Information Visualization in IR

The next step beyond text-based interfaces

- Visualizing Internal Document Substructure
  - Show where the content of interest resides
    - If only a paragraph on page 37 is relevant, then we want to pinpoint this.
  - Distribution of relevant information in a document
    - Aid to selecting documents
    - TextTiling/TileBars
    - DotPlot
- Visualizing Document/WebPage Clusters and Relationships
  - Graphical Elements of Information Browsing Systems

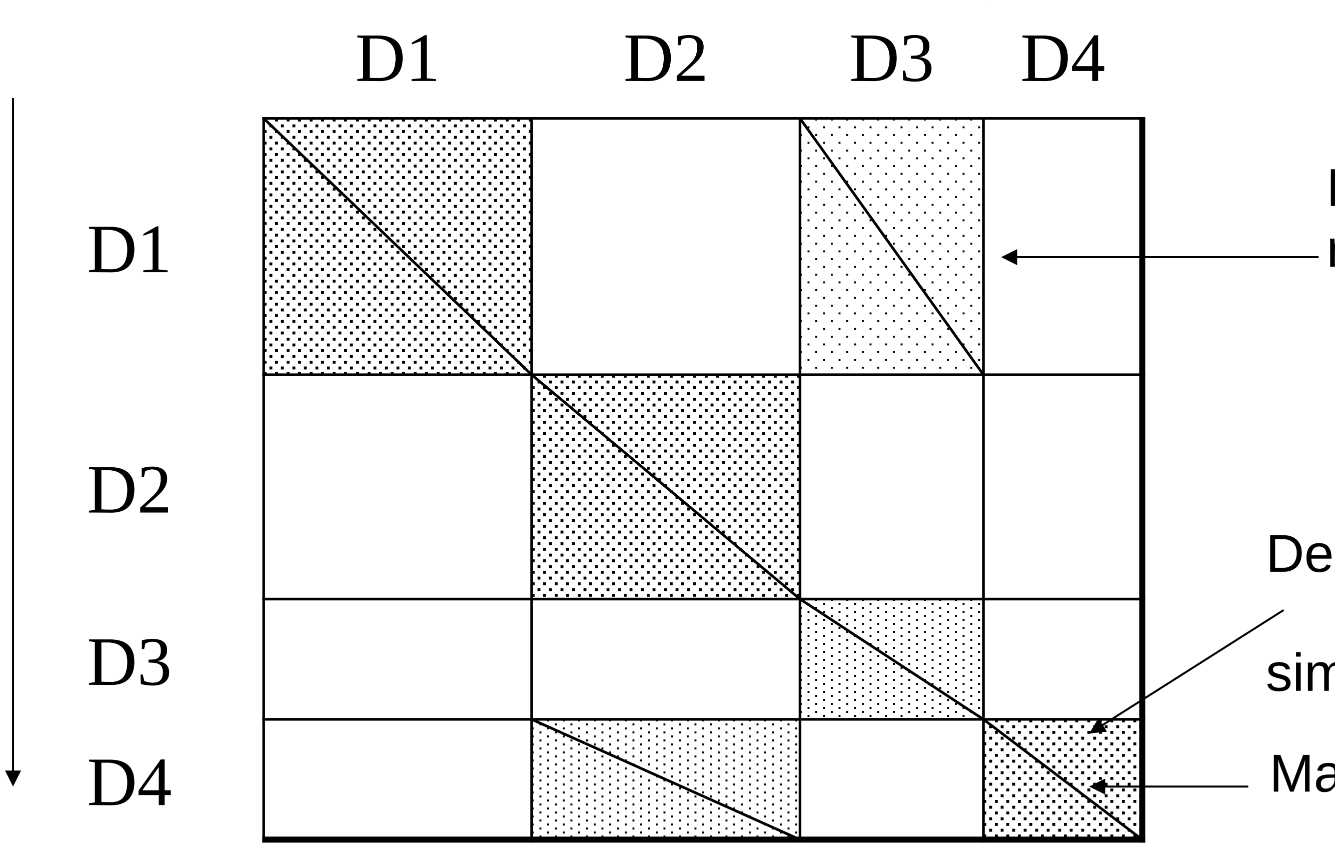
DOT PLOT (addl points)

To Show:

- Repetition
- Regions of self similarity (same **TOPIC** blocks)

