



Lesson2:
**Modelling the Web with Simple Statistical
Descriptive Text Models**
Unit2:
Typical size of a document

Rene Pickhardt

Introduction to Web Science Part 2
Emerging Web Properties





Completing this unit you should

- Be familiar with some basic statistical objects like
 - Median
 - Mean
 - Histograms
- Should be able to relate a histogram to its cumulative distribution function



What is the typical length of a document?

- We saw
 - 16491538 words
 - 119754 documents
- Dividing these numbers makes
 - About 137 words per document on average
- Lets have a closer look!



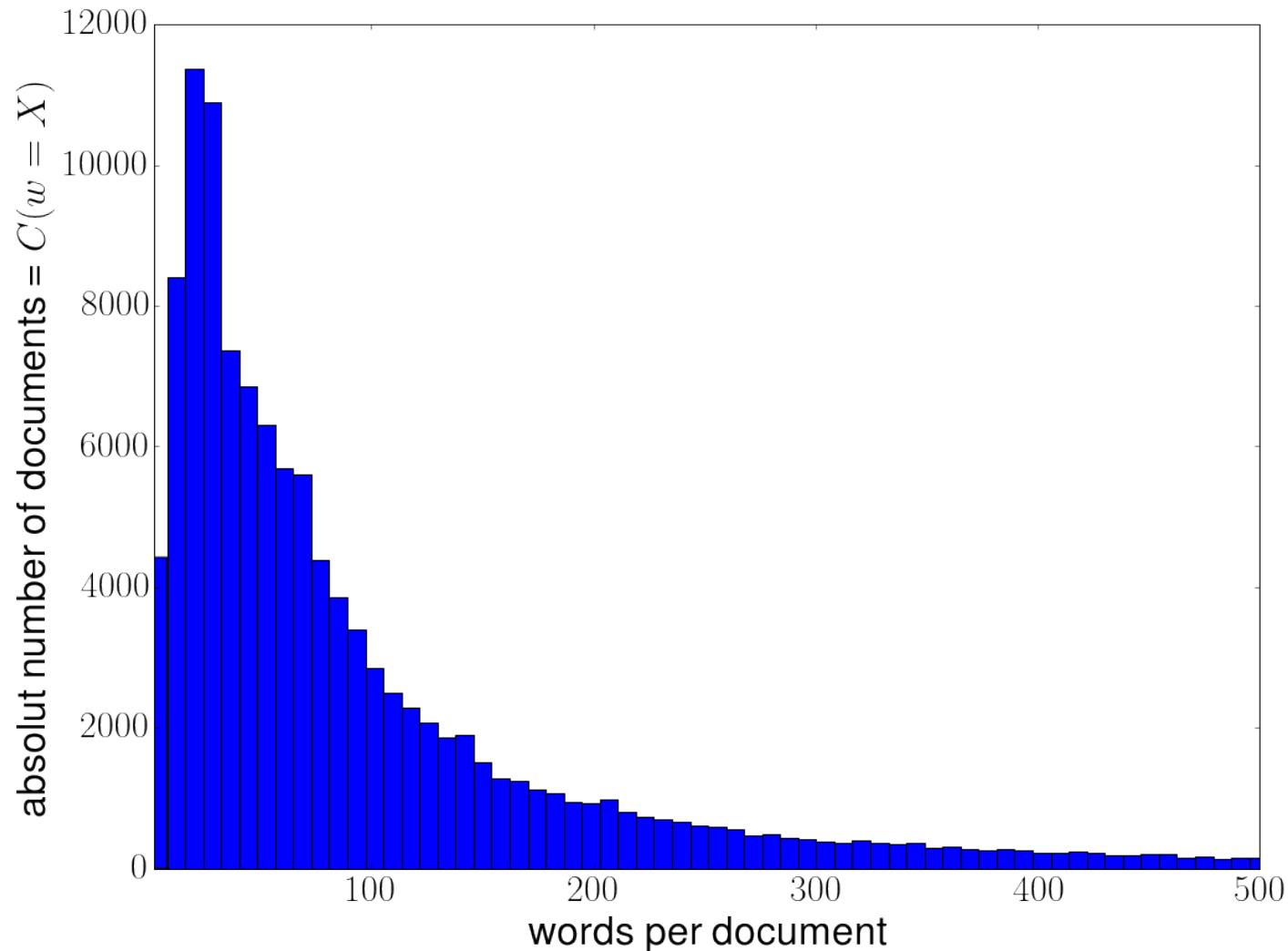
What is the typical length of a document?

