دانشکده مهندسی کامپیوتر

دانشگاه صنعتی شریف

استاد درس: دکتر حمیدرضا ربیعی

بهار ۱۴۰۰

تمرین در خانه پنجم

درس یادگیری ماشین آماری

نام و نام خانوادگی: امیر پورمند

شماره دانشجویی: ۹۹۲۱۰۲۵۹

آدرس ایمیل : **pourmand1376@gmail.com**

۱ سوال ۱

۱.۱ ضرب ماتریس در بردار

خب قسمت اول این سوال ضرب ماتریس در بردار است که باید به دو بخش map و reduce آن فکر شود. فرض کنیم M یک ماتریس $n \times n$ است که سطر i ام و ستون j ام آن با m_{ij} نشان داده میشود. بردار را هم به خاطر کلمه vector با v نشان میدهیم که v_i به معنای المان i ام بردار است.

مشخص است که خروجی ضرب ما به این شکل خواهد بود:

$$x_i = \sum_{j=0}^{n} m_{ij} v_j$$

تابع map : این تابع در واقع کار ضرب سطر i ام ماتریس در المان j ام بردار را برعهده دارد که میتوان هر المان آن را با $m_{ij}v_i$ نشان داد. یعنی در واقع تابع map زوج مرتب هایی به شکل $(i, m_{ij}v_i)$ را تولید خواهد کرد.

تابع reduce : این تابع در واقع تولید عنصر x_i را بر عهده دارد که به سادگی میتواند عناصری که در روش قبل با کلید i مشخص شده اند را با یکدیگر جمع کند تا خروجی مطلوب معادل فرمول بالا حاصل شود. البته لازم به ذکر است که اگر ماتریس اولیه را در حالت کلی تر یک ماتریس مستطیلی فرض کنیم فرق خاصی ندارد و همه عبارتها برقرار هستند.

۲.۱ ضرب ماتریس در ماتریس

خب ابتدا دو ماتریس M و N را داریم که مثلا ماتریس P ضرب آنها می شود که آن را به شکل زیر نمایش میدهم.

$$P = MN$$

حال اگر هر المان ماتریس m را با m_{ij} نشان دهیم و هر المان ماتریس N با n_{jk} نشان دهیم میدانیم هر المان ماتریس P به شکل زیر عبارت خواهد بود با:

$$p_{ik} = \sum_{j=0}^{n} m_{ij} n_{jk}$$

خب پس میتوانیم توابع مورد نیاز را تعریف کنیم. البته دقت داشته باشیم این سوال را میتوان با دو بار map و reduce هم حل کرد ولی روش ساده تر آن است که با استفاده از تکنیکی دو بار را به یک بار تقلیل دهیم که به شرح زیر است.

تابع map : ابتدا به ازای هر المان ماتریس M تمام جفت های $((i,k),(M,j,m_{ij}))$ را به ازای k های از ۰ تا تعداد ستون های ماتریس N تولید کن. سپس به ازای هر المان ماتریس N یعنی n_{jk} باید تمام زوج های $((i,k),(N,j,n_{jk}))$ را تولید کرد که مشخص است که N و M در اینجا خود ماتریس ها نیستند بلکه صرفا یک فلگ هستند که مشخص شود که المان مورد نظر به کدامین ماتریس تعلق دارد.

تابع reduce : در این تابع ابتدا یک مرتب سازی انجام میشود که مقادیری که مقدار j یکسان دارند با فلگ های یکسان M و N کنار هم قرار گیرند و در اینجا با هم ضرب شده و سپس با همه جمع شوند که نتیجه نهایی به ازای سطر و ستون مورد نظر بدست آید.

۲ سوال دوم

اگر بخواهم دو مثال را بررسی کنم اولین مثال در مورد کاربرد این مسئله در گراف است که تابع map reduce قادر به حل کردن آن نیست. علت این امر آن است که این روش نمیتواند به صورت خیلی موثر داده های وابسته را مدل سازی کند پس از حل اینگونه مسائل بر نمی آید.

و دومین مسئله نیز آن است که این روش توابع iterative را به خوبی مدل سازی نمیکند زیرا در هر مرحله فقط قسمتی از دیتا نیاز به محاسبه دارد که قطعا سربار زیادی دارد.

۳ سوال سوم

گرفتن داده های سایتی مانند دیجی کالا که یک تسک ماهانه است میتواند با هر دو انجام شود. البته به شرطی که scheduler هدوپ را داخل خود آن در نظر بگیریم در صورتی که اکثر سایت ها گفته اند هدوپ زمان بند داخلی ندارد و باید از زمان بند خارجی استفاده شود. به هر حال اگر موضوع بالا را در نظر نگیریم تفاوت زیادی بین دو روش وجود ندارد.

پردازش داده های توییتر نیز بستگی به نوع پردازش دارد و این که شبکه را از چه نوعی در نظر بگیریم اگر صرفا جنبه متنی توییت ها مد نظر باشد روش هدوپ بهتر است زیرا در batch processing بهتر کار میکند ولی اگر جنبه گرافی توییتر مد نظر باشد قطعا اسپارک خیلی بهتر خواهد بود زیرا هدوپ نمیتواند با داده های گرافی به خوبی کار کند.

در حالت سوم نیز اسپارک بهتر است زیرا پردازش های real-time که با سرعت بالا قابل انجام باشد را اسپارک خیلی بهتر از هدوپ میتواند انجام دهد.