Plot

Increasing corpus position

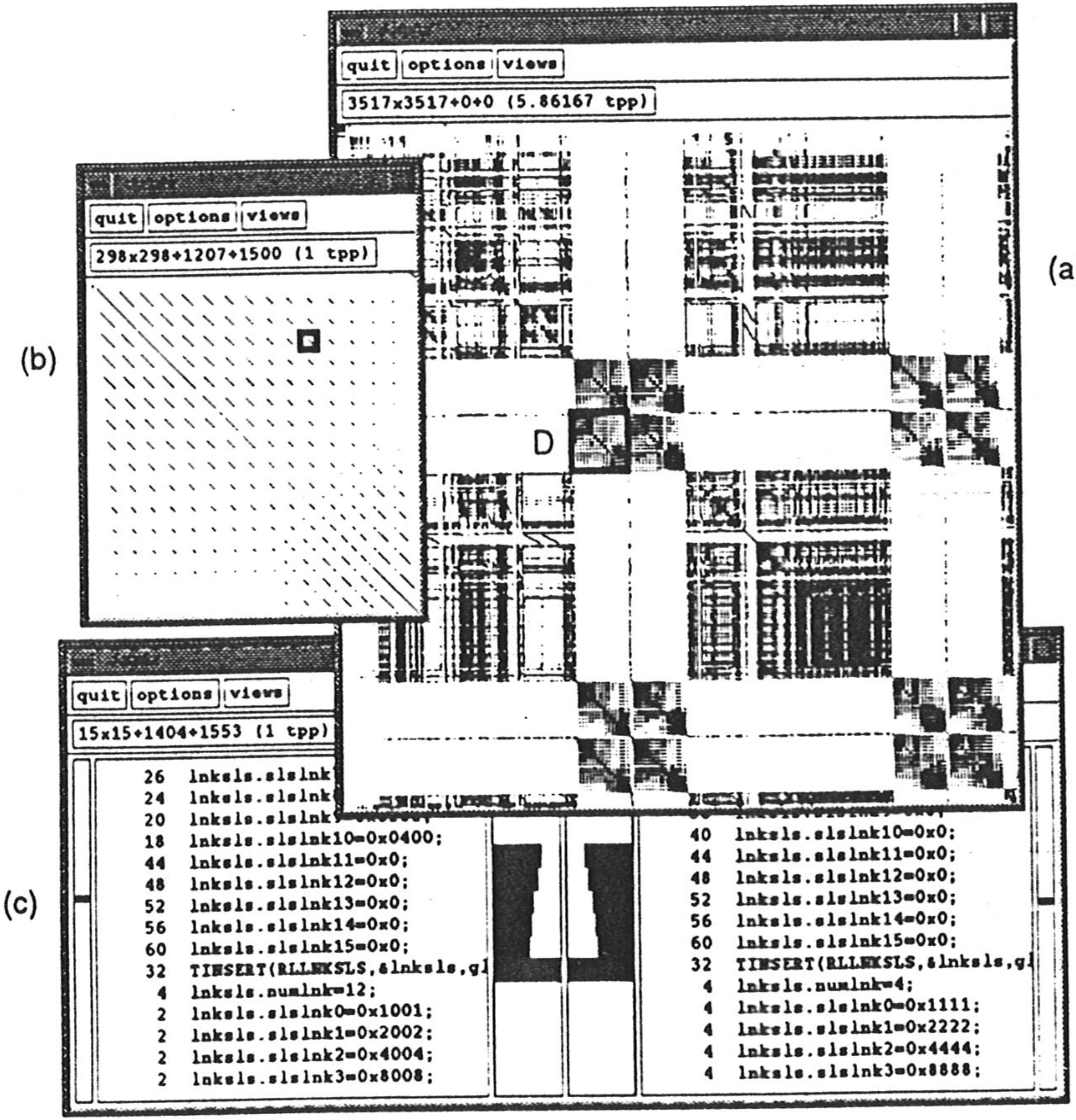


Increasing Corpus Position

Diagonal = repetition

Dense Blocks = Self similarity

Main Diagonal

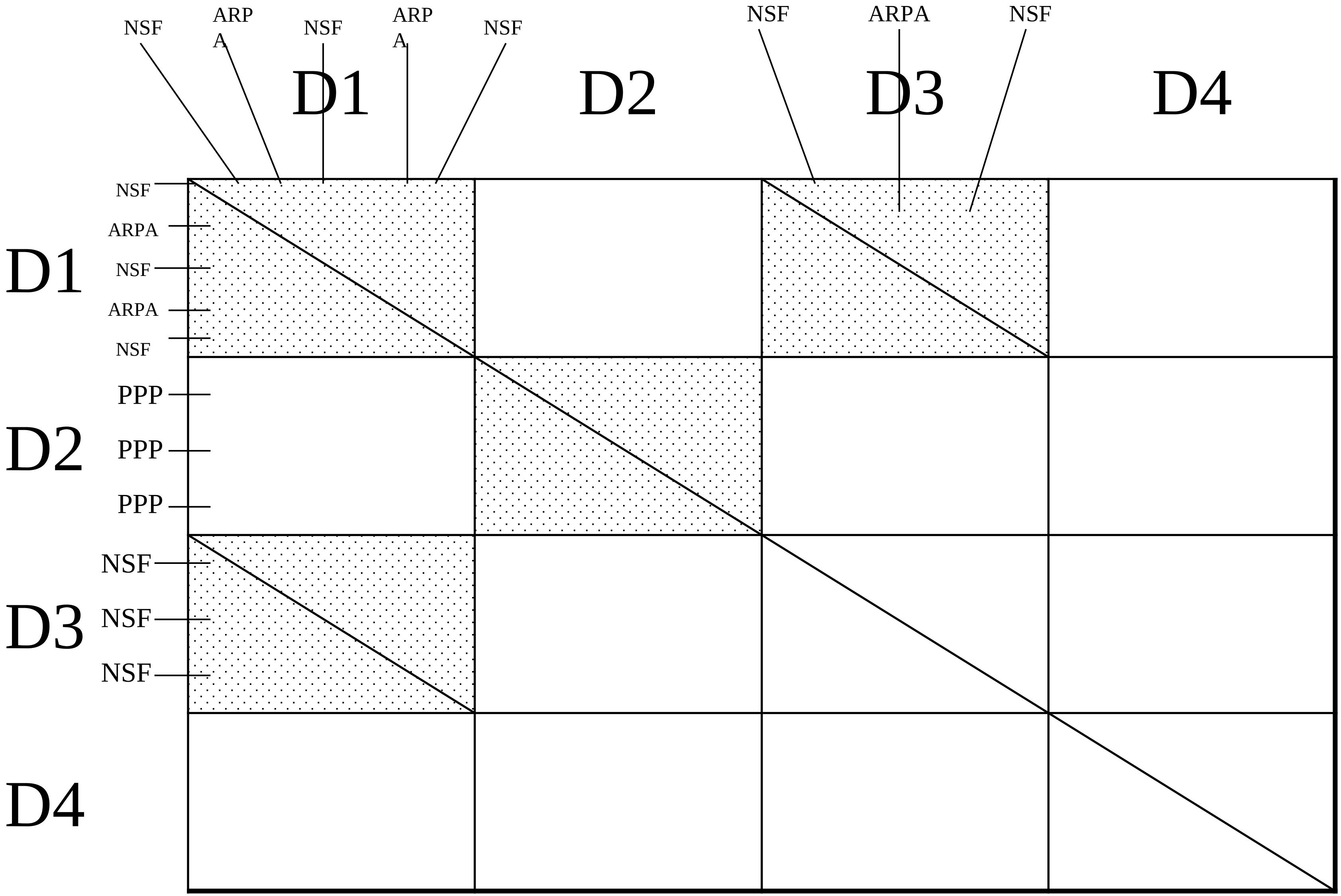


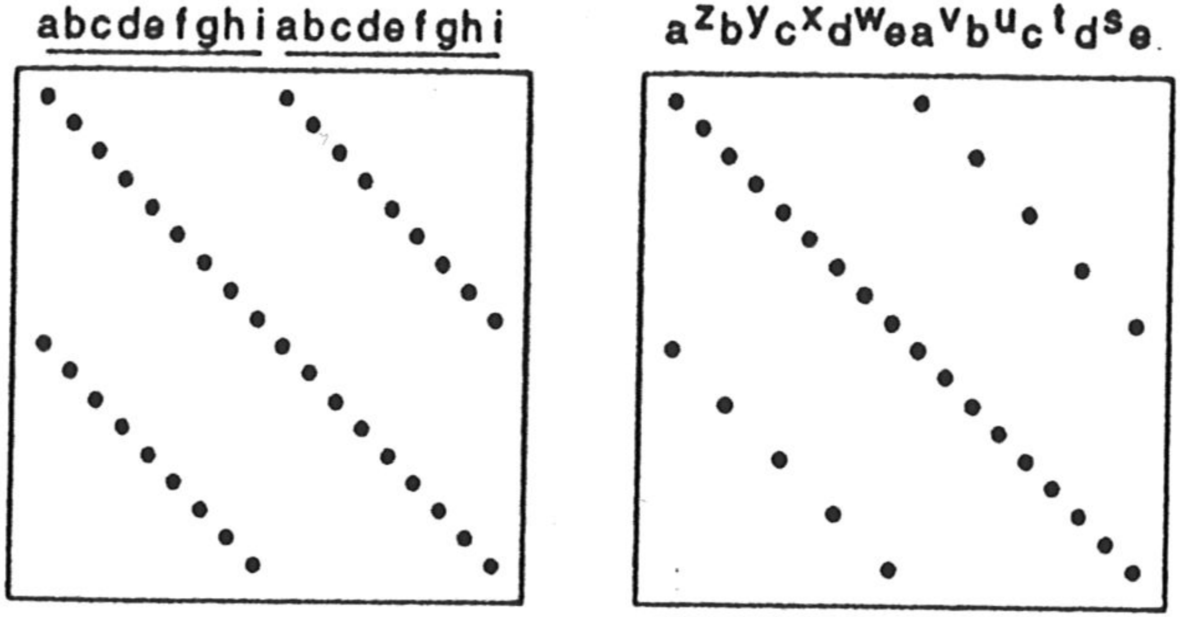
```
26  lnksls.slslnk
24  lnksls.slslnk
20  lnksls.slslnk
18  lnksls.slslnk10=0x0400;
44  lnksls.slslnk11=0x0;
48  lnksls.slslnk12=0x0;
52  lnksls.slslnk13=0x0;
56  lnksls.slslnk14=0x0;
60  lnksls.slslnk15=0x0;
32  TINSERT(RLLNKSLS,&lnksls,gl
 4  lnksls.numlnk=12;
 2  lnksls.slslnk0=0x1001;
 2  lnksls.slslnk1=0x2002;
 2  lnksls.slslnk2=0x4004;
 2  lnksls.slslnk3=0x8008;
```

```
40  lnksls.slslnk10=0x0;
44  lnksls.slslnk11=0x0;
48  lnksls.slslnk12=0x0;
52  lnksls.slslnk13=0x0;
56  lnksls.slslnk14=0x0;
60  lnksls.slslnk15=0x0;
32  TINSERT(RLLNKSLS,&lnksls,gl
 4  lnksls.numlnk=4;
 4  lnksls.slslnk0=0x1111;
 4  lnksls.slslnk1=0x2222;
 4  lnksls.slslnk2=0x4444;
 4  lnksls.slslnk3=0x8888;
```

Figure 1. Dotplot Browser.

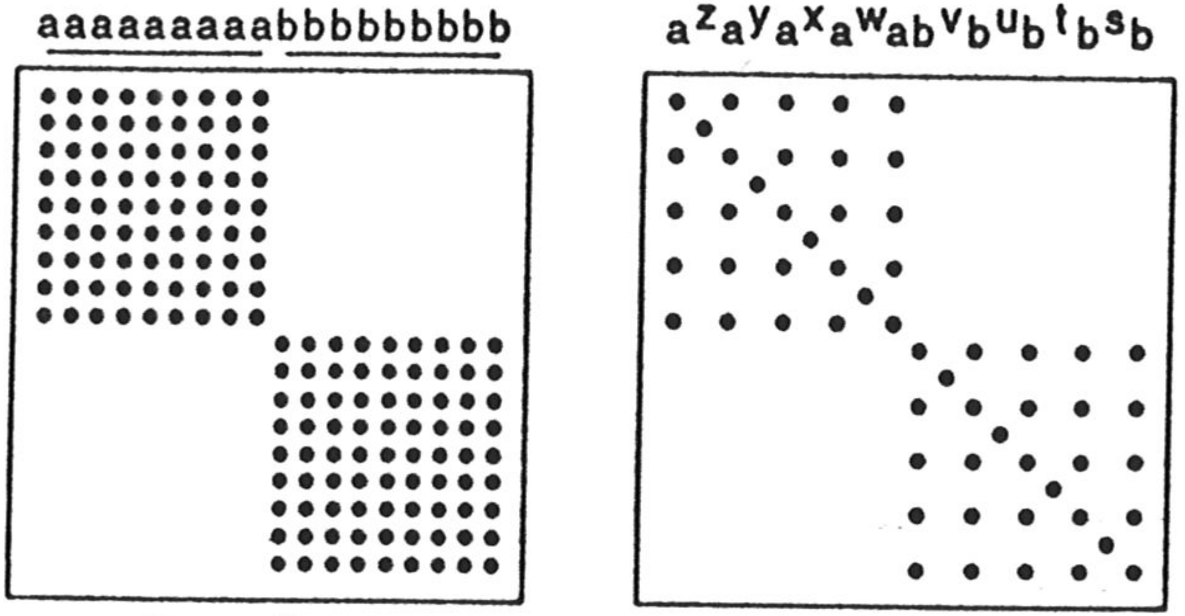
Motivation for Block Formation





a. Dense Diagonals

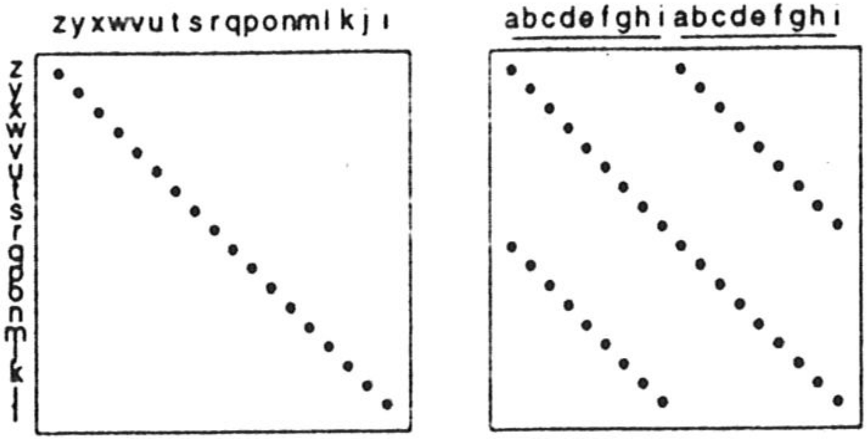
b. Sparse Diagonals



c. Dense Squares

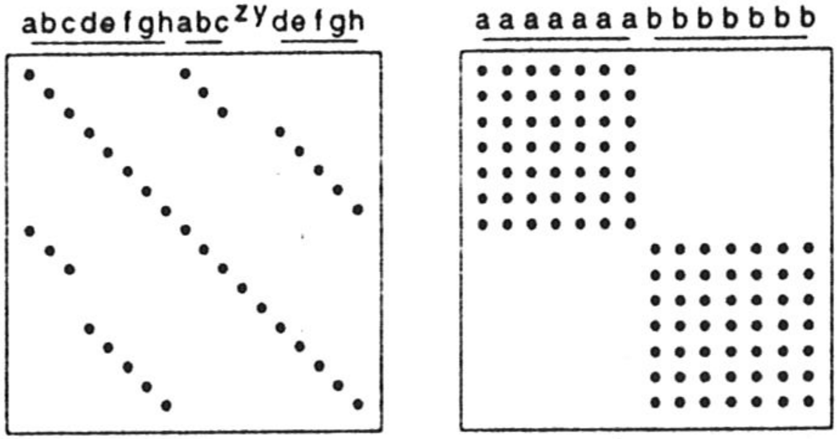
d. Sparse Squares

Figure 6. Dense Versus Sparse Features.



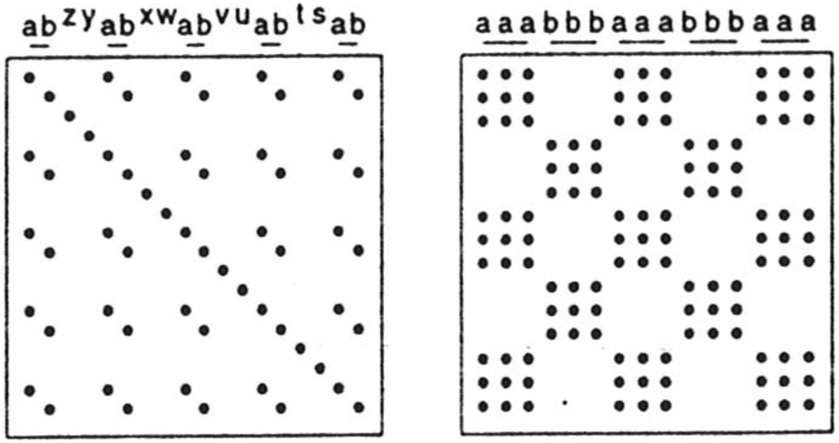
a. No Features

b. Diagonals



c. Broken Diagonals

d. Squares



e. Diagonal Texture

f. Square Texture

Figure 2. Features in Synthesized Dotplots.

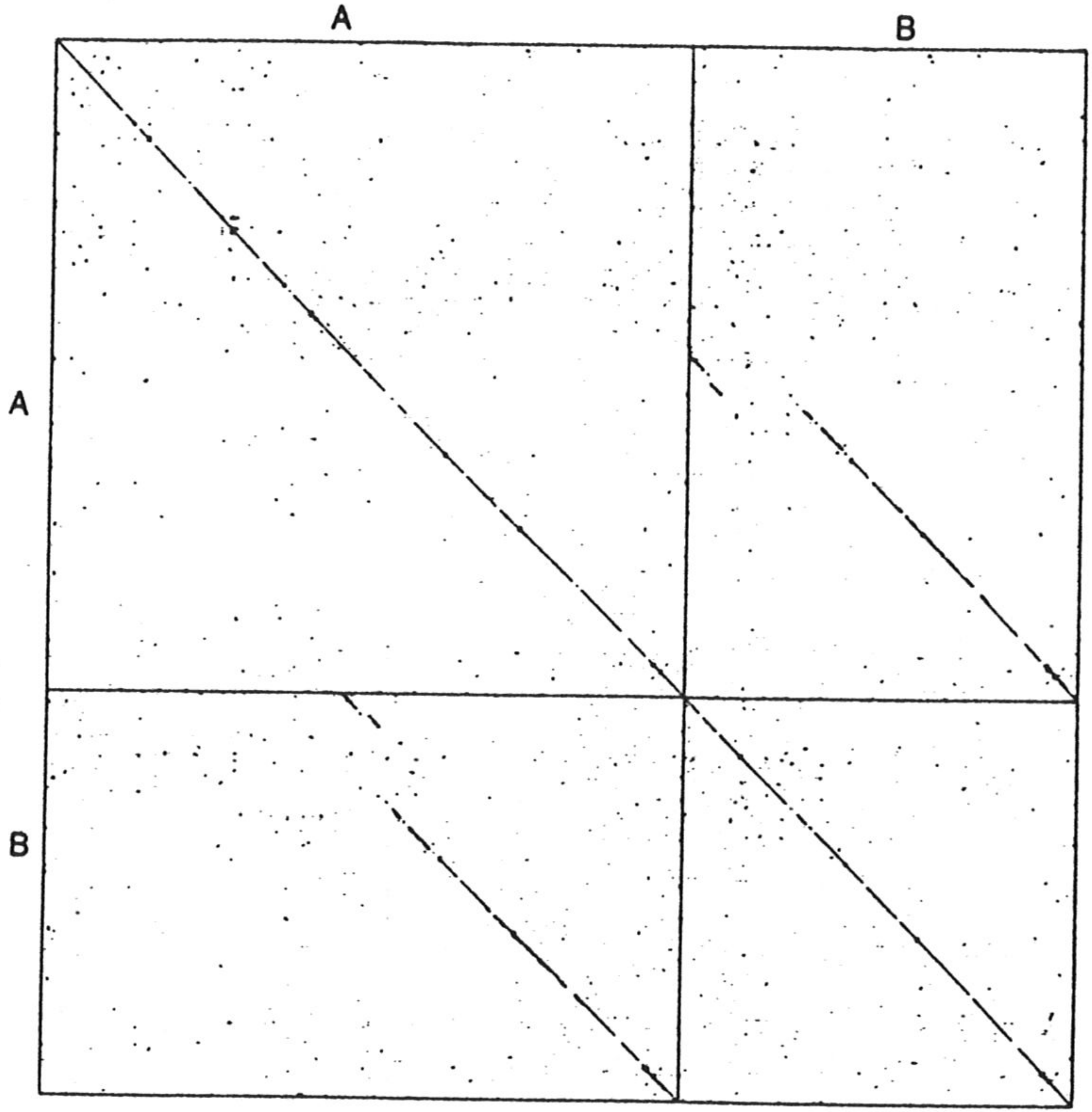


Figure 3. Dotplot of Two DNA Sequences (7,000 Nucleotides).

