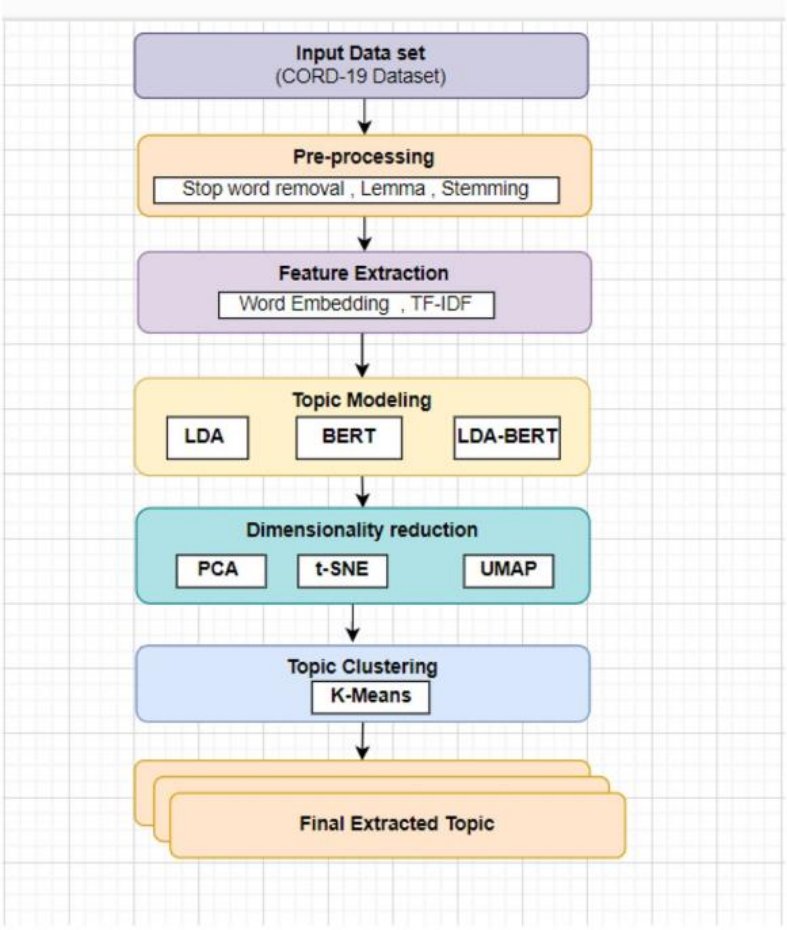
به نام خدا

پروژه ۳ داده کاوی

عاطفه نادری

ابتدا با توجه به مقاله https://link.springer.com/article/10.1007/s41870-023-01268-w

مراحل را انجام دادیم:



Pre process:

برای حذف Stop words ابتدا از spacy bio parser رفتم. این کتابخانه لغات مخصوصی که مرتبط با مقالات علمی حوزه بایو است را جمع آوری کرده است و وابسته به آن stopword هایی دارد. اما در تست ها متوجه شدم خوب نیست و کار درستی را انجام نمیدهد. و به صورت عادی حذف این واژه ها را انجام دادم.

مراحل کار به صورت کلی به شکل زیر است:

- Lower casing
- Removal of punctuations
- Removal of numbers
- Tokenization
- Lemmatization
- stemming
- Removal of stopwords, including a custom list of stopwords

feature Extraction

برای این قسمت، از مدل bow و tf-idf استفاده میکنیم.

topic modeling

برای اهداف topic modeling، از LDA استفاده کردیم. چیزی به عنوان بردار lda وجود ندارد، فقط به عنوان احتمال تعلق به یک موضوع را نشان می دهد.

- اسناد با موضوعات مشابه از گروه های مشابهی از کلمات استفاده می کنند

- موضوعات اسناد، که موضوعات پنهان نامیده می شوند را می توان با جستجوی گروه هایی از کلمات که اغلب با هم در اسناد در سراسر مجموعه وجود دارند، پیدا کرد.

BERT

هم چنین از مدل BERT نیز برای topic mpdeling استفاده کردم که ان را به دلیل ساعت زیادی که برای fit شدن میگرفت فقط توانستم روی k -means اجرا کنم و نتیجه آن را بر روی هر سه روش کاهش ابعاد تست کردم و در یک جدول قرار دادم.

dimension reduction

برای dimension reduction از سه روش PCA و UMAP و TSNE نیز استفاده نمودم.

PCA : Linear dimension reduction
T-SNE : Non-linear dimensional reduction, preserves local structure in the data.
Umap : Non-linear dimensional reduction, preserves both local and most of the global structure

نتایجی که از umap بدست آمد در برخی از آزمایش ها بهتر از سایرین در ترکیب با tf-idf بود.

ماتریس خروجی tf-idf و bow ماتریس اسپارسی است و ما روی آن کاهش ابعاد میزنیم.

Topic clustering

برای این قسمت از ۴ الگوریتم k-means, mean-shift, DBSCAN, OPTICS استفاده کردیم.

معیارهای ارزیابی

silhouette_score

که این معیار نرخ چسنبدگی اعضا به کلاستر خود و جدایی از کلاسترهای دیگر را نشان میدهد. مقدار های نزدیک به یک بهتر است. مقادیر منفی نشان دهنده کلاسترینگ بد است.

برای بدست آوردن k از دو روش elbow و _silhouette استفاده شده است. که در silhouette مقداری که بیشترین را داشته است به عنوان k استفاده شده است. چون نتایج elbow قابل اعتماد نیست.

davies_bouldin_score

این شاخص توسط دیویس و بولدین دو دانشمند در رشته برق در سال 1979 معرفی شد و وابسته به تعداد خوشه‌ها و یا الگوریتم خوشه‌بندی نیست. برای محاسبه این شاخص از دو معیار اندازه پراکندگی(Dispersion measure) و عدم شباهت بین خوشه‌ها(Cluster dissimilarity) استفاده میکند.

مقدار آن هر چه کمتر باشد بهتر است.

تحلیل:

برای تحلیل با توجه به notebook هایی که قرار داده شده است که همه چیز را تقریبا همان جا توضیح داده ام. ابتدا ما در lda یک عددی را برای استخراج topic ها قرار میدهیم. مثلا در ازمایش اخر عدد ۴۰ را هم برای bow و هم برای tf-idf قرار دادیم. و با استفاده از ویژگی features که روش vectorization برای bow و tf-idf وجود دارد. لغات مرتبط با هر topic را پیدا نمودیم و طبق آن ۴۰ تاپیک یک عنوانی را به هر کدام از آن ها اختصاص دادیم. در مراحل بعدی و بعد از کلاسترینگ نیز، لغات پر تکرار هر کلاستر را استخراج میکنیم و similarity آن را با هر یک از ۴۰ topic که داشتیم مسنییج و بیشترین شباهت را گزارش میکنیم که هر کلاستر به کدامtopic مرتبط است. و آن را همانجا چاپ کردیم.

هم چنین با استفاده از بیشترین کلماتی که در هر کلاستر تکرار شده است نیز آن را به صورت ابر کلمات نشان دادیم.

جدول زیر برخی از اجراهایی است که در طول پروژه انجام شده است. و همانطور که قابل مشاهده است هر چقدر در tf-idf مقدار topic ها در lda بزرگتر بود، نتیجه بهتر میشد. شاید به دلیل آن است که ابتدا تمایز خوبی از هر عنوان در هر سند ایجاد میشود و سپس با pca ابعاد به صورت معناداری کاهش می یابد و در نهایت clustering مطلوبی را ایجاد میکند.

در کل طی آزمایشات مختلف و متعدد نتایج حاصل از tf-idf بهتر از bow بوده است. چون که tf-idf اهمیت لغات را بهتر از bow نشان میدهد. بنابراین نتایج بهتری را نیز گرفتیم.

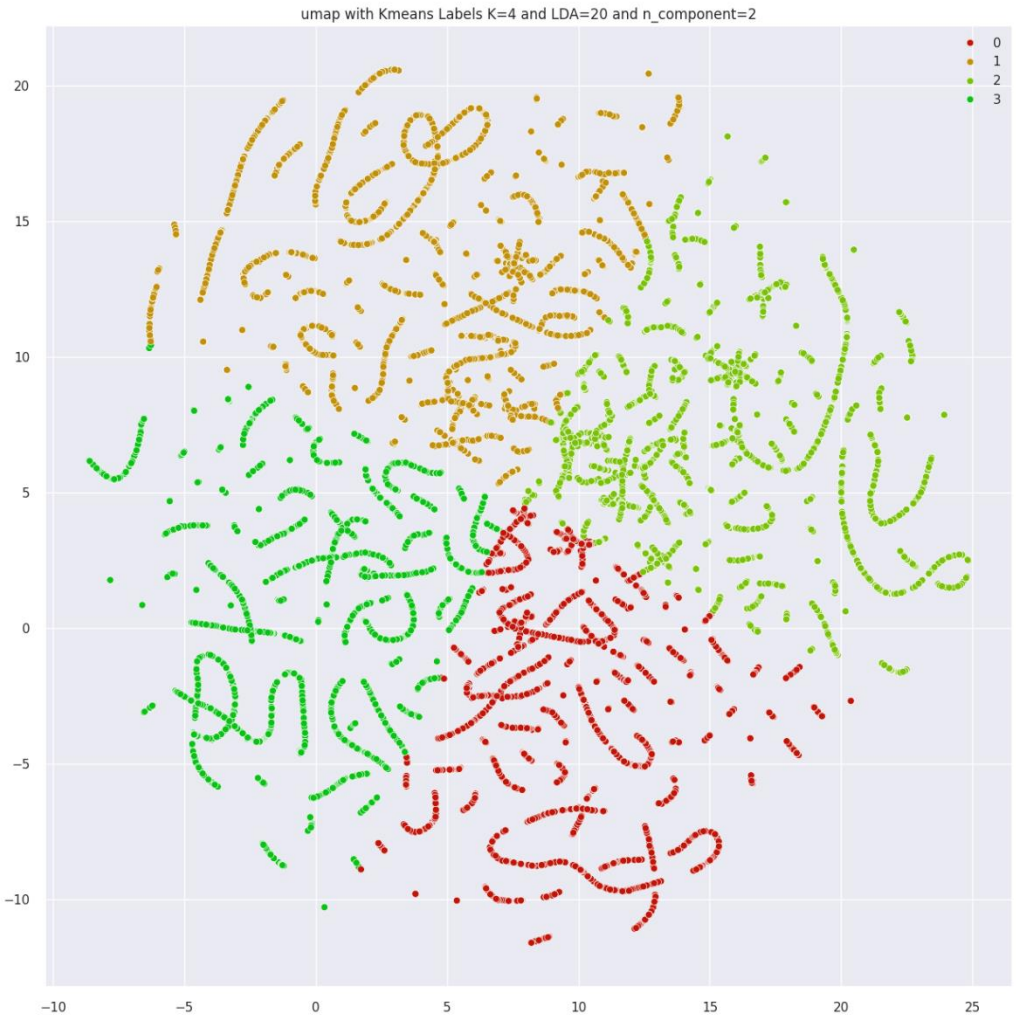
Tf-idf And k-means	K=3, pca=2, lda=20	K=4, pca=2, lda=25	K=3, pac=2, lda=40	K=3, umap, lda =40	K=4 Umpa Lda=20	K=3 Umpa Lda=20	
silhouette	0.877057912562349	0.8749292371478099	0.9270455732377547 0.9293217942529081 0.9063675566175831	0.37730643	0.3527427	0.37235686	
Davies-			0.19922373674285676				

