Lecture 30

408/508 *Computational Techniques for Linguists*

Today's Topics

- Last Lecture today!
- Reminder: term project, email me end of next week
- Reminder: course survey please! Response rate: 44.44%
- Stylometry: (Mendenhall 1887)

Stylometry

- The basic idea:
  - looking for commonalities between works using statistics on (easy-to-retrieve) **stylistic features**.
- **Adversarial stylometry**:
  - hiding authorship by alterations, or perhaps by using ChatGPT?
- *Would have been a good topic for a term project also*

Stylometry

Who wrote *Wuthering Heights*?

Rachel McCarthy and James O'Sullivan
Digital Humanities, University College Cork, Ireland

Abstract

Emily Brontë published *Wuthering Heights* in 1847 under the pseudonym Ellis Bell. It was not until the later second edition, published after Emily's death, that she was credited as the novel's author. Those Victorian attitudes towards women which compelled Brontë to publish as Bell have not been wholly eradicated, with her legitimacy as the sole author being called into question by male commentators at several junctures since. Their claim is that Emily's brother Branwell is the real author of *Wuthering Heights*. Using stylometry, a computer-assisted technique which meas-

nce:

Digital Scholarship in the Humanities, Volume 36, Issue 2, June 2021, Pages 383–391

Stylometry

- p384:
  - Stylometry is a statistical technique which indicates likely authorship, forming an ‘impression’ of how a particular author writes by counting the frequency of words across sample texts. While the specific techniques differ across the iterative stages of this study, the analysis is always conducted using the 100 most frequent words[2] from the chosen samples, with the similarity between styles measured using Support Vector Machine (SVM) classification, **Burrows’ Delta** and Cosine Delta.

2. The authors […] have consistently used no more than 100 most frequent words because they subscribe to the theoretical view that results become less indicative of authorial fingerprint as the number of features is increased.

`nltk.corpus.stopwords`

When stylometry is conducted using a small sample of high-frequency words, typically function words, the analysis is conducted using words, which are ‘especially resistant to intentional authorial manipulation’ (Hoover, 2009, p. 35), and thus suited to determining subconscious authorial fingerprints rather than content distinct to the particular narrative.

Stylometry

Delta: *a Measure of Stylistic Difference and a Guide to Likely Authorship*. (Burrows 2002)

