LING 388: Computers and Language

Lecture 18

Today's Topics

- Homework 7
    - Parts 1, 2 and 3
- Last Time:
    - we did Mendehall (1887) live in class
    - **idea**: use word length statistics on *Oliver Twist* by Charles Dickens

Last Time

- We confirmed Mendenhall's 5,000 word plots

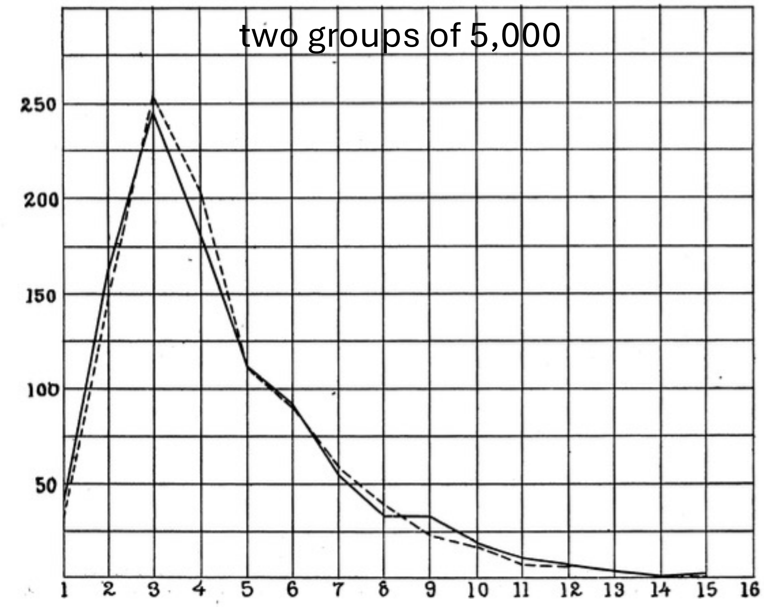
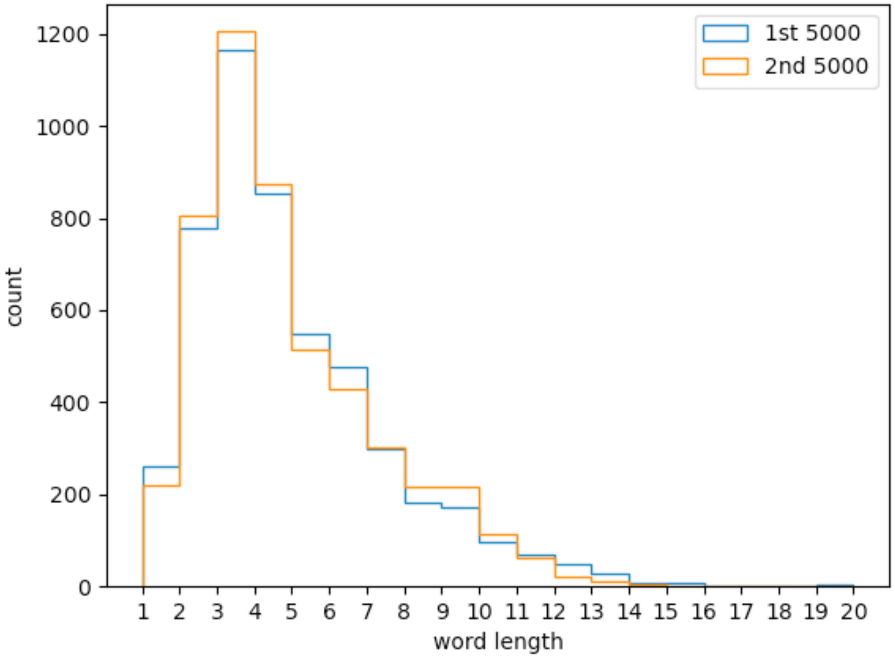


FIG. 4.—TWO GROUPS, OF FIVE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'



Code Recap

```
$ python
Python 3.9.16
>>> raw = open('oliver_twist.txt').read()
>>> import nltk
>>> words = nltk.word_tokenize(raw)
>>> words2 = [word for word in words if any(c.isalpha() for c in word)]
>>> len1 = [len(word) for word in words2[0:5000]]
>>> len2 = [len(word) for word in words2[5000:10000]]
>>> mx = max(max(len1),max(len2))
>>> mx
20
```

may need encoding='utf8'