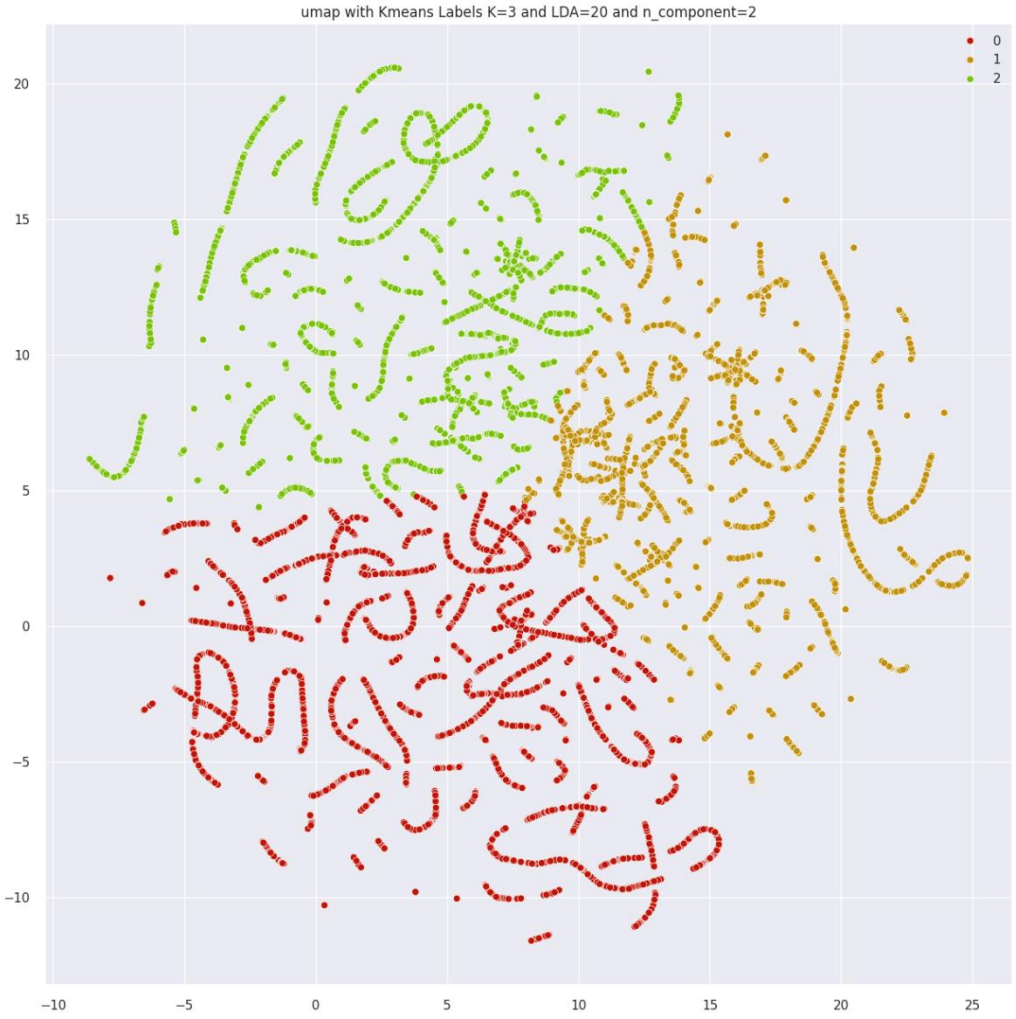
Bouldin index:							
			شکل ۳				

K=4

Umpa

Lda=20

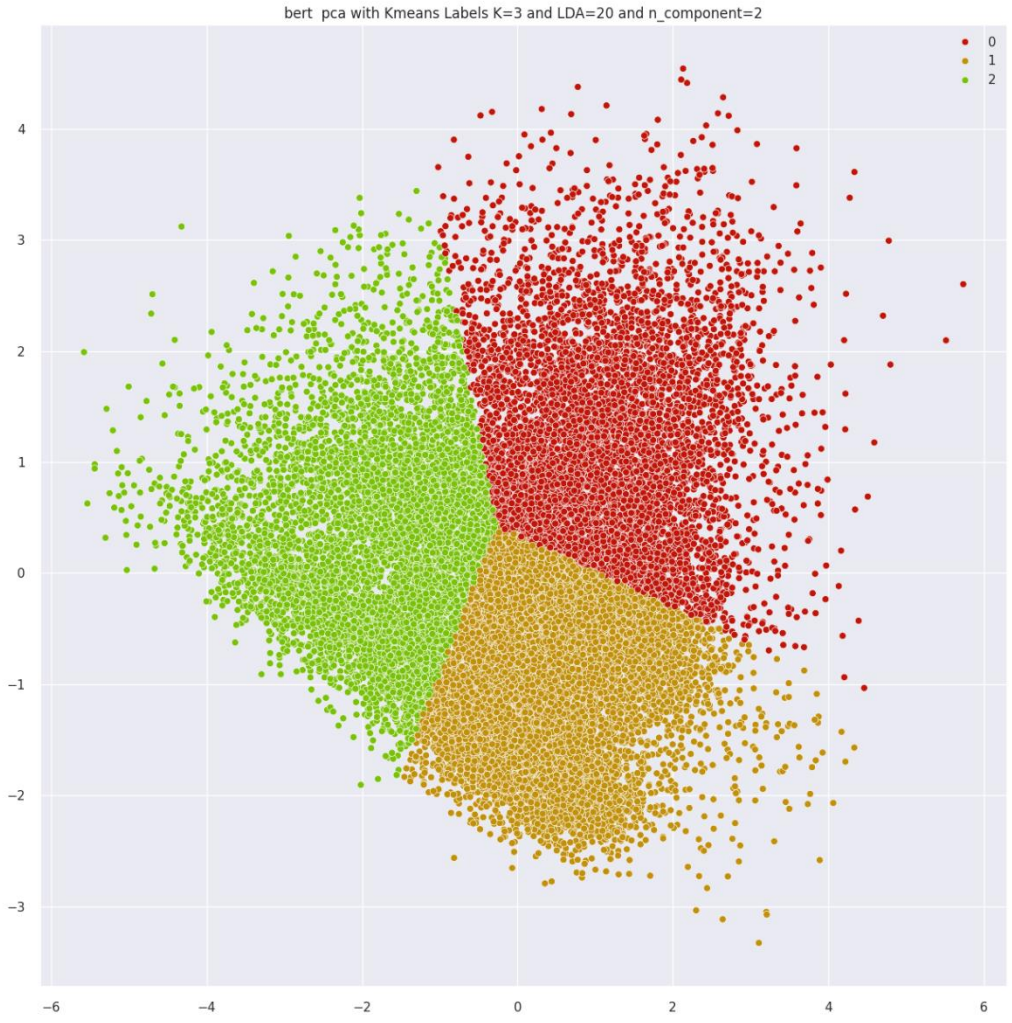

umap with Kmeans Labels K=4 and LDA=20 and n_component=2
0
1
2
3
20
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20
25



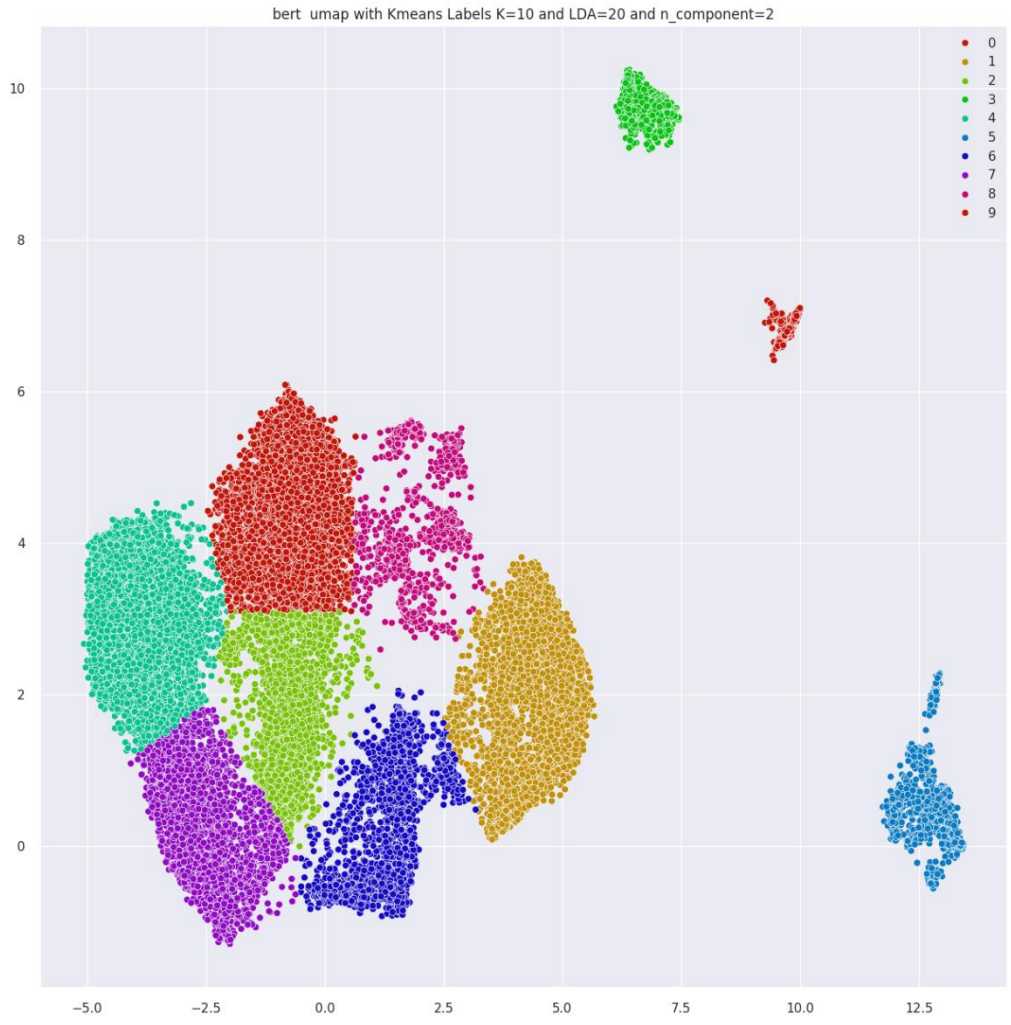
Bert

نمایی از نتایجی که با مدل bert ایجاد شده است با استفاده از هر سه روش کاهش ابعاد تست شده است و مانند نتیجه مقاله با نتیجه بهتری را طبق معیار سیلوئت داشتیم.

