Carnegie Mellon University
HeinzCollege

95-865 Unstructured Data Analytics

Lecture 2: Basic text analysis (cont'd)

Slides by George H. Chen

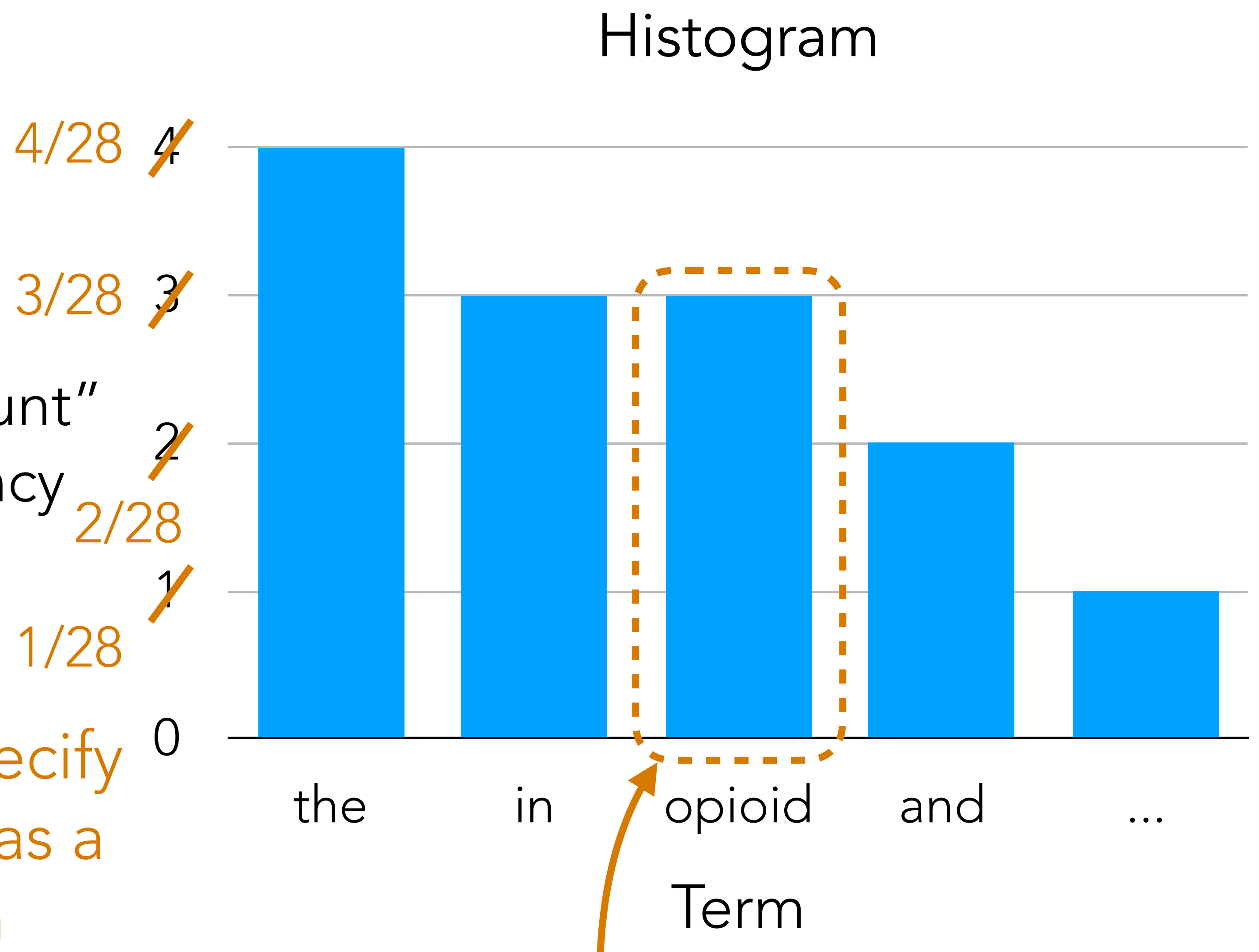
(Flashback)

Term frequencies

The: 1	/28
opioid: 3	/28
epidemic: 1	/28
or: 1	/28
crisis: 1	/28
is: 1	/28
the: 4	/28
rapid: 1	/28
increase: 1	/28
in: 3	/28
use: 1	/28
of: 1	/28
prescription: 1	/28
and: 2	/28
non-prescription: 1	/28
drugs: 1	/28
United: 1	/28
States: 1	/28
Canada: 1	/28
2010s.: 1	/28

