داده کاوی

تمرین چهارم- بخش پیاده سازی

آرش حریرپوش
9731505

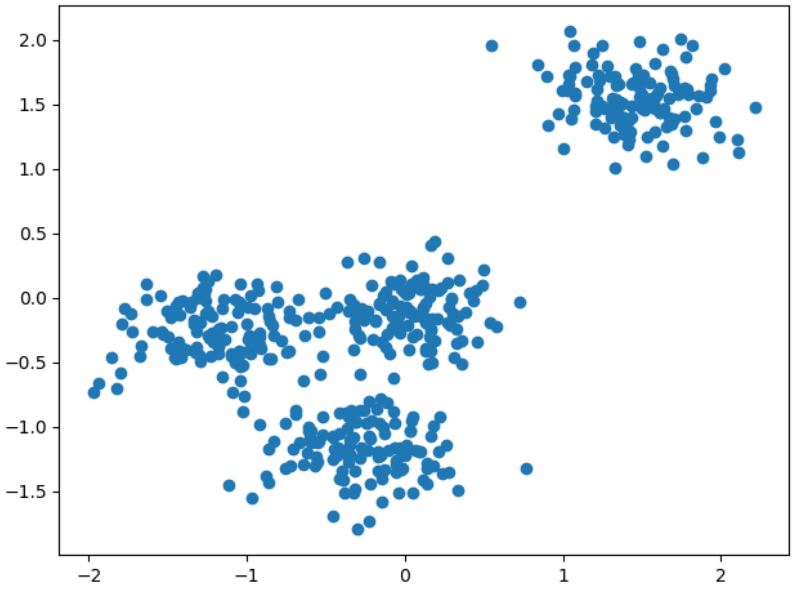
آرش حریرپوش
9731505

.1

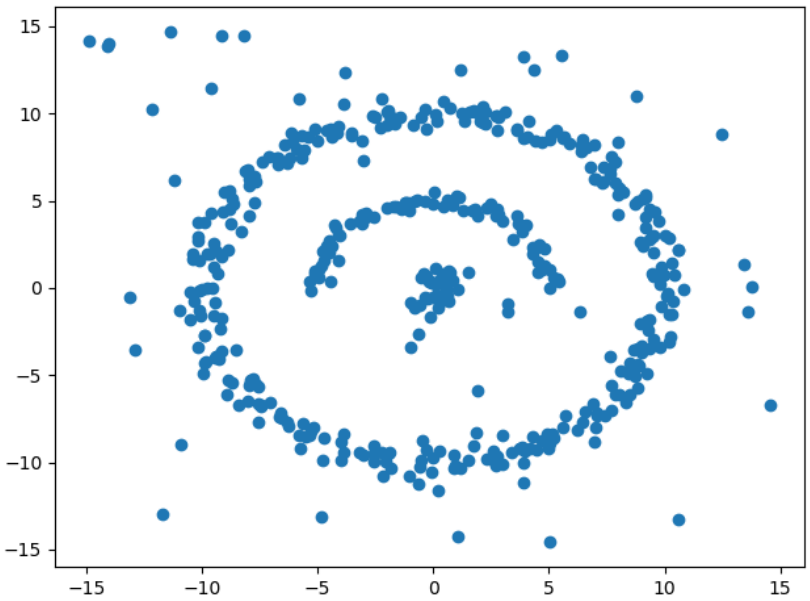
الف)

Scatter Plot مجموعه داده های dataset1 و dataset2 به صورت زیر رسم می شود:

dataset1



dataset2



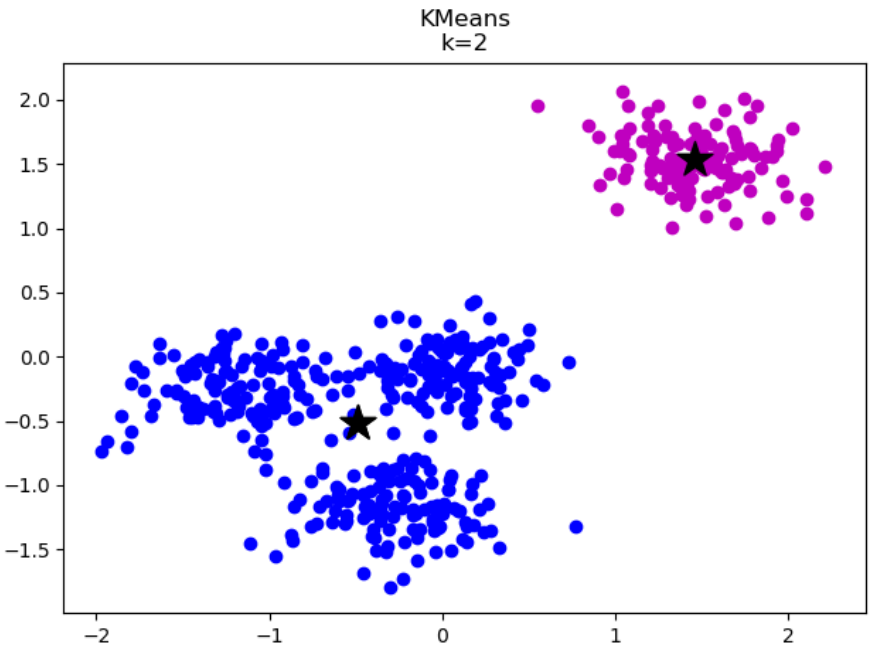
(بخش A1 فایل Q1.ipynb)

آرش حریرپوش
9731505

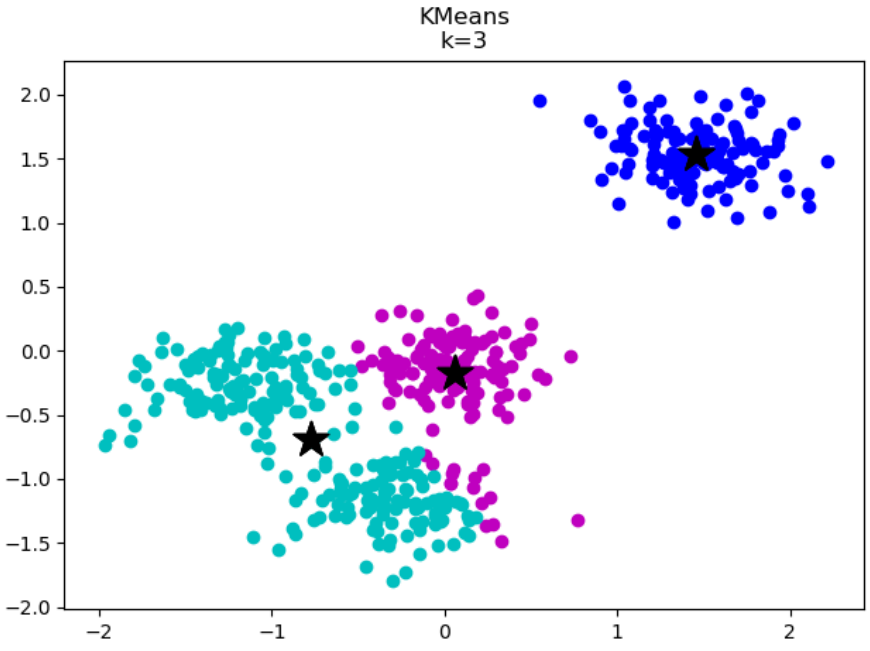
الف)

نتایج این خوشه بندی به ازای k های عنوان شده بر روی dataset1 برابر است با:

K = 2

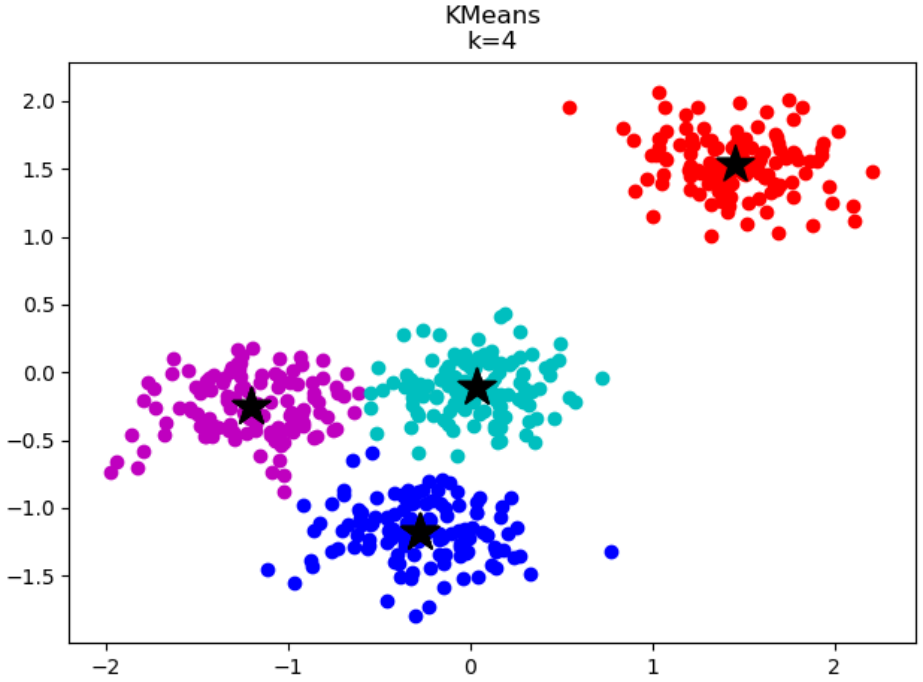


K = 3



آرش حریرپوش
9731505

K = 4



با توجه به اینکه نقاط اولیه در این خوشه بندی به صورت تصادفی انتخاب می شود ممکن است به ازای اجرا های متفاوت خوشه های متفاوت تری بدست بیاید؛ تصاویر بالا نمونه ای از خوشه بندی های مطلوب می باشد که توسط این الگوریتم بدست آمده است.

(بخش A2 فایل Q1.ipynb)

ب)

میانگین فاصله مرکز خوشه با نقاط موجود در آن خوشه برای K = 4 برابر است با:

Cluster 0 Error Rate : 0.2863247821239086

Cluster 1 Error Rate : 0.3135014574774217

Cluster 2 Error Rate : 0.3212871171781822

Cluster 3 Error Rate : 0.3333119438685841

(بخش B فایل Q1.ipynb)

ج)

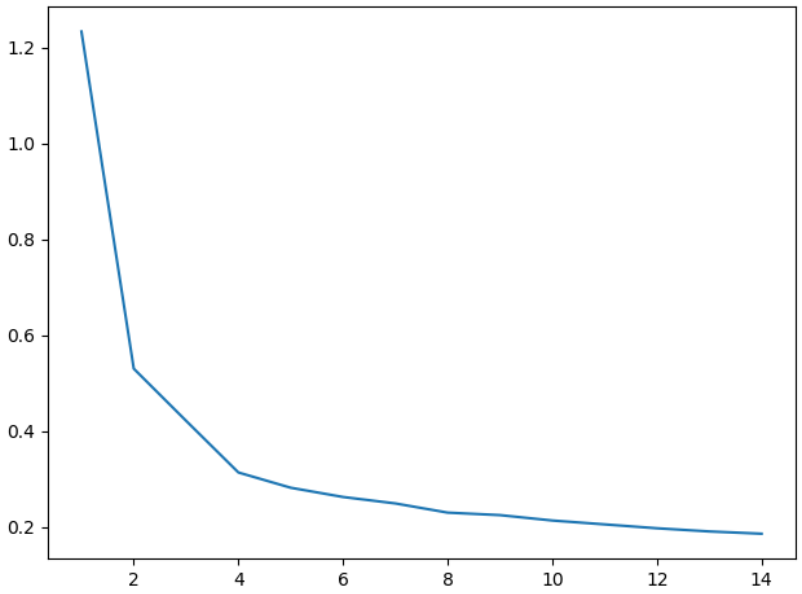
میانگین خطای بدست آمده در قسمت قبل برابر است با:

Average Cluster Error : 0.31360632516202414

(بخش C فایل Q1.ipynb)

د)

نمودار خطا های خوشه بندی به ازای k بین 1 تا 15 برابر است با:



با توجه به نقاط اولیه متفاوت در هر اجرا ممکن است این نمودار اندکی در هر اجرا متفاوت شود.