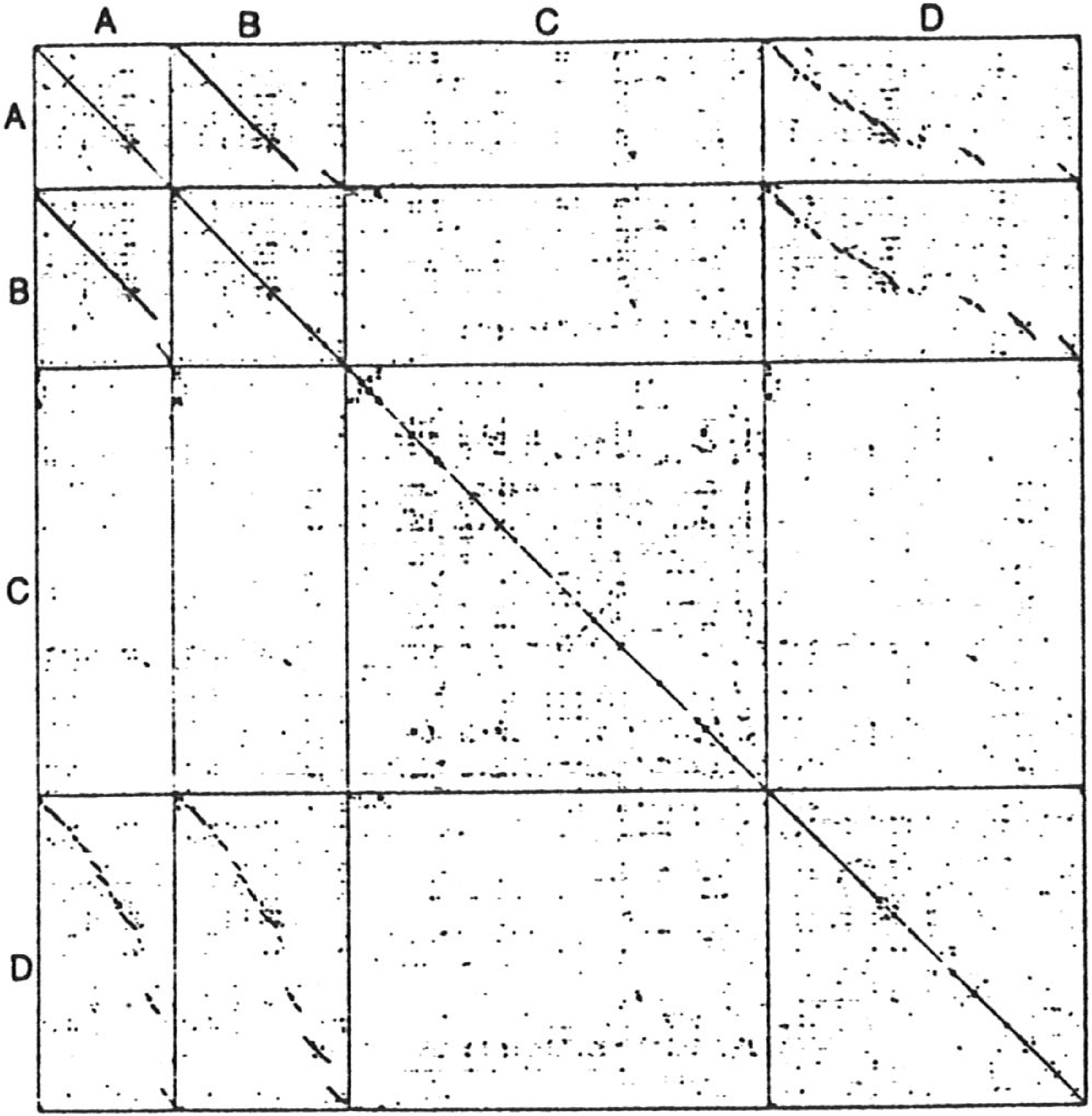
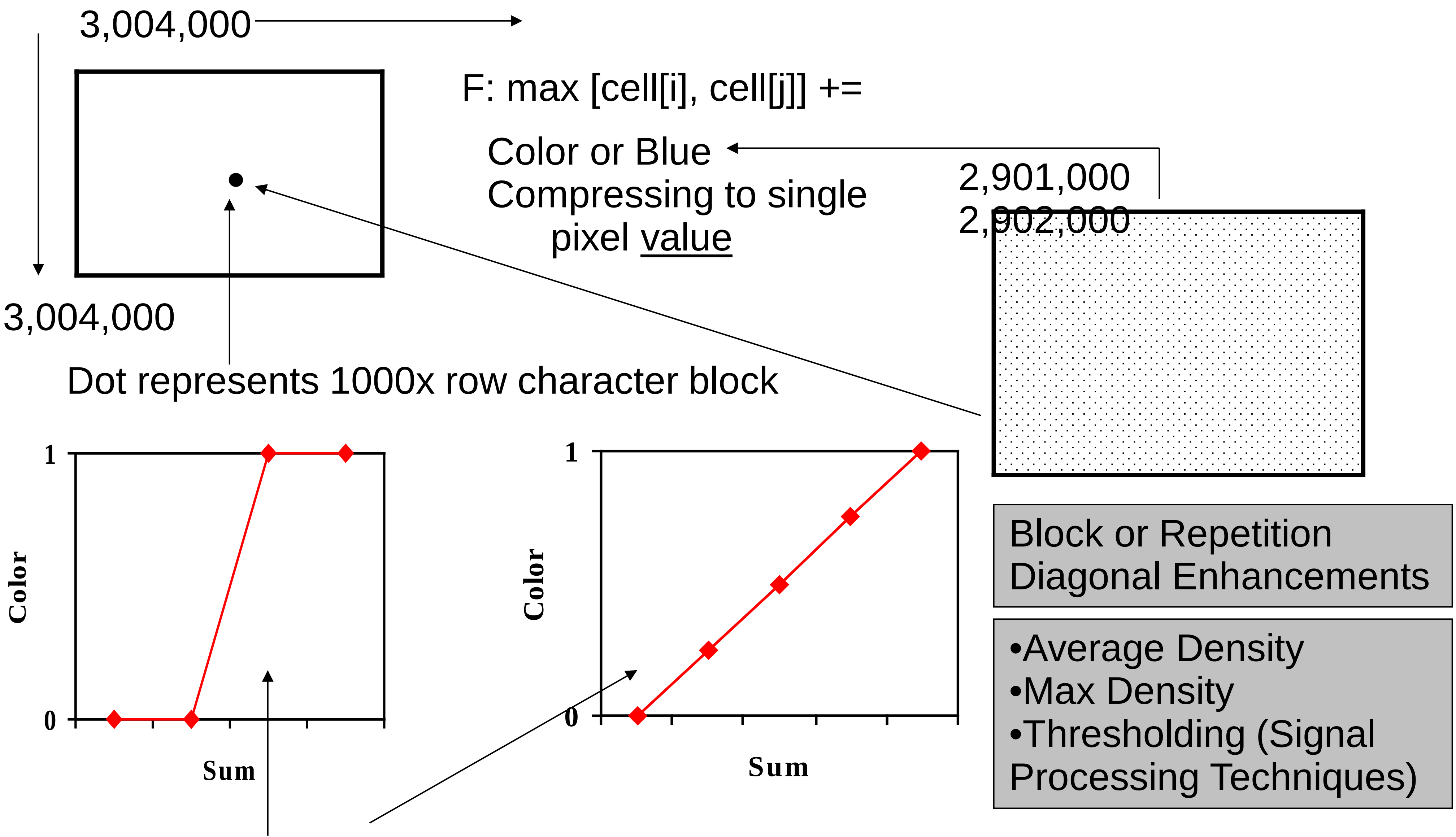


Figure 4. Four AP News Stories (3,000 words).

Region Mapping (Image Compression)



3,004,000

F: max [cell[i], cell[j]] +=

Color or Blue
Compressing to single
pixel value

2,901,000
2,902,000

3,004,000

Dot represents 1000x row character block

Block or Repetition
Diagonal Enhancements

- Average Density
- Max Density
- Thresholding (Signal Processing Techniques)

Cut off threshold for Display if Black or White

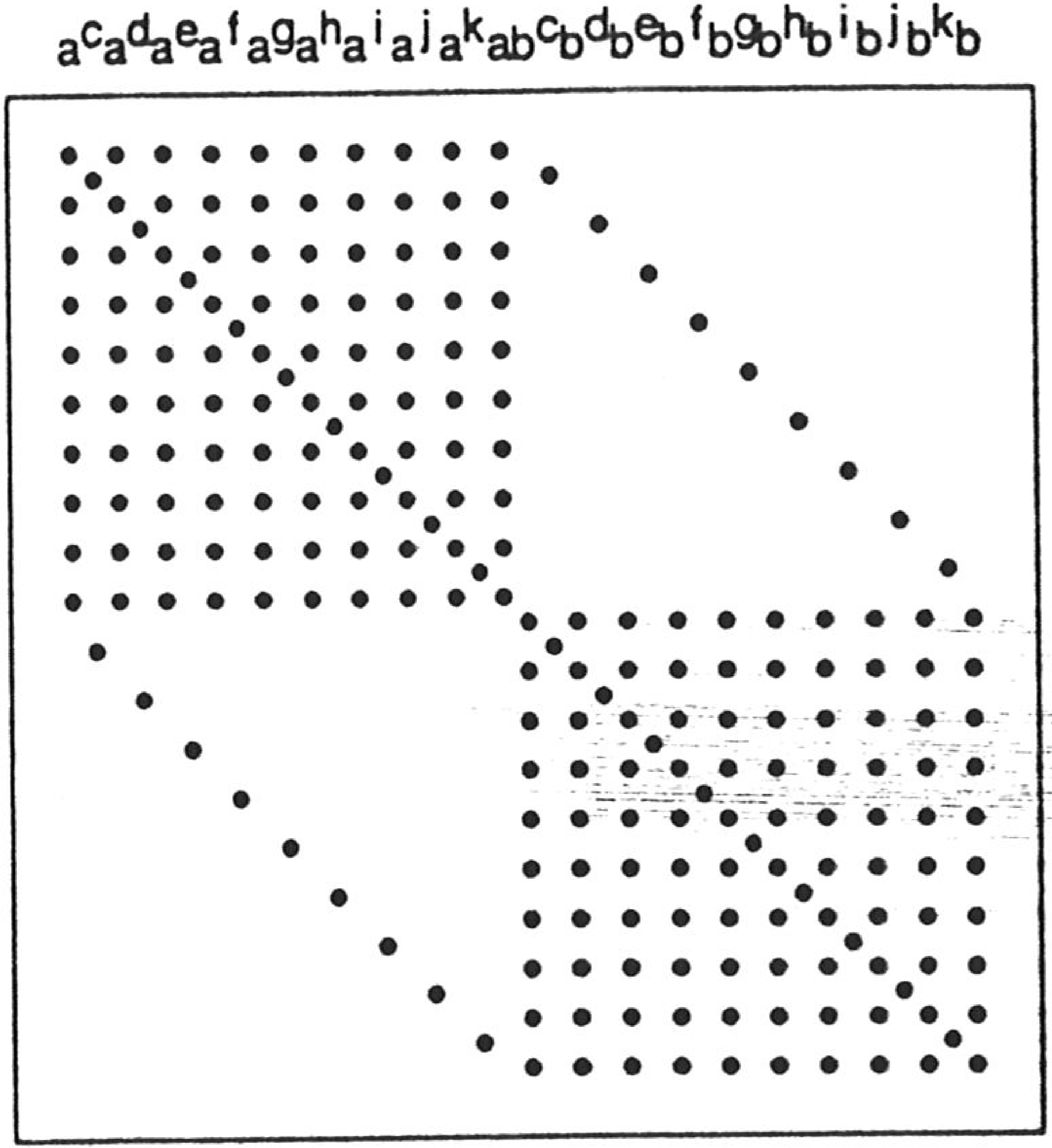


Figure 7. Combination of Diagonals and Squares.