Code Recap

Plotting:

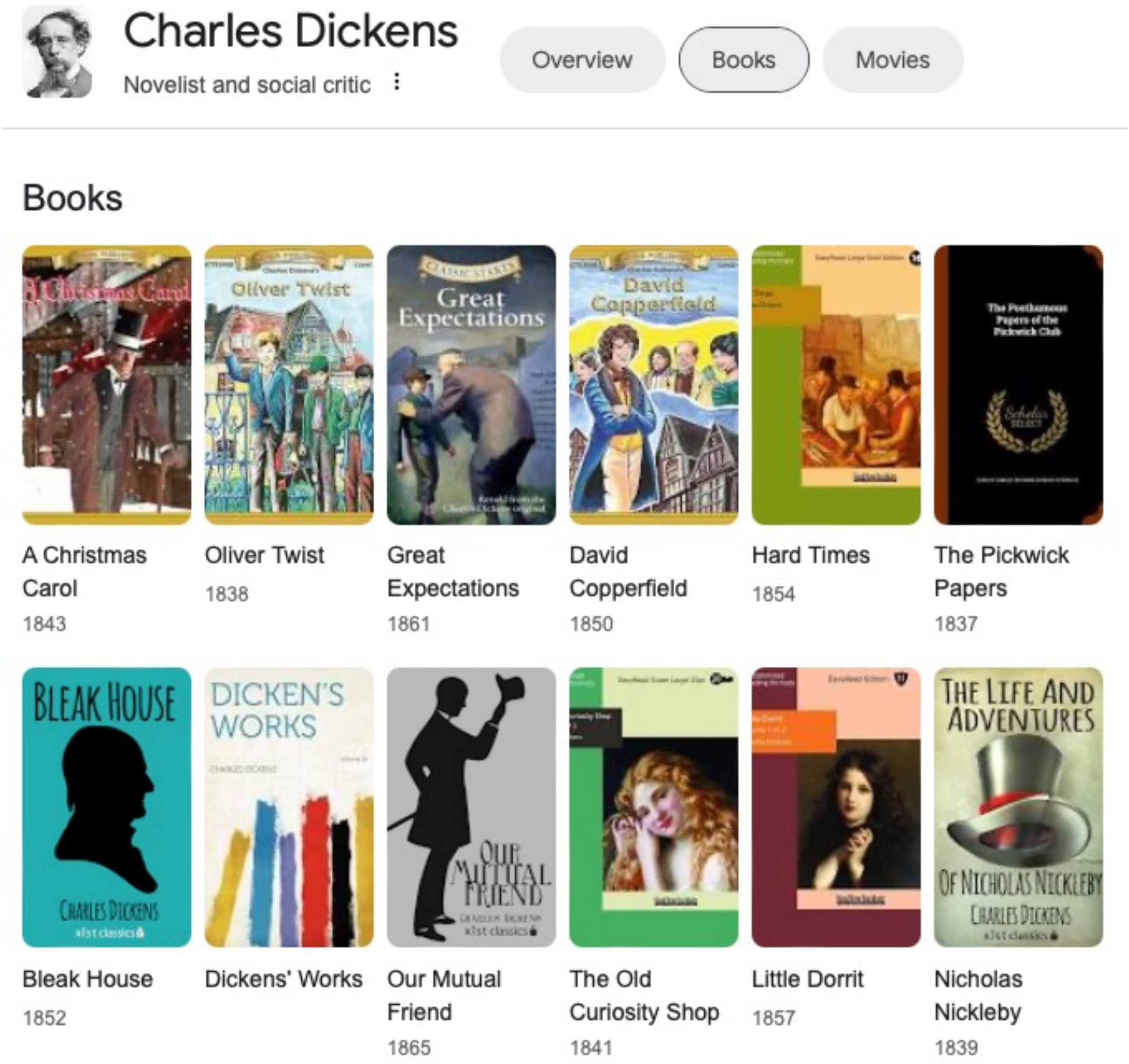
```
>>> import matplotlib.pyplot as plt
>>> plt.hist(len1, range(1,mx+1), histtype='step', label='1st 5000')
>>> plt.hist(len2, range(1,mx+1), histtype='step', label='2nd 5000')
>>> plt.xticks(range(1,mx+1))
>>> plt.xlabel('word length')
>>> plt.ylabel('count')
>>> plt.legend()
>>> plt.show()
```