(بخش D فایل Q1.ipynb)

ه)

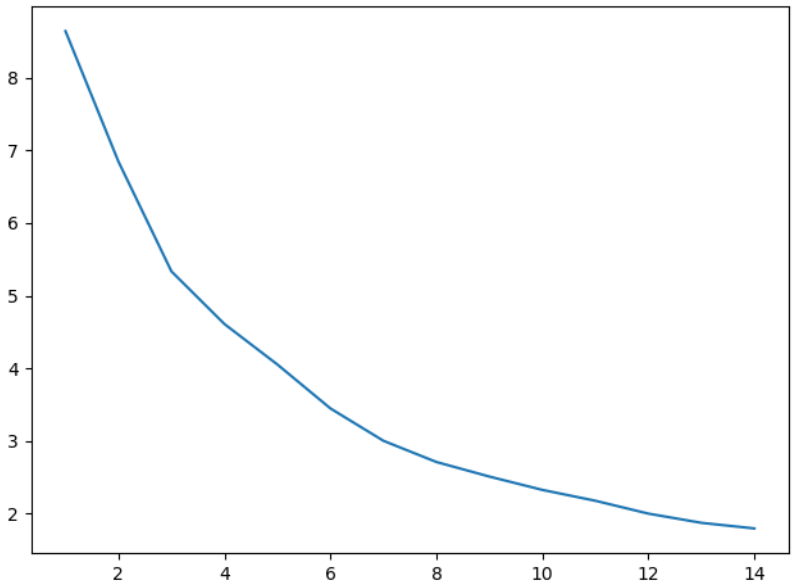
با استفاده از روش elbow مقدار K بهینه برای خوشه بندی dataset1 برابر **4** می باشد. با توجه به اینکه بعد از K = 4 مقدار خطای خوشه بندی به صورت تقریبا خطی کاهش می یابد مقدار بهینه برای خوشه بندی این داده ها برابر K = 4 می باشد.

(منبع: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/)

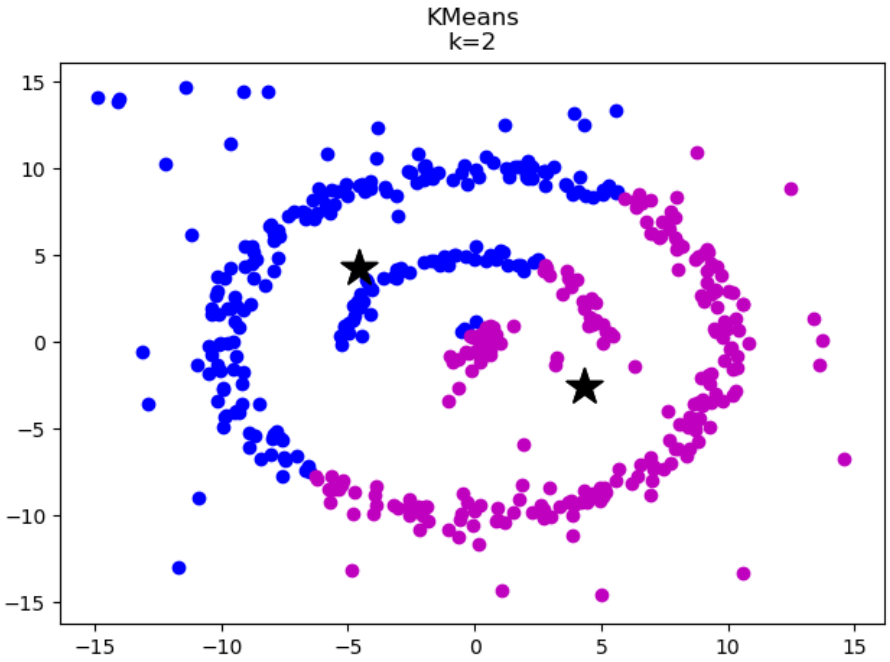
ی)

با اجرای این الگوریتم به ازای k های بین 1 تا 15 خطا های بدست آمده برای خوشه بندی این داده ها برای dataset2 برابر است با:

آرش حریرپوش
9731505



همانطور که مشاهده می شود خطا های بدست آمده در خوشه بندی این الگوریتم بزرگ می باشند؛ که این نشان می دهد که این الگوریتم داده های dataset2 را به خوبی خوشه بندی نمی کند؛ به عنوان مثال خروجی این الگوریتم به ازای K = 2 برابر است با:



که این به این دلیل است که داده های dataset2 دارای چگالی های یکسان نمی باشند و همچنین به صورت کروی پخش نشده اند برای همین این الگوریتم این داده ها را به خوبی خوشه بندی نمی کند. (برای خوشه بندی این داده ها می توان از الگوریتم DBSCAN استفاده کرد)

آرش حریرپوش
9731505

(بخش E فایل Q1.ipynb)

.2

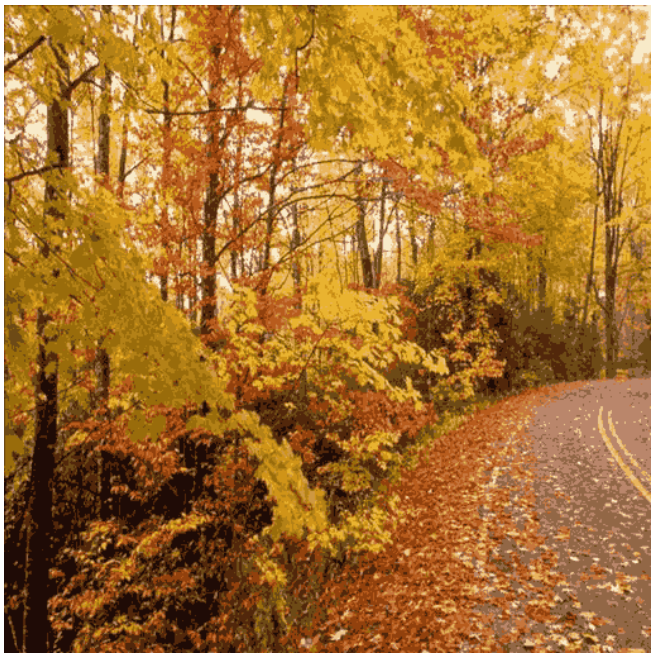
برای انجام این کار ابتدا ماتریس 3 بعدی تصویر را به یک ماتریس 2 بعدی تبدیل می کنیم بدین صورت که طول و عرض تصویر را در یک بعد و کانال های RGB نیز در یک بعد دیگر قرار داده می شود؛ برای دو بعدی کردن تصاویر ماتریس اولیه را reshape می کنیم و بعد اول را برابر حاصل ضرب طول در عرض و بعد دوم را برابر تعداد کانال های تصویر قرار می دهیم؛ سپس الگوریتم خوشه بندی را به ازای K های مختلف بر روی ماتریس 2 بعدی بدست آمده اجرا می کنیم؛ نتایج بدست آمده به ازای K های مختلف در این الگوریتم برابر است با:

K = 2

آرش حریرپوش
9731505

K = 4

K = 16



آرش حریرپوش
9731505

K = 32

K = 64

همانطور که مشاهده می شود، با بیشتر شدن مقدار K کیفیت عکس خروجی نیز بهتر می شود؛ اما تصاویری که از خوشه بندی با مقدار K بزرگ تر از 16 ایجاد می شود؛ تغییرات کمتری با یک دیگر دارند و جزئیات تصاویر تنها بهبود می یابد.

