

95-865 Unstructured Data Analytics

Recitation: Array & Vector
Representations, Python Review

Slides by George H. Chen

Today

1. Revisit last lecture's demo (basic text analysis) but using arrays/vectors instead of dictionaries/Counter objects

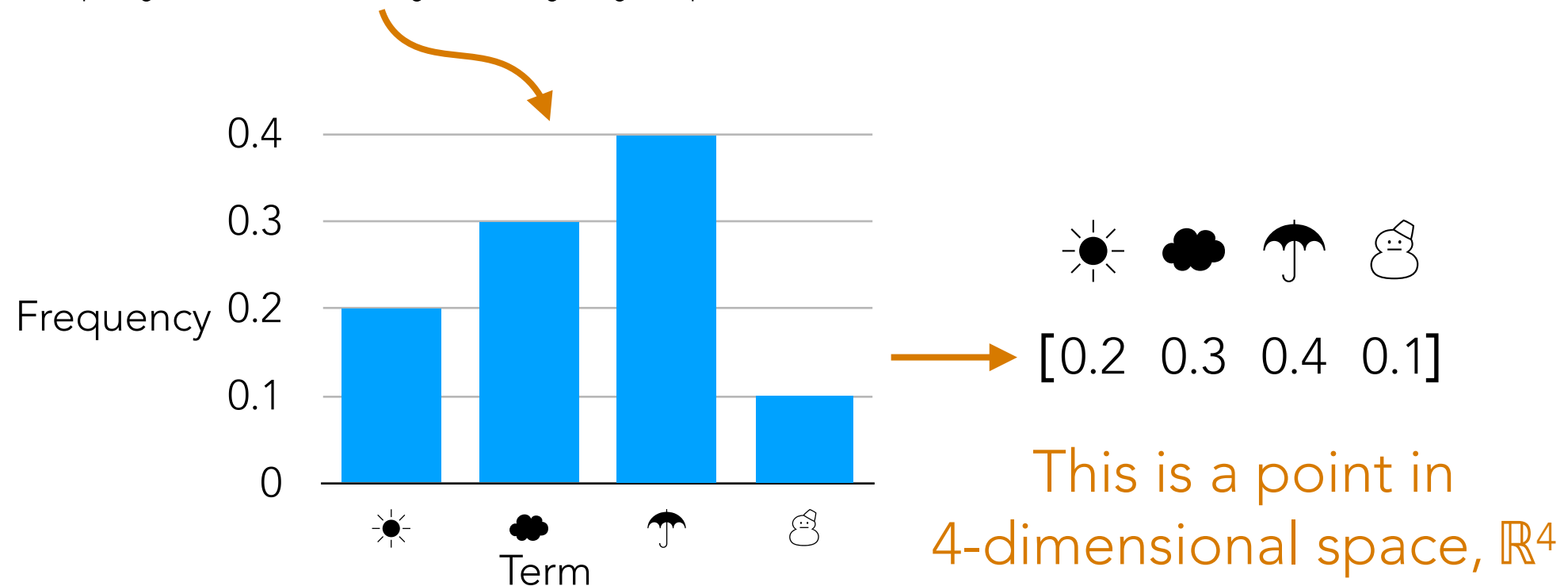
When we get to neural nets: array/vector representations are much more commonly used in practice than dictionaries/Counter objects