Tf-idf and k-means	Pca, k=3	Umap, k=10	Tsne, k=5	
bert	0.37791735	0.5401364	0.42491844	


bert pca with Kmeans Labels K=3 and LDA=20 and n_component=2
0
1
2
4
3
2
1
0
−1
−2
−3
−6
−4
−2
0
2
4
6

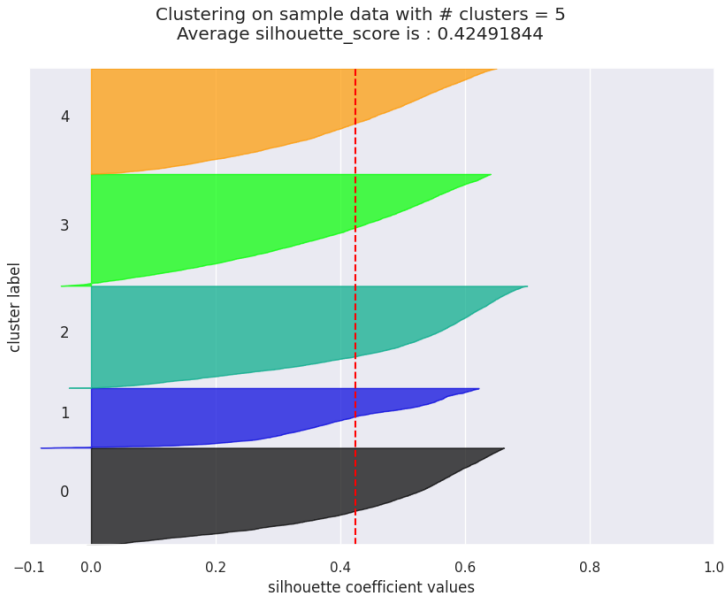
Bert and umap:



Bert & TSNE:


bert tsne with Kmeans Labels K=5 and n_component=2
0
1
2
3
4
80
60
40
20
0
−20
−40
−60
−80
−80
−60
−40
−20
0
20
40
60
80

میانگین امتیاز سیلوئتی که برای k بدست آمد:

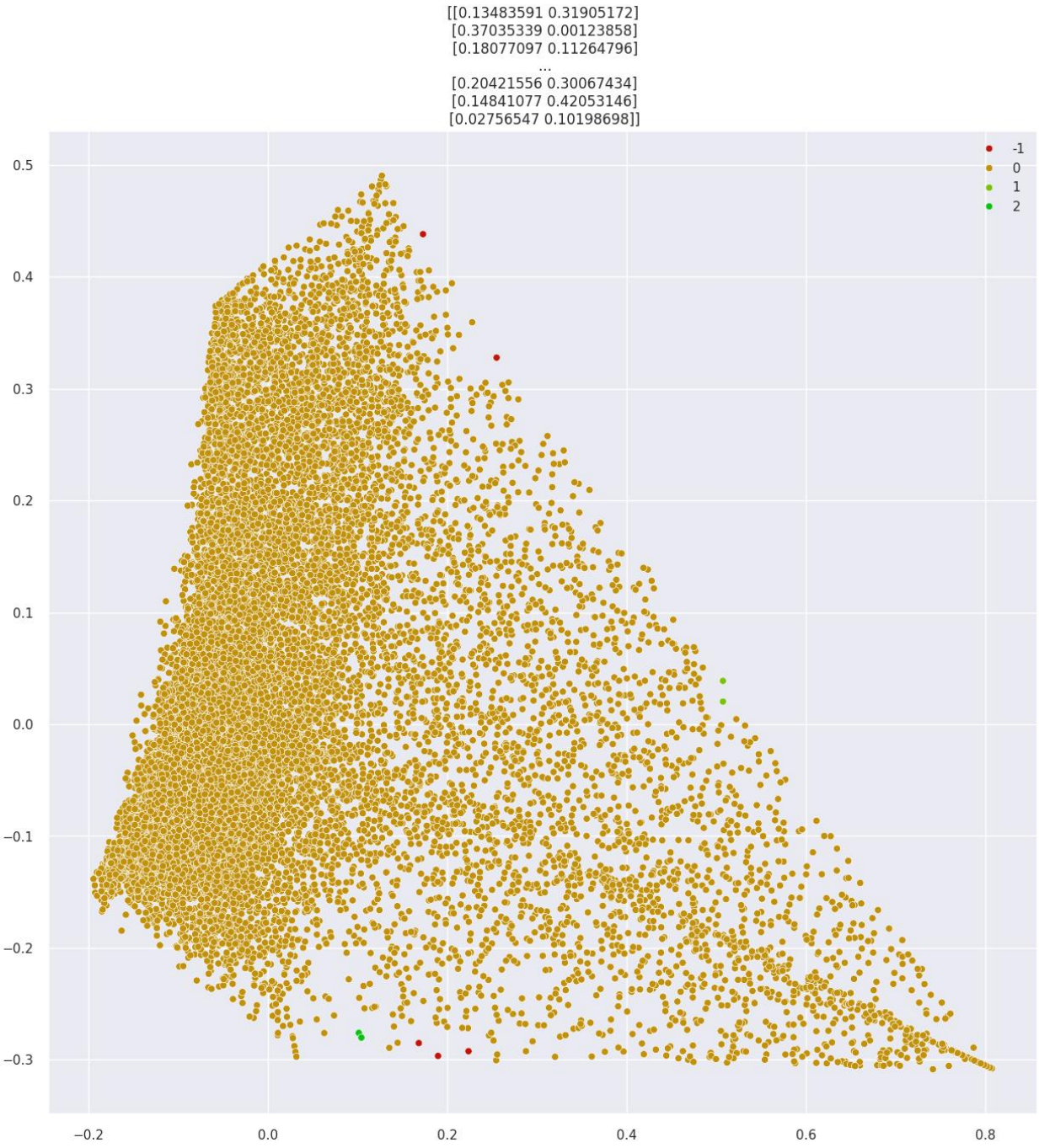


الگوریتم DBSCAN و TF-IDF:

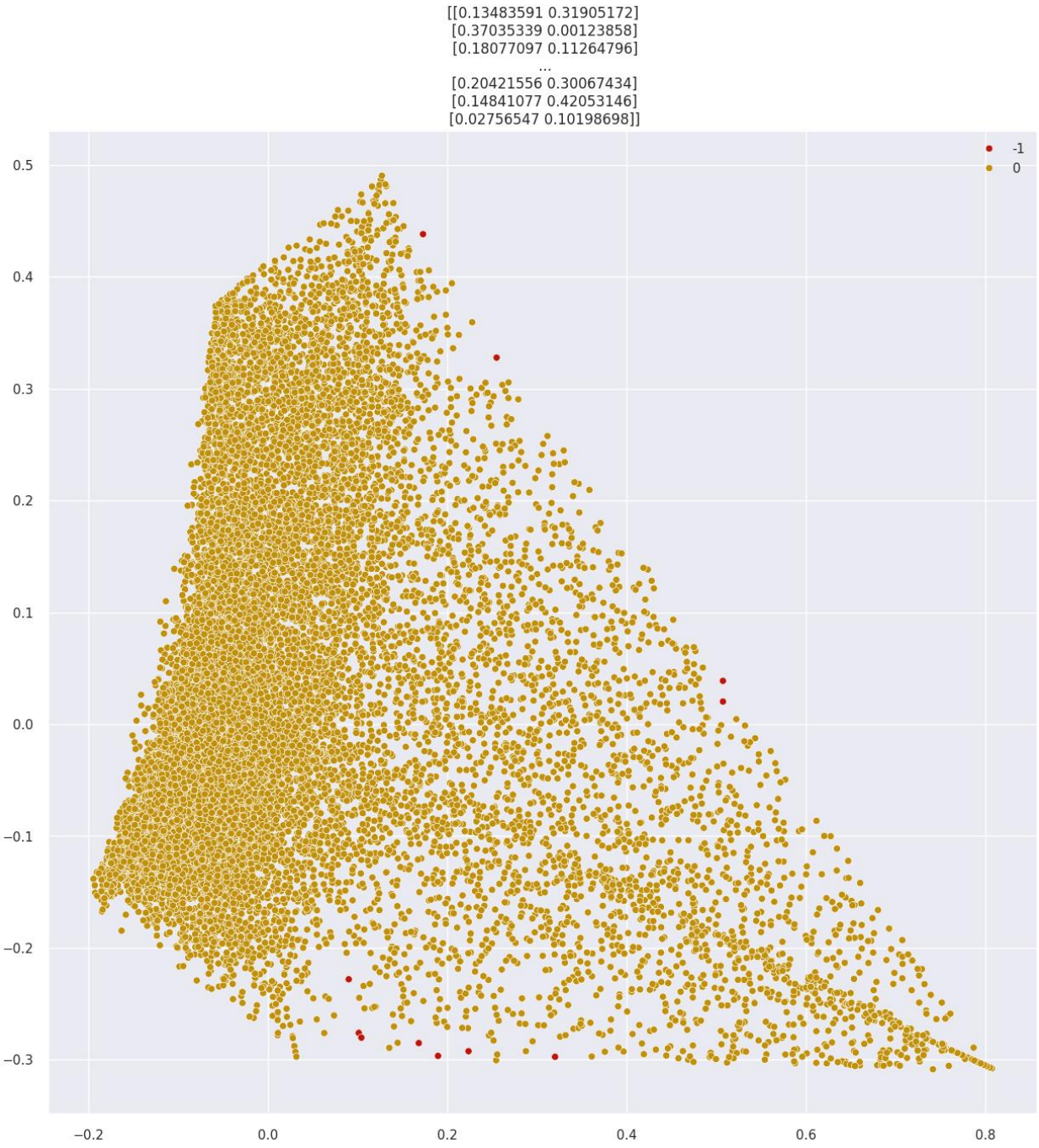
DBSCAN	Pca =2 Lda=30 Cluster=2 Tf-idf eps=0.1, min_samples=2	Pca =2 Lda=30 Cluster=3 Tf-idf eps=0.1, min_samples=5	Pca =2 Lda=30 Cluster=3 Tf-idf eps=0.1, min_samples=5	Pca =2 Lda=40 Cluster=15 Tf-idf eps=0.02, min_samples=2	Pca =2 Lda=40 Cluster=6 Tf-idf eps=0.02, min_samples=5	Pca =2 Cluster=4 BoW eps=0.02, min_samples=2	Pca =2 Cluster=2 BoW eps=0.02, min_samples=5
silhouette	0.678897098895 2632	0.657760853855 7453	0.676104 6648666 912	0.500917 9046544 635	0.579432 7779184 27	0.217003 2558881 6796	0.412463 0679549 918
Davies-Bouldin			0.54761 1165396 1602	2.45509 2773064 7163	4.5689 209560 25978	1.81296 7652864 6153	1.73871 2442679 1212

