Applied Text Analytics & Natural Language Processing

with Dr. Mahdi Roozbahani
& Wafa Louhichi

BoW

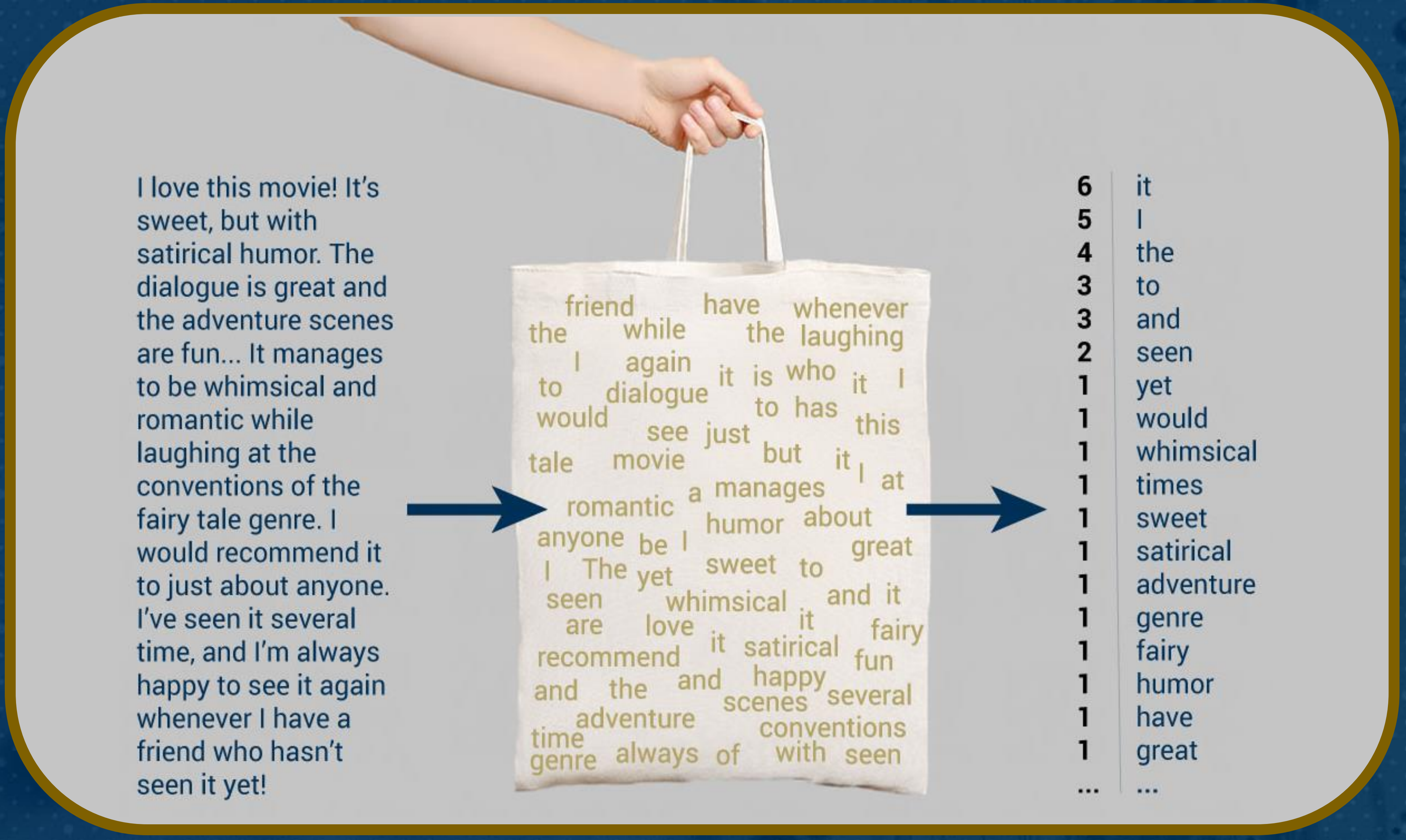
Learning Objectives

In this lesson, you will learn another discrete text representation

- Explain **text data** into numerical format using Bag of Words
- Understand the advantages and disadvantages of Bag of Words

Bag of Words Model

- Represent each **document** as a **bag of words**, ignoring words' ordering for **simplicity**.



I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several time, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

6	it
5	I
4	the
3	to
3	and
2	seen
1	yet
1	would
1	whimsical
1	times
1	sweet
1	satirical
1	adventure
1	genre
1	fairy
1	humor
1	have
1	great
...	...

