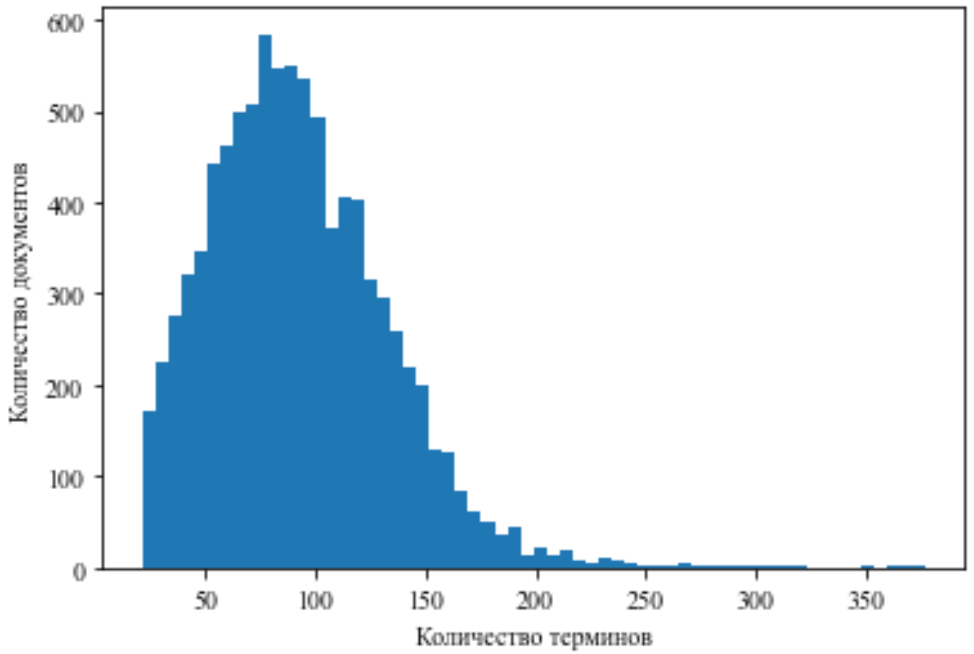
Автоматическое выделение терминов в тематическом моделировании



Распределение количества терминов в документе в датасете ACL RD-TEC

Для решения задачи используется алгоритм выделения коллокаций TopMine в сочетании с модульной технологией тематического моделирования.

Аддитивнаря регуляризация тематических моделей основана на максимизации логарифма правдоподобия и регуляризаторов $R_i(\Phi, \Theta)$, требующихся для корректной постановки задачи разложения матрицы частот термов в документах на произведение матрицы термов тем и матрицы тем документов:

$$P = \Phi\Theta$$

$$P = (p_{\omega d})_{W \times D} \quad \Phi = (\varphi_{\omega t})_{W \times T} \quad \Theta = (\theta_{td})_{T \times D}$$

$$\sum_{d \in D} \sum_{w \in d} \ln \sum_{t \in T} \varphi_{\omega t} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}; \qquad R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta)$$

$$\sum_{w \in W} \varphi_{\omega t} = 1; \quad \varphi_{\omega t} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geqslant 0$$