به نام او

تمرین سری اول درس پردازش داده‌های حجیم

تاریخ تحویل: ۱۸ آبان

۱- نگاشت-کاهش[1] یک مدل برنامه‌نویسی برای پردازش کلان‌داده‌ها به شکل موازی و توزیع‌یافته است که با استفاده از آن بسیاری از الگوریتم‌ها به شکل مقیاس‌پذیر پیاده‌سازی شده‌اند. البته مدل نگاشت-کاهش محدودیت‌هایی هم دارد و با آن نمی توان همه‌ی مسائل را به صورت بهینه حل کرد.

مدل نگاشت-کاهش برای کدام یک از مسائل زیر راه‌حلی بهینه است؟ توابع نگاشت و کاهش را برای آن مسائل تعریف کنید.

- دنبال کردن عملکرد یک ماشین حالت
- برنامه‌ای که یک متن بگیرد و مشخص کند به ازای طول‌های مختلف چه تعداد کلمات با آن طول در متن هستند.

مثال: ورودی: it is rainy

خروجی: یک کلمه با طول ۵ و دو کلمه با طول ۲

- مساله فروشنده دوره گرد
- محاسبه‌ی مکعب یک ماتریس n*n

۲- الگوریتم خوشه‌بندی DBSCAN را با مدل برنامه‌نویسی نگاشت-کاهش پیاده‌سازی کنید.

- برای این منظور از کتابخانه ای که DBSCAN را در اسپارک یا هدوپ پیاده کرده باشد استفاده نفرمایید.
- الگوریتمی را که استفاده نموده اید کمی توضیح دهید (شبه کد تابع map و reduce را در فایل گزارش بیان نمایید.)
- ** (اگر ایده ای برای بهبود آن دارید به صورت واضح بیان فرمایید. – ایده خود را پیاده سازی نمایید و نتایج آن را با روش اصلی مقایسه نمایید.) –این بخش نمره اضافه دارد-
- مقادیر مناسبی برای پارمترهای DBSCAN انتخاب نمایید و نحوه انتخاب آنها را توضیح بفرمایید.
- پیچیدگی زمانی الگوریتم را تا حد ممکن (بر اساس پارامترهایی نظیر تعداد نمونه ها،تعداد توابع map و reduce) تحلیل نمایید.
- نمودار Spead up (زمان اجرا بر حسب تعداد توابع map) را رسم نمایید.
- NMI خوشه بندی خود را گزارش نمایید.

الگوریتم را روی مجموعه داده Letter Recognition آزمایش نمایید.

https://archive.ics.uci.edu/ml/datasets/letter+recognition

---

[1] Map-Reduce

۳- هدف این تمرین پیاده سازی KNN با استفاده از LSH است.

مجموعه داده patches را در نظر بگیرید که شامل 59500 تصویر ۲۰*۲۰ است. با استفاده از LSH تنها می توانید ۲۰ بار داده ها را hash نمایید. (نحوه باند بندی و استفاده از and و or به عهده شماست تا بهترین جواب را به دست آورید.)

سپس از این hash ها استفاده نمایید و K نزدیکترین تصویر به هر تصویر را بیابید.

در این تمرین از فاصله اقلیدسی استفاده بفرمایید.

- الگوریتم خود را مختصرا توضیح دهید.
- مدت زمان اجرای الگوریتم را برای KNN با استفاده از LSH را با KNN ساده مقایسه نمایید.
- برای K=5 و K=10، میزان خطای روش KNN-LSH (نسبتی ازk همسایه که با روش KNN برابر نیستند) را به دست آورید.
- برای دو مورد از تصاویر (شماره ۲۰۴۶ و ۱۱۷۱) ۵ تصویر نزدیک با Knn و ۵ تصویری که با KNN-LSH یافته اید را رسم نمایید.

۴- به یکی از دو سوال زیر را پاسخ دهید.

۱- Kernelized locality sensitive hash چیست؟
۲- یک تابع locality sensitive hash مناسب برای فاصله مربع کای ارائه نمایید.

* فایل pdf مربوط به پاسخ تمرین را به همراه کدهای مربوطه در یک فایل فشرده با نام شماره دانشجویی خود قرار داده و در مودل بارگزاری کنید.