بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

بازیابی پیشرفته‌ی اطلاعات

نیم‌سال دوم تحصیلی ۰۰-۰۱

تمرین چهارم: خوشه‌بندی نوشته‌های مرتبط با سلامت/بیو

محمدمهدی ابوترابی ۹۸۱۰۵۵۵۷
یاسمن زلفی موصلو ۹۸۱۰۵۷۹۵
فاطمه عسگری ۹۸۱۰۵۹۲۱

ساختار کلی پروژه:

در این تمرین، پوشه‌ها و فایل‌های لازم به پروژه‌ی مربوط به تمرین سوم اضافه شده است که در ادامه آن‌ها را توضیح خواهیم داد:

پوشه‌ی crawlers_classification: از آنجایی که داده‌هایی که در تمرین سوم جمع‌آوری شده بود دسته‌ی متون را مشخص نمی‌کرد و فقط دارای عنوان و محتوای متن بود، داده‌های این تمرین همراه با دسته‌ی هر متن جمع‌آوری شده‌اند. کراولر نوشته شده برای سایت نمناک است و از هر دسته، یا دیتای ۱۰۰ صفحه و یا اگر کم‌تر بوده، تمام دیتای موجود کراول شده است. این دیتای جمع‌آوری‌شده در کنار دیتایی که متعلق به مخزن گیتهاب ترم پیش بود، تمام دیتا را تشکیل داده‌اند. این مجموعه داده شامل ۷۰۵۳ متن همراه با عنوان و دسته‌ی مربوطه است که حدودا ۵۰ مگابایت حجم دارد و در داخل پوشه‌ی new-dataset-with-category ذخیره شده است.

پوشه‌ی notebooks: داخل این پوشه سه نوت‌بوک جدید اضافه شده است که به تفصیل به توضیح هر یک خواهیم پرداخت:

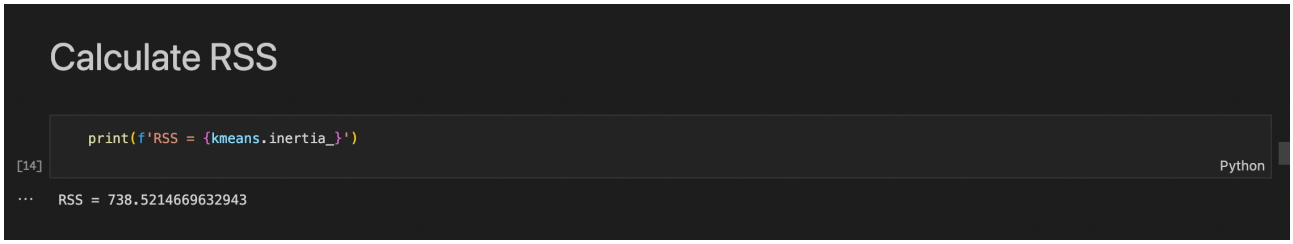
نوت‌بوک main-clustering-wth-fasttext: در این نوت‌بوک ابتدا داده لود شده است. کلاس FastTextEmbedingModel به این شکل است که ابتدا یک داکیومنت گرفته و امبدینگ fasttext آن را برمی‌گرداند. روش کار آن اینگونه است که از عنوان و متن سند استفاده می‌کند. متن سند را توکنایز کرده و امبدینگی که در نظر می‌گیرد، میانگین امبدینگ کلمات توکنایز شده است. برای عنوان هم دقیقا به همین روش امبدینگ به دست می‌آید و در آخر متن و عنوان سند را با نسبت ۱۹ به ۱ میانگین وزن‌دار می‌گیرد تا امبدینگ سند به دست آید. توجه شود که مدل fasttextای که در این کلاس استفاده شده، همان مدل fasttextای است که در تمرین سوم train شده بود.
۷ دسته‌ی اصلی داده‌ها عبارت‌اند از: سلامت روان، دهان و دندان، پوست و مو، تغذیه، سلامت خانواده، سلامت جنسی، و پیشگیری و بیماری‌ها. برای هر دسته ۳۵۰ (یا کم‌تر) سند متعلق به آن دسته برداشته شده و امبدینگ آن‌ها با استفاده از مدل fastext پیاده‌سازی‌شده، به دست آمده است.
سپس به روش Kmeans و با استفاده از ۷ دسته، روی این ۲۴۳۴ سند خوشه‌بندی انجام گرفته است.
سپس (در قسمت Samples from clusters) از دوتا از خوشه‌ها (اولی و دومی به ترتیب نزدیک به سلامت جنسی و دهان و دندان) سمپل گرفته شده است که خروجی آن در تصاویر زیر قابل مشاهده است و خروجی قابل قبول و مربوطی نیز هست:

```
[42]  samples_from_cluster(doc_titles, Y_predicted, 1)                    Python

...  ['اتفاق های خوب بعد از رابطه جنسی!',
      'عوارض رابطه جنسی ناقص برای زن و مرد | رابطه جنسی!',
      'رابطه جنسی سالم چه شکلیه ؟',
      'مضرات زیاده روی در رابطه زناشویی ',
      'کاری که زنان بعد از رابطه جنسی باید انجام دهند 8']
```

```
[47]  samples_from_cluster(doc_titles, Y_predicted, 6)                    Python

...  ['دلیلی اصلی برتری لمینت دندان به کامپوزیت ونیر 5',
      'علل و علائم پری ایمپلنتایتیس + درمان و ..',
      'دلیل اصلی بدرنگی و زرد شدن دندان ها چیست؟',
      'اثر باورنکردنی کلاژن روی دندون هامون',
      'ایمپلنت کاشت دندان , همه چیز درباره بریج دندانی خوب']
```