شماره شکل	1	2	15	14		3	4

تصاویری از اجراهای الگوریتم DBSCAN

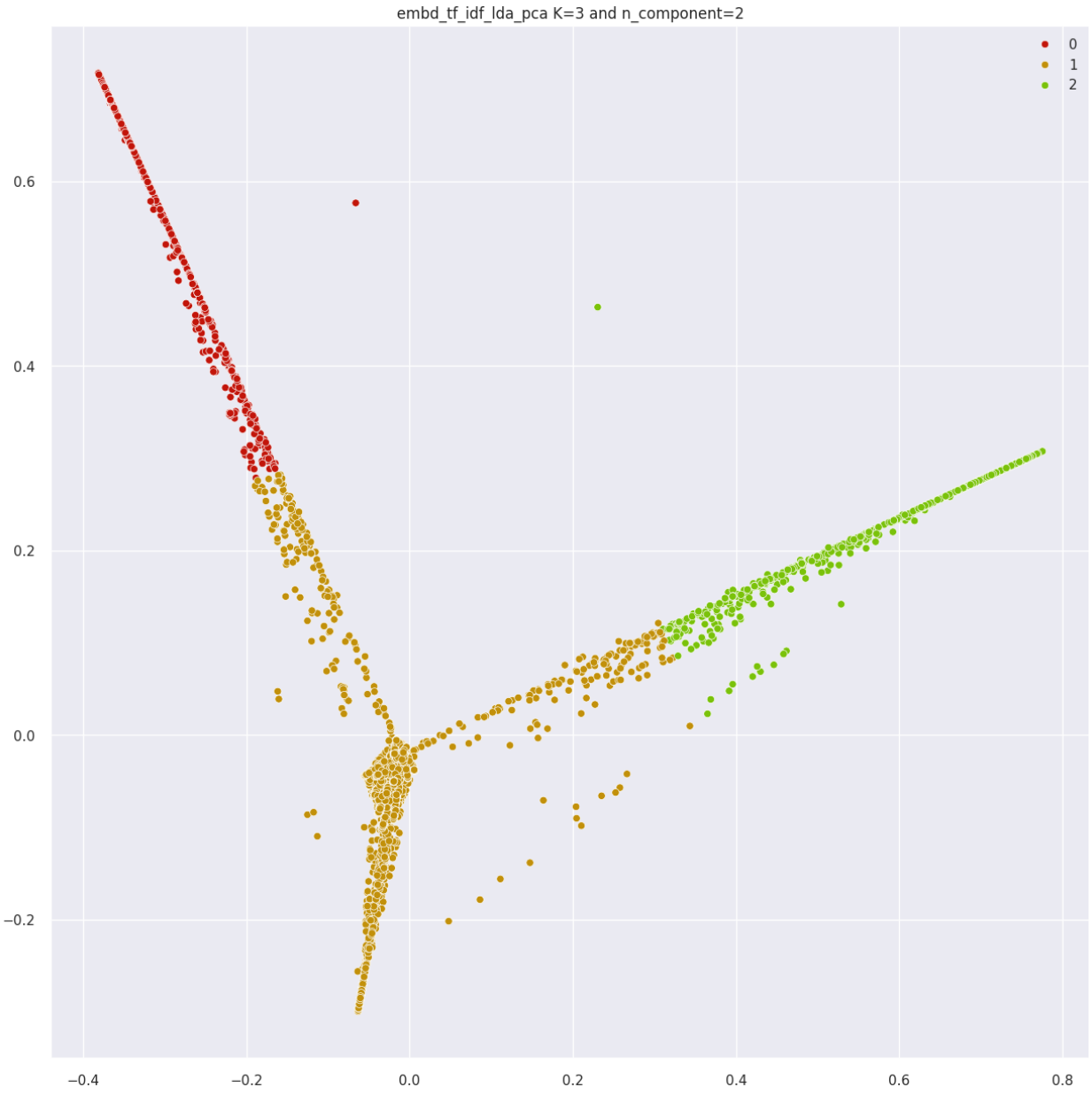

[[0.13483591 0.31905172]
[0.37035339 0.00123858]
[0.18077097 0.11264796]
...
[0.20421556 0.30067434]
[0.14841077 0.42053146]
[0.02756547 0.10198698]]
-1
0
1
2
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
0.0
0.2
0.4
0.6
0.8

شکل ۳

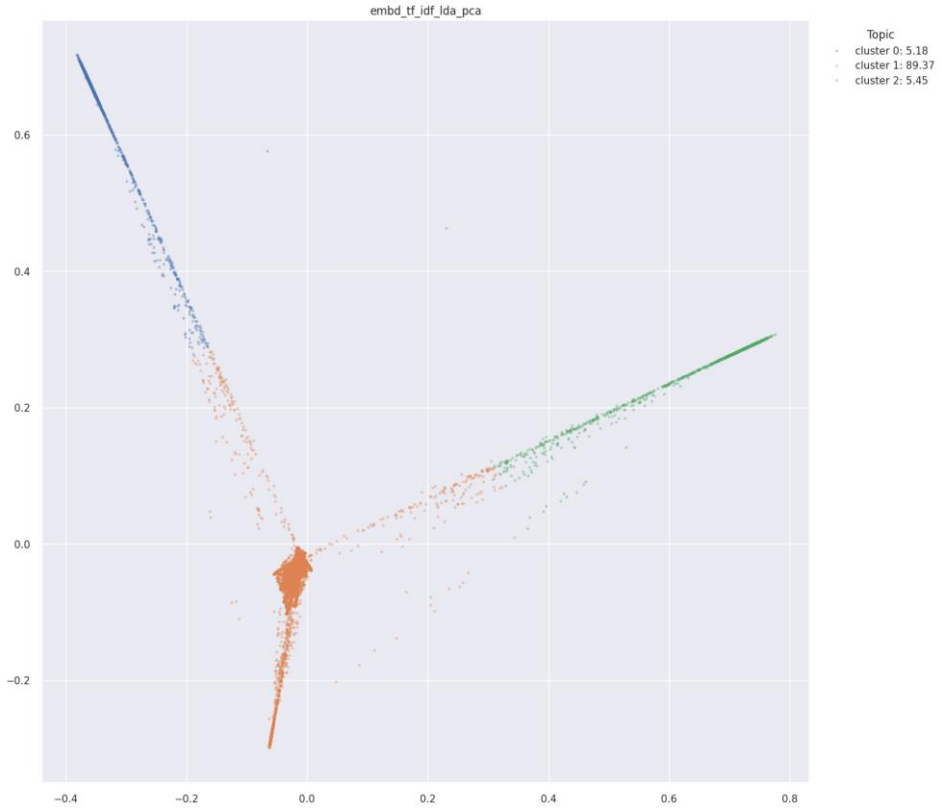

[[0.13483591 0.31905172]
[0.37035339 0.00123858]
[0.18077097 0.11264796]
...
[0.20421556 0.30067434]
[0.14841077 0.42053146]
[0.02756547 0.10198698]]
-1
0
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
0.0
0.2
0.4
0.6
0.8

شکل ۴

حالا به توضیح بهترین حالت در tfidf میپردازیم. این حالت بهترین امتیاز سیلوئت را داشته است. به تحلیل آن نیز میپردازیم.



شکل ۳ از جدول tf-idf-pca



نمایی دیگر از شکل ۳

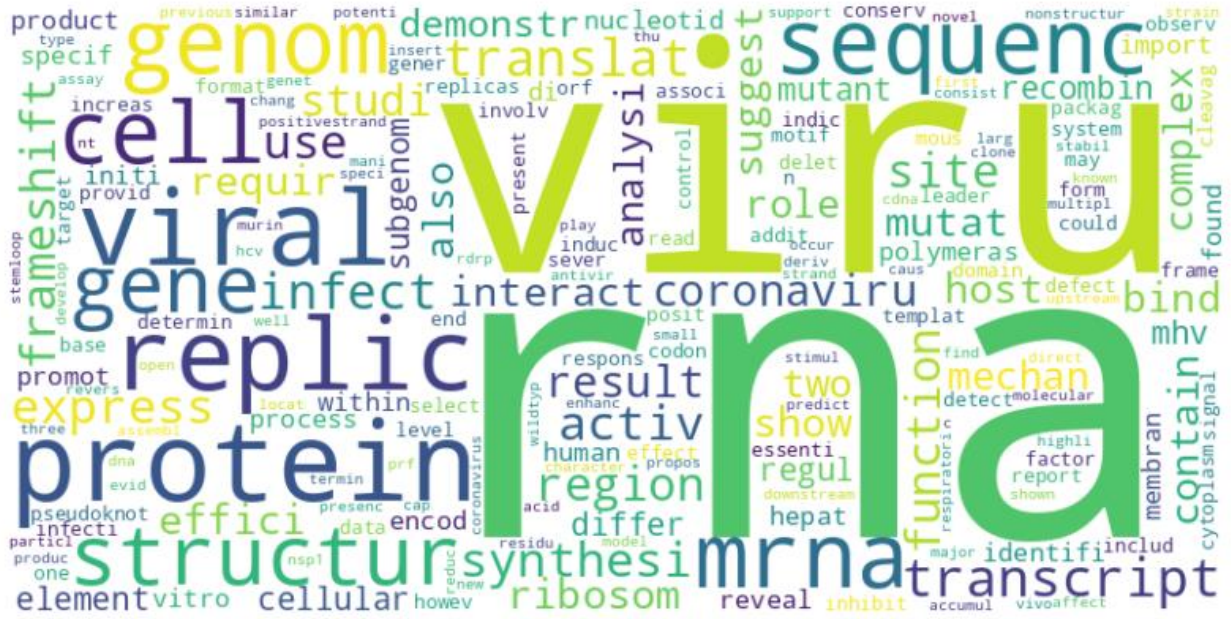
با توجه به رخداد کلمات در هر کلاستر میتوان گفت برای مثال کلاستر اول مرتبط با بحث تجزیه و تحلیل بیان ژن و پاسخ های ایمنی است. کلاستر دوم مرتبط با بحث پاسخ بافت تنفسی به عفونت آنفولانزا است و کلاستر سوم نیز مرتبط با بحث مدل های اپیدمیولوژیک برای گسترش و کنترل بیماری است.

```
Most Similar List of Words: ['rna', 'mrna', 'replic', 'protein', 'viru',
'genom', 'sequenc', 'transcript', 'viral', 'structur', 'gene', 'site',
'cell', 'region', 'express', 'effici', 'function', 'activ', 'mutat',
'contain']
Index of Maximum Similarity: 31
Topic 31: Gene Expression Analysis and Immune Responses

Most Similar List of Words: ['lung', 'infect', 'cell', 'respiratori',
'mous', 'acut', 'respons', 'tissu', 'diseas', 'sever', 'viru', 'immun',
'express', 'pneumonia', 'influenza', 'viral', 'model', 'activ', 'human',
'role']
Index of Maximum Similarity: 11
Topic 11: Respiratory Tissue Response to Influenza Infection

Most Similar List of Words: ['epidem', 'transmiss', 'estim', 'case',
'outbreak', 'model', 'china', 'number', 'diseas', 'spread', 'data',
```

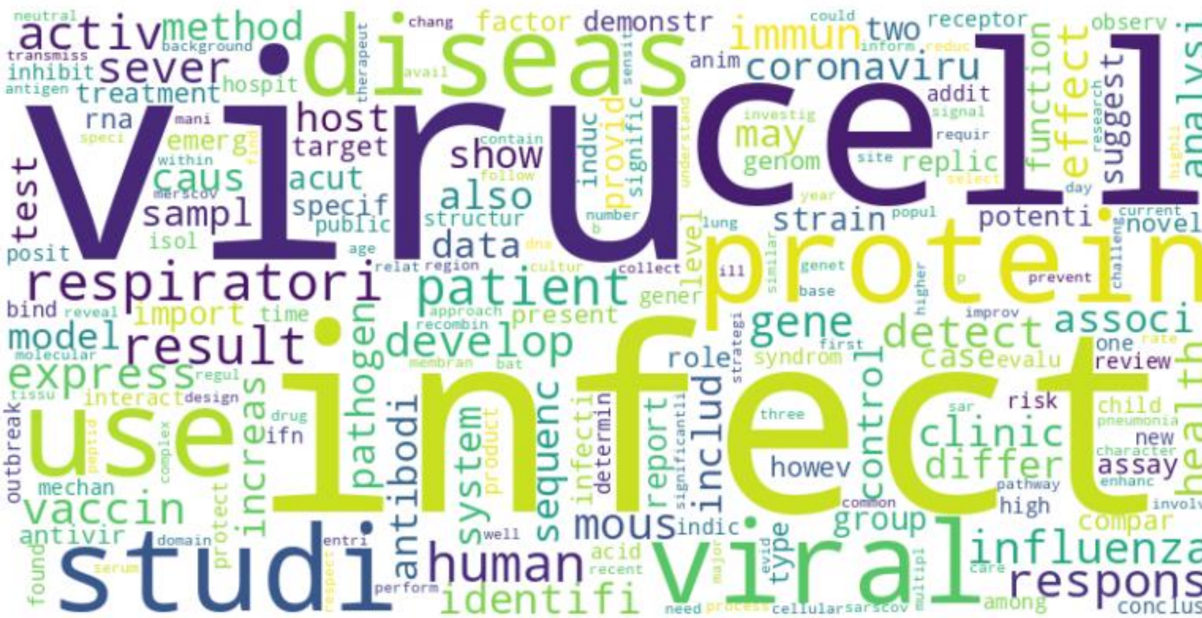
```
'infect', 'control', 'use', 'time', 'epidemiolog', 'measur', 'popul',
'coronaviru', 'infecti']
Index of Maximum Similarity: 7
      Topic 7: Epidemiological Models for Disease Spread and Control
```

از منظر دیگری نیز میتوان به ابر کلمات نگاه کرد و موضعات مرتبط در هر کلاستر را تشخیص داد

کلاستر اول که مرتبط با مباحث ریز تر مثل سلول، ویروس و rna است.