z-score = (x – mean) / stdev

270 Literary and Linguistic Computing, Vol. 17, No. 3, 2002

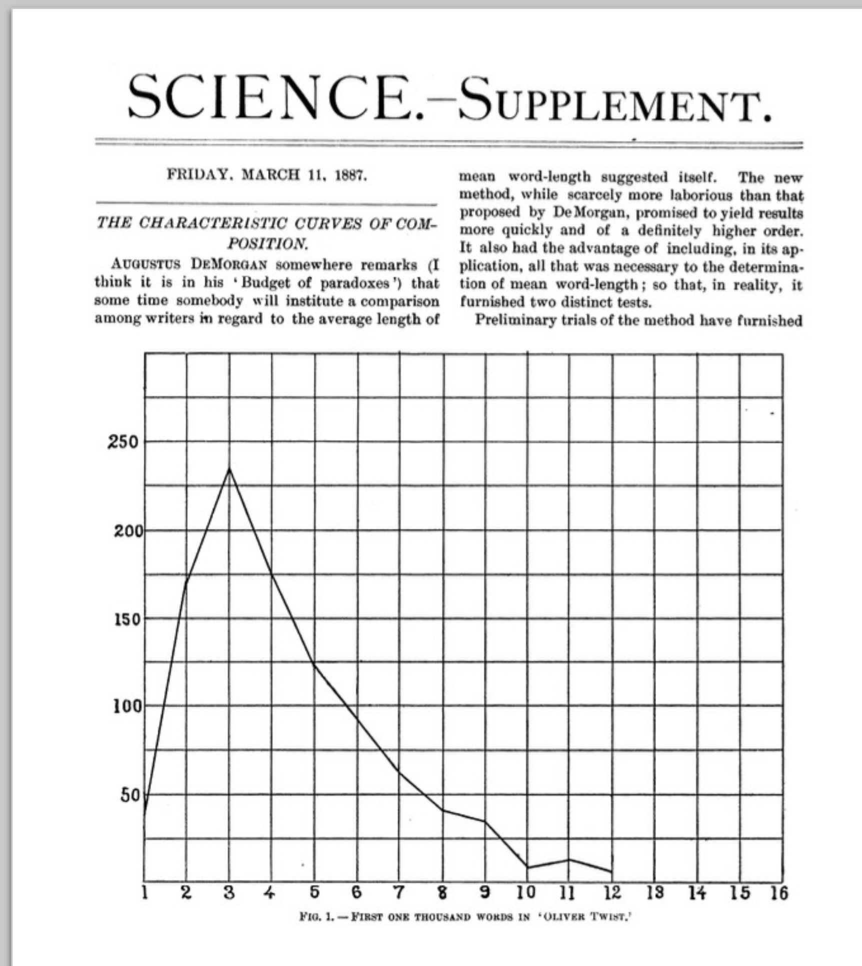
Table 1 Specimen of procedure

	A	B	C	D	F	G	I	J	K	L
1			Main set		Milton		*Paradise Lost*			
2							count			30
3							sum			31.489
4							mean (= "delta")			1.050
5							stdev			0.770
6			Mean	Stdev	Scores	z-scores	Scores	z-scores	Diff.	Abs. diff.
7	1	the	4.242	0.630	4.719	0.757	4.091	–0.239	–0.996	0.996
8	2	and	3.770	0.501	4.407	1.272	4.165	0.789	–0.483	0.483
9	3	of	1.821	0.315	2.420	1.905	2.769	3.015	1.110	1.110
10	4	a	1.601	0.430	0.893	–1.645	0.696	–2.103	–0.458	0.458
11	5	to(i)	1.419	0.272	1.247	–0.634	1.289	–0.480	0.154	0.154
12	6	in(p)	1.358	0.189	1.554	1.035	1.720	1.916	0.881	0.881
13	7	his	1.154	0.323	1.062	–0.284	1.532	1.171	1.454	1.454
14	8	with	1.022	0.208	1.480	2.202	1.484	2.224	0.022	0.022
15	9	to(p)	1.014	0.131	0.999	–0.119	1.245	1.761	1.880	1.880
16	10	is	0.938	0.312	0.502	–1.397	0.239	–2.238	–0.841	0.841
17	11	but	0.923	0.195	0.676	–1.268	0.696	–1.167	0.101	0.101
18	12	he	0.803	0.241	0.465	–1.403	0.703	–0.413	0.990	0.990
19	13	all	0.781	0.193	0.518	–1.366	0.836	0.283	1.649	1.649
20	14	I	0.766	0.391	0.882	0.297	0.700	–0.171	–0.467	0.467
21	15	it	0.766	0.239	0.386	–1.591	0.151	–2.575	–0.984	0.984
22	16	as	0.710	0.224	0.618	–0.410	0.737	0.119	0.529	0.529
23	17	their	0.641	0.237	0.513	–0.540	0.795	0.653	1.193	1.193
24	18	her	0.623	0.336	0.851	0.678	0.435	–0.560	–1.237	1.237
25	19	not	0.616	0.174	0.592	–0.138	0.847	1.324	1.462	1.462
26	20	be	0.586	0.167	0.555	–0.187	0.401	–1.109	–0.921	0.921
27	21	you	0.580	0.252	0.174	–1.608	0.037	–2.154	–0.546	0.546
28	22	they	0.564	0.234	0.270	–1.259	0.464	–0.428	0.830	0.830
29	23	for(p)	0.559	0.114	0.270	–2.539	0.000	–4.905	–2.366	2.366
30	24	by(p)	0.555	0.106	0.412	–1.349	0.689	1.260	2.608	2.608
31	25	my	0.512	0.370	0.587	0.201	0.258	–0.687	–0.888	0.888
32	26	we	0.510	0.275	0.159	–1.279	0.265	–0.891	0.388	0.388
33	27	from	0.500	0.127	0.534	0.265	0.884	3.019	2.754	2.754
34	28	that(rp)	0.476	0.228	0.925	1.964	0.313	–0.715	–2.680	2.680
35	29	or	0.471	0.165	0.856	2.333	0.906	2.636	0.302	0.302
36	30	our	0.460	0.268	0.270	–0.711	0.354	–0.397	0.314	0.314

	A	B	N	O	P	Q	S	T	U	V	W	X	Y	Z
1			*World's Infancy*				*Paradise Regained*				*Samson Agonistes*			
2			count			30	count			30	count			30
3			sum			36.164	sum			32.247	sum			33.814
4			mean (= "delta")			1.205	mean (= "delta")			1.075	mean (= "delta")			1.127
5			stdev			1.163	stdev			1.227	stdev			1.087
6			Scores	z-scores	Diff.	Abs. diff.	Scores	z-scores	Diff.	Abs. diff.	Scores	z-scores	Diff.	Abs. diff.
7	1	the	7.866	5.753	4.996	4.996	3.619	–0.988	–1.746	1.746	2.809	–2.274	–3.031	3.031
8	2	and	3.474	–0.590	–1.862	1.862	4.441	1.340	0.068	0.068	3.298	–0.940	–2.212	2.212
9	3	of	2.169	1.106	–0.799	0.799	2.765	3.002	1.097	1.097	2.561	2.353	0.448	0.448
10	4	a	1.296	–0.708	0.936	0.936	0.873	–1.691	–0.047	0.047	1.094	–1.177	0.468	0.468
11	5	to(i)	0.918	–1.846	–1.212	1.212	1.389	–0.111	0.523	0.523	1.824	1.491	2.124	2.124
12	6	in(p)	1.476	0.624	–0.411	0.411	1.536	0.940	–0.095	0.095	1.552	1.028	–0.007	0.007
13	7	his	1.359	0.635	0.919	0.919	1.287	0.413	0.696	0.696	1.009	–0.448	–0.165	0.165
14	8	with	0.972	–0.239	–2.441	2.441	1.141	0.572	–1.630	1.630	1.436	1.991	–0.211	0.211
15	9	to(p)	0.819	–1.493	–1.373	1.373	1.663	4.957	5.077	5.077	1.428	3.161	3.281	3.281
16	10	is	1.233	0.944	2.341	2.341	0.465	–1.515	–0.118	0.118	0.442	–1.588	–0.191	0.191
17	11	but	0.378	–2.801	–1.533	1.533	0.765	–0.814	0.453	0.453	0.916	–0.038	1.230	1.230
18	12	he	0.603	–0.830	0.573	0.573	0.784	–0.079	1.324	1.324	0.435	–1.529	–0.126	0.126
19	13	all	0.720	–0.318	1.048	1.048	0.975	1.003	2.369	2.369	0.830	0.254	1.620	1.620
20	14	I	0.711	–0.142	–0.438	0.438	1.198	1.103	0.806	0.806	1.676	2.326	2.030	2.030
21	15	it	0.558	–0.870	0.722	0.722	0.299	–1.953	–0.361	0.361	0.450	–1.322	0.270	0.270
22	16	as	0.540	–0.760	–0.350	0.350	0.701	–0.041	0.369	0.369	0.722	0.053	0.463	0.463
23	17	their	0.432	–0.880	–0.340	0.340	0.522	–0.498	0.042	0.042	0.761	0.506	1.046	1.046
24	18	her	0.756	0.396	–0.282	0.282	0.312	–0.923	–1.601	1.601	0.287	–0.998	–1.675	1.675
25	19	not	0.432	–1.054	–0.916	0.916	0.841	1.290	1.428	1.428	1.180	3.231	3.369	3.369
26	20	be	0.459	–0.763	–0.576	0.576	0.503	–0.496	–0.309	0.309	0.520	–0.397	–0.209	0.209
27	21	you	0.261	–1.265	0.344	0.344	0.006	–2.275	–0.666	0.666	0.023	–2.208	–0.599	0.599
28	22	they	0.396	–0.719	0.540	0.540	0.370	–0.831	0.427	0.427	0.310	–1.084	0.175	0.175
29	23	for(p)	0.342	–1.903	0.637	0.637	0.280	–2.444	0.095	0.095	0.466	–0.817	1.722	1.722
30	24	by(p)	0.432	–1.162	0.187	0.187	0.822	2.518	3.866	3.866	0.582	0.254	1.603	1.603
31	25	my	0.351	–0.435	–0.636	0.636	0.472	–0.110	–0.311	0.311	1.226	1.928	1.727	1.727
32	26	we	0.468	–0.153	1.126	1.126	0.127	–1.392	–0.113	0.113	0.124	–1.404	–0.125	0.125
33	27	from	0.567	0.527	0.262	0.262	0.771	2.132	1.866	1.866	0.520	0.157	–0.108	0.108
34	28	that(rp)	0.234	–1.061	–3.026	3.026	0.172	–1.333	–3.297	3.297	0.217	–1.135	–3.099	3.099
35	29	or	0.153	–1.929	–4.263	4.263	1.064	3.595	1.261	1.261	0.908	2.648	0.315	0.315
36	30	our	0.558	0.366	1.078	1.078	0.319	–0.528	0.183	0.183	0.225	–0.877	–0.166	0.166