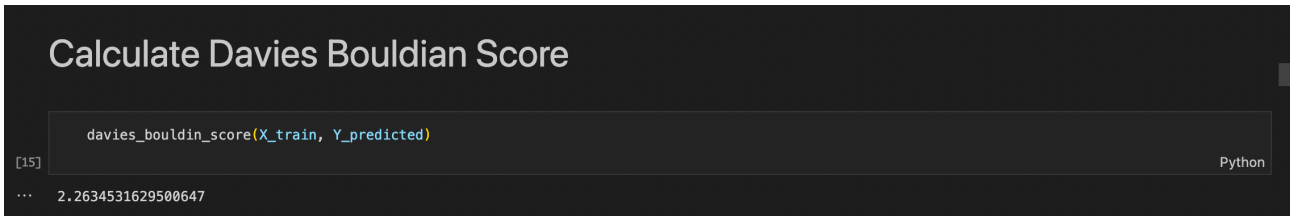
برای ارزیابی خوشه‌بندی انجام گرفته، از ۴ معیار استفاده شده است.

1. معیار RSS: در این معیار، مربع فاصله‌ی نقاط از مرکز خوشه‌هایی که در آن قرار دارند محاسبه می‌شود. هرچه مقدار به دست آمده در این معیار کم‌تر باشد بهتر است. برای خوشه‌بندی ما این مقدار حدودا برابر ۷۳۸/۵۲۱ شده است:



برای اطلاعات بیش‌تر درباره‌ی این معیار می‌توانید به این لینک مراجعه نمایید.

2. Davies Bouldin score: در این معیار، شباهت دو خوشه‌ی نزدیک به هم محاسبه می‌شود. امتیاز محاسبه‌شده در این معیار هرچه به صفر نزدیک‌تر باشد یعنی خوشه‌ها از هم جداتر هستند و بهتر است که برای خوشه‌بندی ما حدودا برابر ۲/۲۶۳ شده است:



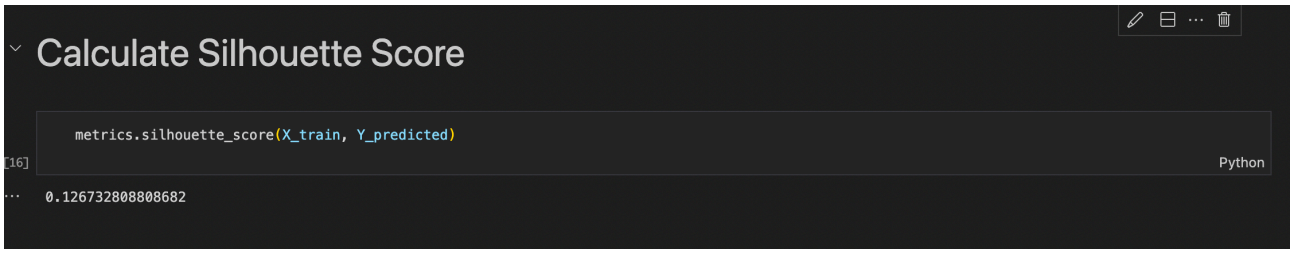
برای اطلاعات بیش‌تر درباره‌ی این معیار می‌توانید به این لینک مراجعه نمایید.