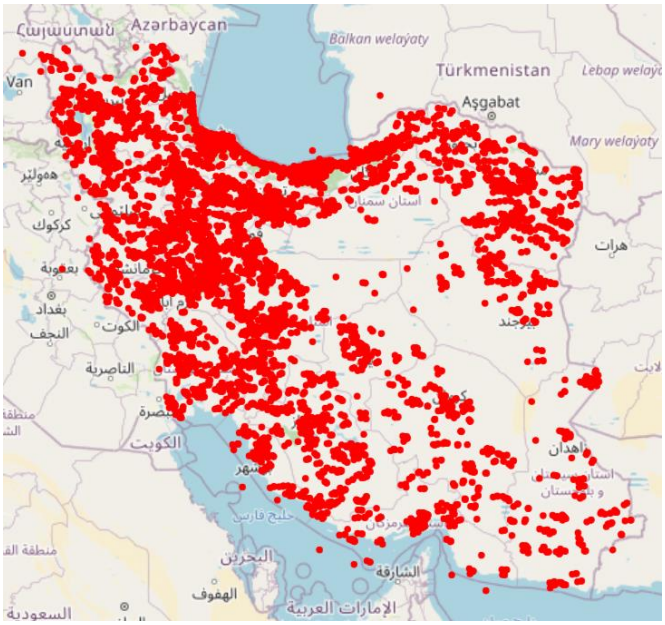
(کد این بخش در فایل ColorReductionWithKMeans.ipynb قرار دارد)

آرش حریرپوش
9731505

.3

الف)

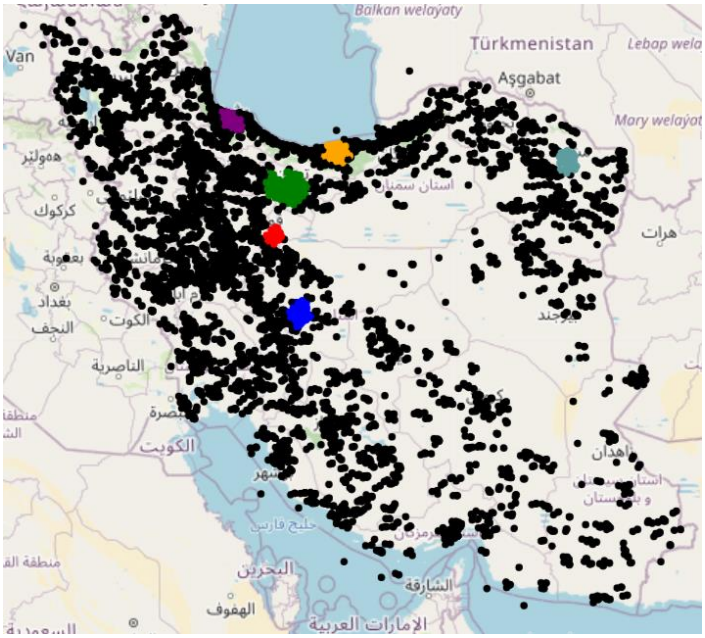
نتیجه بدست آمده از رسم داده ها بر روی نقشه برابر است با:



(بخش A فایل Q3.ipynb)

ب)

نتیجه اجرای الگوریتم DBSCAN به ازای Esp = 0.2 و min_samples = 300 برابر 6 خوشه و 12121 نقطه outlier می باشد، که به صورت زیر مشاهده می شود:



که همانطور که مشاهده می شود شهر های، تهران، قم، گیلان، رشت، اصفهان و مشهد به عنوان شهر های پر تراکم شناخته شده اند.

(بخش B فایل Q3.ipynb)

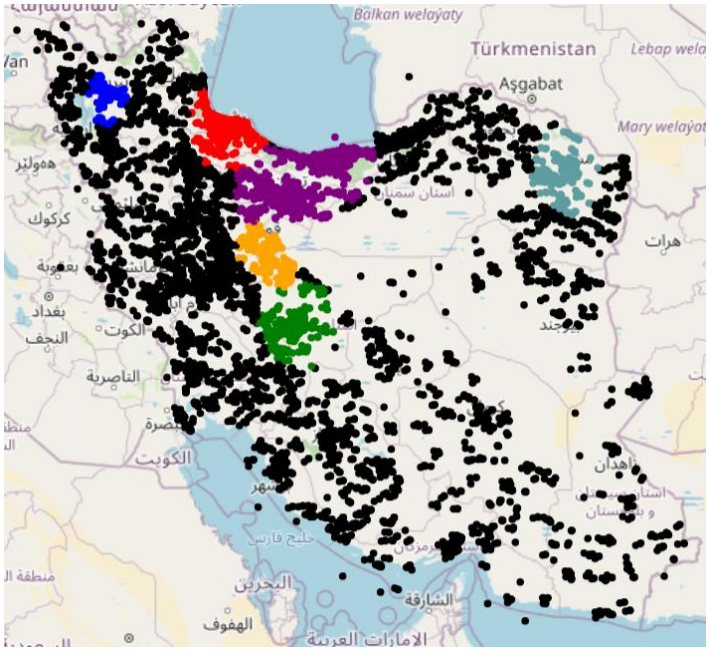
ج)

