سوال ۳

در این مساله ابتدا ۲۰ خط تصادفی در نظر میگیریم سپس هر نقطه را روی آن ها تصویر میکنیم و در نتیجه با فاصله بدست آمده از مبدا باکت مورد نظر برای هر خط را پیدا می کنیم با توجه به این که دیتا داده شده ۴۰۰ بعد دارد این دیتا به ۲۰ بعد کاهش میابد

کافی است برای شباهت سنجی دو نقطه این ۲۰ سطر از ماتریس را باند بندی کنیم که بهترین نتیجه با b= 4, r=5 بدست امد یعنی اگر در هر کدام ( حد اقل) یکی از باند ها همه سطر ها با هم یکسان بود این دو را کاندید شباهت اعلام میکنیم

در روبرو نتیجه به ازای k=5 ,k=10 برای KNN , KNN LSH نشان داده شده است 

تصاویر واقعی : 

در این الگوریتم چون میخواست تمام k نزدیک ترین را پیدا کند کامپیوتر من نتوانست ( عملا نکشید !!) چون O(n^2) هزینه می برد

به علاوه با اینکه ما کاهش بعد دادیم اما پیدا کردن تمامی سطر های یکسان در حد اقل یک باند باز هم به O(n^2) نیاز دارد که من برای این که جواب بگیریم ( من به صورت نرمال گذاشتم ۲ ساعت کامپیوتر ران بود و لی هنوز داشت ران میشد ) به محض این که K تا ی مورد نظر برای یک نقطه را پیدا کردیم break میکنیم و سراغ نقطه بعدی میریم ( کار دیگر این بود که همه سطر های یکسان در حد اقل یک باند را پیدا کنیم و K تای مینیمم را از آنها برداریم ولی برای زمان این خطا را وارد مساله کردیم )

که البته با توجه به تصاویر رسم شده شباهت مناسبی با تصویر اصلی دارند

خطا ها طبق چیزی که سوال خواسته :‌



این خطا ها شامل تعداد تصاویری هستند که الگوریتم KNN LSH با KNN واقعی تفاوت دارد که با توجه به تقریبی که سریع تر پاسخ گرفتن وارد مساله کردیم کاملا طبیعی است

به نظرم اگر بخواهیم خطا را بهتر بسنجیم و این دو الگوریتم را با هم مقایسه کنیم باید معیار های شباهت دیگری بین خود تصاویر تعریف کنیم همان طوری در تصاویر آمده تقریبا با جواب واقعی شبیه اند

فرض کنید معیار شباهت را کانولوشن بگیریم شاید تصویری که ما گزارش میکنیم با آن حالتی که همه شبیه ها را بدست میاوردیم و سپس K تا مینیمم را برمیداشتیم تفاوت خیلی کمی داشته باشد و مقداری ناشبیه تر باشد و در جواب بیاید در این صورت خطایی که سوال میخواسته بالا میرود ولی معیار خطای جدیدی که تعریف کردیم اندکی تغیر میکند و با نتایج عکس ها همسو تر است

یا به تعبیری دیگر معیار خطای سوال پافشاری روی افزایش تعداد نقاط یکسان بین دو الگوریتم میکند و نقاطی که ممکن است مقدار تفاوت کمی از لحاظ شباهت با این نقاط یکسان را داشته باشند را در نظر نمیگیرد و باعث می شود الگوریتم هزینه زمانی بپردازد

اما معیار خطای کانولوشن این نقاط را در نظر میگیرد و زمان پاسخ دهی را بهبود میدهد