Stylometry

- Course website:
  - Mendenhall1887.pdf
  - please read
- Idea:
  - average length of words a guide to authorship
  - easy to compute today with nltk
  - laborious back in 1887



SCIENCE.—SUPPLEMENT.

FRIDAY, MARCH 11, 1887.

THE CHARACTERISTIC CURVES OF COMPOSITION.

AUGUSTUS DEMORGAN somewhere remarks (I think it is in his 'Budget of paradoxes') that some time somebody will institute a comparison among writers in regard to the average length of mean word-length suggested itself. The new method, while scarcely more laborious than that proposed by De Morgan, promised to yield results more quickly and of a definitely higher order. It also had the advantage of including, in its application, all that was necessary to the determination of mean word-length; so that, in reality, it furnished two distinct tests.

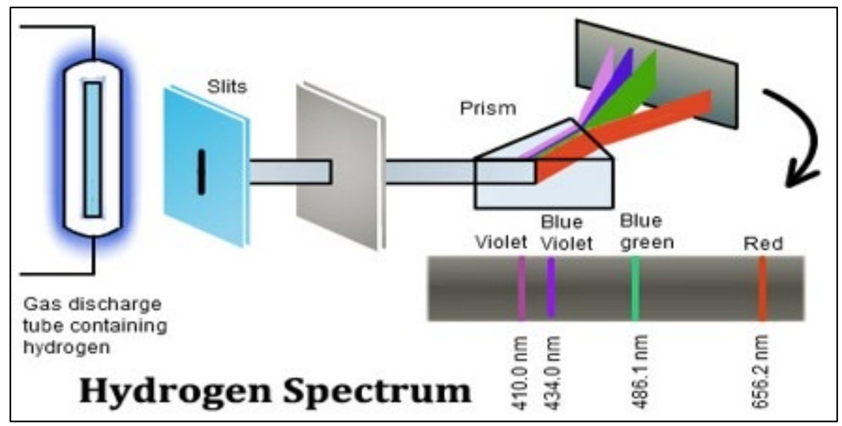
Preliminary trials of the method have furnished

FIG. 1.—FIRST ONE THOUSAND WORDS IN 'OLIVER TWIST.'

Stylometry

By the use of the spectroscope, a beam of non-homogeneous light is analyzed, and its components assorted according to their wave-length. As is well known, each element, when intensely heated under proper conditions, sends forth light which, upon prismatic analysis, is found to consist of groups of waves of definite length, and appearing in certain definite proportions. So certain and uniform are the results of this analysis, that the appearance of a particular spectrum is indisputable evidence of the presence of the element to which it belongs.

- An appeal to physics/science:



Stylometry

In a manner very similar, it is proposed to analyze a composition by forming what may be called a 'word-spectrum,' or 'characteristic curve,' which shall be a graphic representation of an arrangement of words according to their length and to the relative frequency of their occurrence. If, now, it shall be found that with every author, as with every element, this spectrum persists in its form and appearance, the value of the method will be at once conceded. It

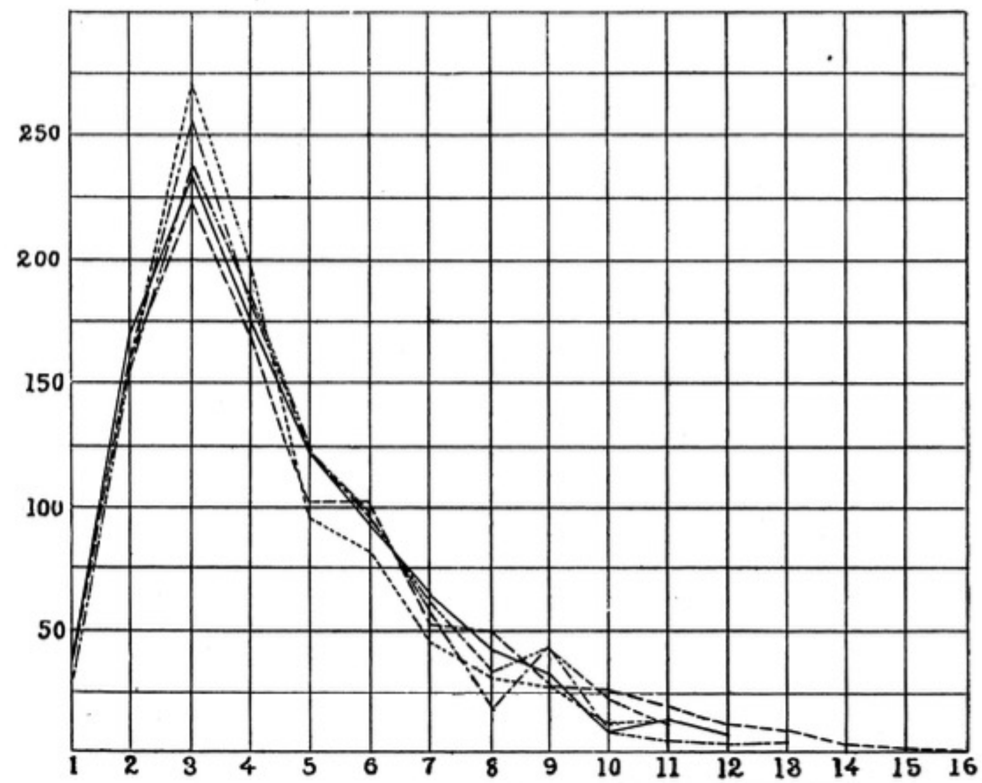


FIG. 2.—SHOWING FIVE GROUPS, OF ONE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'