3. Silhuette score: مقدار محاسبه‌شده در این معیار با استفاده از فرمول زیر به دست می‌آید:

Silhuette score = (b - a) / max(a, b)

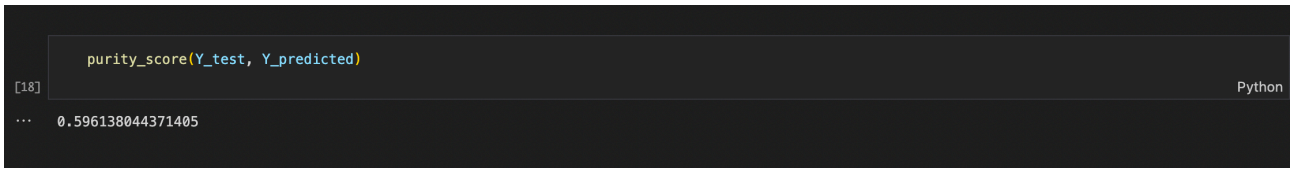
که b فاصله‌ی نقاط از خوشه‌های مختلف و a میانگین فاصله‌ی نقاط داخل یک خوشه است.

هرچه فاصله‌ی نقاط از خوشه‌های دیگر نسبت به فاصله‌ای که نقاط در یک خوشه نسبت به هم دارند بیش‌تر باشد بهتر است. در واقع امتیاز محاسبه‌شده در این معیار عددی بین ۱- و ۱ خواهد بود که هرچه به ۱ نزدیک‌تر باشد بهتر است و برای خوشه‌بندی ما حدودا برای ۰/۱۲۶ شده است:



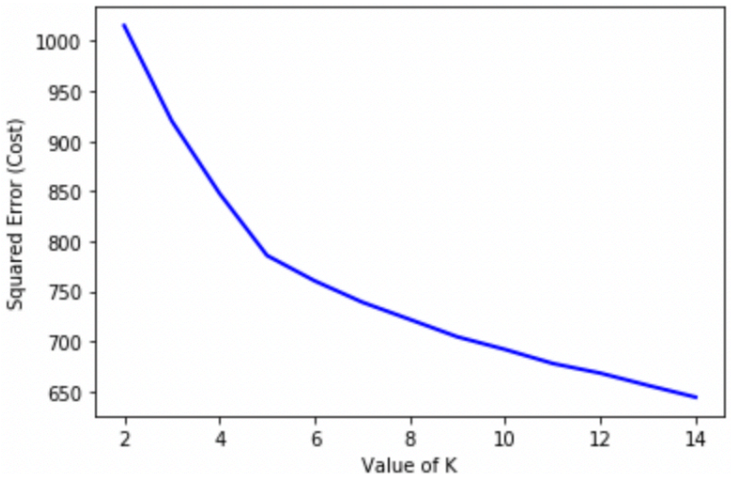
برای اطلاعات بیش‌تر درباره‌ی این معیار می‌توانید به این لینک مراجعه نمایید.

4. Purity score: در این معیار، محاسبه می‌شود که خوشه‌ها به لیبل‌های واقعی چقدر شبیه هستند. این مقدار هرچه به ۱۰۰ درصد (یعنی ۱) نزدیک باشد بهتر است و برای خوشه‌بندی ما حدودا برابر ۶۰ درصد شده است:



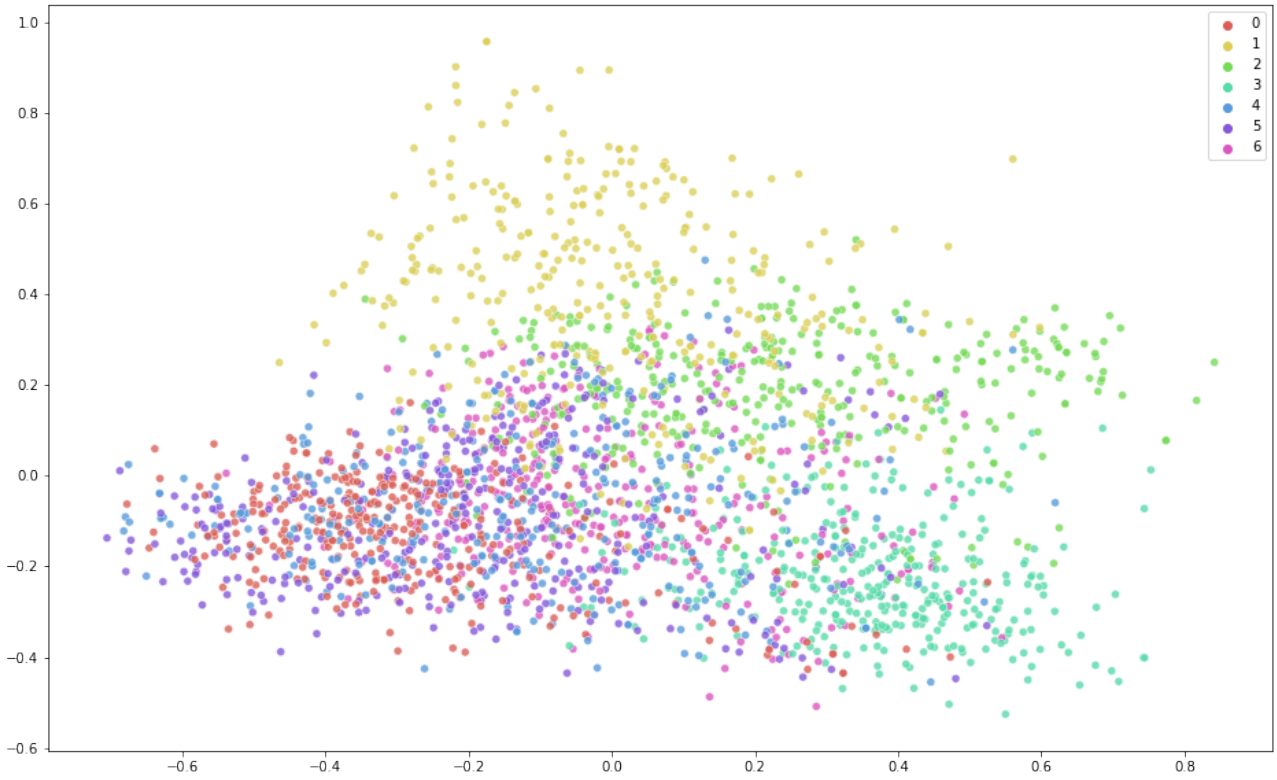
برای اطلاعات بیشتر درباره‌ی این معیار می‌توانید به این لینک مراجعه نمایید.

در قسمت بعدی نوتبوک، به روش elbow مقدار بهینه‌ی K محاسبه شده است که برابر ۵ شده است. این نشان می‌دهد که بعضی از دسته‌ها بسیار به هم شبیه هستند. نمودار squared error بر حسب مقدار k را در تصویر زیر مشاهده می‌کنید:

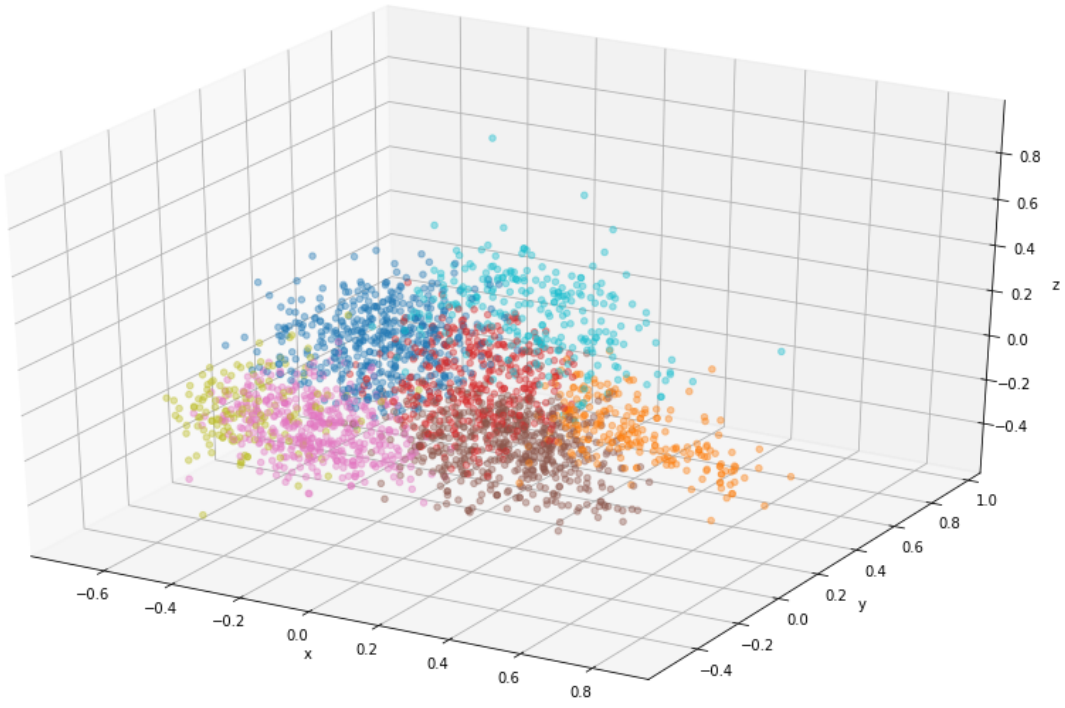


سپس با استفاده از PCA و TSNE که دو الگوریتم برای کاهش ابعاد داده هستند داده‌های خوشه‌بندی‌شده plot شده‌اند. از آنجایی که اندازه‌ی فضای برداری داده‌ها بسیار بزرگ است و visualize کردن آن‌ها عملا غیرممکن است، این دو الگوریتم برای کاهش فضای برداری و نشان دادن داده‌ها در ۲ و ۳ بعد استفاده شده‌اند.
در قسمت Plot Clusters Using PCA از PCA داخل sklearn استفاده شده است و تعداد کامپوننت، همان بعد برداری است که PCA خروجی می‌دهد. سپس داده‌ها فیت شده‌اند. pca_x و pca_y به ترتیب به عنوان بعد اول و دوم به plt داده شده‌اند، لیبل‌های Y_predicted یعنی آن خوشه‌هایی که Kmeans ایجاد کرده به عنوان hue داده شده و نمودار به شکل زیر رسم شده است:

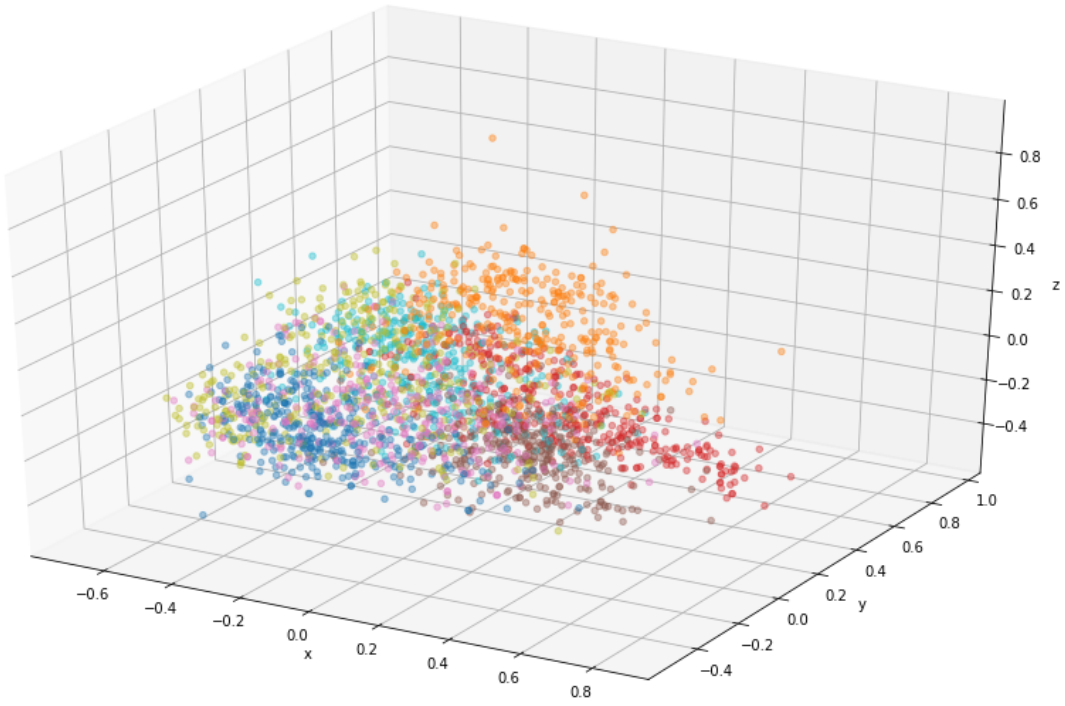
در قسمت بعدی یعنی Documents True Classes لیبل‌هایی که داده‌ها واقعا داشته‌اند نشان داده شده‌اند و Y_test به عنوان hue داده شده است. نمودار به شکل زیر رسم شده است:



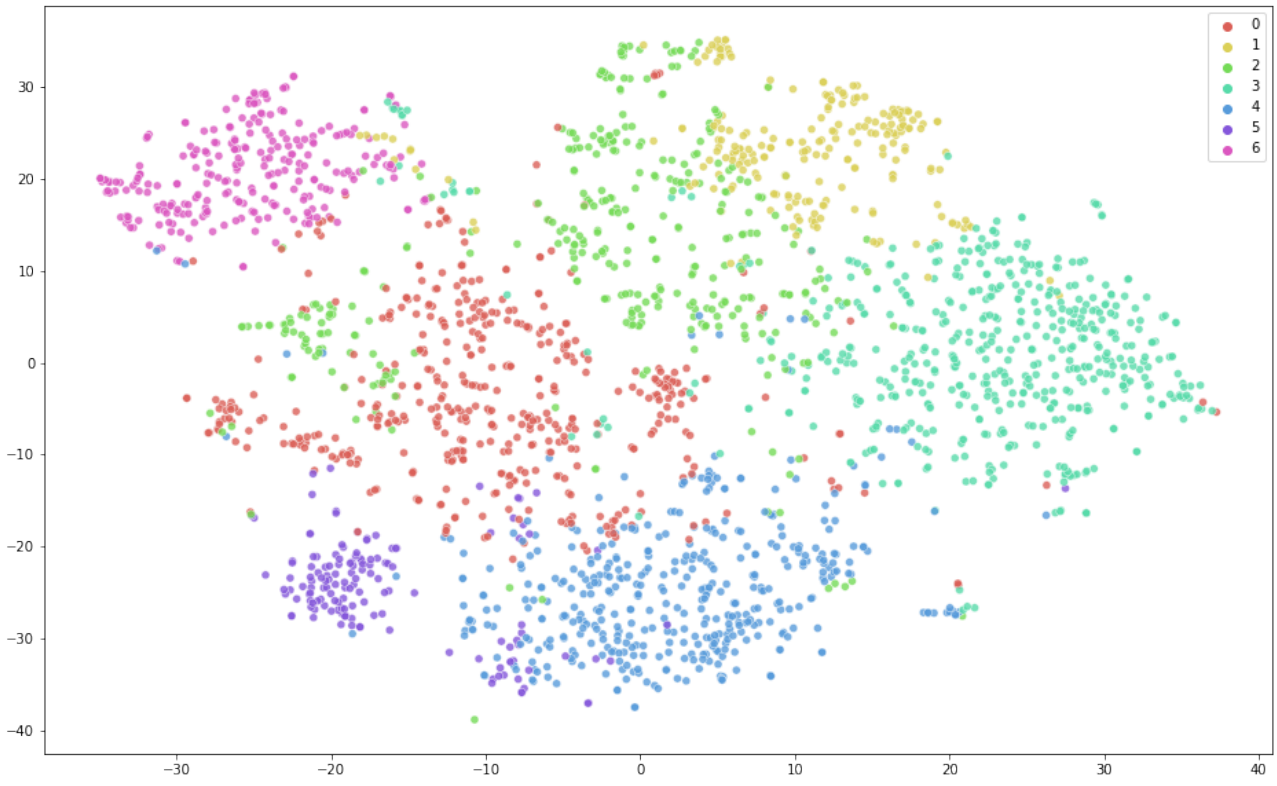
در قسمت بعدی حالت سه‌بعدی نمایش داده‌ها رسم شده است. برای رسم نمودار خوشه‌ها از Y_predicted به عنوان hue استفاده شده و نمودار آن در زیر قابل مشاهده است:

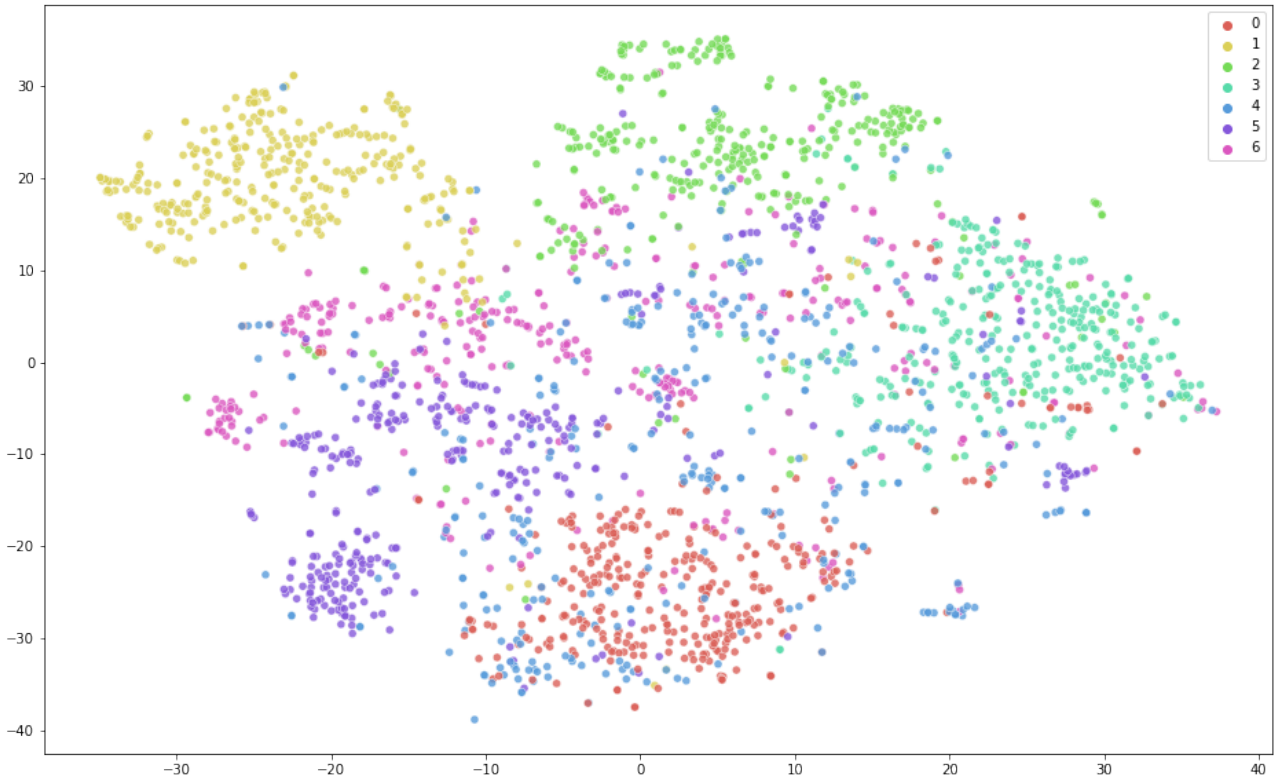


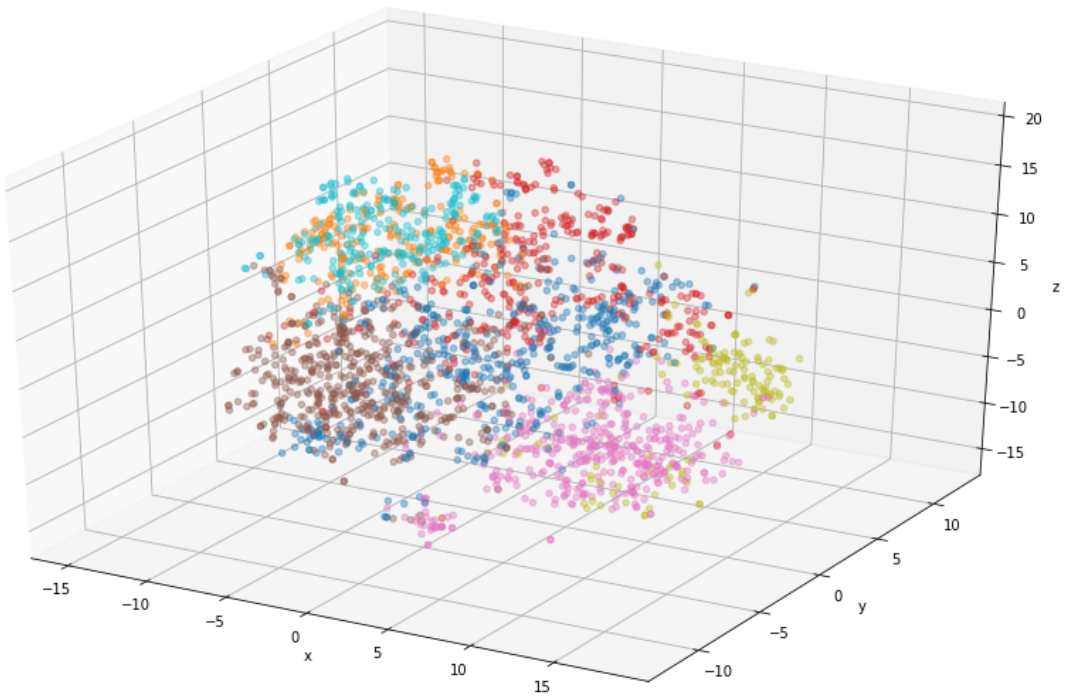
سپس درست مانند حالت دو بعدی، Y_test به عنوان hue داده شده است:

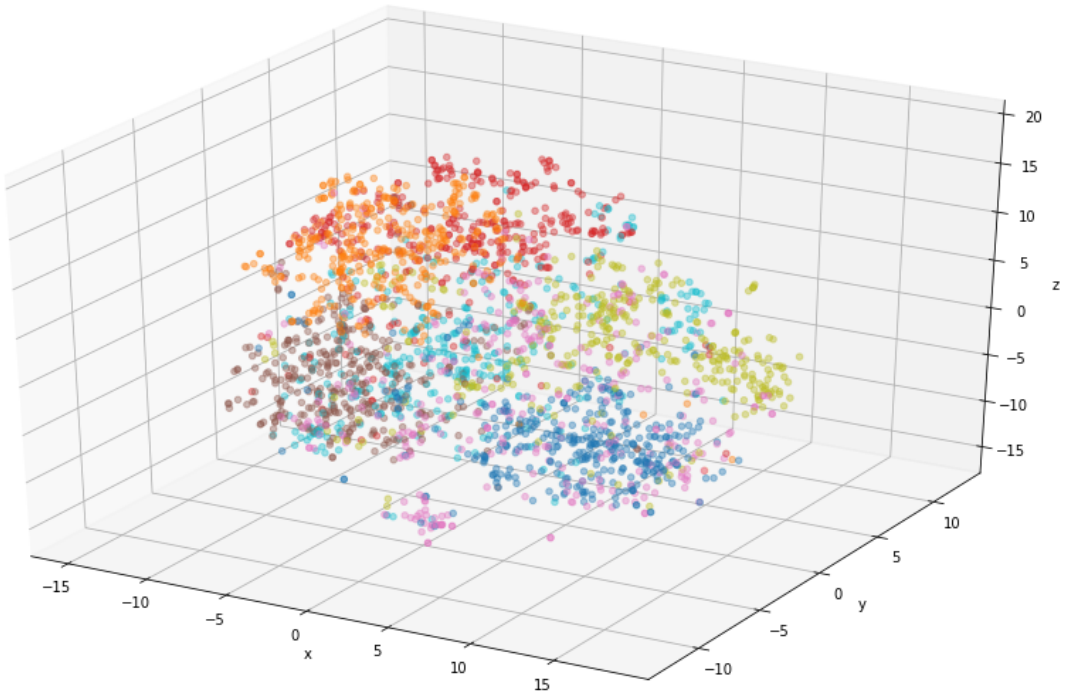


نمودارهای رسم‌شده به کمک TSNE نیز دقیقا مانند روش PCA در دو و سه بعد رسم شده‌اند که خروجی‌های آن‌ها به همان ترتیبی که نمودارهای PCA در بالا قرار داده شده در زیر قابل مشاهده است:




0
1
2
3
4
5
6
30
20
10
0
−10
−20
−30
−40
−30
−20
−10
0
10
20
30
40


20
15
10
5
0
−5
−10
−15
z
−15
−10
−5
0
5
10
15
x
10
5
0
−5
−10
y



پس به طور خلاصه داخل نوتبوک main-clustering-wth-fasttext، خوشه‌بندی، ارزیابی و visualization انجام شده است.

نوتبوک tfidf-clustering: در این نوتبوک به جای fasttext از مدل tfidf استفاده شده و از بردارهای به‌دست‌آمده از مدل tfidf برای بازنمایی سندها استفاده شده است. پس از محاسبه‌های ۴ معیار معرفی شده، purity حدودا برابر ۶۲ درصد شد که از مدل قبلی بهتر است اما در سه معیار دیگر بهبودی حاصل نشده است و به نظر می‌رسد نتایج، مطلوبیت مدل fasttext را ندارند.

نوتبوک clustering-test-tfdif-mix-fasttext: در این نوتبوک از ترکیب دو مدل tfidf و fasttext برای بازنمایی سندها استفاده شده است؛ بدین صورت که از امبدینگ fasttext استفاده شده اما به جای آنکه متن، میانگین امبدینگ کلماتش حساب شود و ضریب‌ها برابر یک باشد، ضریب‌ها برابر idf آن term در نظر گرفته شده است.
با انجام این کار بهبود چندانی در خروجی حاصل نشد و حتی purity score حدود ۵۲ درصد شد.

در نهایت این نتیجه حاصل شد که امبدینگ fasttext بهترین نتیجه را داده است.

توجه: مدل fasttext استفاده‌شده در این تمرین در این لینک بارگذاری شده است.