מעבדה 3: קורס מבוא להנדסת נתונים

מגישים:

דן אמלר 206100224

יעל הרי 316375872

הקדמה:

נניח שאני מעוניינת לנסוע לטיול בפריס. אני רוצה לבקר במקומות תיירותיים, לאכול במאפיות מפורסמות, וללכת לכמה מוזיאונים. בגלל הקורונה אני רוצה להשתמש בתחבורה הציבורית כמה שפחות ולכן אני רוצה לדעת איך אחלק את האטרקציות לפי הימים, ומאיפה כדאי לי להתחיל את הטיול בכל יום, ככה שבכל יום אצטרך ללכת כמה שפחות. קבעתי טיול של 5 ימים (5 קלאסטרים).

יש לי רשימה של 25 מקומות אליהם ארצה להגיע (בניהם הלובר, מגדל אייפל, קונדיטוריה של פייר הרמה ועוד). אספתי אותם כקואורדינטות על המפה (הנקודות האדומות) (הקואורדינטות הועתקו מגוגל מפות לקובץ טקסט, אותו העתקתי כנספח 1.1 בהמשך הקובץ למען הנוחות).



בדקתי איך משפיע בחירת seed התחלתי על השגיאה, כלומר הערכה כללית של כמה אצטרך ללכת בטיול, ועל המיקום של הצנטרואידים.

עשיתי 14 ניסויים. עבור 4 מהם נתתי פירוט גבוה יותר כולל מיקומים על מפה כדי להמחיש את השינוי במיקום הצנטרואידים.

שאלת המחקר: איך שינוי הseed ההתחלתי משפיע על החלוקה לקלאסטרים, ועל השגיאה הכוללת.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Map  (צנטרואידים בכחול) | Total loss | Centroid loss | Centroids |  | Init. seed |
|  | 0.187348 | 0.010384990077930971 | 48.856238744217926, 2.357736410079063 | 0 | 0 |
| 0.004178572103036056 | 48.859476463606825, 2.3505896539297315 | 1 |
| 0.07987097687342612 | 48.864212879719965, 2.3091999149351694 | 2 |
| 0.06819589735457922 | 48.85259283988508, 2.3431109792384404 | 3 |
| 0.02471764809567905 | 48.87690169664441, 2.337675106871038 | 4 |
|  | 0.166895 | 0.012336662452235771 | 48.86087042747559, 2.3171576428433065 | 0 | 11 |
| 0.024113791082464565 | 48.86776088728081, 2.2980694480674764 | 1 |
| 0.07252784806760006 | 48.85412188313276, 2.3516280677819132 | 2 |
| 0.0 | 48.88724022348526, 2.3432240408920544 | 3 |
| 0.057916804440090404 | 48.86058461904531, 2.334310723802885 | 4 |
|  | 0.151897 | 0.02623523103354096 | 48.85612549337385, 2.3340747573798266 | 0 | 4 |
| 0.059755513621577654 | 48.865004703358714, 2.305704725977809 | 1 |
| 0.016794741663642407 | 48.857136289023806, 2.356641878478478 | 2 |
| 0.02471764809567905 | 48.87690169664441, 2.337675106871038 | 3 |
| 0.024393496913761435 | 48.85196873606772, 2.348046774427224 | 4 |
|  | 0.149316 | 0.03906777850439624 | 48.87403999759363, 2.333577454176454 | 0 | 8 |
| 0.0417267492545778 | 48.86489215408808, 2.3018097834490856 | 1 |
| 0.016951034818337773 | 48.85092578225129, 2.347901977534541 | 2 |
| 0.025335432579420843 | 48.85731798401423, 2.3553541580292863 | 3 |
| 0.02623523103354096 | 48.85612549337385, 2.3340747573798266 | 4 |

סך הכול: הניסויים שעשיתי:

|  |  |  |  |
| --- | --- | --- | --- |
| Total loss | Initial seed | Total loss | Initial seed |
| 0.155878 | 7 | 0.187348 | 0 |
| 0.149316 | 8 | 0.162118 | 1 |
| 0.150708 | 9 | 0.151897 | 2 |
| 0.172033 | 10 | 0.158556 | 3 |
| 0.166895 | 11 | 0.151897 | 4 |
| 0.162118 | 12 | 0.152995 | 5 |
| 0.152995 | 13 | 0.151950 | 6 |

מסקנות:

1. השגיאה המינימלית ביותר היא עבור בחירת init. seed = 8.
2. שמתי לב שהשגיאה עבור seed = 2 ועבור seed = 4 זהה, ומבדיקה עלה שהחלוקה לקלאסטרים זהה, אך לא באותו סדר:

|  |  |  |
| --- | --- | --- |
| Points with init. seed = 4 | Points with init. seed = 2 | cluster |
| 10 19 22 7 8 | 1 2 20 21 23 24 3 | 0 |
| 11 12 15 16 18 | 10 19 22 7 8 | 1 |
| 13 25 4 5 6 | 11 12 15 16 18 | 2 |
| 14 17 9 | 13 25 4 5 6 | 3 |
| 1 2 20 21 23 24 3 | 14 17 9 | 4 |

באופן דומה גם השגיאה עבור seed = 5 ועבור seed = 13 זהה, וגם עבור seed = 1 ועבור seed = 12. המצב של כל אחד מהזוגות הללו בדומה לנ"ל (אותה חלוקה בסדר שונה).

לסיכום, אבחר את החלוקה לימי טיול שמציע הניסוי עם init. Seed=8.

נספח

1.1 רשימת קואורדינטות איתה עבדתי של מיקומים תיירותיים ומאפיות בפריז

1 48.85174380888362 2.346829695364803

2 48.85012265963253 2.349214367828763

3 48.849670823578336 2.3490856217997753

4 48.85834041341368 2.356359612640512

5 48.85719531130436 2.3590420702860575

6 48.85763066024796 2.35872875019343

7 48.852551514698376 2.3396095696706327

8 48.85581539303824 2.3333788115404603

9 48.87158461145443 2.3379129641668284

10 48.86025376152619 2.3266758597219734

11 48.87073353456019 2.3045281362736754

12 48.86545490044127 2.3212844960927

13 48.85178859190571 2.3568152071962536

14 48.871880254993535 2.3318883155542323

15 48.85628595450991 2.313030789593913

16 48.874135250676204 2.2951822732153224

17 48.88724022348526 2.3432240408920544

18 48.858413876606036 2.294497934713432

19 48.860409320422335 2.337865177749907

20 48.8523017325953 2.3501876934158203

21 48.84646552230828 2.3468912242567863

22 48.851597477184086 2.332844368216161

23 48.85525014650964 2.345203262541297

24 48.85822645896633 2.348915555783324

25 48.86072646824732 2.352263752076139