پاسخ سوال ۳:

(i)- :تعداد اصطلاحات منحصربه‌فرد در مجموعه:

$M = k \cdot T^b$

در این سؤال:

- تعداد صفحات ۶۰۰,۰۰۰,۰۰۰ است.
- میانگین تعداد توکن‌ها در هر صفحه ۶۰۰ است.
- $T = 600 \times 600{,}000{,}000 = 360{,}000{,}000{,}000$.
- پارامترهای $b = 0.5$ و $k = 100$ داده شده‌اند.

با جای‌گذاری:

$M = 100 \times (360{,}000{,}000{,}000)^{0.5}$

$M = 100 \times 600{,}000 = 60{,}000{,}000$

بنابراین تعداد اصطلاحات منحصربه‌فرد $M = 60{,}000{,}000$ است.

(ii) نسبت اصطلاحاتی که فقط یک بار ظاهر می‌شوند (hapax legomena) :

طبق قانون Zipf:

$cf_i \propto 1/i$

بنابراین برای فراوانی کلی اصطلاحات داریم:

$T = \Sigma(cf_i) = c \cdot \Sigma(1/i)$

که در آن $\Sigma(1/i) \approx \ln(M)$

با جای‌گذاری مقادیر:

- $T = 600 \times 600{,}000{,}000 = 360{,}000{,}000{,}000$
- $M = 60{,}000{,}000$
- $\ln(60{,}000{,}000) \approx 17.9$

بنابراین:

$c = T / \ln(M) = 360{,}000{,}000{,}000 / 17.9 \approx 2 \times 10^{10}$

حال برای محاسبه فراوانی کمترین اصطلاح (با رتبه M):

$cf_M = c / M = (2 \times 10^{10}) / 60{,}000{,}000 \approx 333.33$

این عدد نشان می‌دهد که حتی کم‌ترین اصطلاح نیز بیش از یک بار ظاهر شده است. بنابراین، اصطلاحاتی که فقط یک بار ظاهر می‌شوند (hapax legomena) در این مجموعه وجود ندارند و نسبت آن‌ها صفر است.

(iii)آیا این تخمین درست است؟

این تخمین واقع‌بینانه نیست. در عمل، معمولاً حدود ۵۰٪ از واژگان یک مجموعه شامل اصطلاحاتی است که تنها یک بار ظاهر می‌شوند. این نسبت البته به نوع مجموعه و نحوه توزیع فراوانی اصطلاحات وابسته است.

(iv)دلیل اشتباه بودن این تخمین چیست؟

دلایل احتمالی برای این خطا عبارتند از:

قانون :Heap’s این قانون به‌طور کلی دقیق است و در این مورد عامل خطا نیست.

قانون Zipf: این قانون در قسمت‌های انتهایی توزیع فراوانی (یعنی برای اصطلاحات با فراوانی کم) چندان دقیق نیست. در نتیجه، این قانون عامل اصلی اشتباه در این تخمین است.

.