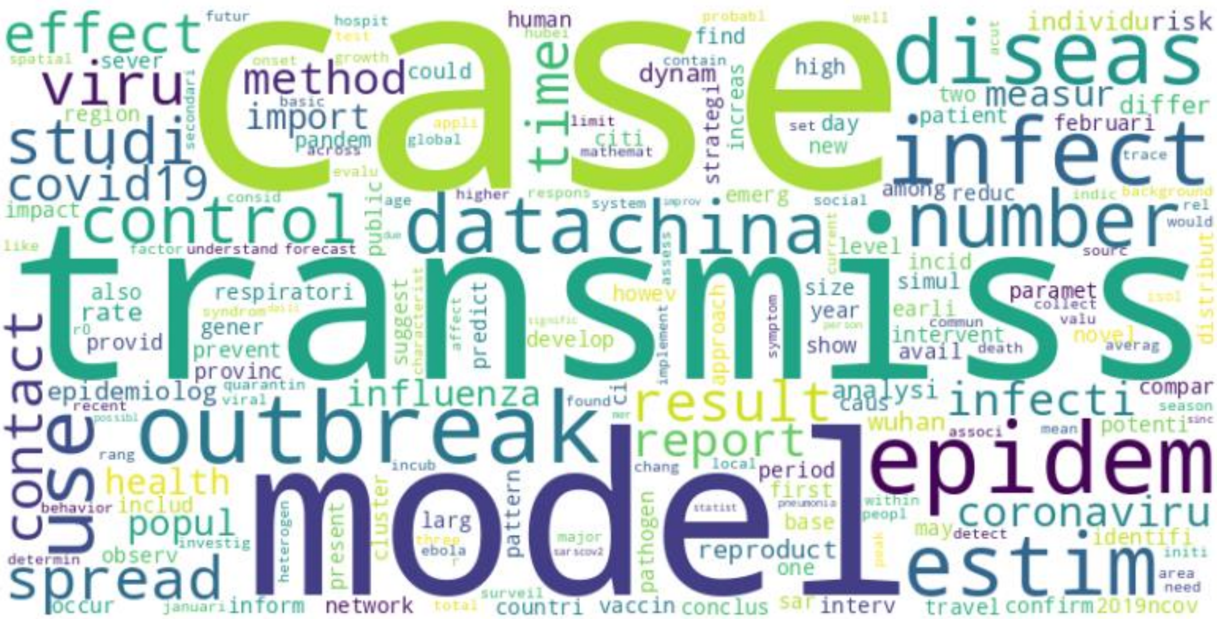


کلاستر دوم: که میتوان در مورد آلودگی و بحث درمان و شیوع بیماری صحبت کرد.



کلاستر سوم:

که میتواند در مورد انتقال بیماری و شیوع گسترده و اپیدمی باشد.



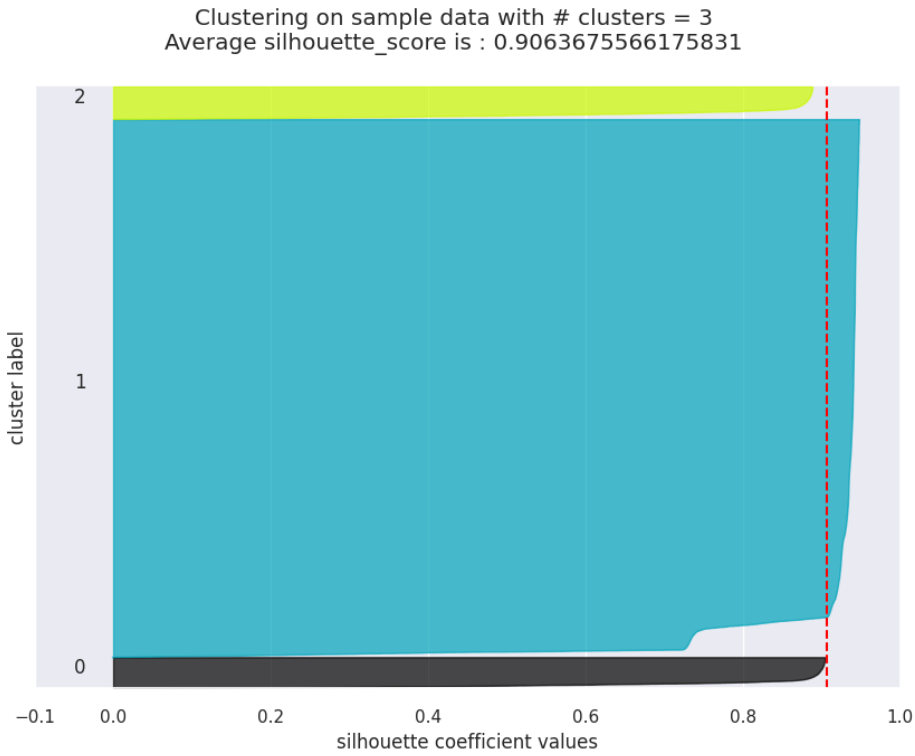
برای BOW در حالت kmeans و با sil = 0.9

```
Most Similar List of Words: ['infect', 'cell', 'viru', 'viral', 'host',
'cultur', 'human', 'replic', 'infecti', 'primari', 'system', 'use',
'studi', 'tissu', 'caus', 'establish', 'howev', 'spread', 'also',
'mechan']

Index of Maximum Similarity: 8
Topic 8: Protein Structure, Binding, and Functional Domains

Most Similar List of Words: ['health', 'public', 'surveil', 'global',
'emerg', 'system', 'countri', 'diseas', 'care', 'respons', 'inform',
'report', 'china', 'develop', 'research', 'need', 'outbreak', 'provid',
'data', 'region']
Maximum Similarity: system
Index of Maximum Similarity: 22
          Topic 22: Viral Receptor Binding, Mutations, and Entry
```

برای هر یک از کلاستر ها بیشترین لغاتی که تکرار شده است نیز جمع آوری شده است. و تاپیکی که برای کلاستر اول و دوم سوم انتخاب شده است ۸و ۲۲و ۳۱ است. و برای مثال نتیجه کلاستر اول مرتبط با ساختار پروتئین، اتصال، و دامنه های عملکردی است.

بطور مثال این ابر کلمه ای است که برای کلاستر اول ایجاد کردیم.

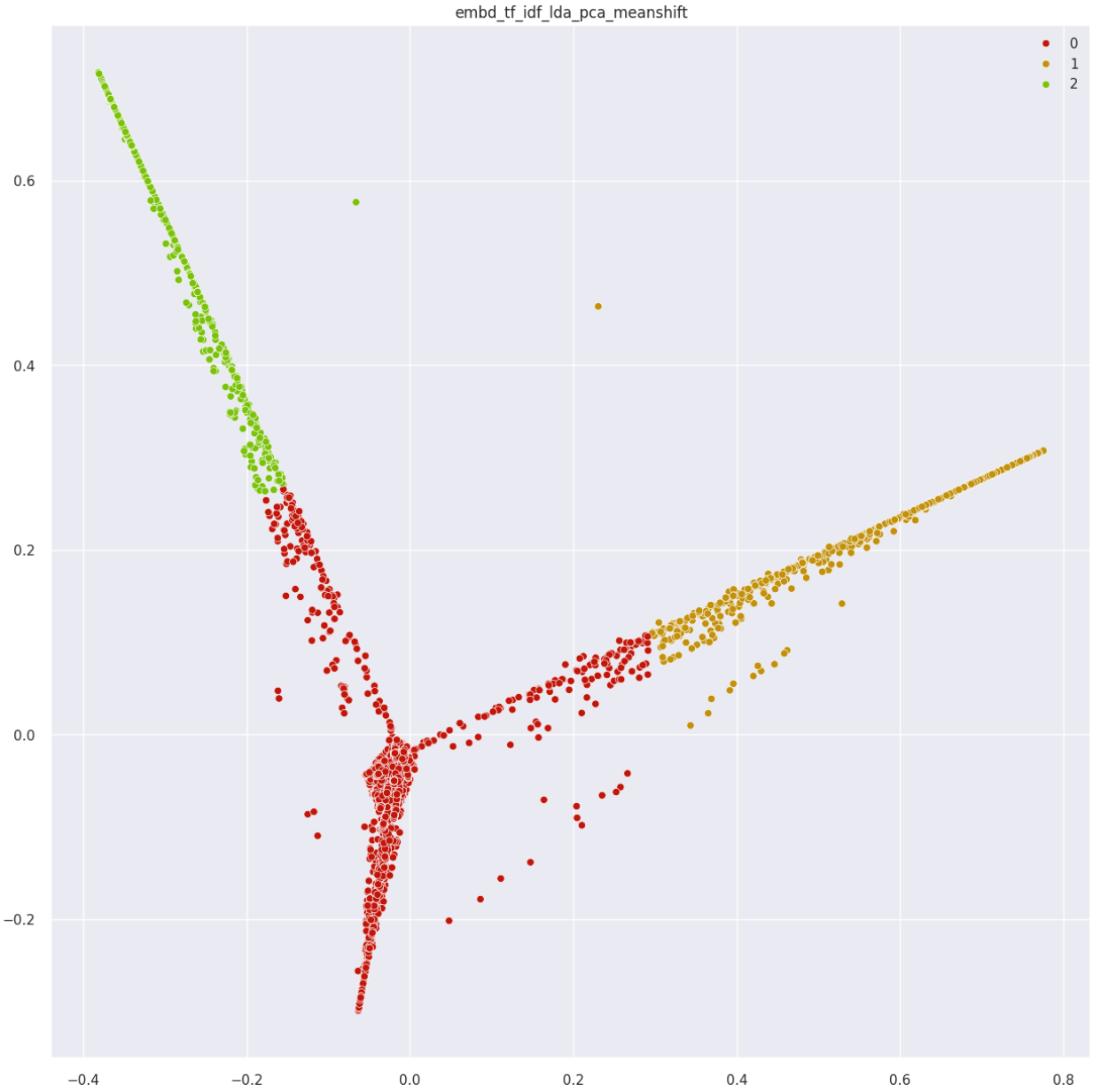
به صورت میانگین ضریب سیلویت برای k ها مختلف در بازه ۰.۸ تا ۱ قرار گرفته است.