The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

Total number of words in sentence: 28

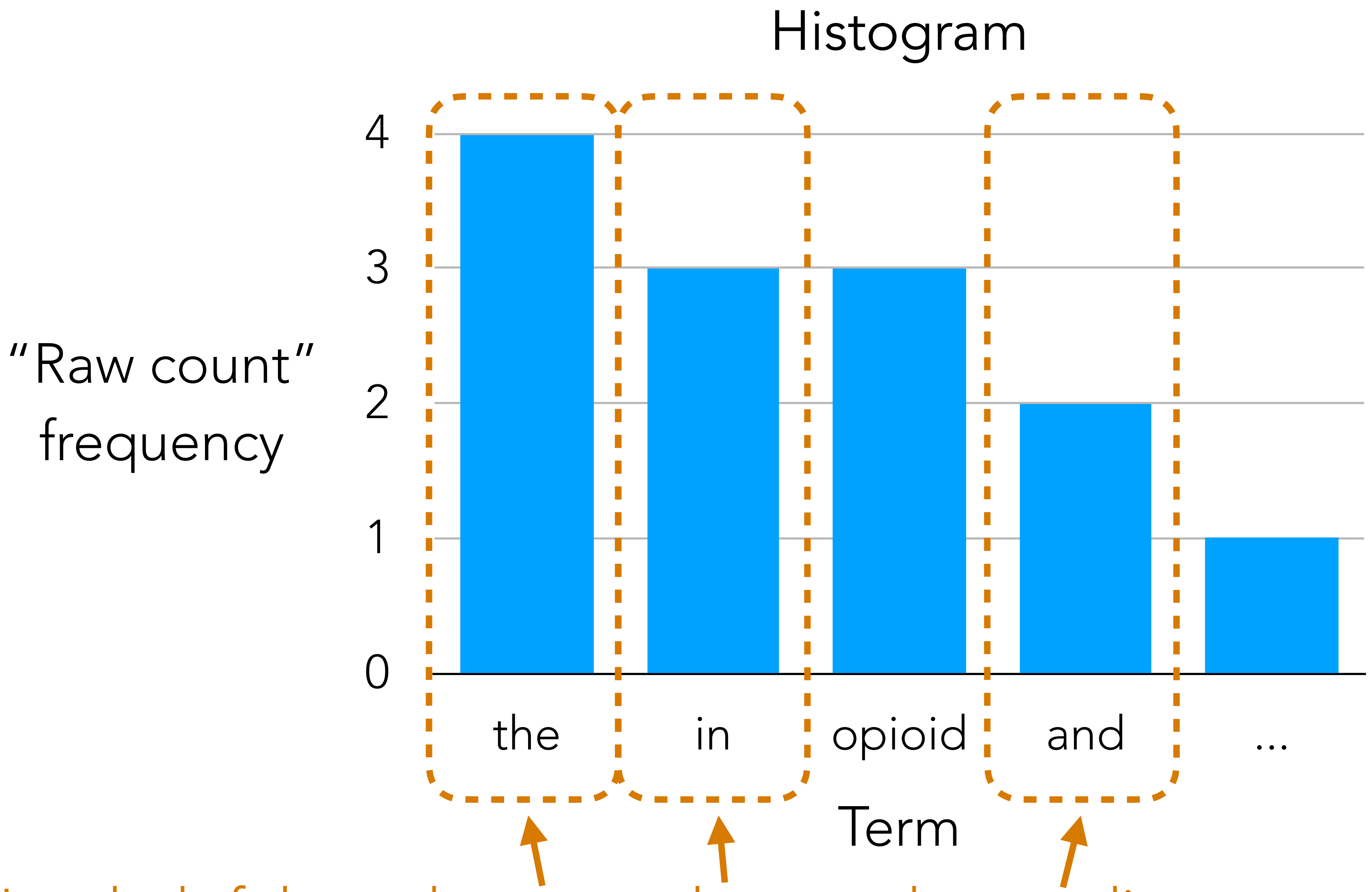


"Raw count" frequency

can also specify frequency as a fraction

Fraction of words in the sentence that are "opioid"

(Flashback) Some Words Don't Help?



How helpful are these words to understanding semantics?

Bag-of-words models: many frequently occurring words unhelpful

We can remove these words first (remove them from the "bag")
➔ words that are removed are called **stopwords**

(determined by removing most frequent words or using curated stopword lists)

(Flashback)

Is removing stop words always a good thing?

“To be or not to be”

Some Words Mean the Same Thing?

