

الگوریتم‌های فشرده‌سازی فایل و کد هافمن

زبان اصلی مورد فهم کامپیوتر ها زبان صفر و یک است به همین علت ذخیره سازی فایل های متنی به این صورت انجام می پذیرد یعنی هر کاراکتر موجود در هر رشته به صورت یک کد (با طول دلخواه) از صفر و یک انکد می شود و پردازش به روی ان انجام می شود و پس از ان با یک دیکدینگ ساده به حالت اول خود باز می گردد.

فشرده‌سازی فایل چیست؟

فشرده‌سازی فایل فرآیندی است که با استفاده از الگوریتم‌های خاص، حجم اطلاعات را کاهش می‌دهد تا فضای کمتری اشغال کنند. دو نوع اصلی فشرده‌سازی وجود دارد:

1. **فشرده‌سازی بدون اتلاف:** اطلاعات اصلی کاملاً قابل بازیابی است. مانند ZIP، PNG، و Huffman Coding.

2. **فشرده‌سازی با اتلاف:** بخشی از اطلاعات حذف می‌شود. مانند MP3، JPEG، MPEG.

کد هافمن چیست؟

الگوریتم هافمن یک الگوریتم فشرده‌سازی بدون اتلاف است که توسط David A. Huffman در سال ۱۹۵۲ معرفی شد. این الگوریتم کاراکترهایی که بیشتر تکرار می‌شوند را با کدهای کوتاه‌تر و کاراکترهایی که کمتر تکرار می‌شوند را با کدهای بلندتر رمزگذاری می‌کند.

مراحل اجرای الگوریتم هافمن

۱. محاسبه فراوانی کاراکترها

• فایل ورودی را خوانده و تعداد تکرار هر کاراکتر را شمارش کنید.

• سپس فراوانی نسبی هر کاراکتر را محاسبه کنید.

۲. ساخت درخت هافمن

درخت دودویی‌ای به نام درخت هافمن ساخته می‌شود:

1. هر کاراکتر را به صورت یک گره (Node) با وزن برابر با فراوانی‌اش در نظر بگیرید.

2. دو گره با کمترین وزن را انتخاب و یک گره جدید با وزن مجموع آن دو ساخته و به عنوان والد آن‌ها قرار دهید.

3. این روند را ادامه دهید تا یک درخت واحد ساخته شود که ریشه آن تمام کاراکترها را پوشش می‌دهد.

۳. اختصاص کدها

- از ریشه درخت، با حرکت به چپ عدد 0 و به راست عدد 1 نسبت داده می‌شود.
- کد نهایی هر کاراکتر مسیر رسیدن از ریشه به آن گره است.
- به این ترتیب، هیچ کدی پیشوند کد دیگری نیست.

مزایای کد هافمن

- فشرده‌سازی موثر برای داده‌هایی با تکرار زیاد.
- الگوریتم ساده و سریع.
- قابل پیاده‌سازی در ساختارهای داده مانند درخت و صف اولویت.

معایب کد هافمن

- کارایی کمتر نسبت به برخی الگوریتم‌های مدرن (مثل Arithmetic coding یا LZ77).
- نیاز به دانستن فراوانی‌ها قبل از شروع فشرده‌سازی (یک‌بار خواندن فایل کامل برای محاسبه لازم است).
- در داده‌هایی با توزیع یکنواخت، کارایی قابل‌توجهی ندارد.

کاربردهای عملی

- فشرده‌سازی فایل‌های متنی (مانند .txt)
- استفاده در فرمت‌های فشرده مثل ZIP، GZIP، و PNG
- استفاده در الگوریتم‌های اولیه فشرده‌سازی تصویر و صوت

خلاصه

کد هافمن یک روش فشرده‌سازی بدون اتلاف است که بر اساس فراوانی کاراکترها، به آن‌ها کدهای دودویی با طول متغیر می‌دهد. این الگوریتم با ساخت درخت دودویی بهینه، اطمینان حاصل می‌کند که هیچ کدی پیشوند کد دیگر نیست، بنابراین فایل فشرده‌شده قابل رمزگشایی است.