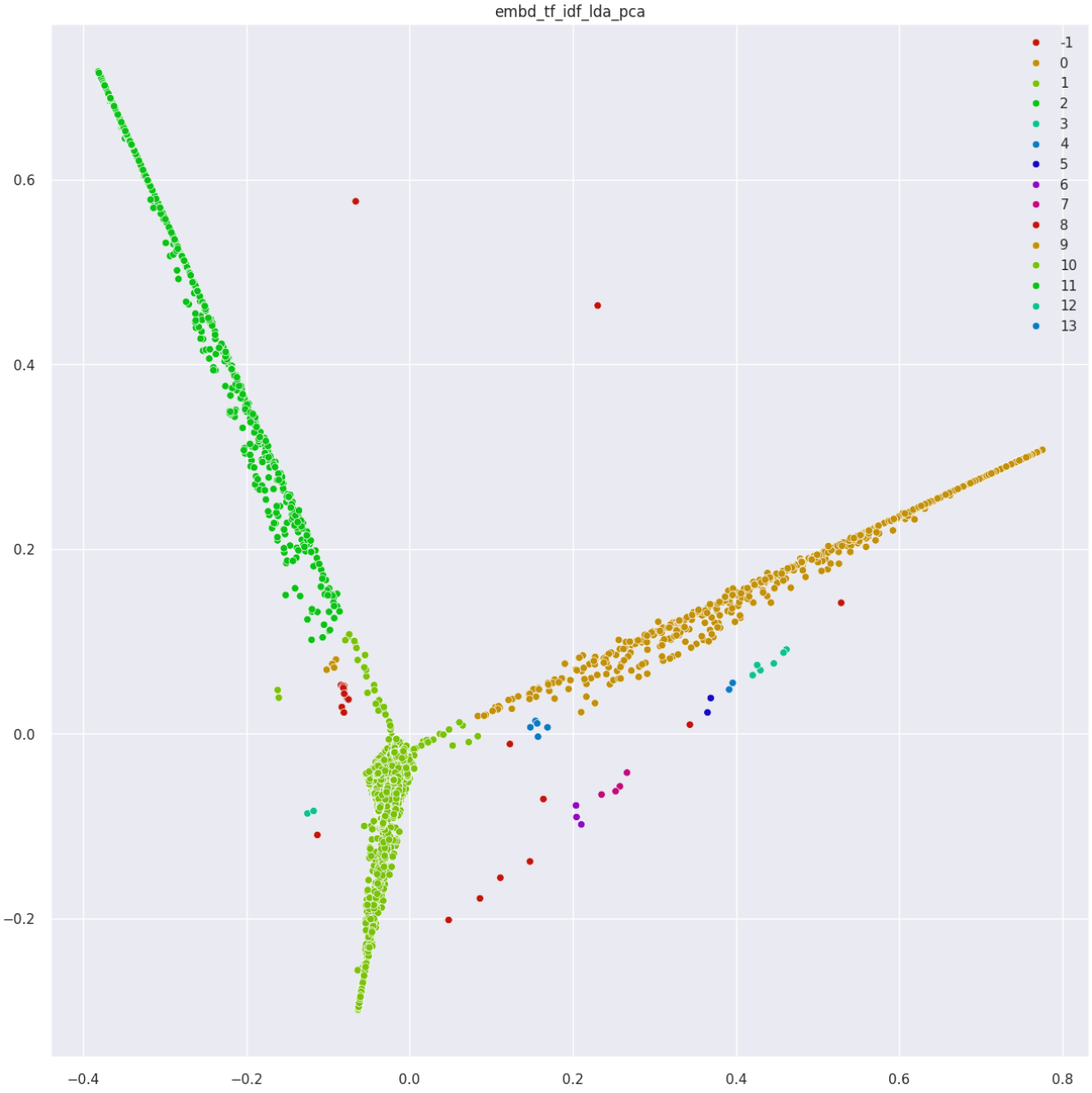
نمایی از اجراهای مختلف Mean_shift که نسبت به دیگر الگوریتم های استفاده شده هم در bow و هم در tf-idf نتایج خوبی را نشان داد.

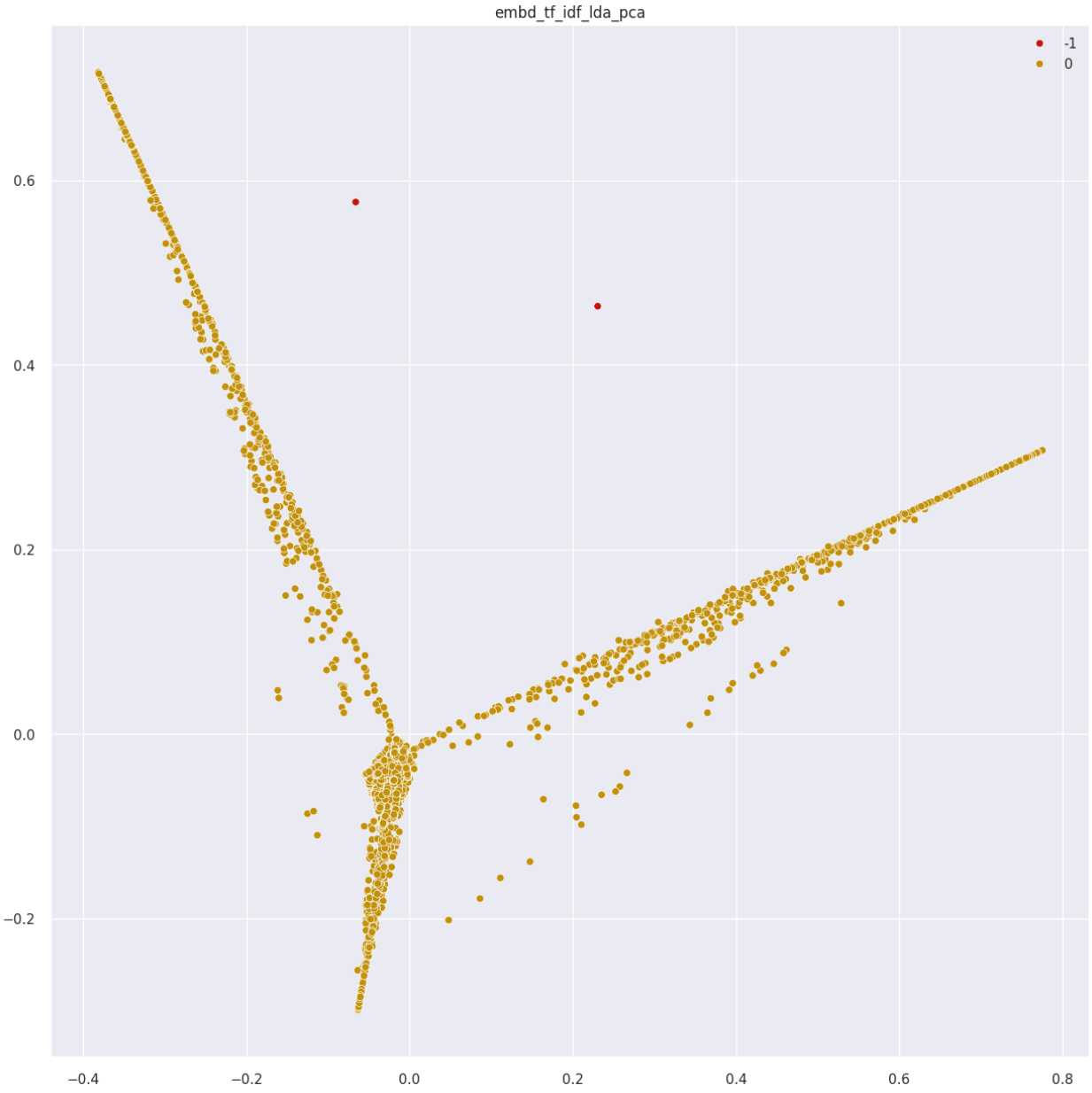
0.9058838612252527

Mean_shift	K=3, pca=2, lda=20	K=4, pca=2, lda=25	K=3, pac=2, lda=40	K=3, umap, lda =40	K=4 Umpa Lda=20	K=3 Umpa Lda=20	
silhouette	0.877057912562349	0.8749292371478099	0.9058838612252527	0.37730643	0.3527427	0.37235686	
Davies-Bouldin index:			0.21145831175495547				
			13				


embd_tf_idf_lda_pca_meanshift
0
1
2
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

شکل ۱۳


embd_tf_idf_lda_pca
-1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8


embd_tf_idf_lda_pca
-1
0
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

Optics and tf-idf

```
min_samples=1000
```

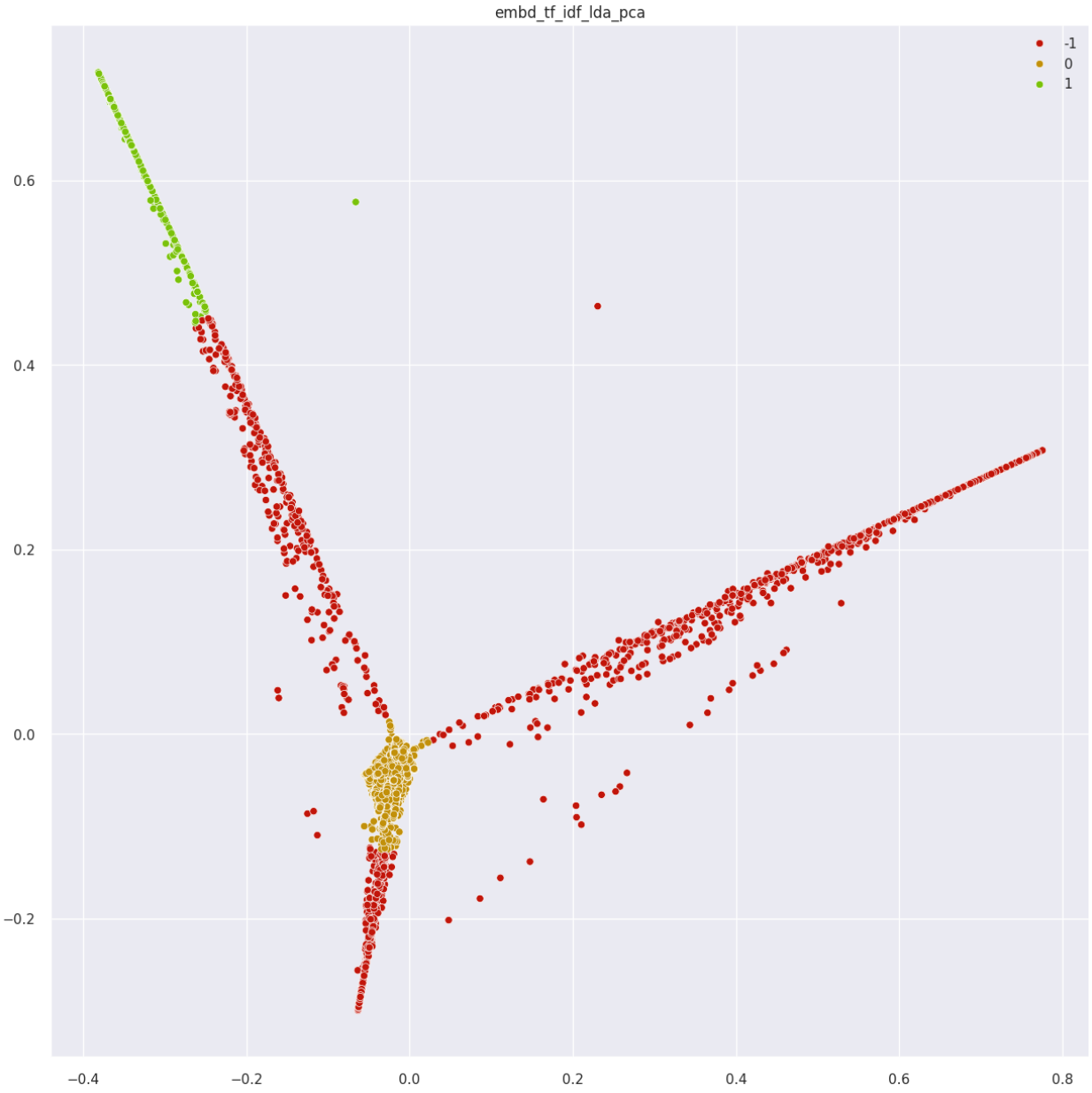
number of cluster = 3

```
silhouette_score
```