Bag of Words Model

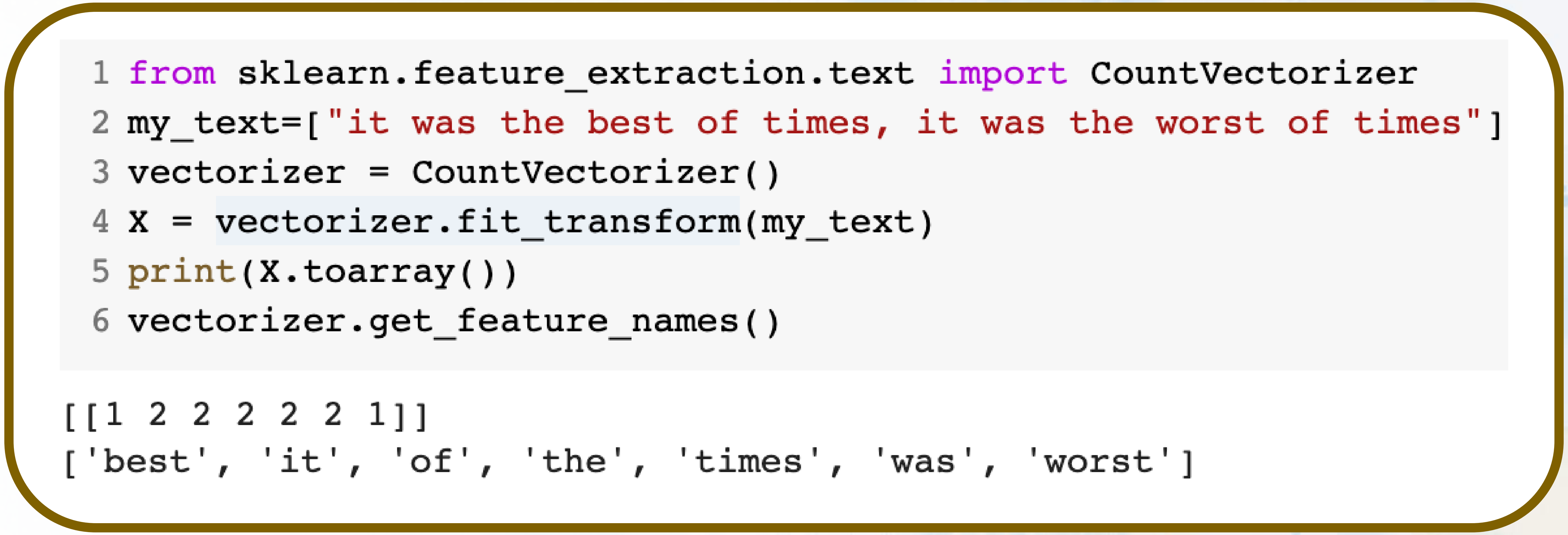
- Represent a document as a column vector X of word counts
- The bag of words is a fixed-length representation, which consists of a vector of word counts:

  Document: $\mathrm{D} = it\ was\ the\ best\ of\ times, it\ was\ the\ worst\ of\ times$

  Vocabulary: $V = [aardvak, \ldots, it, \ldots, best, \ldots, times, \ldots, zyther]$

  Bag of Words: $X = [0, \ldots, 2, \ldots, 1, \ldots, 2, \ldots, 0]$
- The size of X is $1xd$ where d is the size of the vocabulary.
- A collection of n documents is represented with a matrix of size nxd

Example of Bag of Words Using Sklearn



```python
from sklearn.feature_extraction.text import CountVectorizer
my_text=["it was the best of times, it was the worst of times"]
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(my_text)
print(X.toarray())
vectorizer.get_feature_names()
```

```
[[1 2 2 2 2 2 1]]
['best', 'it', 'of', 'the', 'times', 'was', 'worst']
```

Advantages and Disadvantages of Bag of Words

- Advantages of Bag of Words:
  - Simple and easy to implement
- Disadvantages of Bag of Words:
  - Every document is represented as a vector of the size of the vocabulary: not scalable for a large vocabulary (100,000 words)
  - High dimensional sparse matrix which can be memory & computationally expensive
  - The order of words is disregarded and thus the meaning coming from the context is not captured

Summary

- Use Bag of Words as a discrete text representation method to represent documents
- Understand the advantages and disadvantages of this method