تمرین میانترم تحلیل اکتشافی داده‌ها

محمدمهدی خانی

۹۷۲۴۳۰۲۷

توابع

برای نوشتن این برنامه از چندین تابع استفاده شده است و هر بخش با توجه به عملکردی که داشته یک تابع جداگانه شده که در پایین عملکرد هر کدام مختصر توضیح داده شده است.

fileHash

این تابع فایل دریافتی را هش میکند و هش تولیدی را در خروجی برمیگرداند.

makeDate

این تابع تاریخ میلای مربوط به تولید فایل را استخراج میکند و آن را تبدیل به یک رشته با فرمت YYYYMMDD از تاریخ شمسی میکند.

makeName

این تابع نام جدید هر فایل را با توجه به مقدار counter و تاریخ شمسی و اسم قبلی فایل میسازد.

clean

این تابع اسم قبلی فایل را فیلتر میکند و کارکتر های خاص و رقم ها را از آن حذف میکند.

isDuplicate

محتوای دو فایل ورودی را دقیقا چک میکند که آیا با هم برابرند یا خیر.

process

یک آدرس فولدر به عنوان ورودی میگیرد و بعد از استخراج تمامی فایل های دارای فرمت jpg. و png. به صورت بازگشتی، تابع processFile را برای هر یک از فایل های استخراج شده فراخوانی میکند.

ساختمان داده

ساختمان داده های مورد استفاده در این برنامه دو دیکشنری هستند که کلید های یکی از دیکشنری ها یک رشته (تاریخ) است و مقدار های این دیکشنری تعداد عکس های یکتای ثبت شده در آن تاریخ تا قبل از پردازش عکس فعلی است. با این دیکشنری میتوان شماره ٤ رقمی مربوط به هر عکس را ساخت. کلید های دیکشنری دیگر آدرس های هش شده هستند و مقدار های این دیکشنری خود رشته ی آدرس ها هستند. در هنگام پردازش یک عکس، با استفاده از این دیکشنری چک میکنیم که آیا هش عکس فعلی با هش های درون این دیکشنری مطابقت دارد یا خیر و در صورتی که با یکی از هش ها مطابقت داشت از طریق هش به مقدار دیکشنری (آدرس فایل) دسترسی داریم و میتوانیم مقایسه دقیق دو عکس را انجام دهیم.

روند پردازش

ابتدا ساختمان داده ها خالی هستند و بعد از اینکه به صورت بازگشتی تمامی فایل های jpg. و png. در دایرکتوری داده شده پیدا شدند، یکی یکی این فایل ها بررسی میشوند و در صورتی که مشابه فایلی که در حال بررسی آن هستیم قبلا بررسی شده بود (فایل های Duplicate) از کپی کردن این فایل به فولدر مقصد خودداری میکنیم و در غیر این صورت برای فایل یک نام مناسب با فرمت توضیح داده شده درست میشود و در فولدر مقصد کپی میشود.