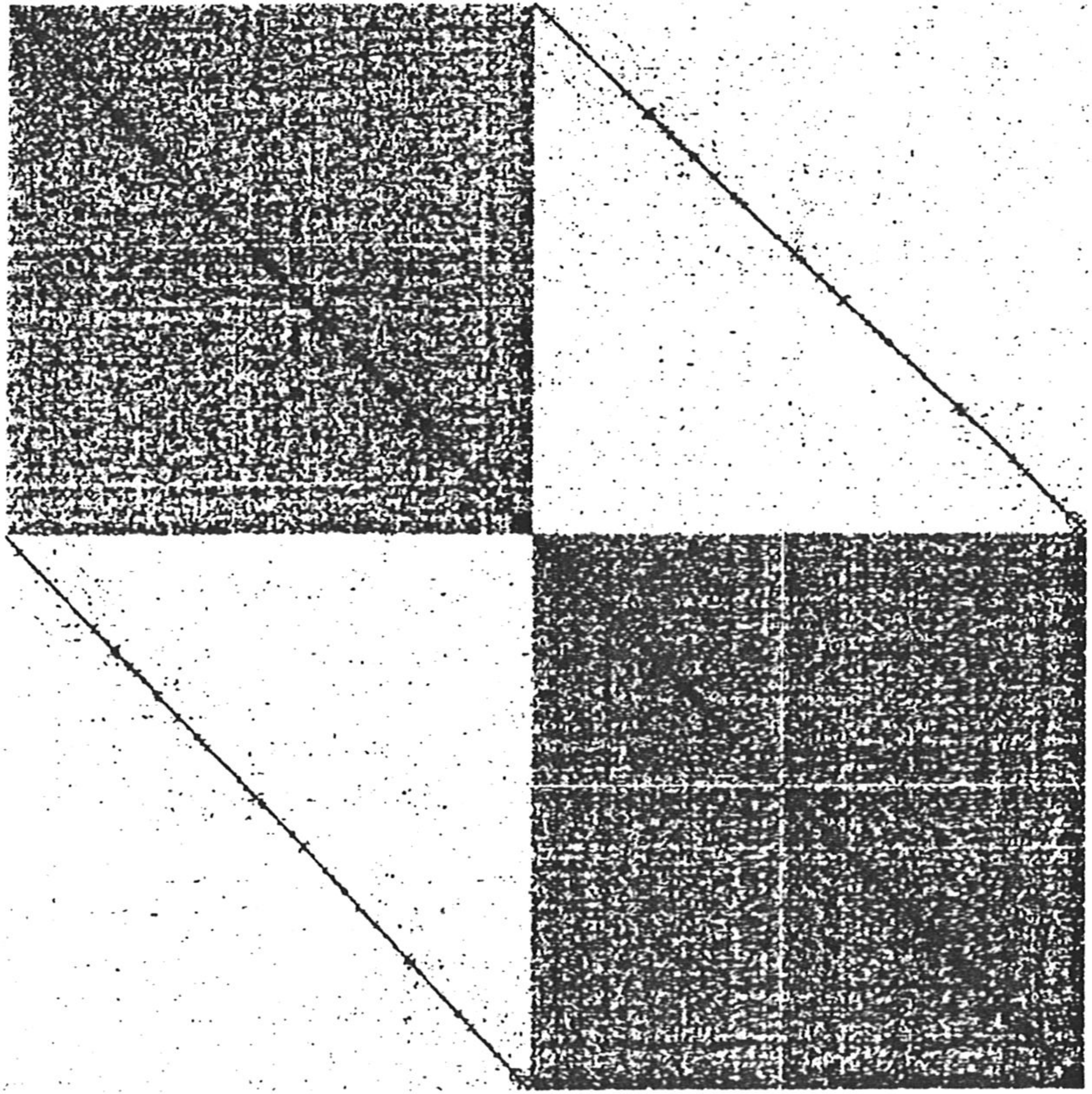


Figure 5. Three Years of Hansards (37 million words).

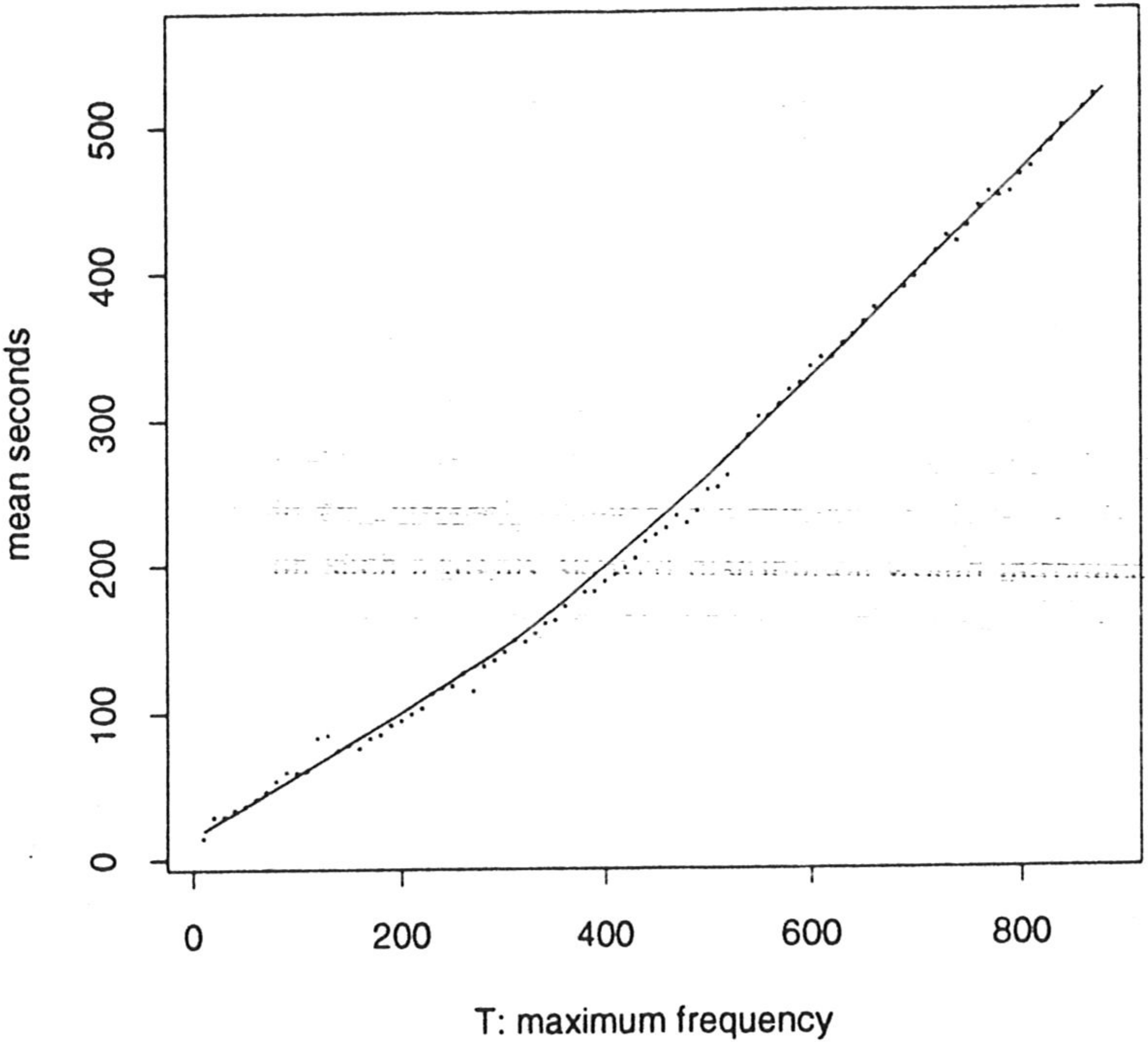


Figure 14. Computing a Hansard Plot With Different Values of T.

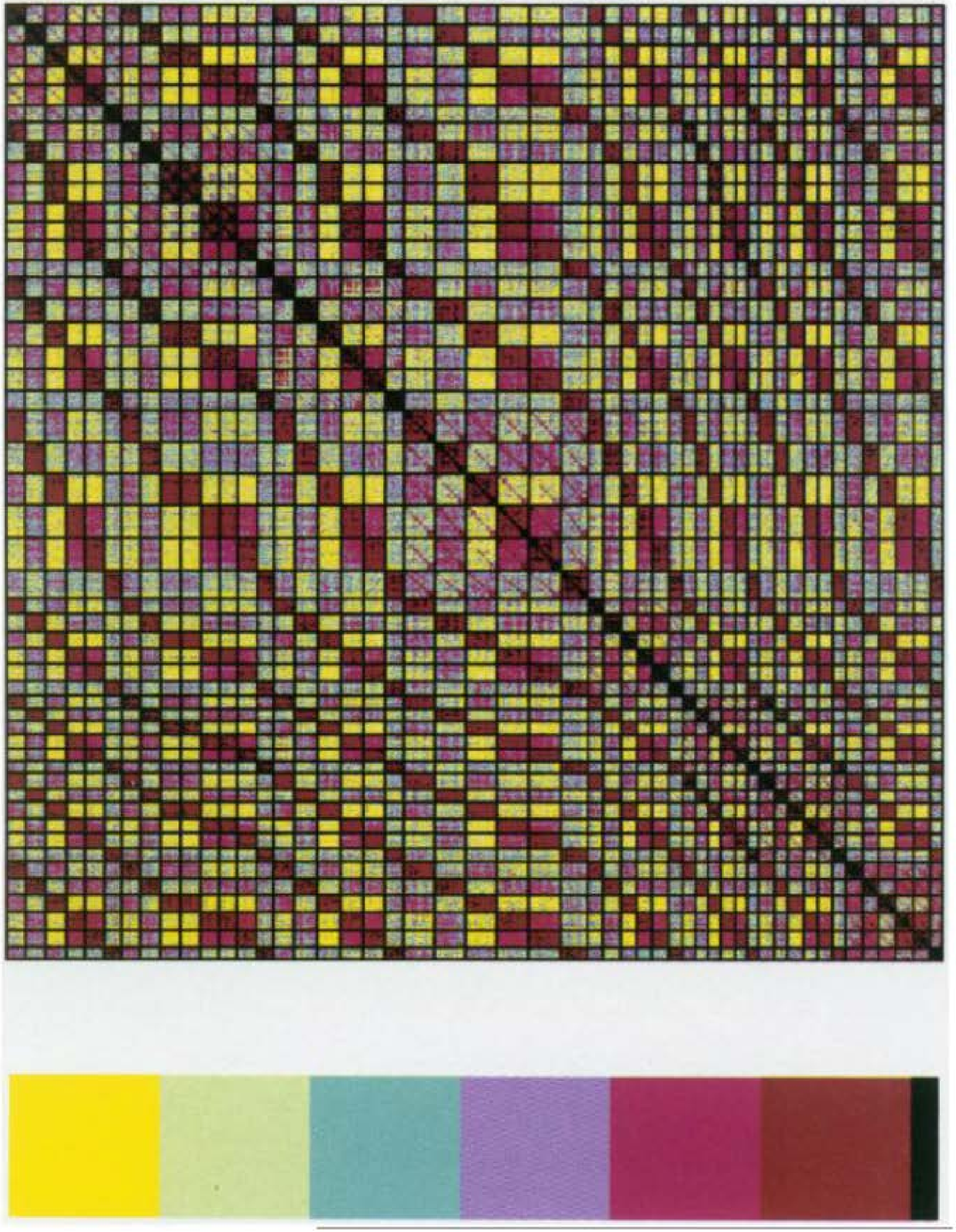


Figure 8. Six Chapters of Microsoft Manuals in Seven Languages (3.3 Million Words) With Color Map.

