chars of extracted text per 1 GB of raw compressed crawls