به نام خالق زیبایی‌ها
فاز اول پروژه درس ذخیره و بازیابی اطلاعات
نیم‌سال دوم سال تحصیلی ۹۴ -۱۳۹۳

تعریف پروژه:

هدف از پروژه درس ذخیره و بازیابی اطلاعات تجربه عملی برخی از روش‌ها و تکنیک‌های ذخیره و بازیابی اطلاعات آموزش داده شده در کلاس است. در اولین فاز این پروژه، هر یک از گروه‌ها می‌بایست برنامه‌ای بنویسند که تعدادی عملیات ساده بر روی یک فایل داده حجیم انجام دهد. این فایل حجیم بر روی سرور Temp دانشکده قرار داده شده و مشخصات آن در پایان این سند ذکر شده است. به دلیل حجم بالای این فایل، استفاده از تکنیک‌های ذخیره و بازیابی اطلاعات به منظور افزایش سرعت انجام عملیات مختلف الزامی است. علاوه بر این، محدودیت‌های خاصی نظیر اندازه حافظه RAM مورد استفاده برنامه، به منظور ارزیابی کارآیی روش‌های توسعه داده‌شده، اِعمال خواهد شد.

کارهایی که در این فاز می‌بایست بر روی فایل ورودی انجام شوند، به شرح زیر است:

۱)خواندن فایل ورودی از ابتدا تا انتها و تعیین اندازه بهینه بافر

در موارد بسیاری لازم است فایل داده از ابتدا تا انتها خوانده شده و پردازش خاصی بر روی محتوای فایل صورت پذیرد. به عنوان مثال، از شما خواسته شده است که فایل داده مورد نظر را از ابتدا تا انتها خوانده و تعداد تکرار هر یک از حروف الفبای فارسی در این فایل را محاسبه نمایید.

در این بخش، فایل ورودی چندین بار از ابتدا تا انتها با اندازه‌های مختلف بافر خوانده شده و پردازش مورد نظر بر روی آن انجام می‌شود. منظور از بافر، حافظه‌ای است که در برنامه برای خواندن از فایل در نظر گرفته می‌شود. به طور مثال در جاوا در هنگام استفاده از کلاس DataInputStream می‌توانید تعیین کنید که در هر بار مراجعه به حافظه، چه مقدار از فایل را بخواند و در اختیار برنامه قرار دهد.

ضمن رسم نمودار زمان خواندن و پردازش کل فایل بر اساس اندازه‌ی بافر، بررسی کنید که زمان خواندن کل فایل علاوه بر اندازه بافر، به چه عوامل دیگری بستگی دارد و چگونه می‌توانید زمان خواندن کل فایل را تا حد ممکن کاهش داد. برای انجام این کار، شما باید اندازه‌ی بافر را از ۱ بایت تا ۱۰۰ مگابایت کم کم افزایش داده، و بهترین اندازه بافر را بر روی سیستم خود پیدا کنید. همین کار را بر روی دو دستگاه از سیستم‌های سایت دانشکده نیز انجام داده و گزارش تمامی کارهای انجام شده را در گزارش خود درج کنید. در گزارش باید مشخصات سخت‌افزاری سیستم خود و دو سیستم سایت را نیز به همراه جمع‌بندی خود از نتایج به دست آمده بنویسید.

۲) روش‌های مختلف خواندن فایل ورودی

با جست‌وجو در اینترنت، روش‌های مختلف خواندن فایل ورودی را در زبان برنامه نویسی مورد نظرتان، به‌دست آورده و زمان خواندن و پردازش فایل را با هرکدام از روش‌ها محاسبه و در گزارش خود گزارش کنید. (به طور مثال در جاوا این روش‌ها شامل استفاده از FileInputStream, DataInputStream, BufferedReader و... می‌باشد.)

۳)جست‌وجوی کلمات قرآن، تعطیل و سریال در فایل ورودی

تعداد زیادی مقاله در فایل ورودی وجود دارد که این مقالات با برچسب <مقاله> از یکدیگر جدا شده‌اند. دو نمونه از مقالات موجود در این فایل به صورت زیر می‌باشند:

<مقاله>

سیف الله داد معاون امور سینمایی وزارت فرهنگ و ارشاد اسلامی، دیروز در جمع خبرنگاران با بیان این مطلب افزود: همزمان با برپایی این جشنواره، سومین بازار بین المللی فیلم فجر نیز با انگیزه معرفی آثار برتر سینمای کشور به بازارهای جهانی از ۱۷ تا ۲۲ بهمن ماه با حضور بیش از ۴۵ شرکت توزیع کننده جهانی فیلم در تهران برپا می شود.

<مقاله>

در میان تمامی اساتید صاحب نام موسیقی ایران، زنده یاد ابوالحسن صبا نامی یگانه است. وی علاوه بر آشنایی عمیق با موسیقی مقامی و ردیفی ایران، در سنت ها اسیر نماند و بر نوآوری های مطابق و همذات با سنت تاکید فراوان نمود، نوشته زیر یادکردی است از زندگی و آثار این مرد عرصه موسیقی ایران که تقدیم شما گرامیان می‌گردد.