0.7916178385946345

```
Davies-Bouldin index
```

1.2208399694536198


embd_tf_idf_lda_pca
-1
0
1
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

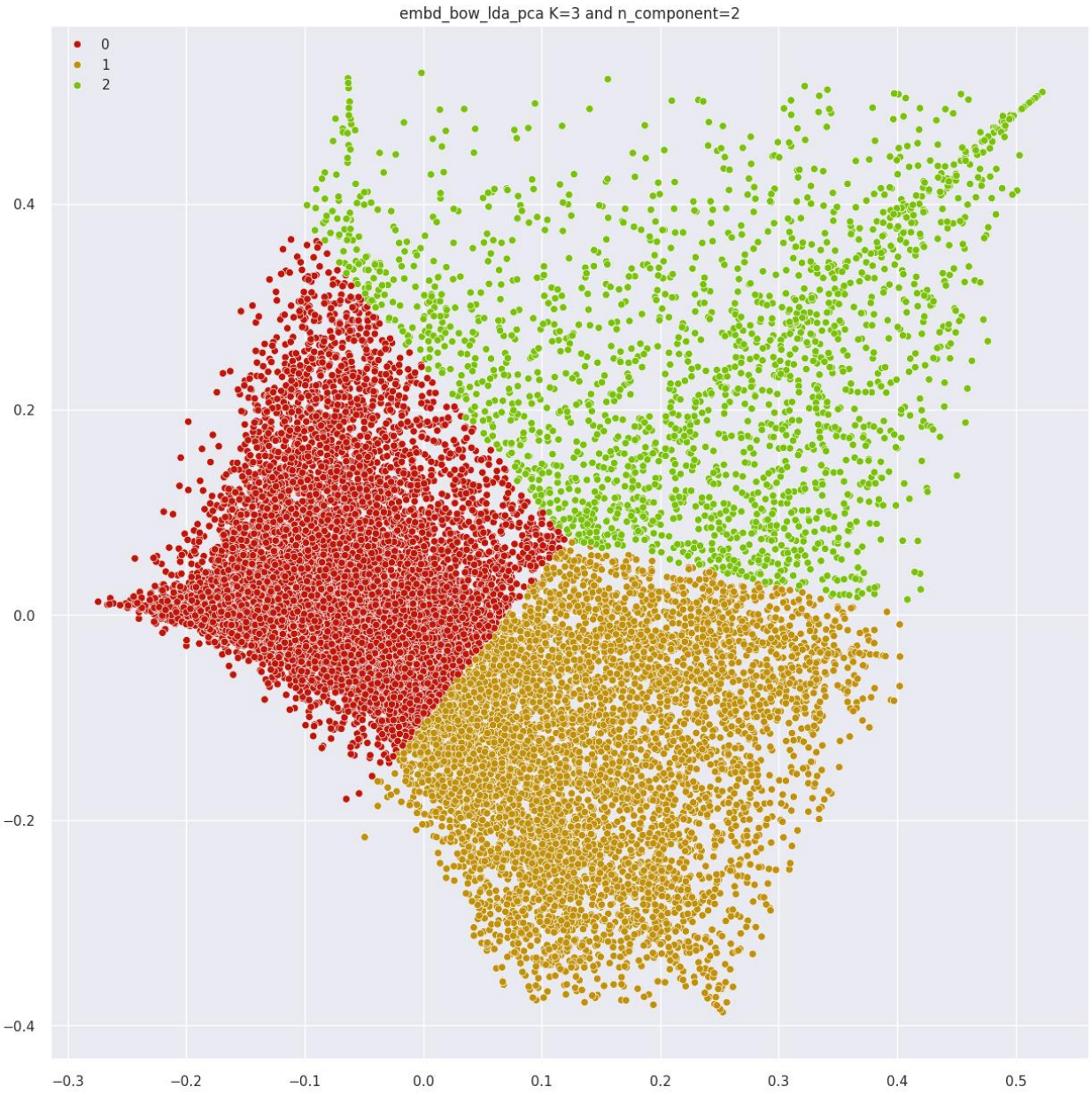
Bow

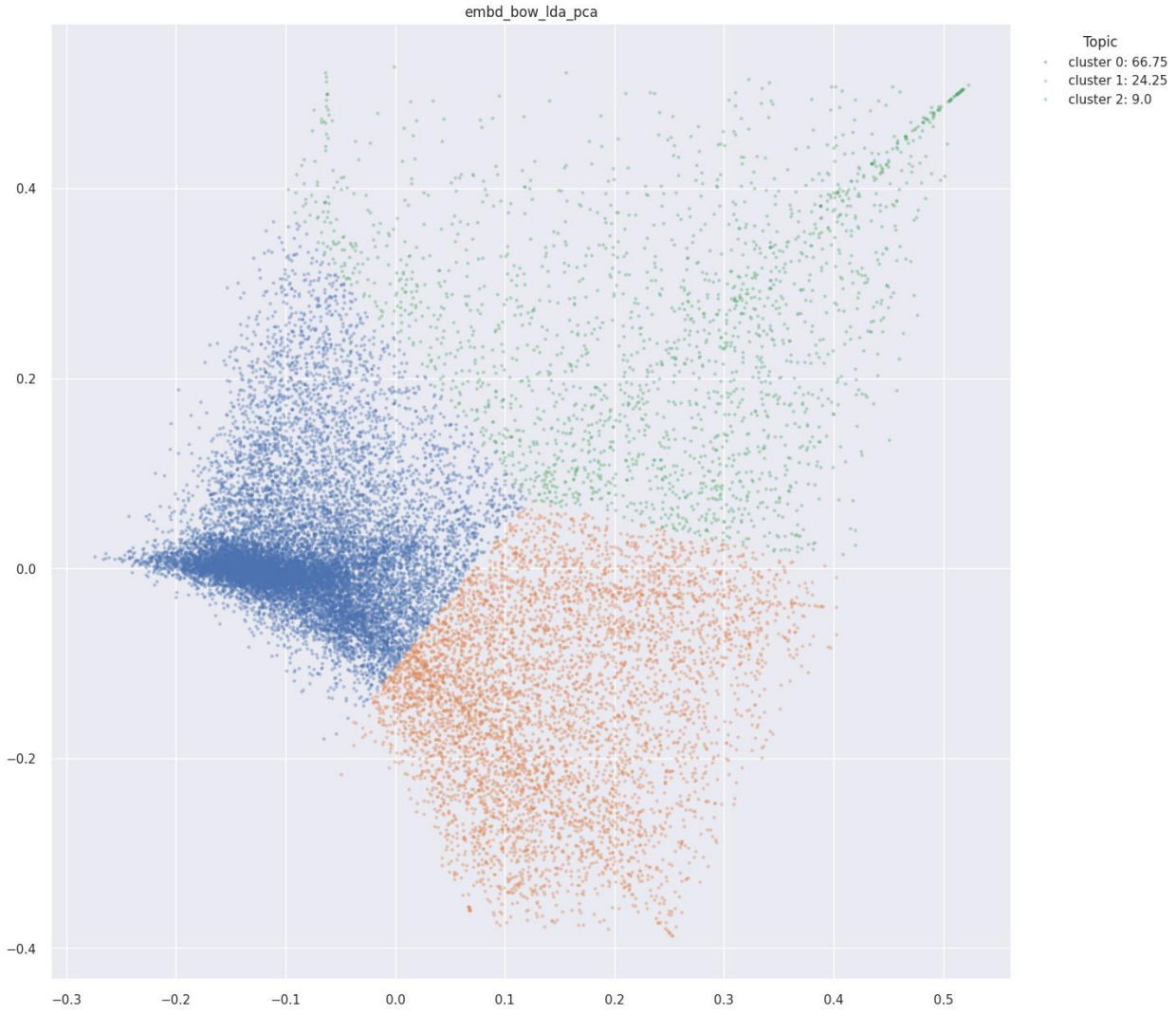
Lda = 40

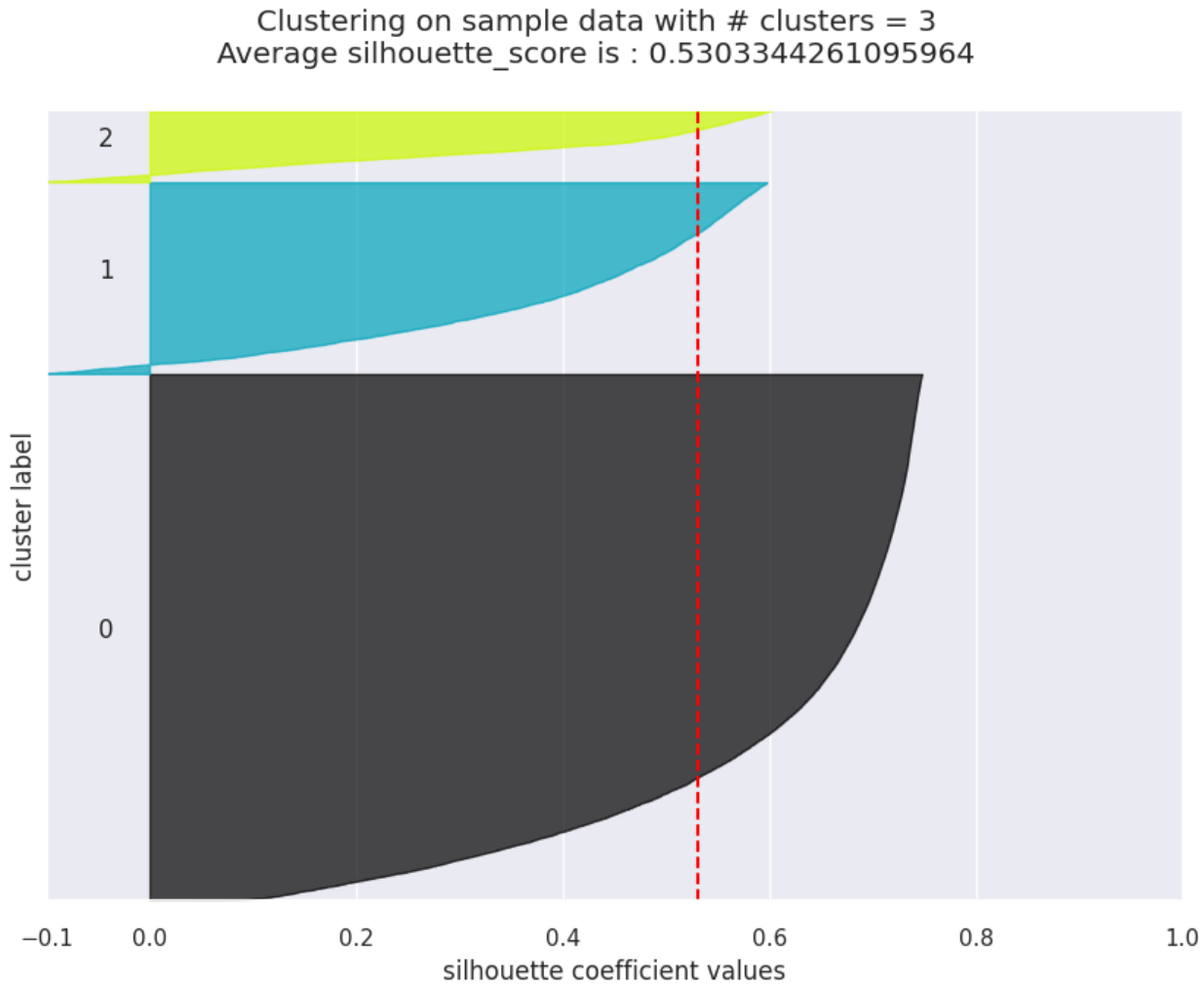
Pca =2

K=3

silhouette_score 0.5303344261095964
Davies-Bouldin index: 0.7447208458599722




embd_bow_lda_pca
Topic
cluster 0: 66.75
cluster 1: 24.25
cluster 2: 9.0
0.4
0.2
0.0
−0.2
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5