Stylometry

a single mean word length statistic is not enough

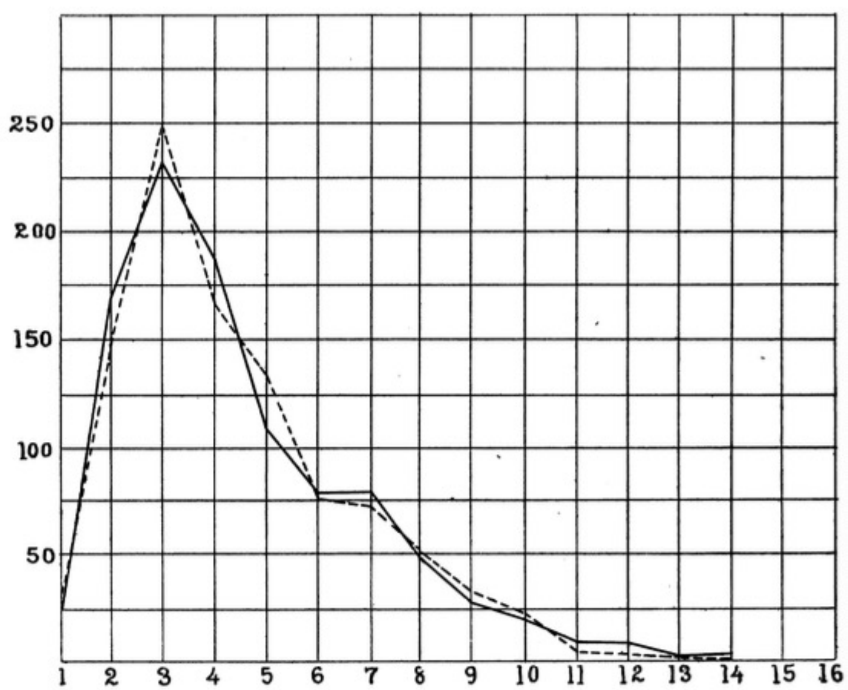


FIG. 3.—TWO CONSECUTIVE GROUPS, OF ONE THOUSAND WORDS EACH, FROM 'VANITY FAIR.' THESE GROUPS SHOW SENSIBLY THE SAME AVERAGE WORD-LENGTHS.

Letters.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Words in 1st group	25	169	232	187	109	78	79	48	28	20	10	10	2	3
Words in 2d group	33	146	248	164	135	76	73	52	35	23	6	5	2	2

It will be seen that the total number of letters in the first group is 4,507, and in the second 4,508, or an average of 4.507 and 4.508 letters to each word in the respective groups. If this average, or 'mean word-length,' be alone considered, the two groups must be regarded as sensibly identical; but an inspection of the diagram shows that they are in reality quite different.

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

Stylometry

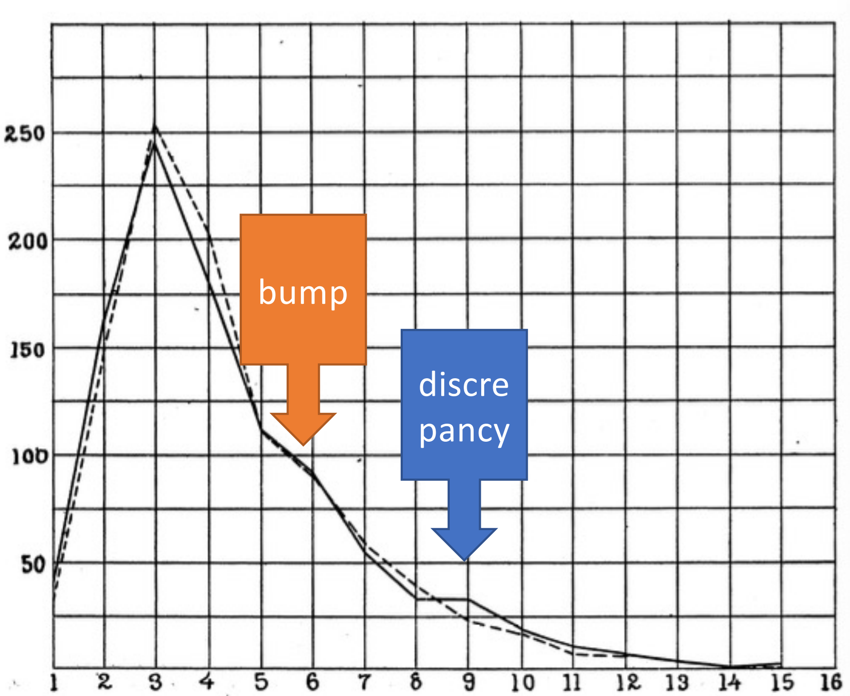


FIG. 4.—TWO GROUPS, OF FIVE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'

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.

"smooth" meaning *montonically decreasing*

Stylometry

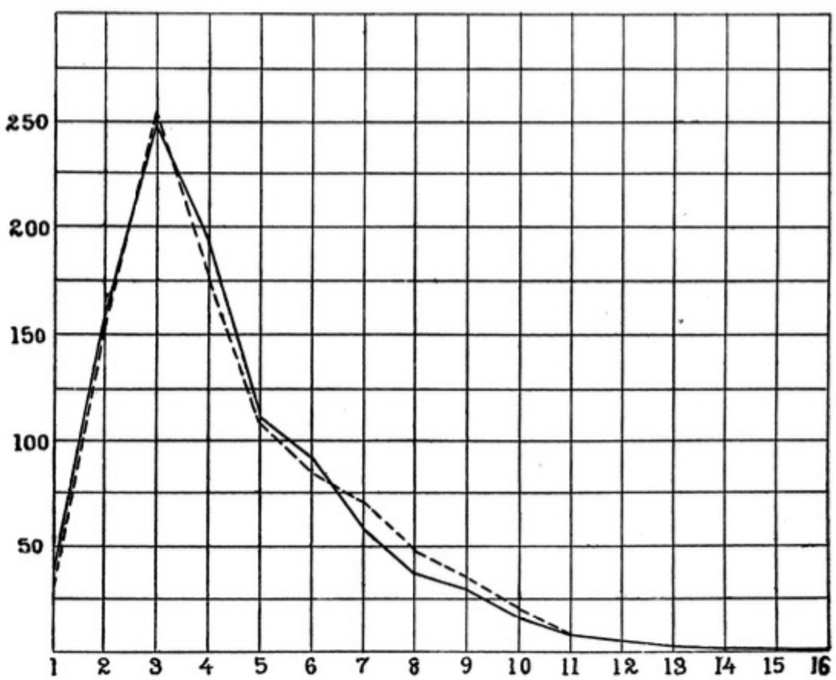


FIG. 7.—TWO GROUPS, OF TEN THOUSAND WORDS EACH, FROM 'OLIVER TWIST,' ——; AND FROM 'VANITY FAIR,' — — — —.