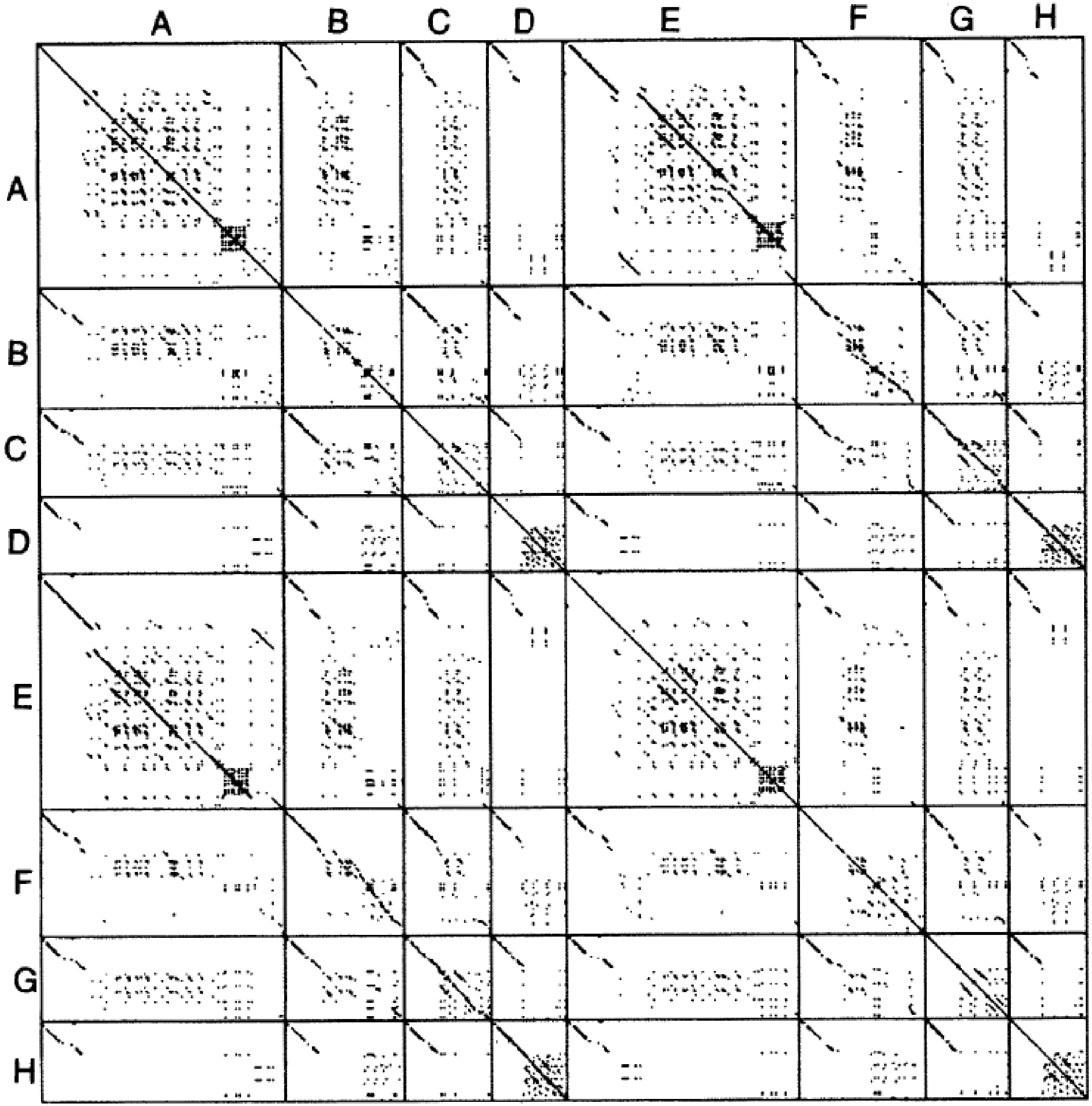


Figure 9. Three Thousand Lines of Code.

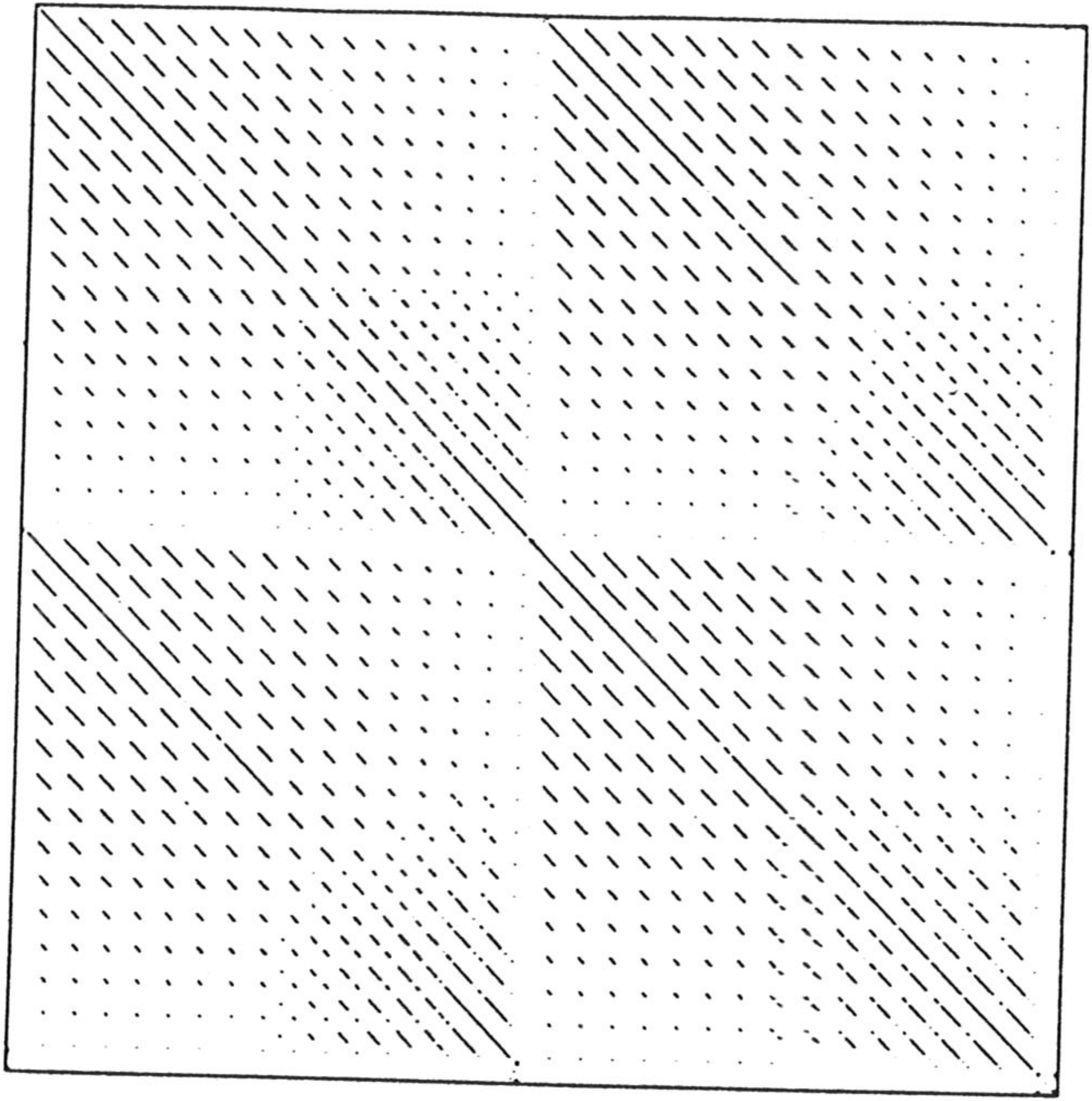


Figure 10. *Six Hundred Lines of Code (detail of Figure 1).*

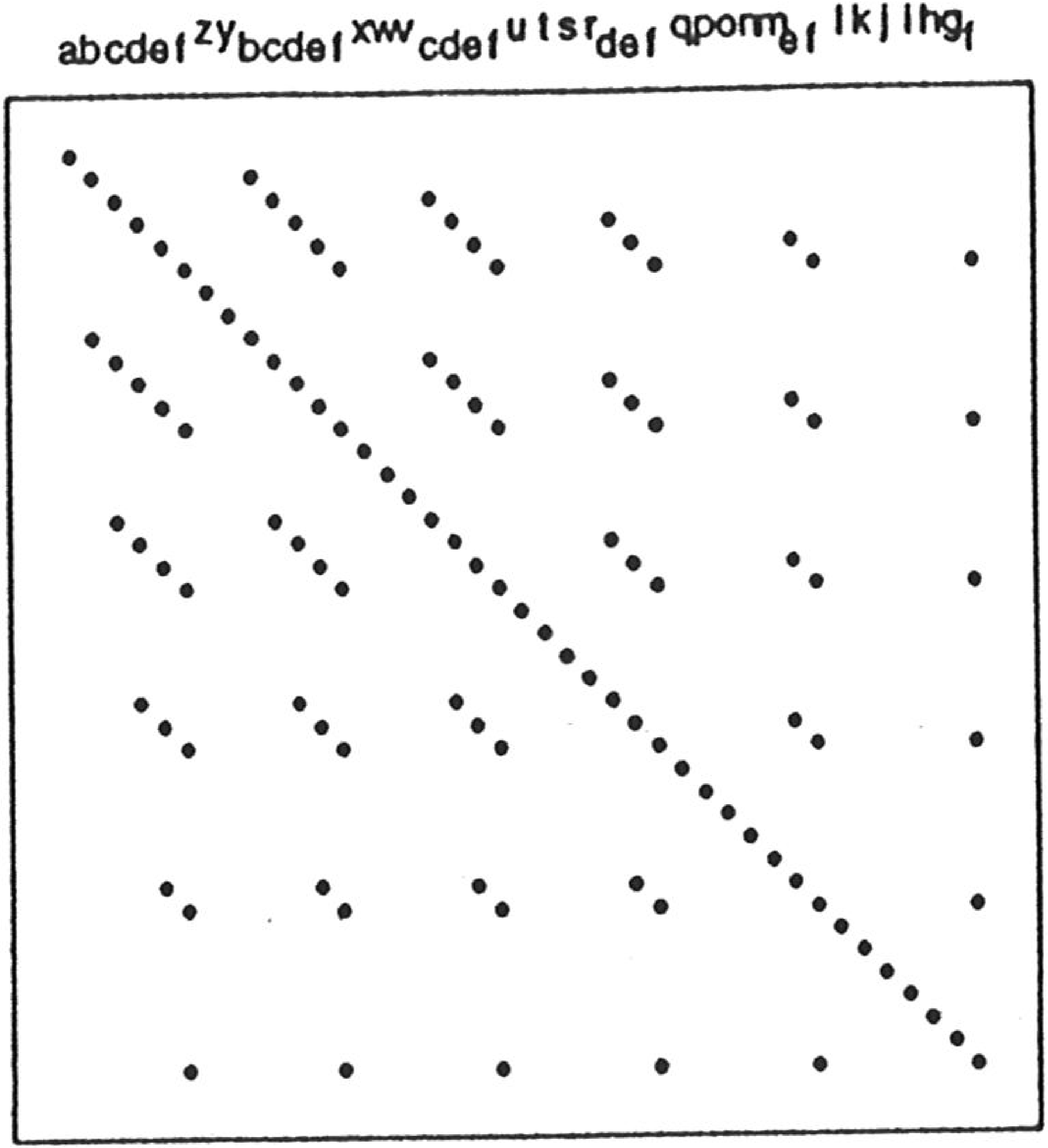


Figure 11. Shrinking Diagonals.

