درس کلان دادگان

ملیحه غزالیان

830402056

گزارش تمرین 1:

سوال 1.

در این سوال ابتدا متن ها از فاصله اضافی و کاراکتر ها و استاپ ورد ها پاک شدند (در فایل preprocess) و سپس بعد و قبل حذف استپ ورد ها تعداد کلمات با مپ ردوس با اسپارک اندازه گرفته شد و 20 کلمه برتر معرفی شدند. همچنین برای محاسبه اندیس وارونه انجام شد. (در فایل wordCount)

سوال 2.

در این سوال برای محاسبه پیج رنک ابتدا در فایل preprocess دیتا خوانده شده و نود هایی که لینک خروجی ندارند (دنگلینگ نود ها) شناخته شد و در نظر گرفته شد و در فایل pagerank تمام نود ها یک مقدار اولیه رنک داده شده و الگوریتم در اسپارک با ساختار مپ ردوس نوشته شده است. الگوریتم با دمپینگ فاکتور یا همان آلفا 0.85 اجرا شده است. و در نهایت رنک نهایی نود ها با تکرار 10 محاسبه شده و در فایل متنی با نام final_ranks ذخیره شدند.

10 نود با بالاترین رنک:

Page: 41909 PageRank: 0.0005089770118389987

Page: 597621 PageRank: 0.0004643397628564387

Page: 504140 PageRank: 0.00045573071171432913

Page: 384666 PageRank: 0.000448578294179169

Page: 537039 PageRank: 0.0004383960561316232

Page: 486980 PageRank: 0.00043640513591067167

Page: 751384 PageRank: 0.00041316451455076917

Page: 32163 PageRank: 0.0004127175181650983

Page: 163075 PageRank: 0.0004087180742310319

Page: 605856 PageRank: 0.00040720779003414307

سوال 3.

در این سوال تصاویر اول فلت شدند که بتوان با هم مقایسه کرد و به صورت یک وکتور ذخیره شدند. بعد در dask تصاویر فلت شده خوانده شده و به 10 پارتیشن تقسیم شده و به هر تصویر در هر پارتیشن به تعداد 9 کلید نصبت داده شد تا بتوان با بقیه تصاویر در پارتیشن های دیگر مقایسه شوند بعد دیتا فریم جدید با datafram.reduction قصد داشتم که با کلید آنها را گروه بندی کنم و هر گروه را با هم دو به دو مقایسه کنم که متاسفانه ارور داشت و نتوانستم ارور را برطرف کنم.

سوال 4.

پیدا کردن دوستان مشترک در اسپارک انجام شد و افراد با بیشترین دوست مشترک برای هر فرد در فایل متنی ذخیره شد

سوال 5.

تکنیک مین هش اجرا شد و برای تعداد 50 و 100 و 200 هش فانکشن کاندیداها مشخص شده و بر اساس آن ها fp و fn با thresholdهای گفته شده بدست آمد برای شباهت بالا 0.6 جفتهای با شباهت دقیق: [(328, 788), (408, 898), (489, 587)]

برای 50 :

False Positive: 4336

False Negative: 7

برای 100:

False Positive: 2269

False Negative: 6

برای 200:

False Positive: 1859

False Negative: 6

تکنیک بندینگ سبب کاهش fp و fn میشود که نتایج بعد از بندینگ:

برای 50 :

False Positive: 4

False Negative: 2

برای 100:

False Positive: 26

False Negative: 2

برای 200:

False Positive: 26

False Negative: 2

افزایش تعداد hf از 50 به 100 سبب افزایش fp شده است.

سوال 6.

الگوریتم simhash برای دو متن دلخواه (متن ها از تمرین اول هستند) اجرا شده و فاصله ی دو متن با فاصله همینگ بدست آمد.

سوال 7.

ابتدا lsh روی یک دیتاست ساختگی در فضای فاصله همینگ اجرا شده (فایل q7) و از کاندیدا های بدست آمده برای یک دیتا پوینت پرسش دلخواه به تعداد 30 عدد انتخاب شد و سه نقطه با کمترین فاصله از آن بدست آمد.