عباس مهیار استاد ابوالحسن صبا فرزند کمال السلطنه بود. ابوالقاسم کمال السلطنه، مردی ادیب و سخندان بود و به موسیقی آشنایی کامل داشت و سه تار نیک می نواخت. درباره او نوشته اند که: طبیبی حاذق بود و از دوستان نزدیک ایرج میرزا - شاعر معروف - به شمار می رفت ایرج در شعرهای خود، بارها از او یاد کرده است.

<مقاله>

در فایل داده مورد نظر، هر یک از مقالات را می‌توان به عنوان یک رکورد با طول متغیر دانست که از یک فیلد تنها حاوی دنباله‌ای درهم (بدون ترتیب خاص) از کلمات و علائم نگارشی تشکیل شده است. در این قسمت از پروژه، کلیه مقالات را از ابتدا تا انتها خوانده و هر یک از موارد زیر را برای سه کلمه قرآن، تعطیل و سریال محاسبه نمایید:

اول) هر کدام از این کلمات در چند مقاله آمده‌اند؟

دوم) هر کدام از این کلمات چند بار تکرار شده‌اند؟

سوم) بیشترین تعداد تکرار هر کدام از کلمات در یک مقاله چند بار بوده است؟

همچنین زمان انجام عملیات فوق را محاسبه کرده و در گزارش خود به طور مشخص و دقیق ارائه دهید. برای خواندن فایل ورودی در این قسمت و قسمت‌های بعدی باید از اندازه‌ی بافر بهینه‌ای که در قسمت اول به‌دست آورده‌اید، استفاده کنید. در تمامی مراحل باید سعی کنید که روشی را پیاده سازی کنید که بهترین زمان پاسخ و کمترین استفاده از حافظه را داشته باشد.

۴)خواندن فایل ورودی و مرتب کردن هر مقاله به صورت مجزا

در بخش‌های قبل، عملیات مورد نظر بر روی فایل صرفا شامل خواندن فایل بود. در این بخش پردازش فایل شامل خواندن فایل و تولید یک فایل جدید است که طبیعتا شامل هر دو عملیات خواندن و نوشتن فایل می‌شود.

فرض کنید قرار است فایل ورودی مورد نظر را از ابتدا تا انتها خوانده و کلمات موجود در هر یک از مقالات را بر اساس حروف الفبا مرتب کنید. در زیر نمونه ای از یک فایل ورودی و خروجی مربوط به آن نمایش داده شده است.

ورودی:

<مقاله>

سیف الله داد معاون امور سینمایی وزارت فرهنگ و ارشاد اسلامی

<مقاله>

خروجی:

<مقاله>

ارشاد اسلامی الله امور داد سیف سینمایی فرهنگ معاون و وزارت

<مقاله>

همانند بخش های قبل زمان انجام عملیات فوق را محاسبه کنید، زمان خواندن، نوشتن و انجام عملیات مرتب‌سازی را جدا گانه محاسبه کرده و در گزارش خود درج کنید. تاثیر اندازه بافرهای ورودی و خروجی را بر

روی زمان انجام عملیات بررسی کنید و روش‌های پیشنهادی خود برای افزایش سرعت انجام عملیات را پیاده‌سازی و تاثیر آنها را بررسی و گزارش نمایید.

مواردی که باید تحویل داده شوند:

۱) کد کامل برنامه
۲) نسخه اجرایی به همراه فایل شرح نحوه اجرای بخش‌های مختلف برنامه
۳) گزارش پروژه که مهم‌ترین بخش پروژه بوده و در صورت ناقص بودن این بخش، از نمره‌ی شما کسر خواهد شد.

نکات مهم در پیاده‌سازی:

- فایل ورودی بر روی Temp در پوشه Information Retrieval قرار گرفته است. حجم این فایل در حدود ۲ گیگابایت است.
- فرمت کد گذاری فایل ورودی UTF8 می‌باشد که به هنگام خواندن فایل در برنامه باید به این نکته توجه کنید.
- هرچه تعداد روش‌های آزمایش‌شده و دقت انجام کار بیشتر باشد، نمره‌ی شما نیز افزایش خواهد یافت.
- در تمامی فازهای پروژه، باید میزان استفاده از حافظه اصلی[1] را کنترل کنید. اجازه استفاده بیشتر از ۱۲۸ مگابایت از حافظه را ندارید.
- هر چه سرعت اجرای پروژه بیشتر باشد، بهتر است و نمره بیشتری خواهید گرفت.
- آخرین مهلت ارسال این فاز پروژه روز دوشنبه ۱۸ اردی‌بهشت ۱۳۹۴ می‌باشد. این تاریخ به هیچ عنوان تمدید نخواهد شد. لذا در زمان‌بندی انجام پروژه دقت لازم را داشته باشید.
- زمان تحویل حضوری این پروژه متعاقباً اعلام خواهد شد.

موفق باشید

[1] RAM