- Count words for every document
- Build mean over all documents
 - 137 words per document
- Have look at the histogram
 - Visualize how many documents have
 - 0-10 words
 - 10-20 words
 - 20-30 words
 - ...



Histogram of words per document

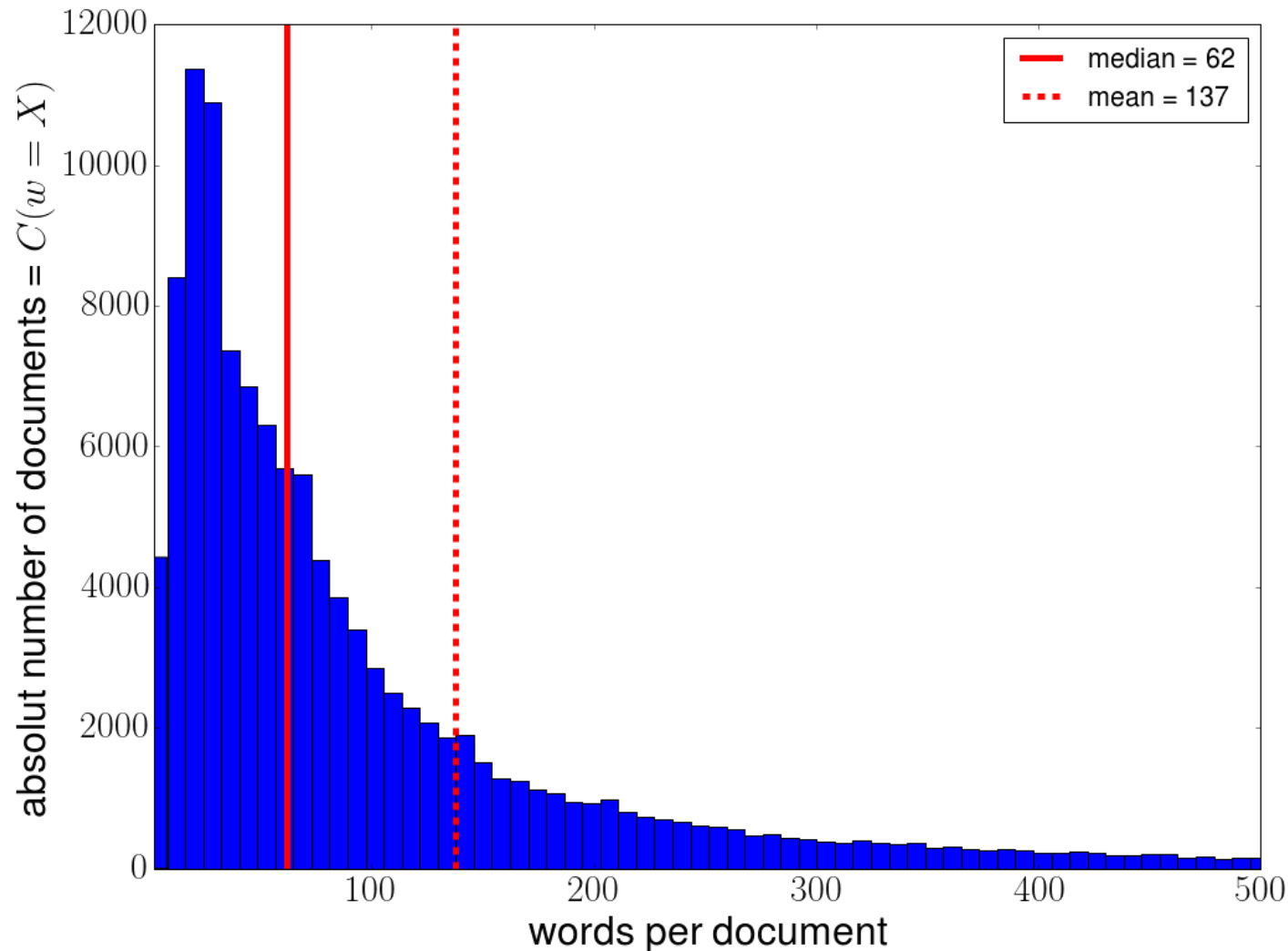
Distribution of article lengths in words of Simple Wikipedia articles