Homework 7

Part 1: Let's compare our two 5,000 word Mendenhall test for *Oliver Twist* (1838)

with

- *Nicholas Nickleby* (1839) and
- *David Copperfield* (1850)



Charles Dickens

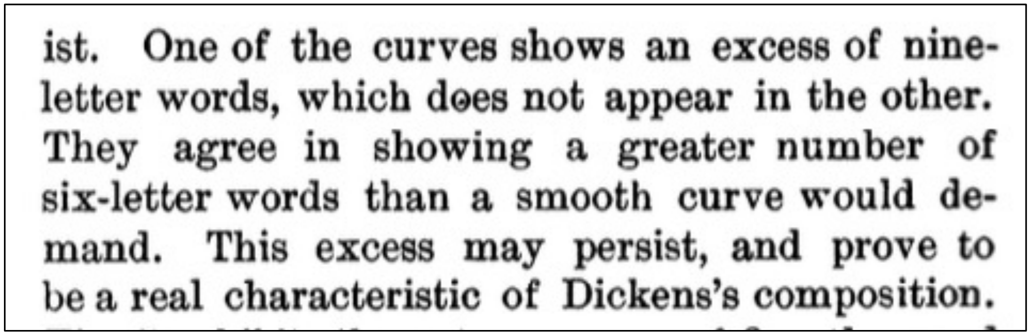
Novelist and social critic

Books

A Christmas Carol	1843
Oliver Twist	1838
Great Expectations	1861
David Copperfield	1850
Hard Times	1854
The Pickwick Papers	1837
Bleak House	1852
Dickens' Works	
Our Mutual Friend	1865
The Old Curiosity Shop	1841
Little Dorrit	1857
Nicholas Nickleby	1839

Homework 7

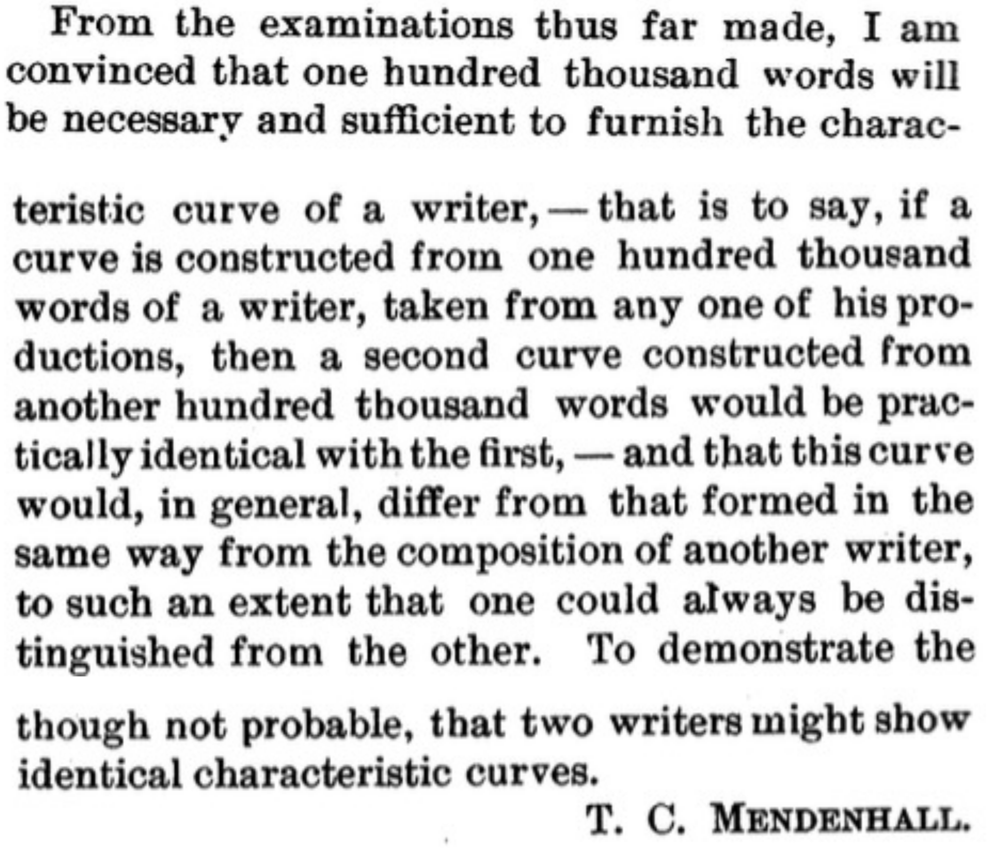
- Part 2:
  - Mendenhall claims something about six-letter words



ist. One of the curves shows an excess of nine-letter words, which does not appear in the other. They agree in showing a greater number of six-letter words than a smooth curve would demand. This excess may persist, and prove to be a real characteristic of Dickens's composition.