fig. 7, two groups of ten thousand each, from 'Oliver Twist' and 'Vanity fair,' are placed side by side for comparison, the former being represented by the continuous line, and the latter by the broken line. Although these curves differ, and while it is believed that the difference will persist with an increased number of words, it is certainly surprising, that in the analysis of ten thousand words from Dickens, and the same number from Thackeray, so close an agreement should be found. This agreement is particularly striking in words of eleven, twelve, and thirteen letters, the numerical comparison of which is as follows :—

Number of letters....................	11	12	13
Number of words in Dickens.......	85	57	29
Number of words in Thackeray....	85	58	29

This closeness to identity must be largely the result of accident, and it would not be likely to repeat itself in another analysis.

Stylometry

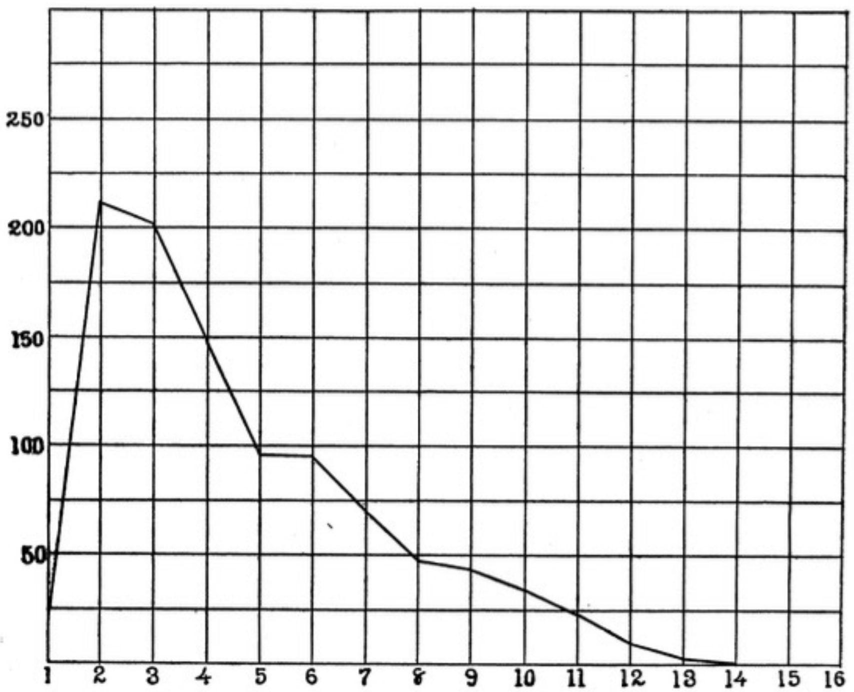


FIG. 8. — CURVE OF FIVE THOUSAND WORDS FROM MILL'S 'POLITICAL ECONOMY.'

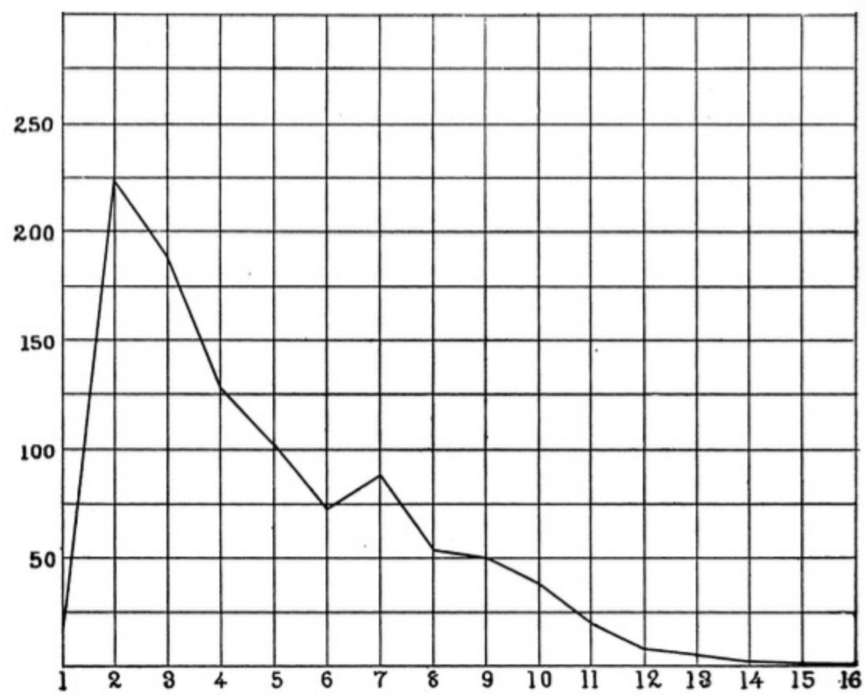


FIG. 9. — CURVE OF FIVE THOUSAND WORDS FROM MILL'S 'ESSAY ON LIBERTY.'