Histogram of words per document

Distribution of article lengths in words of Simple Wikipedia articles





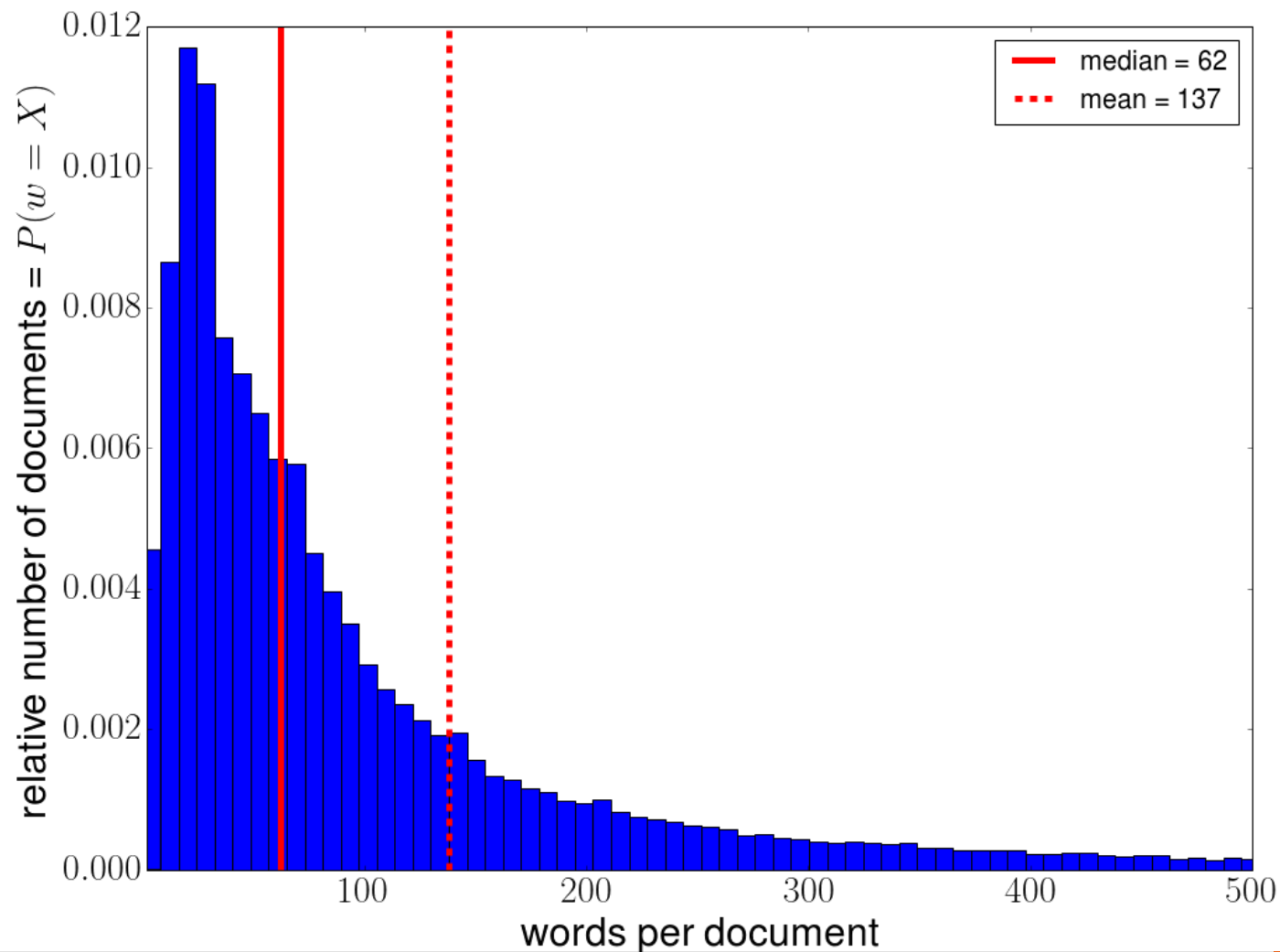
Some facts about the median

- The element which splits a set into halves of equal size
- `wordsPerDoc = [10,11,12, 14, 1000000]`
- `mean(words) = 200'009.4`
- `median(words) = 12`
 - Two documents have less words and two have more
- What if `length(wordsPerDoc)` is even?



