

Na Figura 1 é mostrado um exemplo de uma matriz de similaridade onde a intensidade do ponto (i, j) representa a similaridade entre as sentenças i e j . Observa-se que a matriz é simétrica, assim cada ponto na linha diagonal representa a similaridade quanto $i = j$ (ou seja, com a mesma sentença) e revela quadrados com maior concentração de pontos ao longo da diagonal. Essas regiões indicam porções de texto com maior coesão léxica.

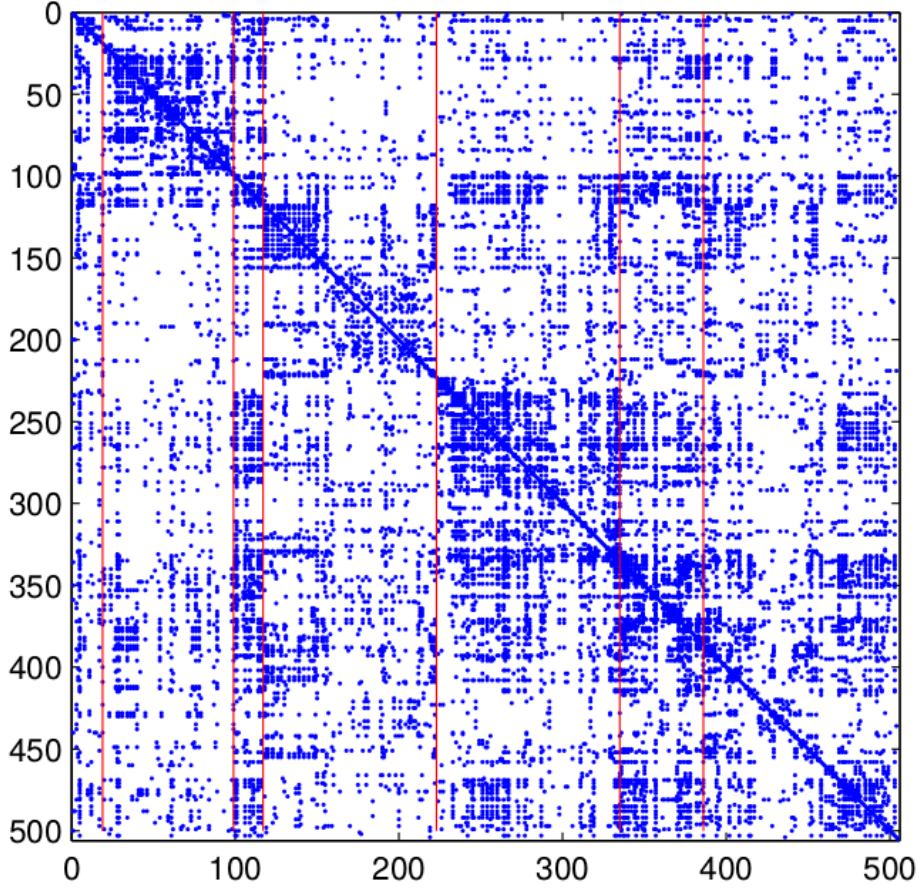


Figure 1: *DotPlot* da similaridade entre sentenças onde as linha verticais representam segmentos reais.

O processo de identificação dos limites é baseado no método DotPlotting [?] que usa regiões com maior densidade em uma matriz de similaridades para determinar como os segmentos são distribuídos. Um segmento é definido por duas sentenças i e j que representam uma região quadrada ao longo da diagonal da matriz. Seja $s_{i,j}$ a somatória dos *rankings* de um segmento e $a_{i,j}$ sua área interior. Seja $B = \{b_1, \dots, b_m\}$ a lista de m segmentos.

Calcula-se a medida de densidade de uma região por

$$D = \quad (1)$$

onde s_k a_k são a somatória dos valores dos rankings e a área de um segmento

k em B .

O processo incia com um único segmento formado por todas as sentenças do documento e o divide recursivamente em m segmentos. Cada passo divide um dos segmentos em B no ponto ij que maximiza D . O processo se repete até que se atinja o número de segmentos desejados ou um limiar de similaridade.