با تغییر مقدار Eps و min_samples برای این الگوریتم و بررسی تعداد خوشه ها و نقاط outlier بدست آمده؛ نتیجه گرفته می شود که مقدار 0.6 برای Eps و 500 برای min_samples می تواند تعداد 6 خوشه و 9174 نقطه outlier پیدا کند؛ که این 6 خوشه مکان های پر تراکم را نشان می دهد.

(بخش C فایل Q3.ipynb)

د)

شکل زیر خوشه های بدست آمده در بخش قبل و نقاط outlier را نشان می دهد:



(بخش D فایل Q3.ipynb)

.4

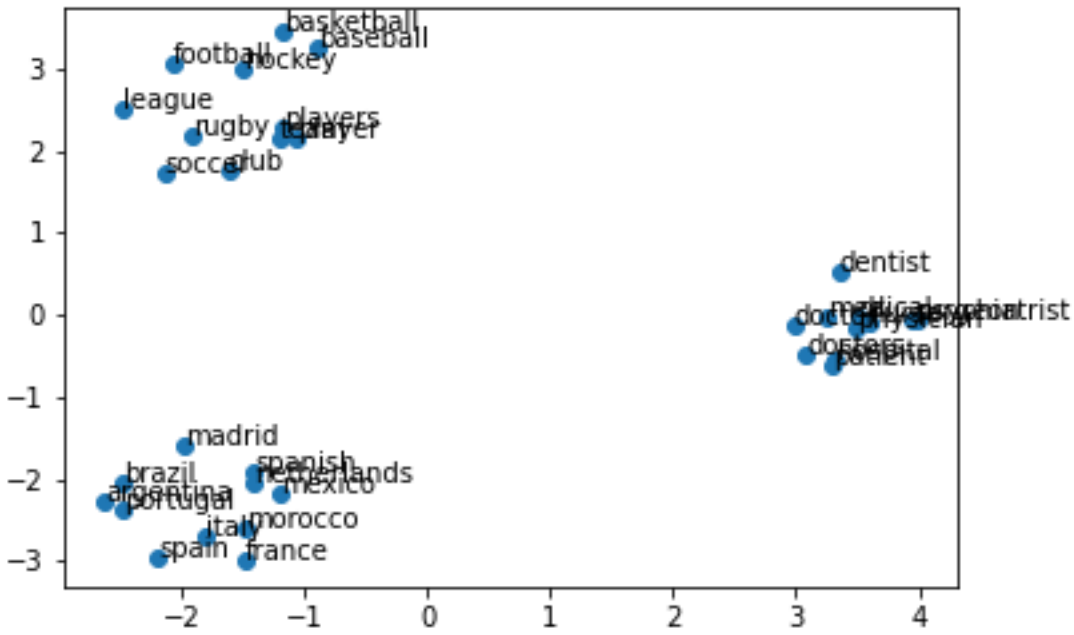
برای کاهش ابعاد word embedding ها ابتدا کلمات از فایل glove.6B.50d.txt خوانده می شود و سپس در یک dictionary ذخیره می شوند، و سپس 33 کلمه بدین صورت که هر 11 کلمه به یک دیگر مشابه باشد به صورت دستی برای کاهش ابعاد انتخاب می گردند؛ این سه دسته از کلمات برابر اند با:

Spain: {Portugal, Argentina, Italy, France, Spanish, Brazil, Mexico, Madrid, Morocco, Netherlands}

Football: {Soccer, Basketball, League, Rugby, Hockey, Club, Team, Baseball, Players, Player}

Doctor: {Physician, Nurse, dr., Doctors, Patient, Medical, Surgeon, Hospital, Psychiatrist, Dentist}

خروجی PCA برای کلمه های عنوان شده برابر است با:



همانطور که مشاهده می شود کلمات در سه خوشه متفاوت نمایش داده شده اند؛ که در هر خوشه کلمات مشابه با یک دیگر قرار دارد؛ خوشه سمت راست برابر کلمات مشابه با کلمه Doctor، خوشه سمت چپ پایین برابر کلمات مشابه با کلمه Spain و خوشه سمت چپ بالا مشابه با کلمه Football می باشند.