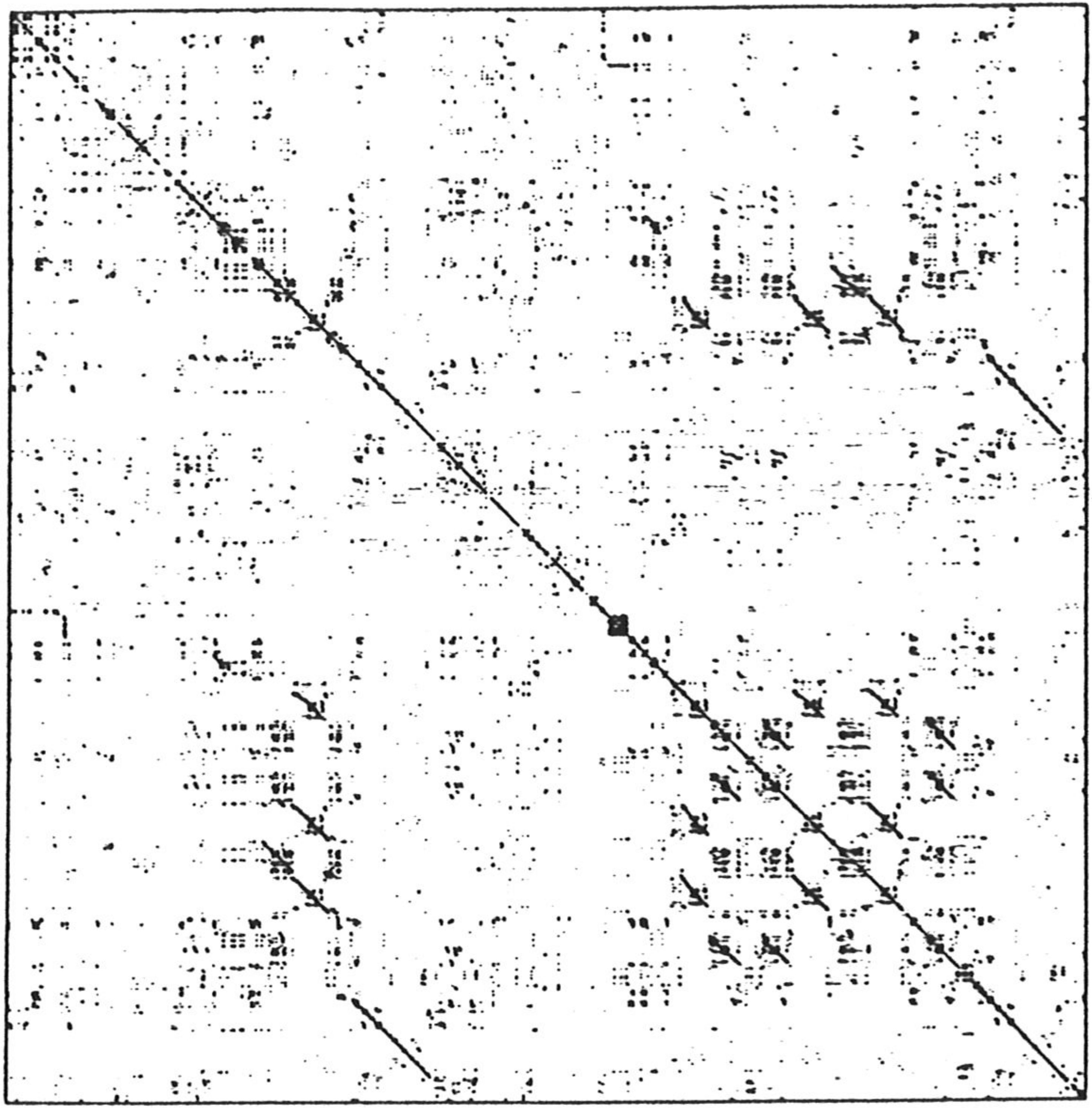


Figure 12. Three Thousand Four Hundred Lines of Code.

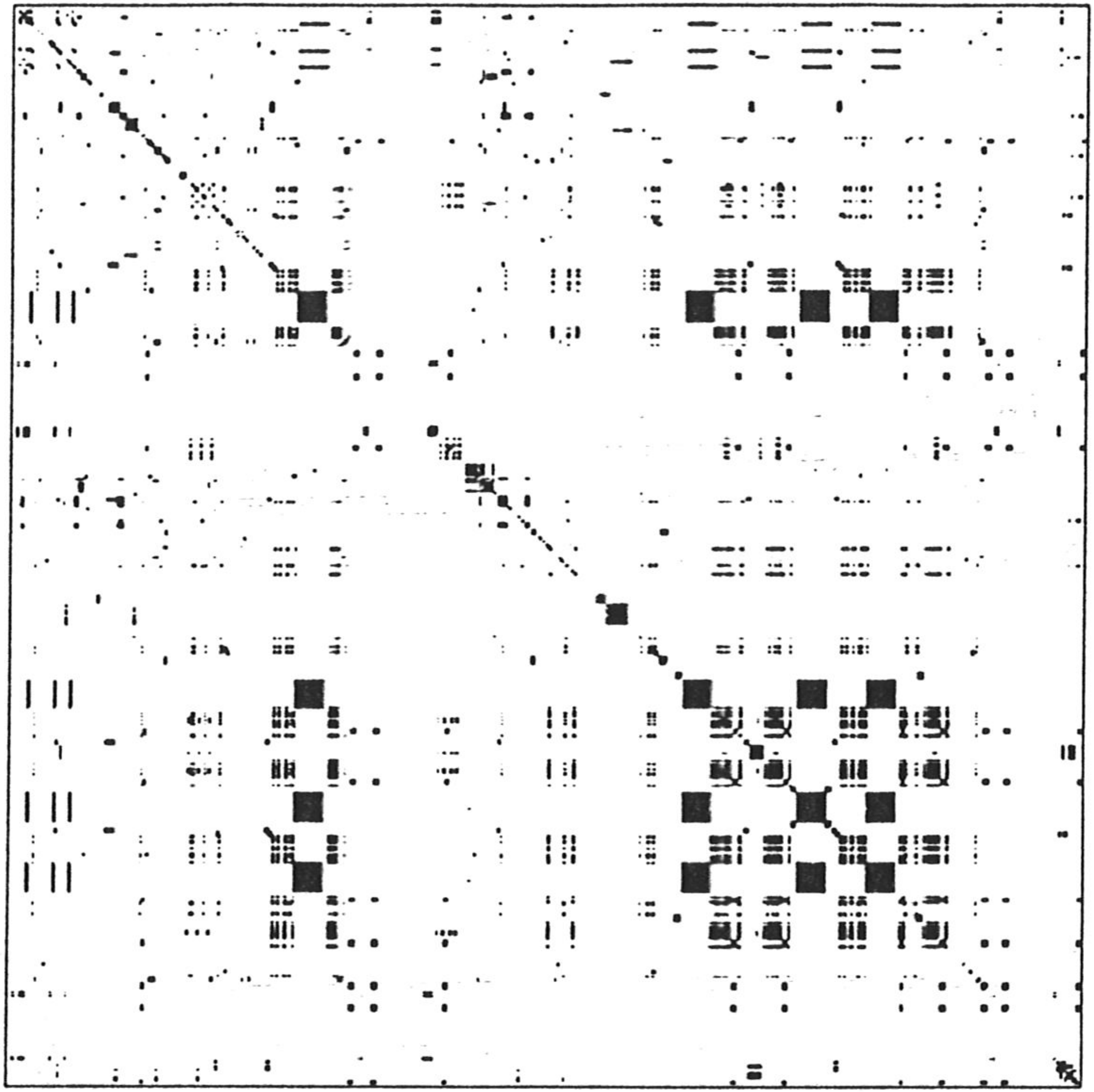


Figure 13. Three Thousand Four Hundred Author Attributes.

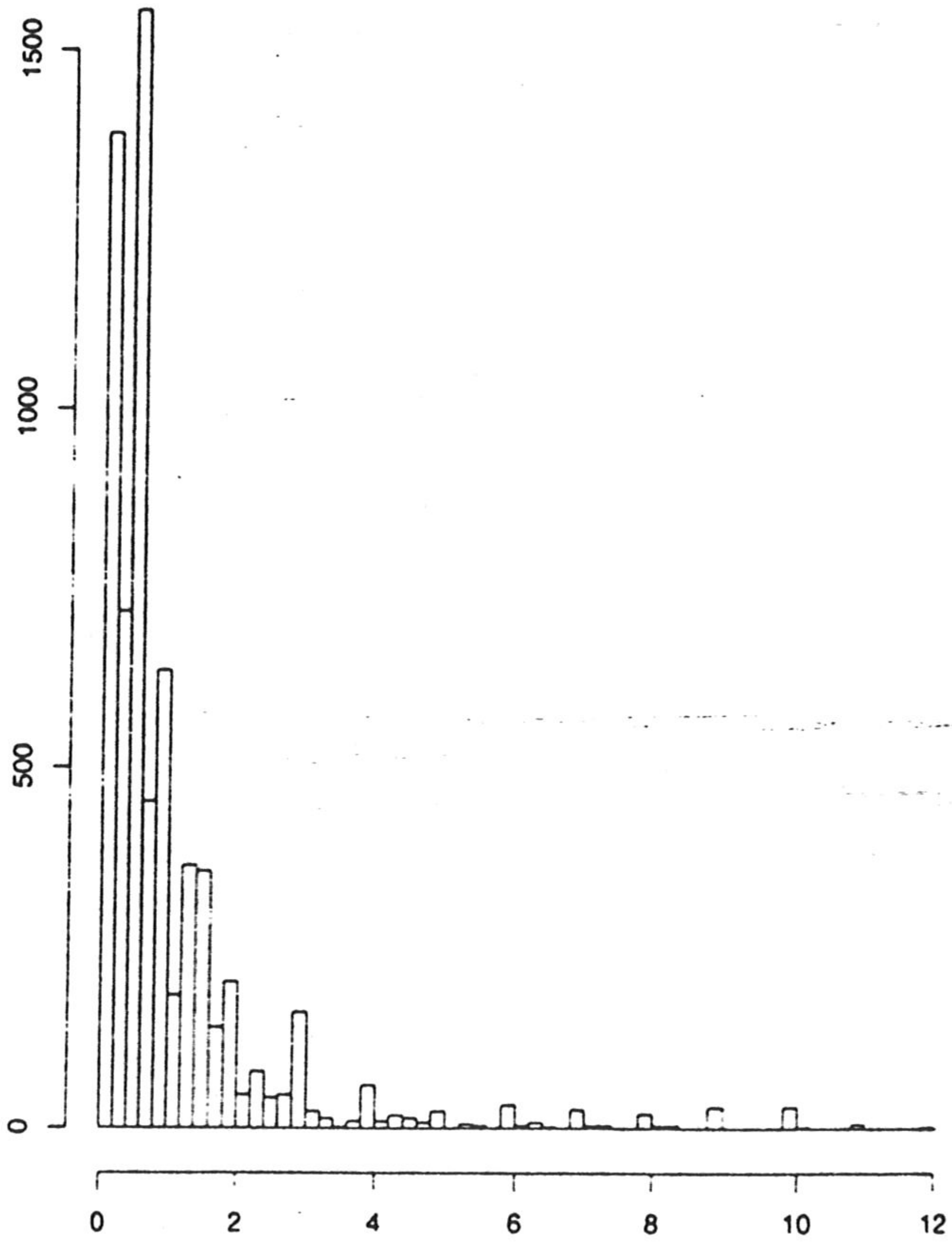


Figure 15. Histogram of Values in Figure 9's f-Image.

