هوالحق

آشنایی با الگوی نگاشت–کاهش و کتابخانه MrJob

هدف از این تمرین، آشنایی با الگوی نگاشت-کاهش (MapReduce) و تجزیه تسک‌های پردازشی توسط این استراتژی می‌باشد. در این تمرین از کتابخانه پایتون MrJob برای توسعه و اجرای برنامه‌های MR (به صورت محلی) استفاده خواهیم کرد.

لطفاً تمامی مراحل در محیط سیستم‌عامل Ubuntu 20.04 انجام گردد. همچنین در این تمرین از زبان برنامه نویسی پایتون با نسخه مفسر Python3.8 و یا بالاتر استفاده کنید.

آشنایی با کتابخانه MrJob

در این بخش لازم است تا با کتابخانه MrJob آشنایی کافی را پیدا کرده و مستندات و نمونه کدهای آن را مطالعه کنید. با مطالعه این صفحه می‌توانید با نحوه کارکرد MrJob آشنا شوید.

در صورتی که با Iteratorها و Generatorها در زبان پایتون آشنایی ندارید، لازم است تا با جستجو در منابع آنلاین از نحوه کارکرد آن‌ها مطلع شوید. به ازای هر برنامه نوشته شده باید قادر باشید تا تمامی بخش‌های آن را تشریح کرده و تسلط کامل بر فرآیند مراحل نگاشت و کاهش را داشته باشید.

سؤال ۱- شمارش کلمات متن

در این قسمت قصد داریم تا کلمات کتاب «جنگ و صلح» اثر تولستوی را بشماریم. فایل WarAndPeace.txt شامل متن کامل این کتاب به زبان انگلیسی می‌باشد.

بخش ۱

با استفاده از الگوی نگاشت-کاهش یک برنامه با زبان پایتون و با استفاده از کتابخانه MrJob بنویسید که ۲۰ کلمه پر تکرار را به همراه تعداد تکرار آن‌ها به صورت نزولی نمایش دهد. دقت کنید که در این بخش نیاز است تا تمامی پیش‌پردازش‌های متنی لازم را انجام دهید (مانند یکسان‌سازی حروف بزرگ و کوچک، حذف علامات گرامری و ...).

بخش ۲

ایست واژه‌ها (StopWords)، کلماتی پر تکرار در متن هستند که دارای معنای خاصی نبوده، به مفهوم و منظور متن چیزی را نمی‌افزایند، ولی جزو ارکان تشکیل دهنده جملات و جزو اصول گرامر می‌باشند. به

همین علت در اکثر پردازش‌های متنی، حذف ایست‌واژه‌ها از کلمات متن مطلوب است. به طور مثال کلماتی مانند «The» و یا «is» ایست‌واژه محسوب می‌گردند.

یک لیست از ایست‌واژه‌های مناسب برای متن کتاب جنگ و صلح تهیه کنید و برنامه نگاشت-کاهش شمارش کلمات که در بخش ۱ نوشته‌اید را به گونه‌ای تغییر دهید که این کلمات از لیست خروجی حذف شوند. نتیجه را با خروجی بخش ۱ مقایسه کنید.

سؤال ۲– پردازش لاگ‌های وب‌سرور Apache

یک وب‌سرور سرویس‌دهنده از نوع Apache یک فایل لاگ شامل اطلاعات درخواست‌های کاربران را ذخیره کرده است. فایل فشرده شده این لاگ با نام access_log.gz در اختیار شما قرار داده شده است. می‌خواهیم اطلاعات آماری مورد نیاز را از این فایل استخراج کنیم؛ برای این کار از الگوی نگاشت-کاهش استفاده خواهیم کرد. استفاده از ایده‌های Top-N، Filtering، Binning و ازین نظیر که در الگوهای کلان داده مطرح شده است می‌تواند در این بخش راهگشا باشد.

بخش ۱

۱۰ پر درخواست‌ترین عکس‌های درخواست شده (به همراه تعداد درخواست آن‌ها)، به ازای هر سال را بیابید.

بخش ۲

آمار ماهیانه بازدید کاربران را بیابید. بازدید را به این صورت تعریف می‌کنیم: «هر بازدید عبارت است حداقل یکبار دسترسی یک کاربر در یک روز».

سپس سعی کنید با استفاده از یک کتابخانه گرافیکی در پایتون، یک نمودار ستونی که نشان‌دهنده بازدیدهای ماهانه کاربران باشد را به تصویر بکشید.

بخش ۳

می‌خواهیم ۱۰ کاربر با بیشترین مدت زمان سپری شده در سایت را به ازای هر سال بیابیم. مدت زمان سپری شده یک کاربر در سایت در لاگ سرور به صورت صریح بیان نشده‌است، برای همین می‌بایست یک روش خلاقانه برای محاسبه میزان زمان سپری شده حضور یک کاربر در سایت را ارائه دهید. بدیهی است که روش شما می‌بایست یک مبنای هیوریستیک واقع‌گرایانه داشته باشد.

سؤال ۳- پردازش فایل XML

یک فایل XML شامل بخشی از پرسش و پاسخ کاربران سایت StackOverflow در اختیار شما قرار داده شده است (فشرده شده با نام Posts.zip). می‌خواهیم با استفاده از الگوی نگاشت-کاهش، موارد زیر را بیابیم.

بخش ۱

۱۰ کاربر با بیشترین تعداد پست شامل کلمه socket را بیابید.

بخش ۲

۱۰ کلمه پر تکرار در تمامی پست‌های حاوی کلمه socket بیابید.

بخش ۳

میخواهیم کاربران پر حرف در سیستم را پیدا کنیم و برای آن‌ها محدودیت طول سؤال بگذاریم. ده کاربر پر حرف سایت را پیدا کنید.

سؤال ۴- پردازش مقدماتی گراف

یک فایل شامل اطلاعات ارسال پیامک بین چند شماره مختلف به شما داده شده است (فشرده شده با نام CDR.zip). اطلاعات زیر را با استفاده از الگوی نگاشت-کاهش از فایل مربوطه استخراج کنید.

بخش ۱

۱۰ شماره با بیشترین تعداد پیامک ارسالی را بیابید.

بخش ۲

۱۰ شماره با بیشترین تعداد پیامک دریافتی را بیابید.

بخش ۳

به ازای هر شماره، تعداد پیامک‌های دریافتی و ارسالی را بیابید (به صورت همزمان).