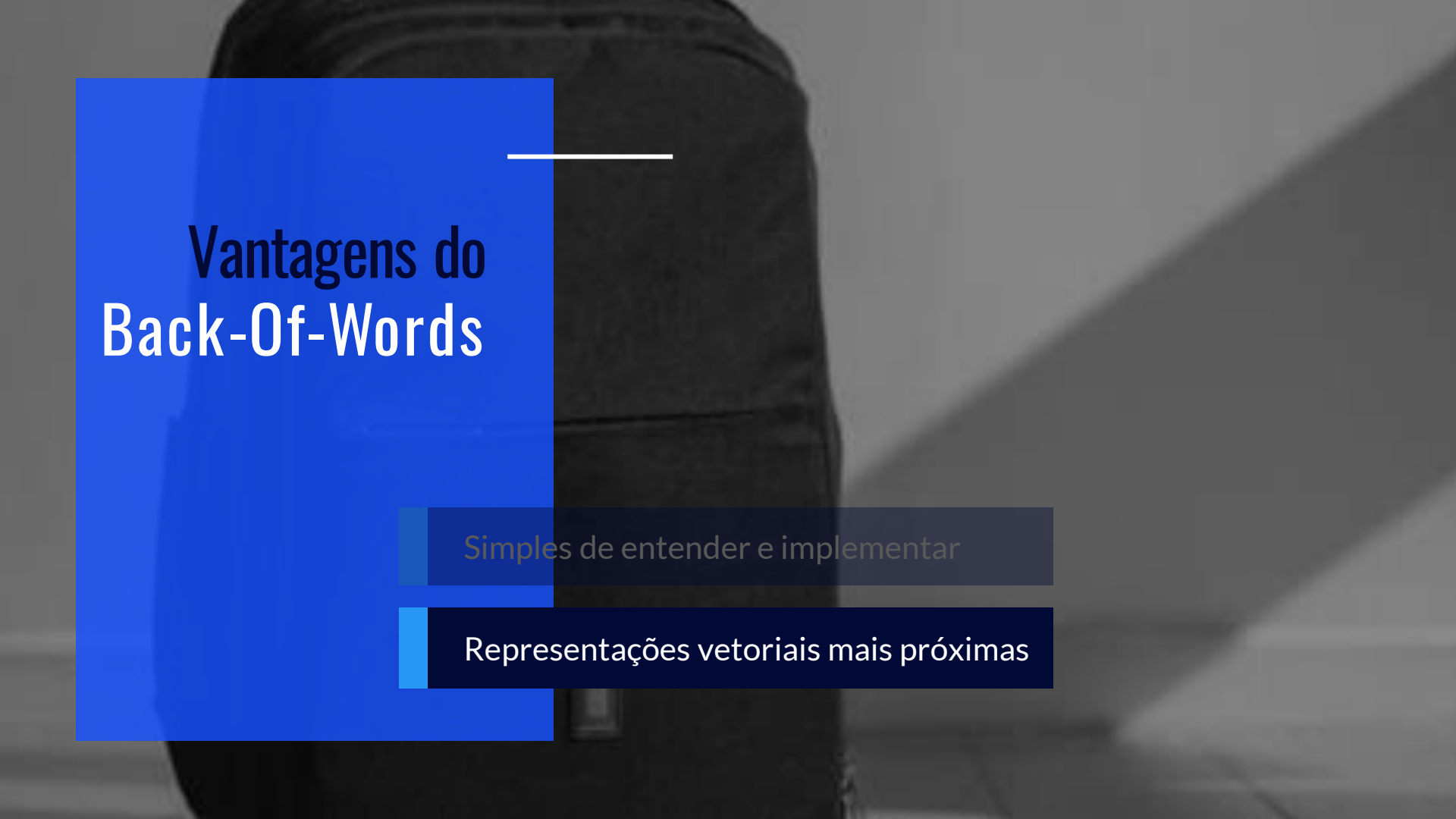
C3 Este filme é assustador e bom

Exemplo de Bag-of-Words

	filme	assustador	longo	lento	bom
C1	1	1	1	0	0
C2	1	1	0	1	0
C3	1	1	0	0	1

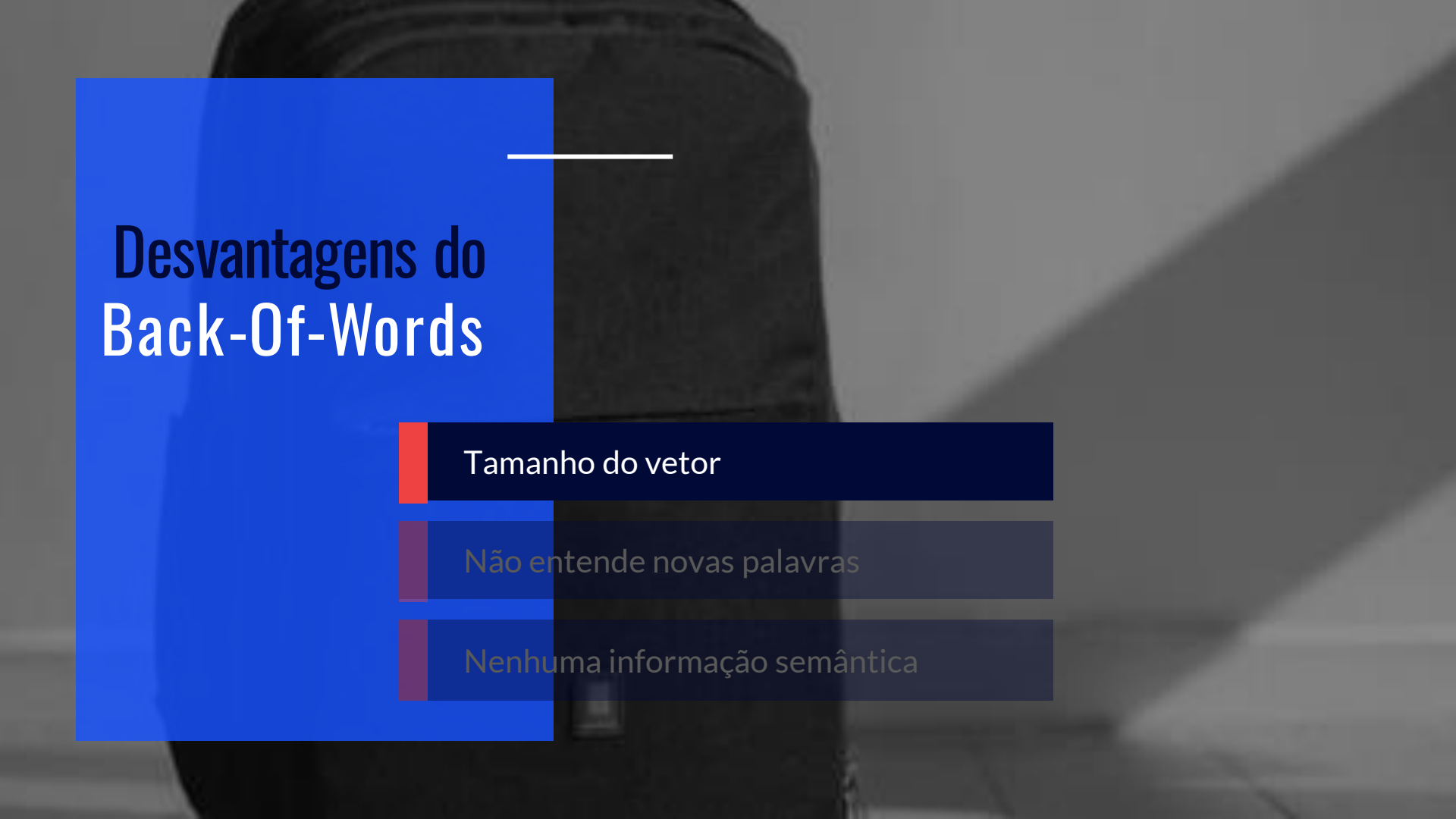
Vantagens do Back-Of-Words

- Simples de entender e implementar
- Representações vetoriais mais próximas



Vantagens do Back-Of-Words

- Simples de entender e implementar
- Representações vetoriais mais próximas



Desvantagens do Back-Of-Words

- Tamanho do vetor
- Não entende novas palavras
- Nenhuma informação semântica

Desvantagens do Back-Of-Words

- Tamanho do vetor
- Não entende novas palavras
- Nenhuma informação semântica

Desvantagens do Back-Of-Words

- Tamanho do vetor
- Não entende novas palavras
- Nenhuma informação semântica

```
1. from sklearn.feature_extraction.text import CountVectorizer

2. # cria a instância do CountVectorizer para converter o
3. # texto em uma matriz binária
4. vectorizer = CountVectorizer(binary=True)

6. # transforma o corpus em uma matriz esparsa de presença
7. de # tokens
8. bag = vectorizer.fit_transform(corpus)
```

Implementação do Bag-Of-Words

PUC Minas
Virtual