به نام خالق یکتا

فاز دوم پروژه درس ذخیره و بازیابی اطلاعات

نیم‌سال دوم سال تحصیلی ۹۴ -۱۳۹۳

تعریف پروژه:

در این مرحله از پروژه فایل را با اندازه‌ی بافر بهینه‌ی به‌دست آمده در مرحله اول پروژه بخوانید و عملیات زیر را بر روی آن انجام دهید. توجه داشته باشید که به دلیل محدودیت اندازه RAM مورد استفاده باید بهینه‌سازی‌های لازم را در برنامه خود اِعمال کنید.

در کلیه بخش‌های پروژه، پارامترهای مهم نظیر زمان انجام عملیات مربوطه توسط برنامه، اندازه حافظه جانبی مورد نیاز برای ذخیره فایل(های) خروجی تولید شده و ایده‌های استفاده شده به منظور کاهش حافظه مورد نیاز و یا افزایش سرعت انجام عملیات را در گزارش خود ذکر نمایید.

۱) حذف کلمات متداول

در این بخش، شما باید کلمات متداول زبان فارسی (Stop Words) را از متن مقالات حذف کنید. فهرستی از کلمات متداول زبان فارسی را می‌توانید در فایل پیوست شده پیدا کنید. برای نگهداری مجموعه کلمات متداول در حافظه از ساختمان‌داده‌های مختلفی نظیر توابع درهم‌سازی (Hash)، مجموعه (Set)، درخت (Tree) و ... استفاده کرده و تاثیر ساختمان‌داده‌ی استفاده شده بر سرعت انجام عملیات مربوطه را بررسی و گزارش نمایید.

همچنین شما باید ۲۵ کلمه متداول دیگر -که مختص این فایل است- را نیز بیابید و آن‌ها را نیز حذف کنید. دلایل انتخاب این کلمات را نیز کاملاً شرح دهید. به عنوان مثال می‌توان کلماتی که در تمامی مقالات وجود داشته و در هر کدام از این مقالات حداقل ۵ بار تکرار شده اند را یک کلمه متداول گرفت و آن را حذف کرد. شما باید تعداد تکرار کلمات در مقالات مختلف را محاسبه کرده و با توجه به تکرار کلمات الگوریتم برای پیدا کردن این ۲۵ کلمه متداول ارائه کنید.

تاثیر حذف کلمات متداول در کاهش اندازه فایل پردازش شده را بررسی و گزارش نمایید. همچنین در گزارش خود زمان مورد نیاز برای انجام عملیات حذف کلمات متداول از فایل ورودی و ایده‌هایی که در این

بخش استفاده کرده‌اید را ذکر نمایید. در کنار این موارد لیست ۲۵ کلمه متداولی که حذف کردید را نیز در گزارش خود قرار دهید.

۲) ریشه‌یابی کلمات

همان‌طور که می‌دانید به منظور کم کردن اندازه فرهنگ لغات و همچنین پاسخ‌گویی بهتر و سریع‌تر به جستجوی کاربر، به جای نگه‌داری کلیه شکل‌های مختلف یک کلمه، سعی می‌شود تمام این کلمات با ریشه اصلی کلمه مربوطه جایگزین شوند.

الگوریتم‌های مختلفی برای ریشه‌یابی کلمات انگلیسی وجود دارد. از جمله این الگوریتم‌ها می‌توان به Porter، Lovins، Paice/Husk، Snowball و... اشاره کرد. اما همچنان الگوریتم مناسب و کاملی برای ریشه‌یابی کلمات فارسی ارائه نشده است. در این مرحله شما باید با استفاده از قواعدی که در زیر ارائه شده است کلمات فارسی (فقط افعال فارسی) را ریشه‌یابی کرده و در انتها به جای خود کلمه ریشه آن کلمه را در فایل مقصد ذخیره کنید.

