

**گزارش پروژه درس داده‏کاوی**

**نام و نام خانوادگی**

نگار فتحی

**شماره دانشجویی**

97723137

**هدف گزارش:** تحلیل تراکم (در کدام طول و عرض جغرافیایی تراکم تاکسی‏ها بیشتر است؟)

**گام 1:** نمونه‏گیری از مجموعه‏داده:

یک نمونه‏ی تصادفی به اندازه‏ی 6277 رکورد، از مجموعه‏داده گرفته شده است و sample.txt نامگذاری شده است.

این نمونه‏ی تصادفی حاصل اجرای برنامه‏ی C# زیر می‏باشد.

using System;

using System.Collections.Generic;

using System.Text;

namespace Sampling

{

public static class Class

{

private static readonly Random rng = new Random();

public static void Shuffle<T>(this IList<T> list)

{

int n = list.Count;

while (n > 1)

{

n--;

int k = rng.Next(n + 1);

T value = list[k];

list[k] = list[n];

list[n] = value;

}

}

}

}

using System;

using System.Collections.Generic;

using System.IO;

using System.Linq;

namespace Sampling

{

class Program

{

static void Main(string[] args)

{

var dest = @"C:\Users\hamed\Desktop\sample.txt";

var folderPath = @"C:\Users\hamed\Desktop\T-drive Taxi Trajectories\release\taxi\_log\_2008\_by\_id";

var allFiles = Directory.GetFiles(folderPath);

var allData = new List<string>();

foreach (var file in allFiles.Take(500))

{

var lines = File.ReadAllLines(file);

var length = lines.Length;

int percent = Convert.ToInt32(0.01 \* length);

lines.Shuffle();

var newList = lines.Take(percent).ToList();

allData.AddRange(newList);

Console.WriteLine(Path.GetFileNameWithoutExtension(file));

}

allData.Shuffle();

File.WriteAllLines(dest, allData);

}

}

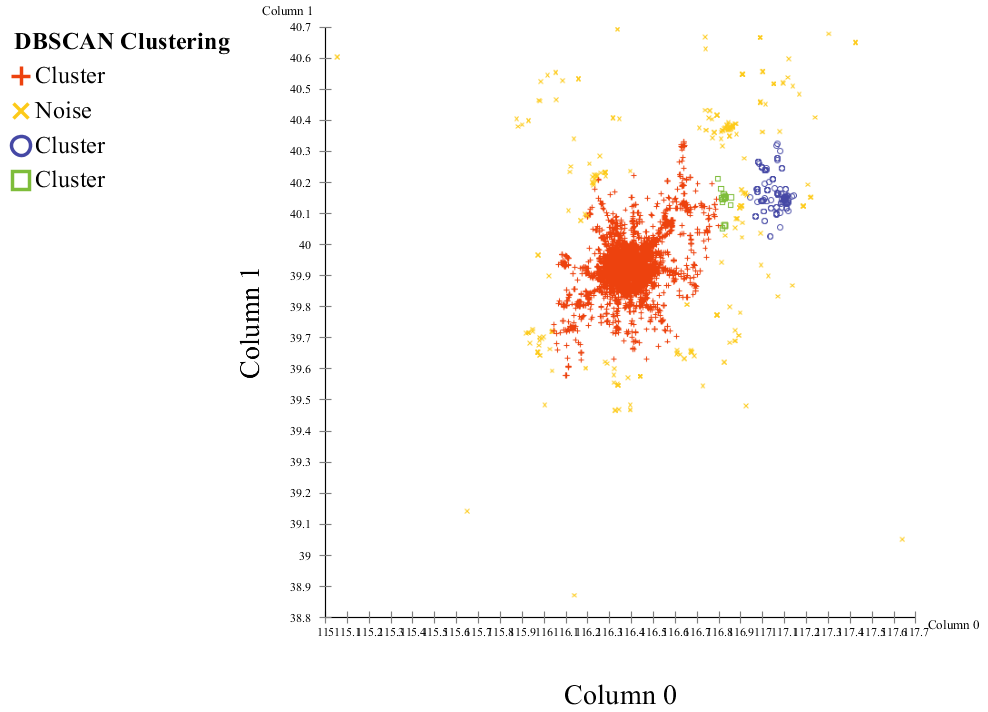
}

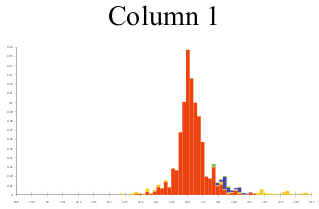
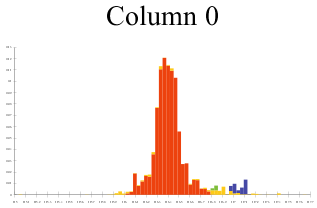
**گام 2:** اجرای sample.txt در ELKI

KDDCLIApplication -dbc.in "C:\\Users\\hamed\\Desktop\\sample.txt" -parser.labelIndices 0 -dbc.filter transform.ProjectionFilter -projection NumericalFeatureSelection -projectionfilter.selectedattributes 2,3 -time -algorithm clustering.DBSCAN -algorithm.distancefunction geo.LatLngDistanceFunction -dbscan.epsilon 5000.0 -dbscan.minpts 50 -resulthandler ResultWriter,AutomaticVisualization -out "C:\\Users\\hamed\\Desktop\\out" -out.silentoverwrite

|  |  |
| --- | --- |
| توضیحات | دستورات |
| مسیر فایل ورودی | -dbc.in "C:\\Users\\hamed\\Desktop\\sample.txt" |
| از آن‏جایی که رکوردهای مجموعه‏داده دارای 4 بعد بودند ولی ما نیازمند تنها 2 بعد longitude و latitude بودیم این دستورات اعمال شدند. | -parser.labelIndices 0 -dbc.filter transform.ProjectionFilter -projection NumericalFeatureSelection -projectionfilter.selectedattributes 2,3 |
|  | -time |
| الگوریتم خوشه‏بندی DBSCAN نیازمند تعیین تعداد خوشه‏ها نبوده (البته نیازمند تعیین دو پارامتر epsilon و minpts است) و خوشه‎هایی با اشکال دلخواه تولید می‏کند و یکی از گزینه‏های خوب برای خوشه‏بندی داده‏های مکانی\_زمانی است. | -algorithm clustering.DBSCAN |
| با توجه به مجموعه‎داده (حاوی داده‏های مکانی\_زمانی)، این تابع فاصله مناسب‏تر از سایرین به نظر می‏آمد. | -algorithm.distancefunction geo.LatLngDistanceFunction |
| پس از دادن مقادیر مختلف به epsilon و مشاهده نتایج، این مقدار برای epsilon مناسب‏تر از سایرین به نظر می‏آمد. | -dbscan.epsilon 5000.0 |
| پس از دادن مقادیر مختلف به minpts و مشاهده نتایج، این مقدار برای minpts مناسب‏تر از سایرین به نظر می‏آمد. | -dbscan.minpts 50 |
| دو مدل برای نمایش نتیجه | -resulthandler ResultWriter,AutomaticVisualization |
| مسیر فایل نتیجه | -out "C:\\Users\\hamed\\Desktop\\out" |
|  | -out.silentoverwrite |

**گام 3:** خروجی اجرای sample.txt در ELK

****



**گام 4:** نتیجه

تراکم تاکسی‏ها در مختصات جغرافیایی زیر بیشتر از سایرین‏ است.

longitude = 116 تا 116.8

latitude = 39.6 تا 40.3