Homework 7

- Part 3:
  - Mendenhall claims something about 100,000 words



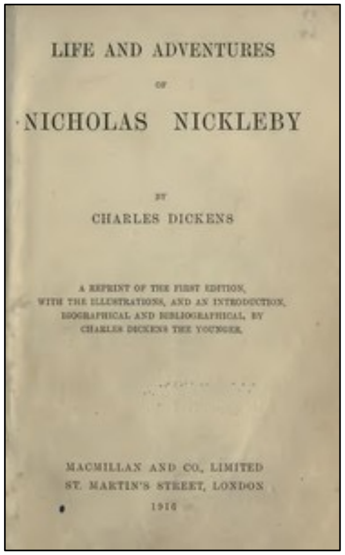
From the examinations thus far made, I am convinced that one hundred thousand words will be necessary and sufficient to furnish the characteristic curve of a writer, — that is to say, if a curve is constructed from one hundred thousand words of a writer, taken from any one of his productions, then a second curve constructed from another hundred thousand words would be practically identical with the first, — and that this curve would, in general, differ from that formed in the same way from the composition of another writer, to such an extent that one could always be distinguished from the other. To demonstrate the

though not probable, that two writers might show identical characteristic curves.

T. C. Mendenhall.

Homework 7

- Part 1 details:
  - Step 1: grab txt files for *Nicholas Nickleby* (1839) and *David Copperfield* (1850) from https://www.gutenberg.org



LIFE AND ADVENTURES

OF

NICHOLAS NICKLEBY

BY

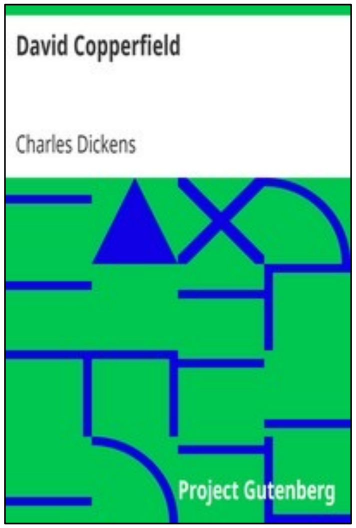
CHARLES DICKENS

A REPRINT OF THE FIRST EDITION,
WITH THE ILLUSTRATIONS, AND AN INTRODUCTION,
BIOGRAPHICAL AND BIBLIOGRAPHICAL, BY
CHARLES DICKENS THE YOUNGER.

MACMILLAN AND CO., LIMITED
ST. MARTIN'S STREET, LONDON
1916

```
pg967.txt
raw: 1,848,364
words: 396,970
```

*after editing

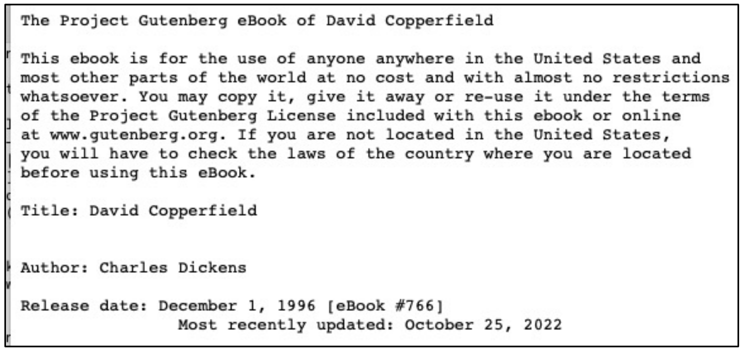


```
pg766.txt
raw: 1,934,660
words: 443,615
```

* after editing

Homework 7

- Part 1 details:
  - Step 2: edit `pg967.txt` and `pg766.txt` to remove the Project Gutenberg boilerplate.
    - You may want to save the edited versions under new names, e.g. `nn.txt` and `dc.txt`.