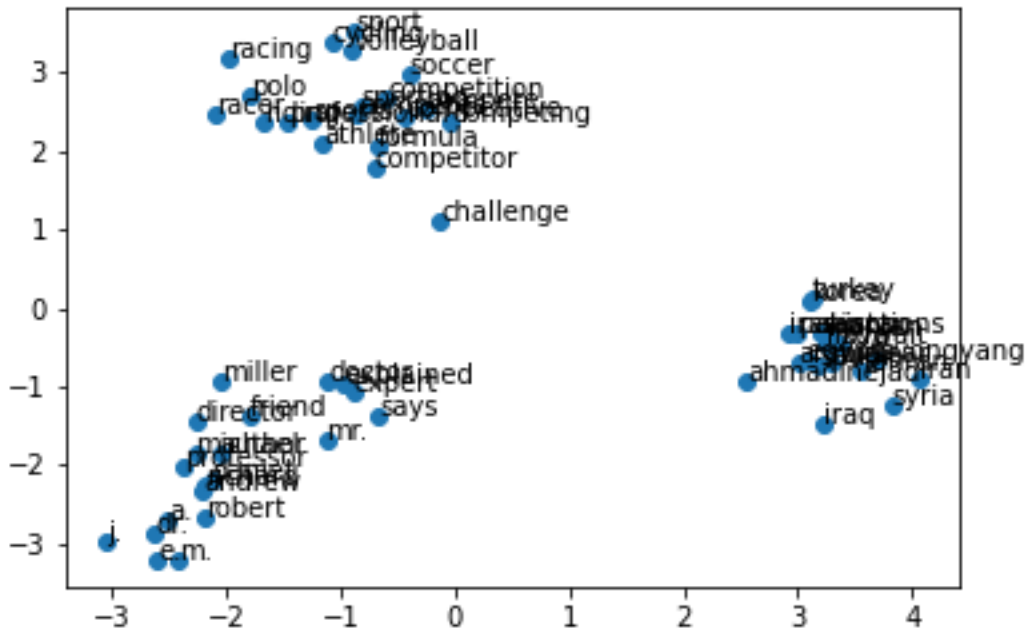
(کلمات فوق بر مبنای سایت https://medium.com/analytics-vidhya/getting-started-with-nlp-word-embeddings-glove-and-classification-5b38c5e9a45 نوشته شده است)

سپس در بخش بعد تعدادی کلمه انتخاب می شود و سپس بر اساس فاصله اقلیدسی آن ها با بقیه کلمات سعی می شود تا 20 کلمه مشابه دیگر برای هر یک از آن کلمه ها انتخاب شود؛ بدین صورت که کلماتی که فاصله اقلیدسی آن ها کمتر 4.25 باشد به عنوان کلمه مشابه در نظر گرفته می شود؛ برای مثال برای تعداد کلمات مشابهی که برای هر یک از سه کلمه Sport، Iran و dr. پیدا شده است؛ برابر است با:

آرش حریرپوش
9731505

{'dr.': 20, 'iran': 18, 'sport': 20}

خروجی PCA برای کلمه های پیدا شده برابر است با:



همانطور که مشاهده می شود کلمات در سه خوشه متفاوت نمایش داده شده اند؛ که در هر خوشه کلمات مشابه با یک دیگر قرار دارد؛ خوشه سمت راست شامل کلمات مشابه با کلمه Iran، و خوشه سمت چپ پایین شامل کلمات مشابه با کلمه .dr و خوشه سمت چپ بالا مشابه با کلمه Sport می باشند.

(کد این بخش در فایل PCA_GloveDataSet.ipynb یا google colab برابر https://colab.research.google.com/drive/15_rjCNEHJin53d2dWfW6WxNSMmlhcWDR#scrollTo=Ijlw6_TYFRBk قرار دارد)