Stylometry

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

Stylometry

- On the course website:
  - 53 chapters, no chapter headings, no titles etc.
  - `oliver_twist0-53.txt`
- Python:

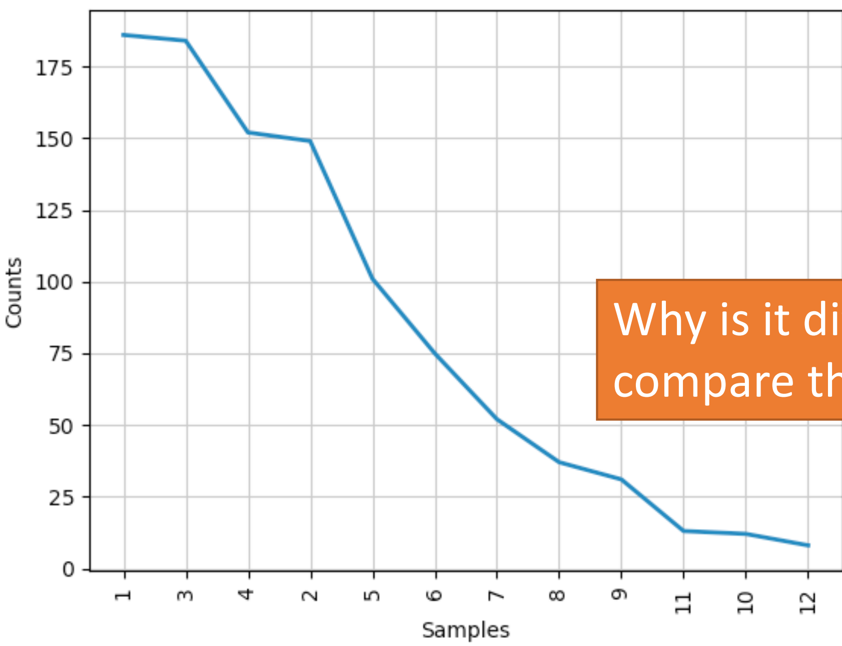
```
>>> raw = open('oliver_twist0-53.txt', encoding='utf-8', errors='ignore').read()
>>> len(raw)
882296
>>> import nltk
>>> words = nltk.word_tokenize(raw)
>>> len(words)
197947
>>> vocab = set(words)
>>> len(vocab)
12379
```

Stylometry

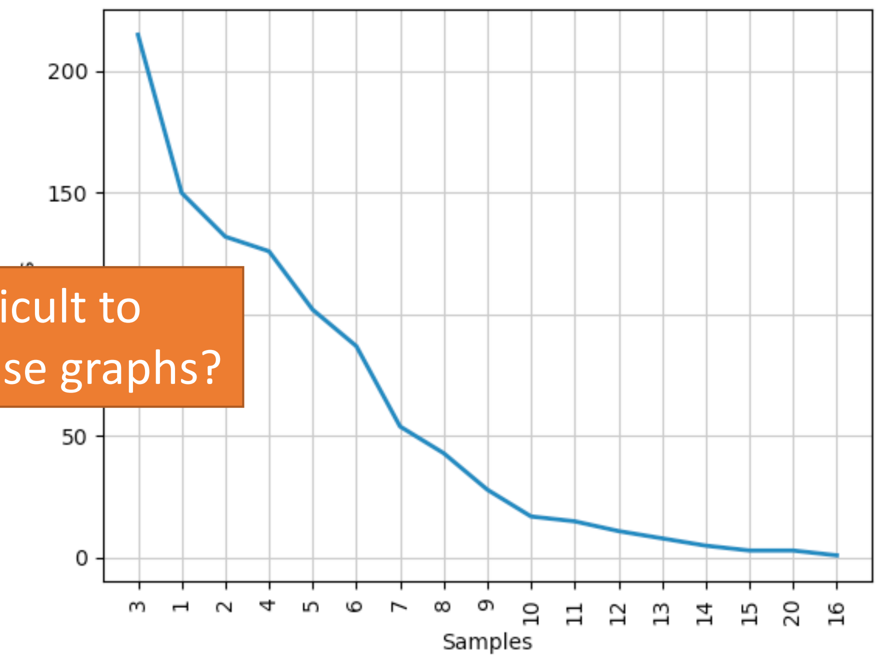
- Let's take the (*unmodified*) text a thousand words at a time:
  - `words1 = words[0:1000]`
  - `words2 = words[1000:2000]`
  - *etc.*
- Mendenhall's *word-spectrum* based on word length:
  - `len1 = [len(word) for word in words[0:1000]]`
  - `len2 = [len(word) for word in words[1000:2000]]`
- Frequency distribution of the *word-spectrum*:
  - `fd1 = nltk.FreqDist(len1)`
  - `fd2 = nltk.FreqDist(len2)`

Stylometry

`fd1.plot()`



`fd2.plot()`



Why is it difficult to compare these graphs?

Stylometry

```
>>> fd1
FreqDist({1: 186, 3: 184, 4: 152, 2: 149, 5: 101, 6: 75, 7: 52, 8:
37, 9: 31, 11: 13, ...})
>>> fd2
FreqDist({3: 215, 1: 150, 2: 132, 4: 126, 5: 102, 6: 87, 7: 54, 8:
43, 9: 28, 10: 17, ...})
>>> max(fd1)
12
>>> max(fd2)
20
```

histogram plot

- https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html

matplotlib.pyplot.hist

```
matplotlib.pyplot.hist(x, bins=None, range=None, density=False, weights=None, cumulative=False, bottom=None, histtype='bar', align='mid', orientation='vertical', rwidth=None, log=False, color=None, label=None, stacked=False, *, data=None, **kwargs)
```

[source]

Compute and plot a histogram.

This method uses `numpy.histogram` to bin the data in *x* and count the number of values in each bin, then draws the distribution either as a `BarContainer` or `Polygon`. The *bins*, *range*, *density*, and *weights* parameters are forwarded to `numpy.histogram`.

histogram plot

data columns

`matplotlib.pyplot.hist(x, bins)`

Parameters:

x : ***(n,) array or sequence of (n,) arrays***

Input values, this takes either a single array or a sequence of arrays which are not required to be of the same length.

bins : ***int or sequence or str, default:*** `rcParams["hist.bins"]` ***(default:*** `10`***)***

If *bins* is an integer, it defines the number of equal-width bins in the range.