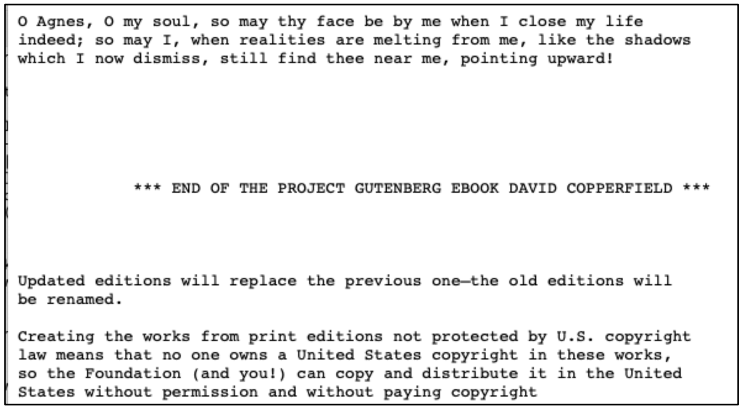
```
The Project Gutenberg eBook of David Copperfield

This ebook is for the use of anyone anywhere in the United States and
most other parts of the world at no cost and with almost no restrictions
whatsoever. You may copy it, give it away or re-use it under the terms
of the Project Gutenberg License included with this ebook or online
at www.gutenberg.org. If you are not located in the United States,
you will have to check the laws of the country where you are located
before using this eBook.

Title: David Copperfield


Author: Charles Dickens

Release date: December 1, 1996 [eBook #766]
                Most recently updated: October 25, 2022
```



```
O Agnes, O my soul, so may thy face be by me when I close my life
indeed; so may I, when realities are melting from me, like the shadows
which I now dismiss, still find thee near me, pointing upward!







            *** END OF THE PROJECT GUTENBERG EBOOK DAVID COPPERFIELD ***




Updated editions will replace the previous one—the old editions will
be renamed.

Creating the works from print editions not protected by U.S. copyright
law means that no one owns a United States copyright in these works,
so the Foundation (and you!) can copy and distribute it in the United
States without permission and without paying copyright
```

Homework 7

- Part 1 details:
  - Step 3: put them in the right directory, start `python`. Read in the raw files and `nltk.word_tokenize()` them.
  - Step 4: remove the punctuation, see conditional list comprehension in Lecture 17 using condition:
    - `any(c.isalpha() for c in word)`
  - Step 5: slice the corpus into 5,000 word chunks
    - `[0:5000]` and `[5000:10000]`
  - Step 6: use a list comprehension to grab the word lengths
    - `len = [len(word) for word in chunk]`
  - Step 7: histogram plot them with overlay
    - `plt.hist(len, range(1,mx+1), histtype='step', label='1st 5000')`

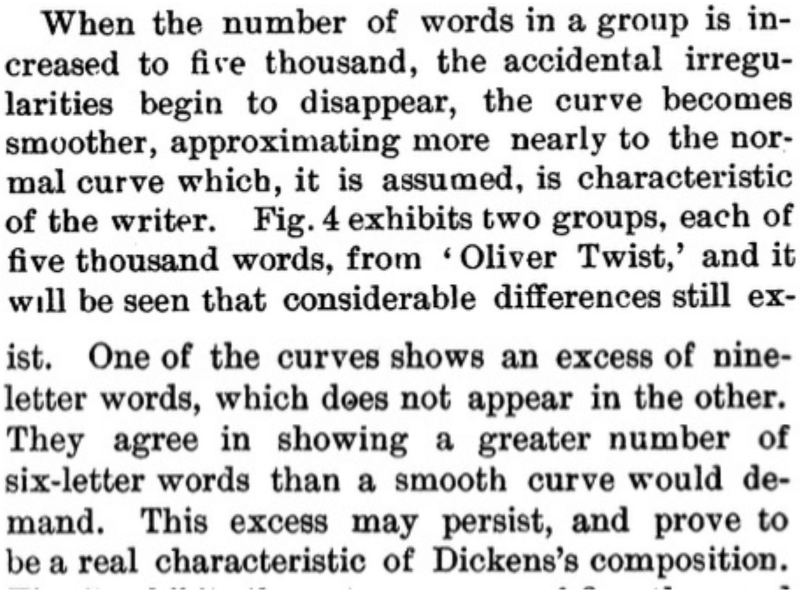
Homework 7

Part 1:

- Let's compare our two 5,000 word Mendenhall test for *Oliver Twist* (1838) with *Nicholas Nickleby* (1839) and *David Copperfield* (1850).
- Submit your histograms and python code
- What do you think? E.g.
  - Do you think they are comparable?
  - Or are there significant differences?
  - Do you think it's reasonable to think they are written by the same author?