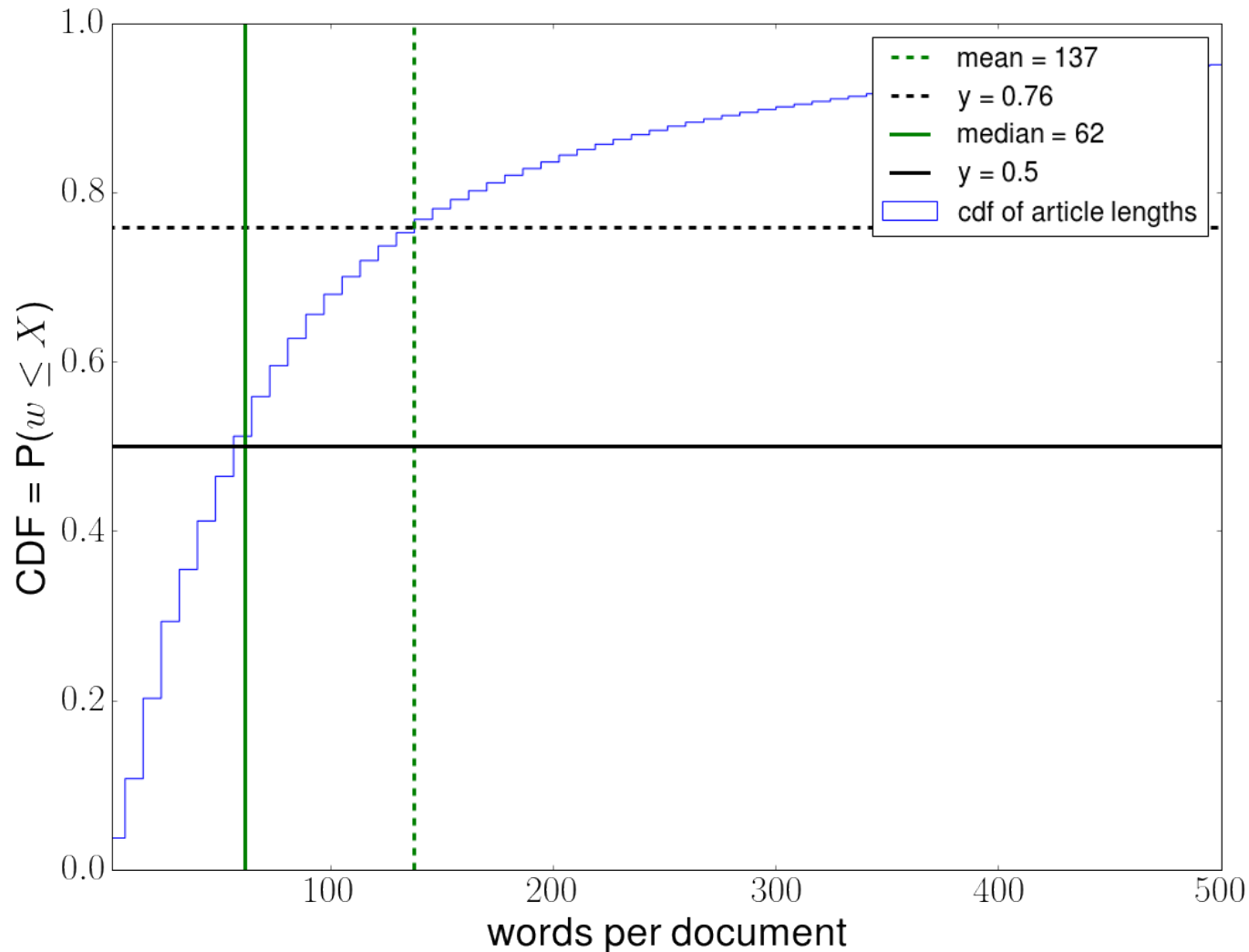
Normalize the histogram

Normed Distribution of article lengths in words of Simple Wikipedia articles



3 out of 4 articles are shorter than average!

CDF of article lengths in words of Simple Wikipedia articles





Thank you for your attention!



Contact:

Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de

WeST 
People and Knowledge Networks



Copyright:

- This Slide deck is licensed under creative commons 3.0. share alike attribution license. It was created by Rene Pickhardt. You can use share and modify this slide deck as long as you attribute the author and keep the same license. All graphics have been self made (unless otherwise stated)