قواعدی که برای ریشه‌یابی افعال فارسی وجود دارد به شرح زیر می‌باشد:

ماضی ساده: بن ماضی+ {م,ی؛یم,تد,ند}

ماضی استمراری: می + ماضی ساده

ماضی نقلی: بن ماضی + ه + {ام, ای, است, ایم, اید, اند }

ماضی بعید: بن ماضی+ه+بود+ {م,ی,یم,ید,ند}

ماضی التزامی: بن ماضی+ ه+باش + {م,ی,یم,ید,ند}

مضارع اخباری: می + بن مضارع + {م,ی,د,یم,ید,ند}

مضارع ساده: ب + بن مضارع + {م,ی,د,یم,ید,ند}

مضارع ساده: منفی ن + بن مضارع + {م,ی,د,یم,ید,ند}

فعل امر: ب + بن مضارع+ {یم,ید}

فعل امر: منفی ن + بن مضارع+ {یم,ید}

دیگر افعال منفی: ن{+ ماضی ساده ,ماضی استمراری ,ماضی نقلی,ماضی بعید ,ماضی التزامی,مضارع اخباری}

به عنوان یک بخش امتیازی، می‌توانید برای کلمات فارسی (غیر از فعل‌ها) الگوریتمی را ارائه و پیاده‌سازی کنید.

۳) ایجاد فرهنگ لغات

در این قسمت از پروژه شما باید ساختمان‌داده‌ای پیاده سازی کرده که در آن بتوانید کلمات فارسی را ذخیره‌سازی کنید. این ساختمان‌داده باید از لحاظ زمان و حجم حافظه مصرفی در بهینه‌ترین حالت ممکن باشد. برای ایجاد فهرست کلمات می‌توانید از ساختمان‌داده‌هایی مانند Set، Hash، Trie و... استفاده کنید. فرهنگ لغت پیاده‌سازی شده باید قابلیت جست‌وجوی سریع کلمات را داشته باشد و در صورتی که کلمه جستجو شده در فرهنگ لغت وجود نداشته باشد، باید قابلیت اضافه کردن کلمه به فرهنگ لغت وجود داشته باشد. فرهنگ لغت پیاده‌سازی شده باید اطلاعات مربوط به هر کلمه را نیز درون خود داشته باشد. این اطلاعات عبارتند از شماره شناسه یکتایی که به هر کلمه اختصاص داده می‌شود و تعداد تکرار کلمات در تمامی اسناد. از ساختمان‌داده‌های مختلف برای ایجاد فرهنگ لغت استفاده کرده و کارایی آن‌ها را از نظر حافظه مورد نیاز، سرعت ساخت (درج لغات) و سرعت استفاده (جستجوی لغات) با یکدیگر مقایسه کنید و تمامی مراحل و کارهای انجام شده را در فایل گزارش خود درج کنید.

در این بخش، برای تست این فرهنگ لغات پیاده‌سازی شده‌ی شما، تنها با جستجوی کلمات در آن‌ها کار می‌شود و باید فرهنگ لغات شما در صورت وجود کلمه جست‌وجو شده، تعداد تکرار آن در کل فایل را بازگرداند.

۴) ایجاد شاخص سند-لغت

در هنگام پیمایش هر سند، شناسه لغات آن سند و تعداد تکرار هر کلمه در آن سند را به‌دست آورده و به صورت برداری برای آن سند ذخیره کنید.

مواردی که باید تحویل داده شوند:

۱) کد کامل برنامه
۲) نسخه اجرایی به همراه فایل شرح نحوه اجرای بخش‌های مختلف برنامه (در فایلی با نام Readme.txt)
۳) گزارش پروژه

نکات مهم در پیاده‌سازی:

- در تمامی بخش‌های پروژه باید میزان استفاده از حافظه اصلی استفاده شده کمتر از ۱۲۸ مگابایت باشد. هر چه میزان بیشتری از این مقدار استفاده کنید نمره کمتری خواهید گرفت و استفاده کمتر از حافظه نمره اضافی دارد.
- هر چه سرعت اجرای پروژه شما بیشتر باشد بهتر است و نمره بیشتری خواهید گرفت.
- به ازای هر روز تاخیر ۱۵ درصد از نمره پروژه کسر خواهد شد. به تاخیر بیشتر از ۷ روز نمره‌ای تعلق نخواهد گرفت.

مهلت ارسال:

- مهلت ارسال این مرحله از پروژه روز ۸ خرداد ۱۳۹۴ خواهد بود.
- تاریخ ذکر شده به هیچ عنوان تمدید نخواهد شد و زمان تحویل حضوری آن متعاقبا اعلام خواهد شد.

موفق باشید