Homework 7

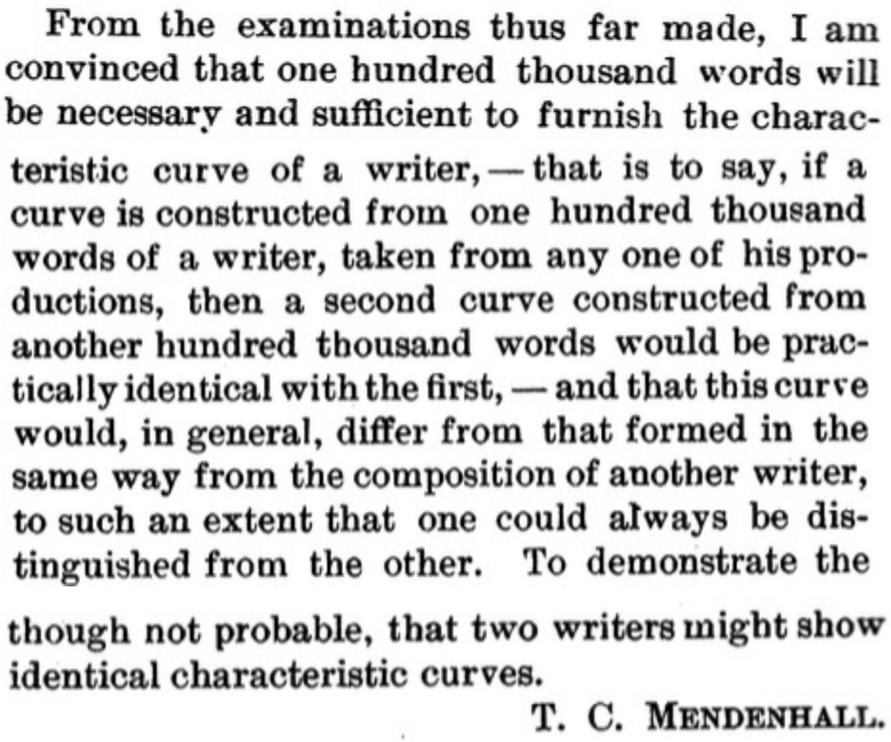
- Part 2:
  - Based on your three-way comparison, what do you think about Mendenhall's claim about six-letter words for Charles Dickens? Is it justified? Explain.



When the number of words in a group is increased to five thousand, the accidental irregularities begin to disappear, the curve becomes smoother, approximating more nearly to the normal curve which, it is assumed, is characteristic of the writer. Fig. 4 exhibits two groups, each of five thousand words, from 'Oliver Twist,' and it will be seen that considerable differences still exist. One of the curves shows an excess of nine-letter words, which does not appear in the other. They agree in showing a greater number of six-letter words than a smooth curve would demand. This excess may persist, and prove to be a real characteristic of Dickens's composition.

Homework 7

- Part 3:
  - Now take (slice) the first 100,000 words for each of *Oliver Twist* (1838) with *Nicholas Nickleby* (1839) and *David Copperfield* (1850).
  - Plot them over one another.
  - Submit your histogram and code.
  - What do you think of Mendenhall 100,000 word claim?



From the examinations thus far made, I am convinced that one hundred thousand words will be necessary and sufficient to furnish the characteristic curve of a writer, — that is to say, if a curve is constructed from one hundred thousand words of a writer, taken from any one of his productions, then a second curve constructed from another hundred thousand words would be practically identical with the first, — and that this curve would, in general, differ from that formed in the same way from the composition of another writer, to such an extent that one could always be distinguished from the other. To demonstrate the

though not probable, that two writers might show identical characteristic curves.

T. C. Mendenhall.

Homework 7

- One PDF file!
- Submit to sandiway@arizona.edu
- SUBJECT: 388 Homework 7 *YOUR NAME*
- One PDF file only
  - include Python terminal and histogram screenshots in your answer
- Deadline:
  - midnight Monday
  - we will review the homework on Tuesday