Document Segmentation via DotPlot Approaches

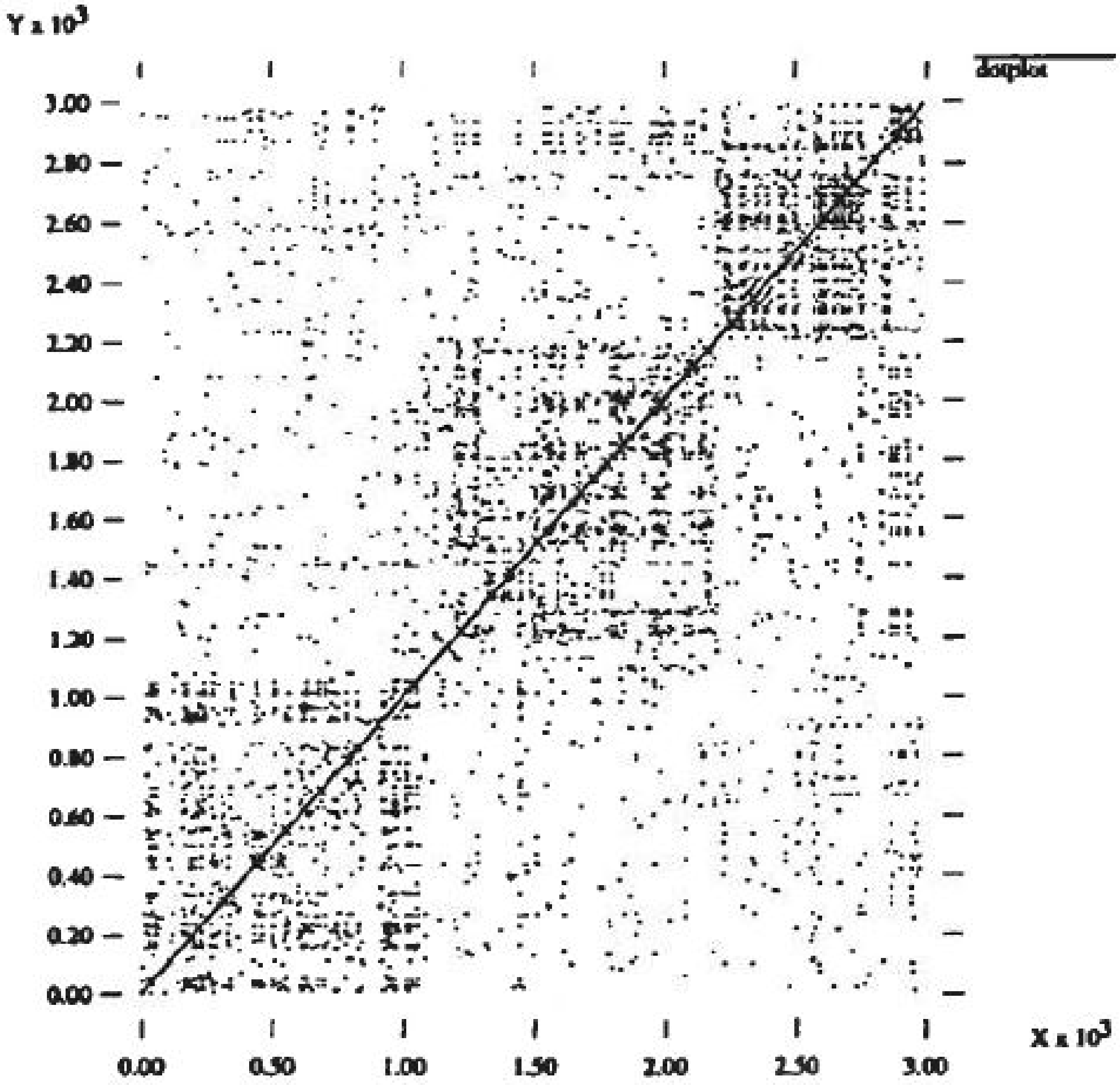


Figure 1: The dotplot of four concatenated *Wall Street Journal articles.*

Document Segmentation via DotPlot Approaches

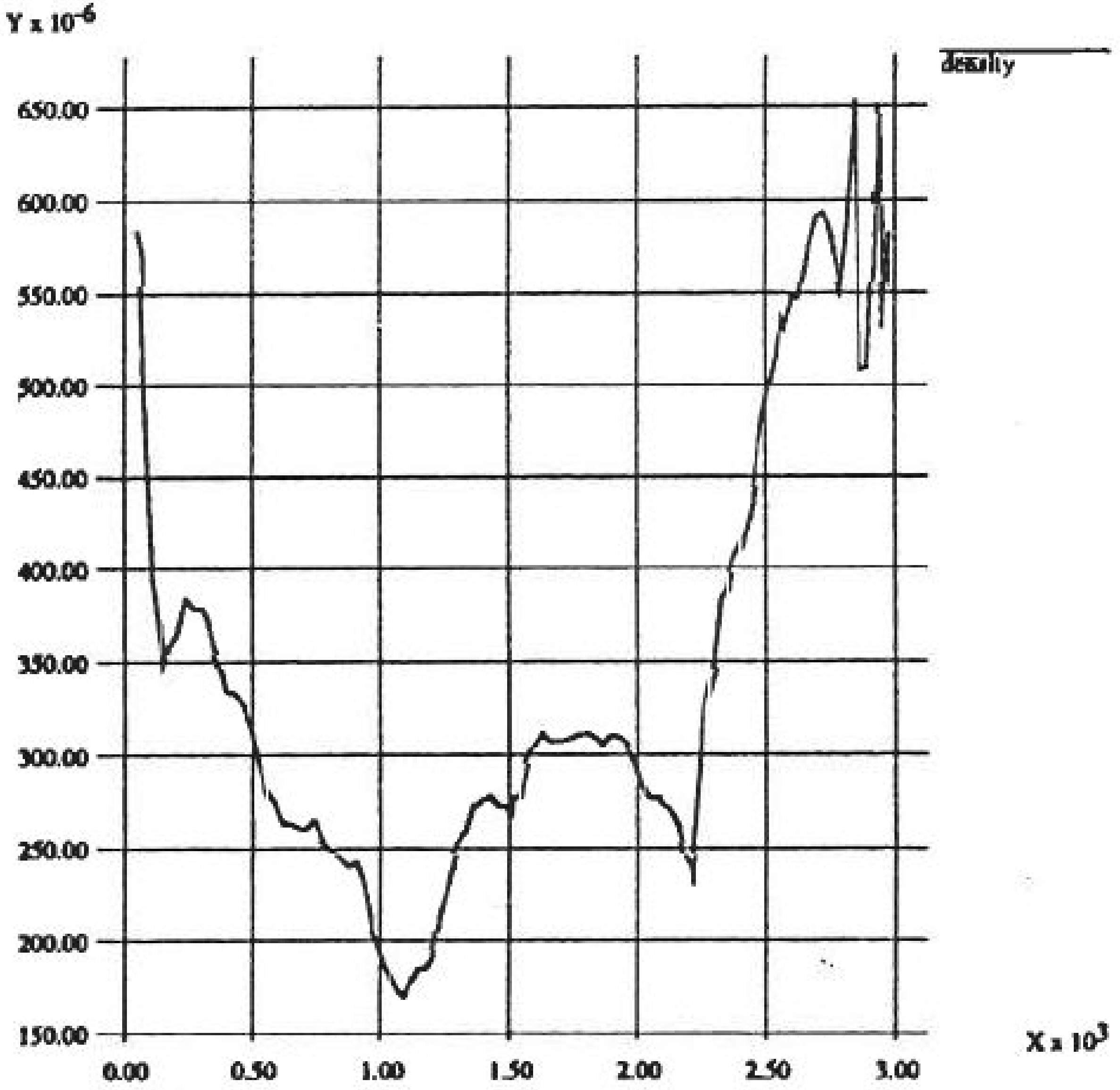


Figure 2: The outside density plot of the same articles.

Document Segmentation via DotPlot Approaches

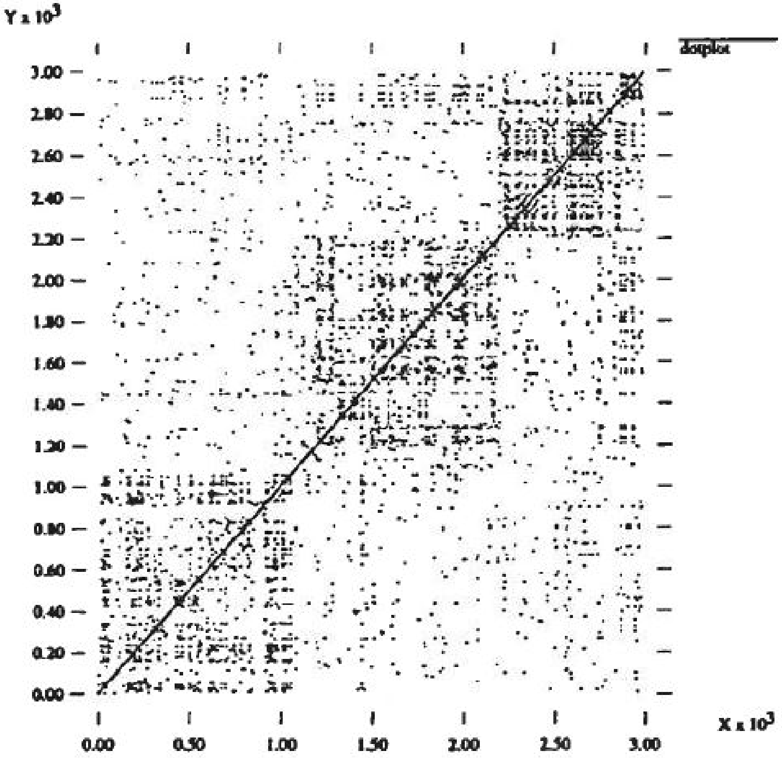


Figure 1: The dotplot of four concatenated *Wall Street Journal articles.*

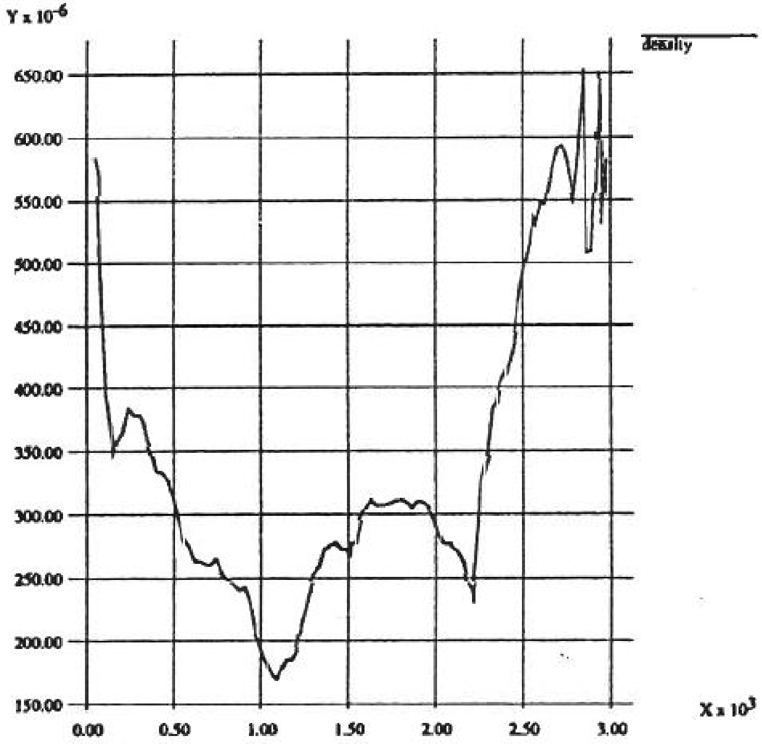


Figure 2: The outside density plot of the same articles.

Document Segmentation via DotPlot Approaches

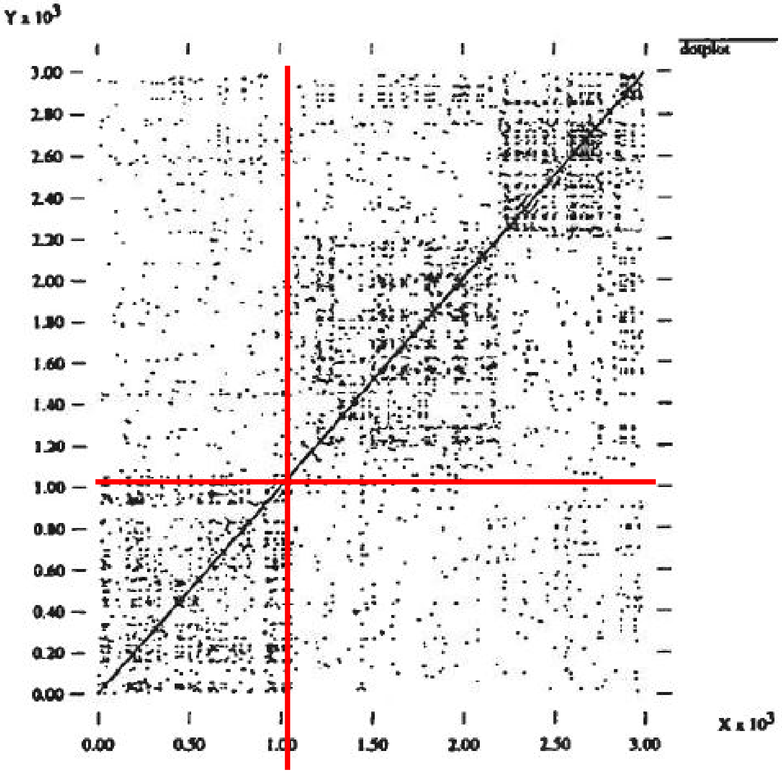


Figure 1: The dotplot of four concatenated *Wall Street Journal articles.*

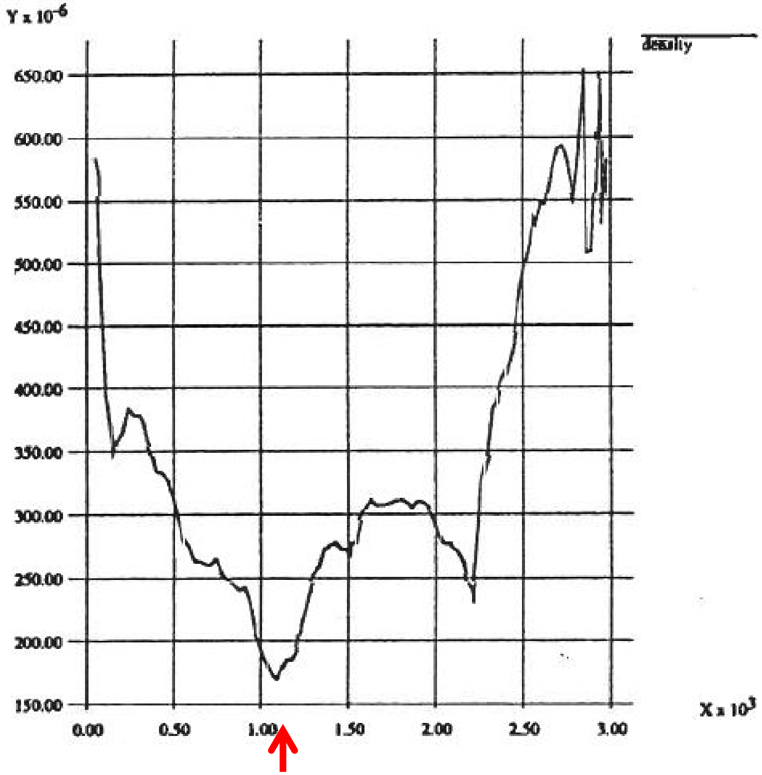


Figure 2: The outside density plot of the same articles.