سپس فایل lsh.py تکمیل شد و جهت بدست آوردن موارد سوال اجرا شد:

3 تا از نزدیک ترین همسایه ها برای ردیف های گفته به قرار زیر است:

Nearest Neighbors for row 100 in LSH: [7551, 8196, 28351]

Nearest Neighbors for row 100 in linear search: [7551, 8196, 28351]

Nearest Neighbors for row 200 in LSH: [91, 604, 52040]

Nearest Neighbors for row 200 in linear search: [91, 604, 1888]

Nearest Neighbors for row 300 in LSH: [15818, 26600, 10830]

Nearest Neighbors for row 300 in linear search: [15818, 22057, 9006]

Nearest Neighbors for row 400 in LSH: [28676, 33010, 5875]

Nearest Neighbors for row 400 in linear search: [28676, 33010, 5875]

Nearest Neighbors for row 500 in LSH: [1178, 557, 35904]

Nearest Neighbors for row 500 in linear search: [1178, 557, 35904]

Nearest Neighbors for row 600 in LSH: [49309, 373, 44503]

Nearest Neighbors for row 600 in linear search: [49309, 373, 44503]

Nearest Neighbors for row 700 in LSH: [41352, 44006, 36422]

Nearest Neighbors for row 700 in linear search: [41352, 44006, 36422]

Nearest Neighbors for row 800 in LSH: [30478, 44743, 34353]

Nearest Neighbors for row 800 in linear search: [30478, 44743, 34353]

Nearest Neighbors for row 900 in LSH: [29023, 15184, 20405]

Nearest Neighbors for row 900 in linear search: [29023, 15184, 20405]

Nearest Neighbors for row 1000 in LSH: [54473, 20929, 7396]

Nearest Neighbors for row 1000 in linear search: [27042, 24630, 33514]

Nearest Neighbors for row 1100 in LSH: [35654, 1114, 640]

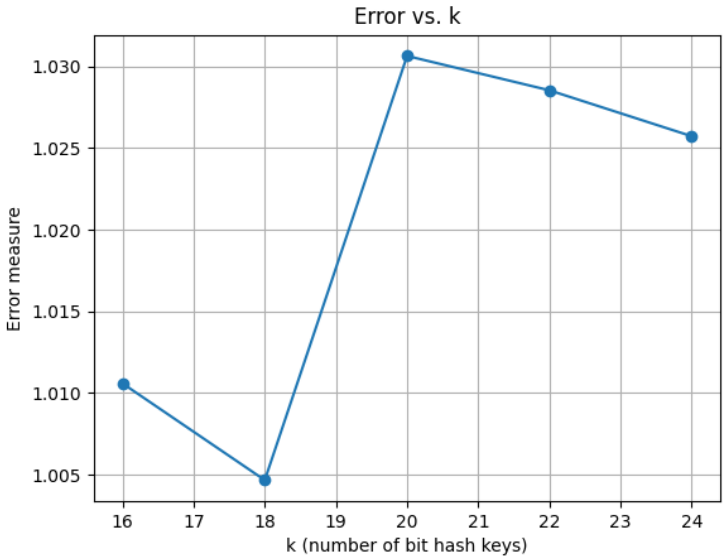
Nearest Neighbors for row 1100 in linear search: [35654, 1114, 640]

زمان اجرا LSH بسیار کمتر از جستجو خطی است:

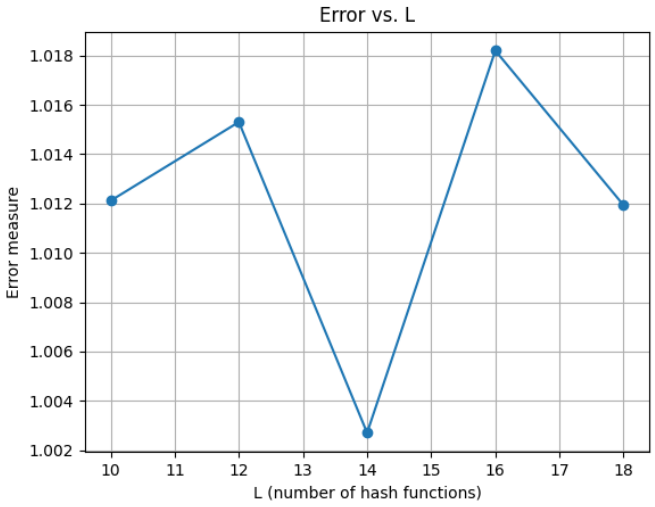
Average Time Running LSH: 0.3178043148734353

Average Time Running linear Search: 5.975934852253307

سپس مقدار خطا برای k ها و L های مختلف محاسبه شد و به شرح نمودار زیر است:



نمودار بالا که برای k رسم شده نشان میدهد که با افزایش بیت هش ابتدا کاهش ولی بعد از 18 افزایش در خطا داریم.



نمودار بالا برای L های مختلف نشان میدهد که یک نقطه بهینه برای خطا نسبت به تعداد هش فانکشن وجوددارد.

و در نهایت تصویر ردیف 100 و 10 تا از نزدیک ترین همسایه هایش در سرچ خطی و روش LSH بدست امد:

تصویر 100:

تصاویر 10 همسایه از سرچ خطی:

تصاویر 10 همسایه از LSH :

تصاویر به نظر شبیه می آیند.