2. Time-permitting: some additional Python review

Recap: Basic Text Analysis

We represent each document as a histogram/probability distribution

Document: ☀️☔️☁️☁️☁️☔️☷☔️☔️☀️



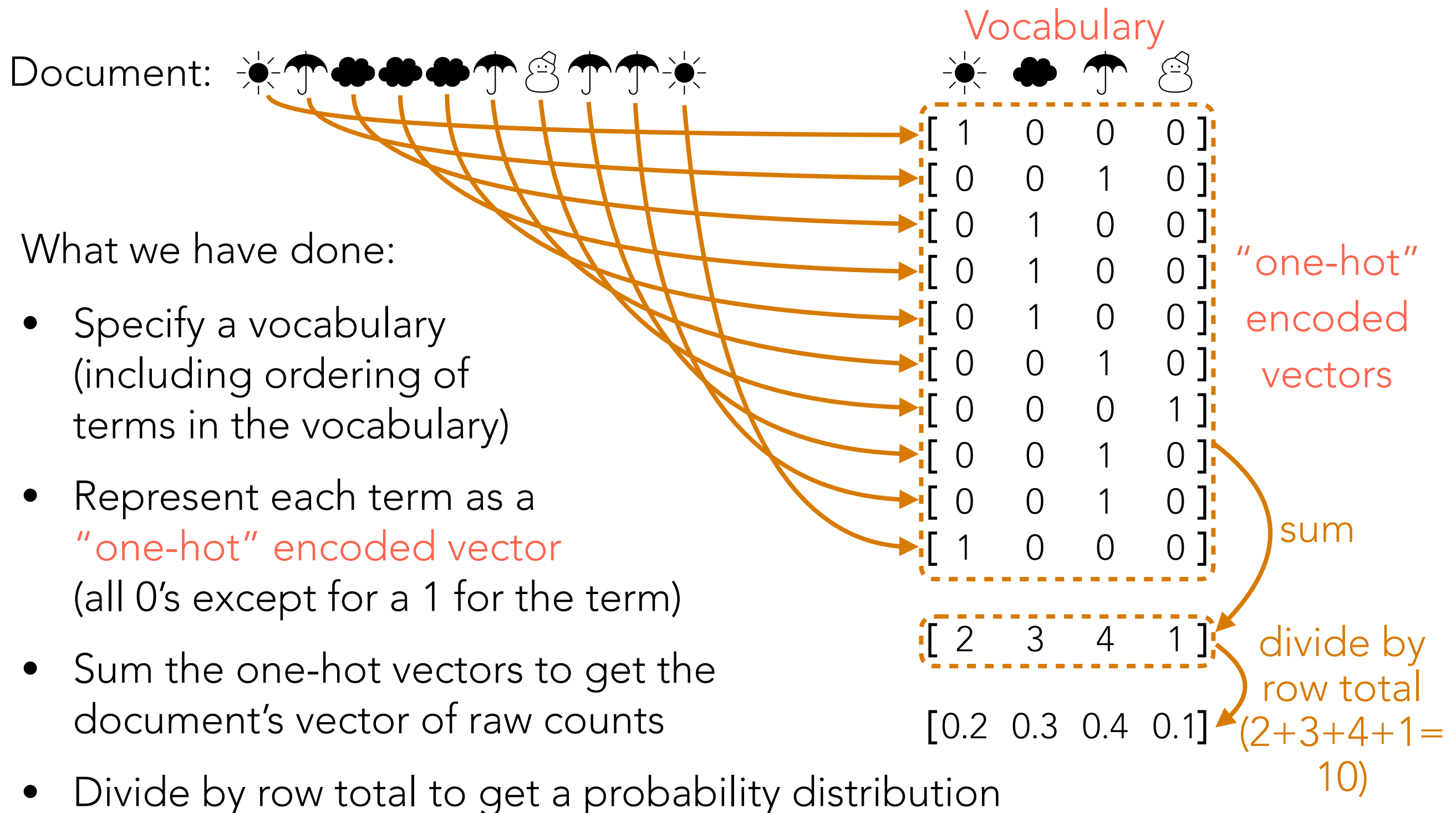
We refer to a vector representation of the document as a feature vector

dimensions = number of terms

If there are lots of terms $\Rightarrow$ feature vectors are high-dimensional

Another Viewpoint

We represent each document as a histogram/probability distribution



Vector Version of Previous Code Demo

- We first build the vocabulary of the Wikipedia article (by finding all unique tokens present in the article)

Index: 0 1 2 3 1 4 5 6 7
The opioid epidemic or opioid crisis is the rapid
 8 9 6 10 11 12 13 14 15
 increase in the use of prescription and non-
 12 1 16 9 6 17 18
prescription opioid drugs in the United States ...

```
word_to_idx = {'The': 0,  
               'opioid': 1,  
               'epidemic': 2, ...}
```

maps each word to its index; this is *not* a term frequency table

```
vocab = ['The', 'opioid',  
         'epidemic', ...]
```

listed in the same order as the word indices

Index	Word
0	The
1	opioid
2	epidemic
⋮	⋮

- We can then represent every word as a one-hot encoded vector
- We sum the one-hot encoded vectors to get a raw counts vector

Vector Version of Previous Code Demo

Demo