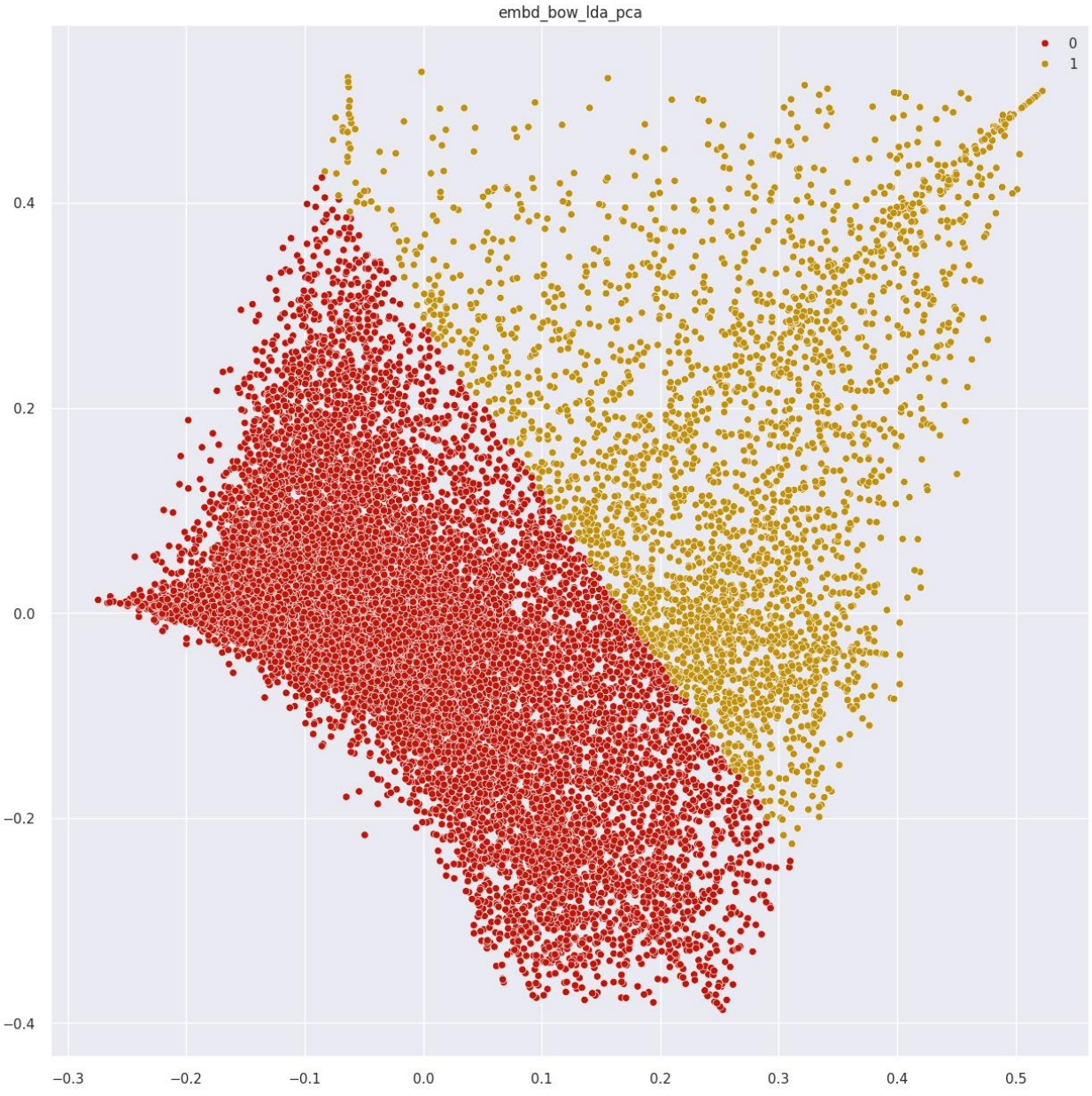


Mean- shift

```
bandwidth=0.1
silhouette_score
```

0.5226578911593829

```
Davies-Bouldin index: 0.9092823906391729
```

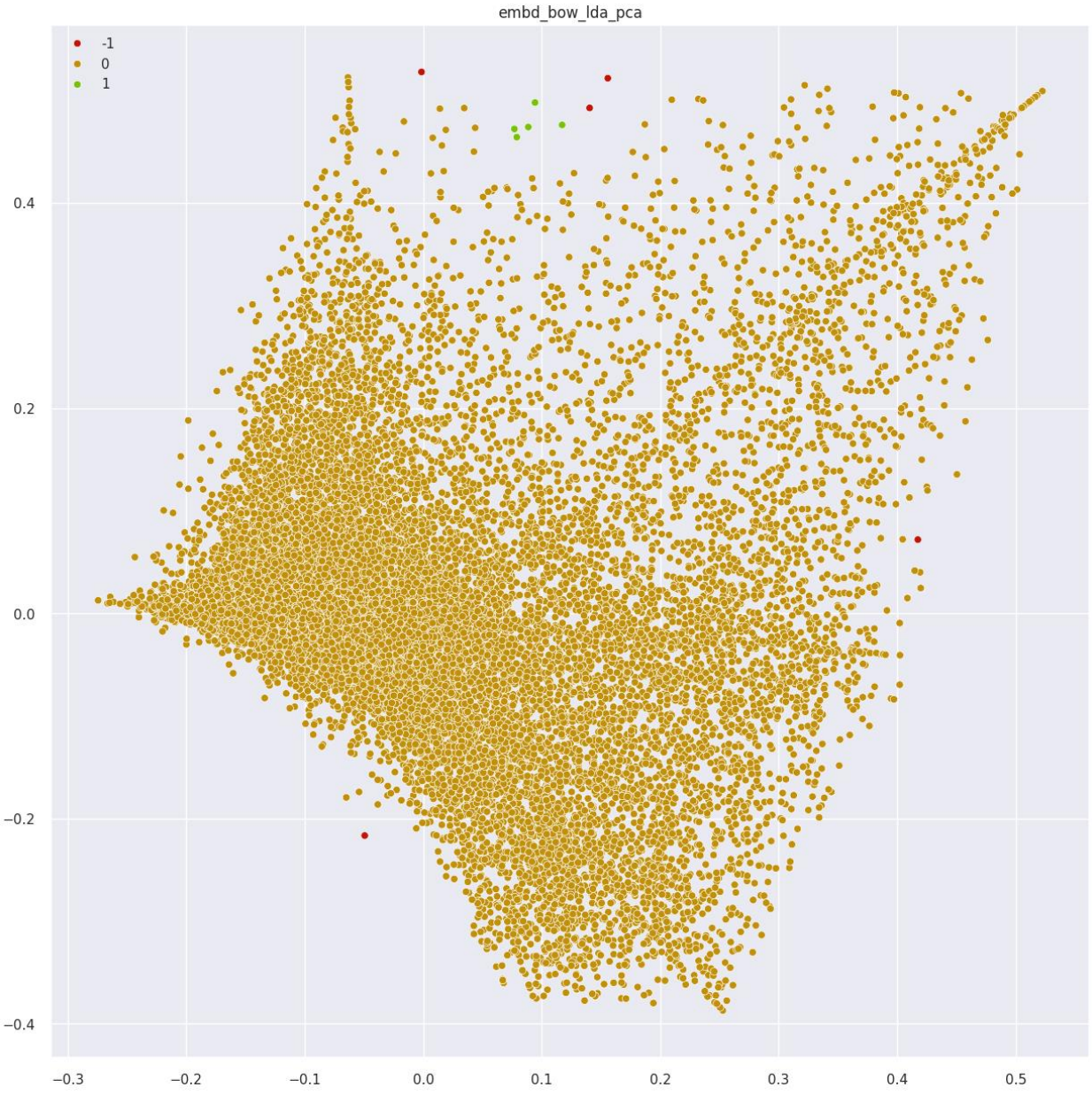


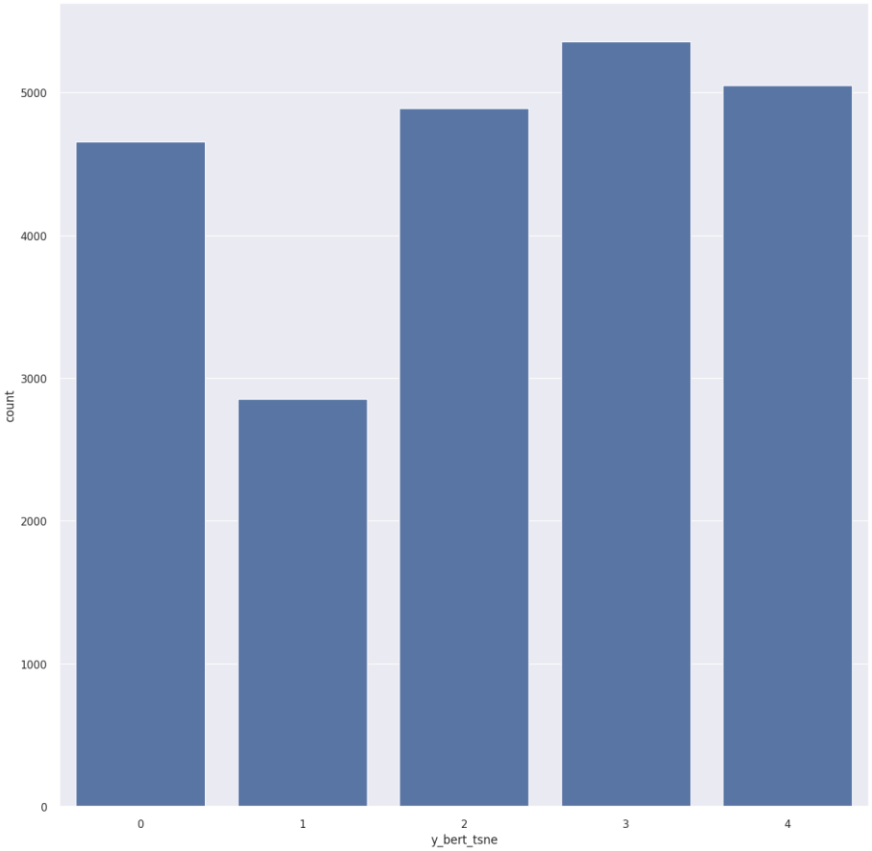
Dbscan

```
eps=0.03, min_samples=5
silhouette_score
```

0.4867697732251089

```
Davies-Bouldin index: 1.650833686858342
```


embd_bow_lda_pca
-1
0
1
0.4
0.2
0.0
−0.2
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5



در آخر هم نموداری داشتیم که میتوانست مقدار و بالانس بودن هر یک از کلاسترها را نشان دهد. که به دلیل جلوگیری از زاید شدن مطالب صرفا در انتها آورده شد که در صورت لزوم بتوان از آن استفاده نمود.

نتیجه گیری:

بنظرم از میان همه روش های vectoriczation بین tf-idf و bow، tf-idf نتیجه بهتری داشته است. چون بردارهای آن بر اساس وزن کلماتی است که تاثیر بیشتری دارند. و بنابراین به نسبت عملکرد بهتری داشته است.

در مورد الگوریتم ها وابسته به مقادیر داده شده، دو الگوریتم k-means و mean shift در هر دو حالت bow و tf-idf عملکرد بهتری را با توجه به معیار سیلوئت داشته اند. اما به نسبت باز نتایج در tf-idf بهتر شده است.

در مورد دیگر الگوریتم های optics و DBscan مجددا نتایج در tf-idf همانطور که به دلیل آن بیان شد، نتایج بهتری داشته است.

اما نتایج k-means و mean-shift بهتر بود آن ممکن است به خاطر جنس داده های متنی و اثری که وکتوری شدن روی آن ها گذاشته است باشد. احتمال دارد که ساختار کروی شکل تاثیر بهتری را گذاشته باشد.

در کل از روش DB SCAN راضی نبودم واینجا خوب جواب نداد و همچنین optics. شاید در این جور مسایل از روش های چگالی محور استفاده نشود بهتر باشد.