Term frequencies

The: 1
opioid: 3
epidemic: 1
or: 1
crisis: 1
is: 1
the: 4
rapid: 1
increase: 1
in: 3
use: 1
of: 1
prescription: 1
and: 2
non-prescription: 1
drugs: 1
United: 1
States: 1
Canada: 1
2010s.: 1

Should capitalization matter?

What about:

- walk, walking
- democracy, democratic, democratization
- good, better

Merging modified versions of "same" word to be analyzed as a single word is called lemmatization *(we'll see software for doing this shortly)*

What about a word that has multiple meanings?

Challenging: try to split up word into multiple words depending on meaning (requires inferring meaning from context)

This problem is called word sense disambiguation (WSD)

Treat Some Phrases as a Single Word?

Term frequencies

The: 1
opioid: 3
epidemic: 1
or: 1
crisis: 1
is: 1
the: 4
rapid: 1
increase: 1
in: 3
use: 1
of: 1
prescription: 1
and: 2
non-prescription: 1
drugs: 1
United: 1
States: 1
Canada: 1
2010s.: 1

First need to detect what are "named entities": called named entity recognition
(we'll see software for doing this shortly)

Treat as single 2-word phrase "United States"?

Some Other Basic NLP Tasks

- **Tokenization:** figuring out what are the atomic "words" (including how to treat punctuation)

- **Part-of-speech tagging:** figuring out what are nouns, verbs, adjectives, etc

- **Sentence recognition:** figuring out when sentences actually end rather than there being some acronym with periods in it, etc

Bigram Model

The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

The opioid	opioid epidemic	epidemic or
or opioid	opioid crisis	crisis is

Ordering of words now matters (a little)

…

# unique cards changes dramatically

If using stop words, remove any phrase with at least 1 stop word

1 word at a time: unigram model

2 words at a time: bigram model

3 words at a time: trigram model

n words at a time: *n*-gram model

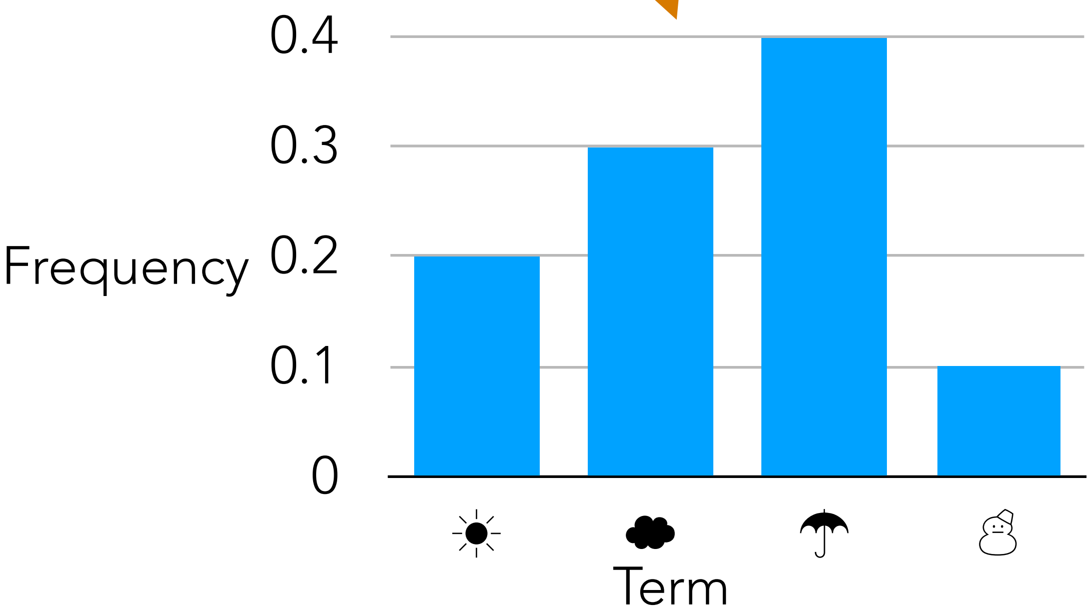
The spaCy Python Package

Demo

Recap: Basic Text Analysis

We represent each document as a histogram/probability distribution

Document: ☀☂☁☁☁☂☃☂☂☀



☀ ☁ ☂ ☃

[0.2 0.3 0.4 0.1]

This is a point in 4-dimensional space, $\mathbb{R}^4$

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

# dimensions = number of terms

If there are lots of terms ⇒ feature vectors are high-dimensional