If *bins* is a sequence, it defines the bin edges, including the left edge of the first bin and the right edge of the last bin; in this case, bins may be unequally spaced. All but the last (righthand-most) bin is half-open. In other words, if *bins* is:

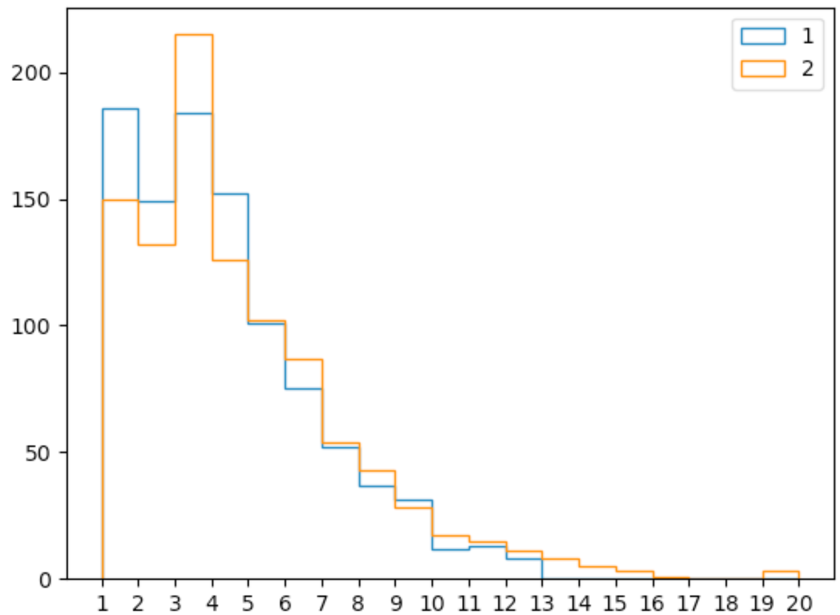
```
[1, 2, 3, 4]
```

then the first bin is `[1, 2)` (including 1, but excluding 2) and the second `[2, 3)`. The last bin, however, is `[3, 4]`, which *includes* 4.

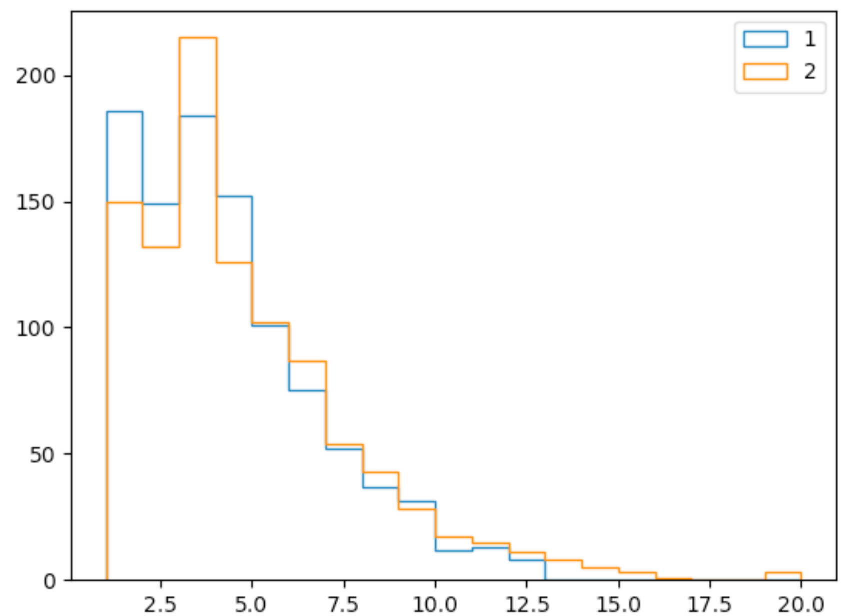
histogram plot

- Let's use matplotlib.pyplot to plot them together.
  - `import matplotlib.pyplot as plt`
  - `mx = max(max(len2),max(len1))`
  - `mx` (*longest lword across the two chunks*)
  - `20`
  - `plt.hist(len1, range(1,mx+1), histtype='step', label='1')`
  - `plt.hist(len2, range(1,mx+1), histtype='step', label='2')`
  - `plt.xticks(range(1,mx+1))`
  - `plt.legend()`
  - `plt.show()`

histogram plot



`plt.xticks(range(1,mx+1))`



without specifying xticks()

histogram plot

- What about punctuation - what does the graph look like without punctuation?
    - we can extract the 1-letter long vocab as follows:

```
>> a = [word for word in words if len(word) == 1]
>>> set(a)
{'D', ']', '?', ';', 't', 'l', '(', 'e', 'T', 'r', 'o', '.', 'P', ':',
'p', '‘', 'n', ')', '_', '“', 'd', 's', 'I', '!', ',', ''', 'a', '[',
'”', 'A', 'b', 'S', 'm'}
```

- we can remove them from the corpus as follows:

```
>>> words2 = [word for word in words if not (len(word) == 1 and not word.isalnum())]
>>> len(words2)
158950
>>> len(words)
197947
```

> `str.isalnum()`
>
> Return `True` if all characters in the string are alphanumeric and there is at least one character, `False` otherwise. A character `c` is alphanumeric if one of the following returns `True`: `c.isalpha()`, `c.isdecimal()`, `c.isdigit()`, or `c.isnumeric()`.

histogram plot

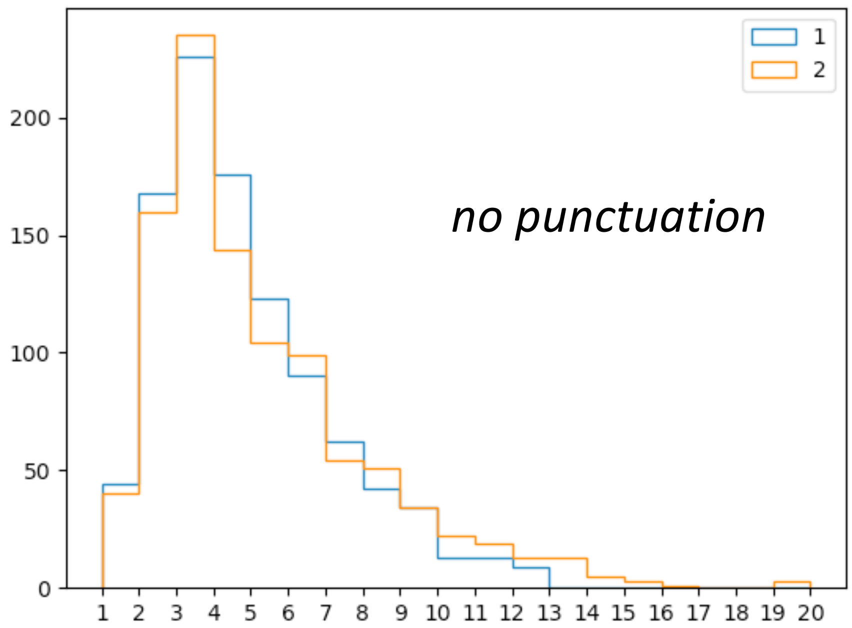
- Check to see whether we've filtered out the punctuation:

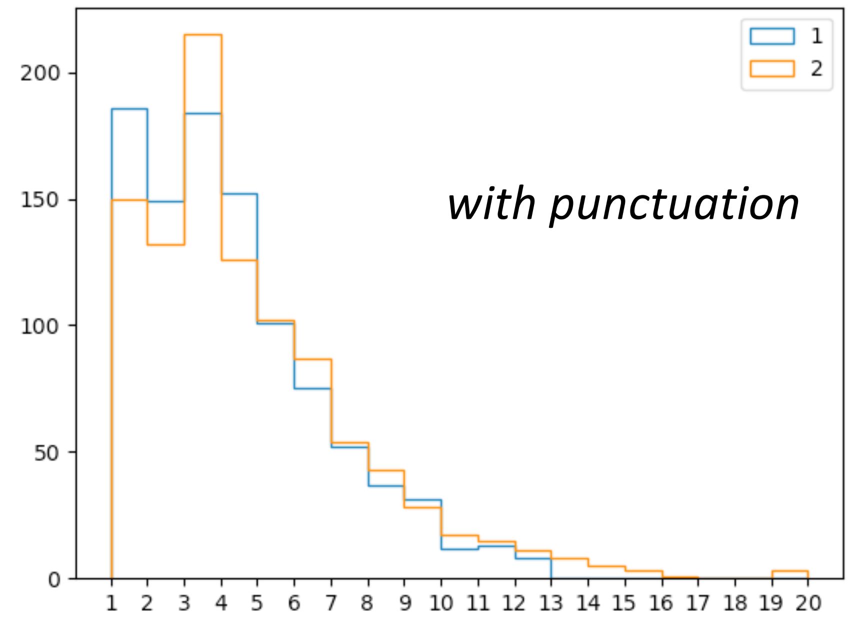
```
>>> a = [word for word in words2 if len(word) == 1]
>>> set(a)
{'p', 'I', 'D', 'P', 't', 'a', 'n', 'A', 'l', 'e', 'd', 'T', 'r',
'o', 'b', 'S', 's', 'm'}
```

- cf. `in words`

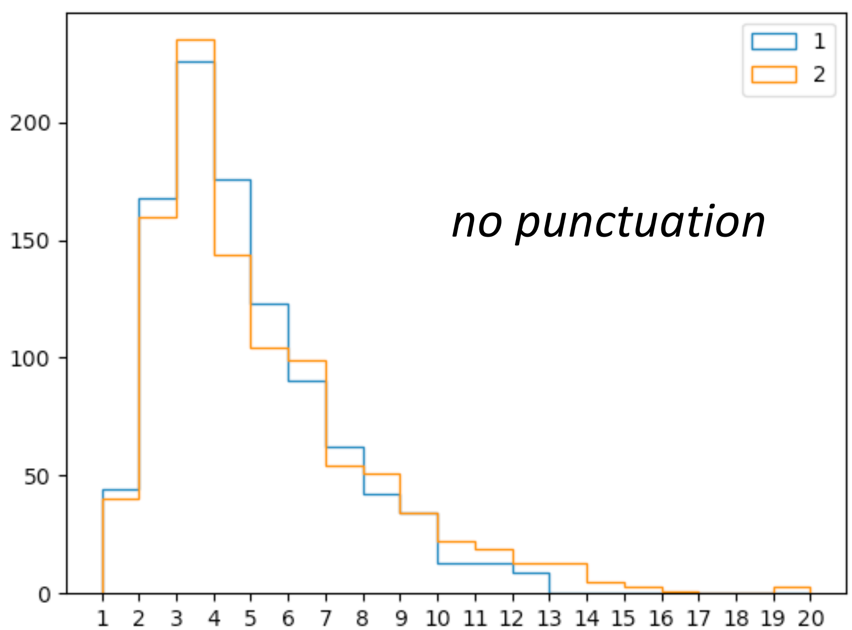
```
{'D', ']', '?', ';', 't', 'l', '(', 'e', 'T', 'r', 'o', '.', 'P',
':', 'p', '‘', 'n', ')', '—', '“', 'd', 's', 'I', '!', ',', '’',
'a', '[', '”', 'A', 'b', 'S', 'm'}
```

histogram plot





histogram plot



first two 1,000 word chunks, i.e. groups, from *Oliver Twist*

Mendenhall

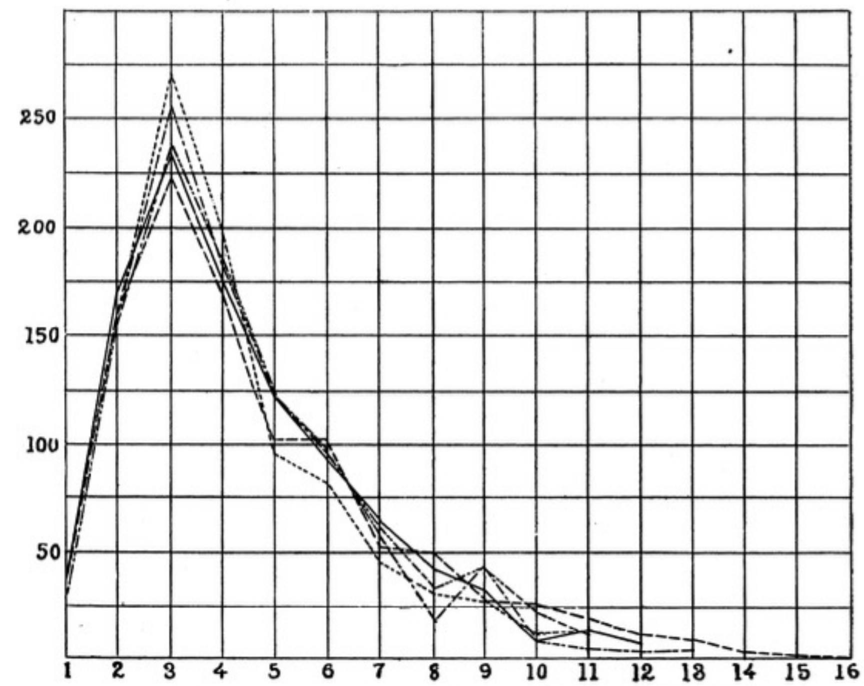


FIG. 2.—SHOWING FIVE GROUPS, OF ONE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'