Document Segmentation via DotPlot Approaches

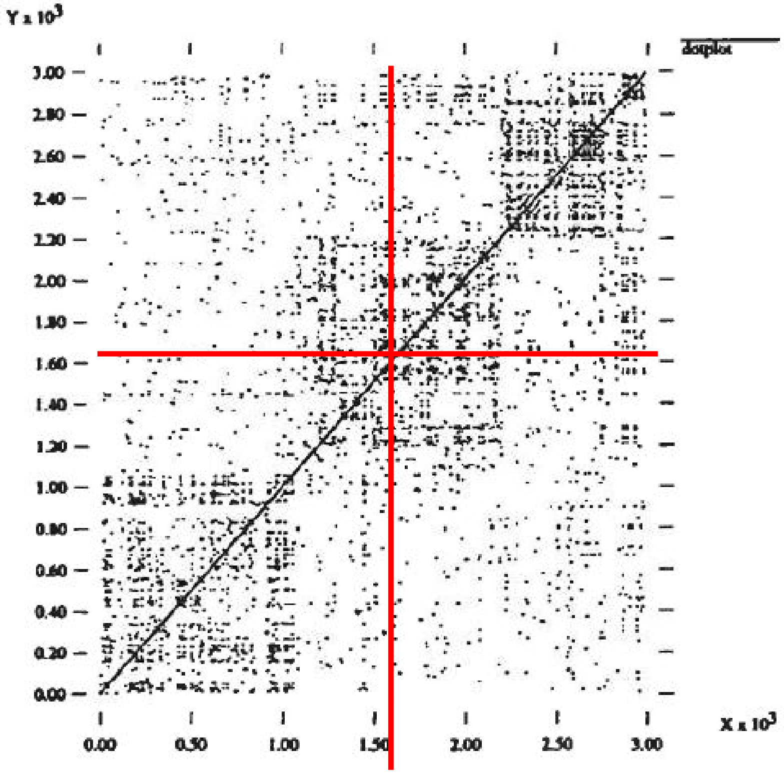


Figure 1: The dotplot of four concatenated *Wall Street Journal articles.*

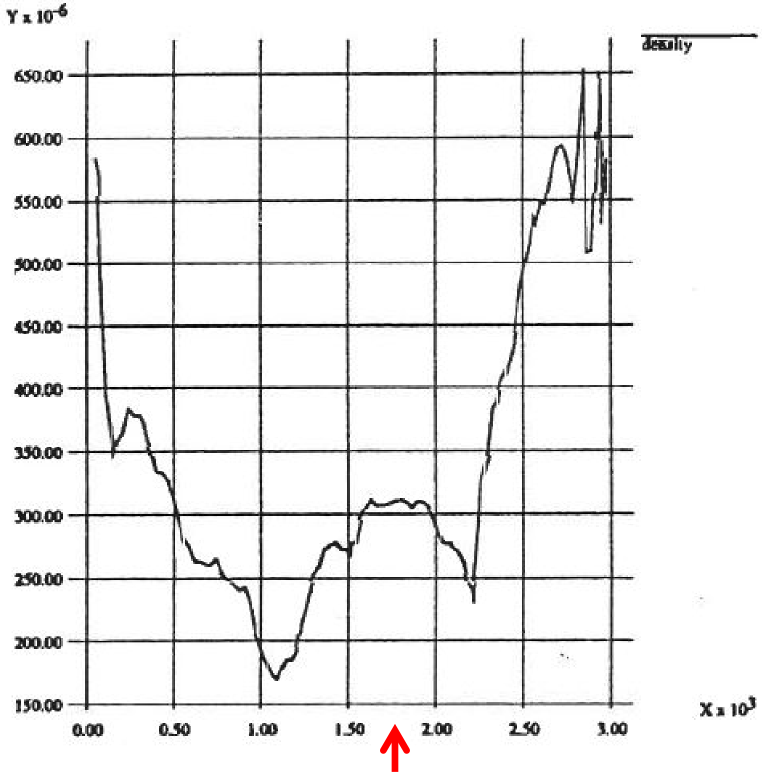


Figure 2: The outside density plot of the same articles.

Document Segmentation via DotPlot Approaches

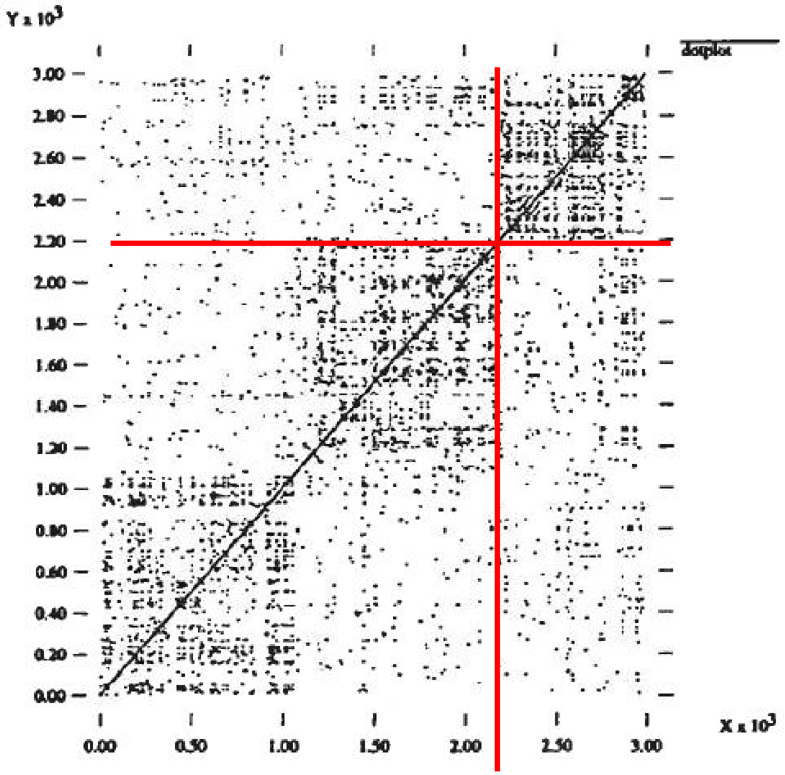


Figure 1: The dotplot of four concatenated *Wall Street Journal articles.*

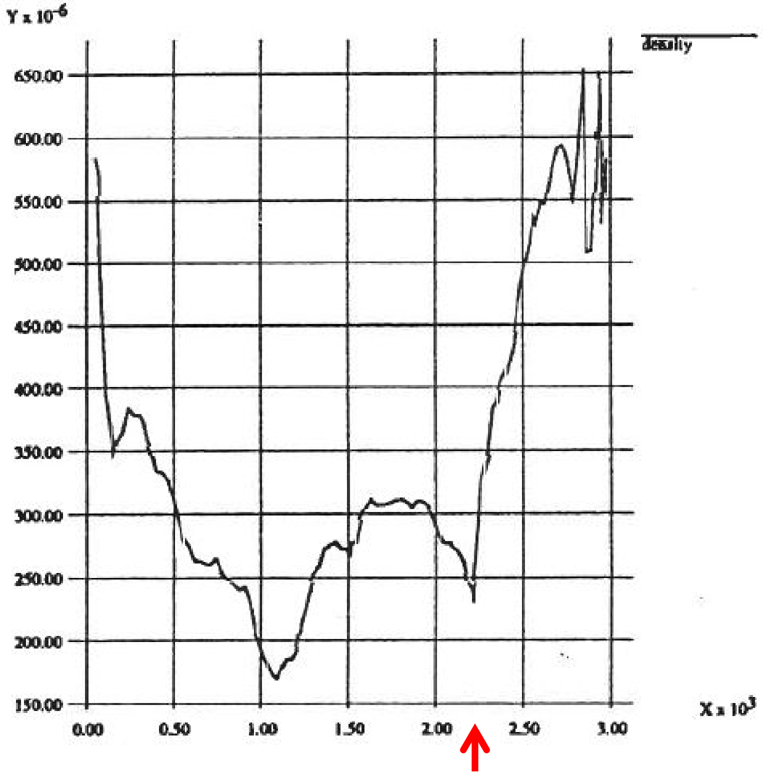


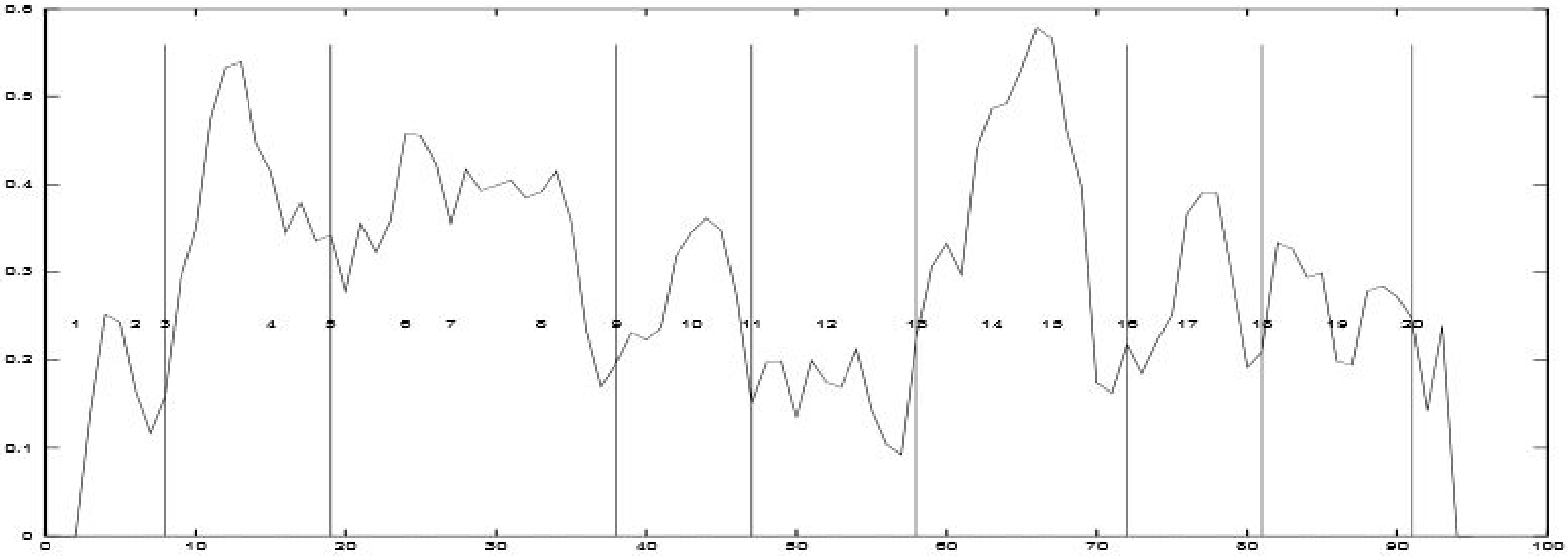
Figure 2: The outside density plot of the same articles.

Document Segmentation via DotPlot Approaches

Predicted and actual document region boundaries.

Predicted = local minima

Actual = vertical lines.


0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20