به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

جبر خطی

پروژه شماره 2

نام و نام خانوادگی	محمد عسکری
شماره دانشجویی	810198441
تاریخ ارسال گزارش	1403/04/21

سوال 1

برای رتبه بندی اسناد به نسبت ارتباطشان به یک عبارت خاص از روش TF-IDF استفاده میکنیم که هر قسمت از این روش را به تفصیل توضیح میدهیم.

Term frequency: به معنی فراوانی اصطلاح است و بررسی میکند در مجموعه اسناد موجود اصطلاح مورد نظر ما در کدام سند ها با فراوانی بیشتری تکرار شده است اما برای یک رتبه بندی صحیح به کمک متد IDF نیاز داریم.

Inverse document frequency: به معنی معکوس فراوانی سند است که این متد بیان می کند که اگر یک اصطلاح در تعداد زیادی از اسناد تکرار شده است احتمالا ارزش کمی برای رتبه بندی به ما می دهد و باید وزن این اصطلاح در نظام رتبه بندی کاهش یابد و برعکس وزن اصطلاحات کم تکرار تر را افزایش دهد.

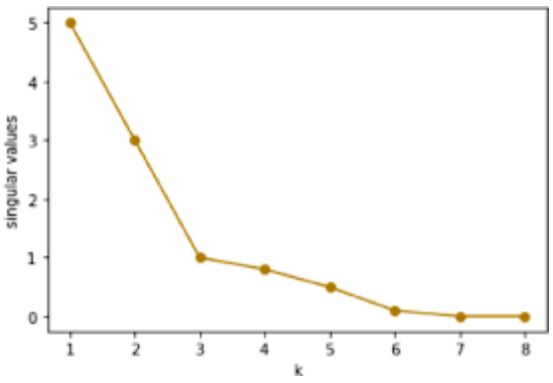
سوال 2

در Truncated SVD از نقطه زانویی برای انتخاب آستانه کاهش بعد مقدار ویژه ها استفاده میکنیم.

نقطه زانویی به این دلیل انتخاب می شود که توازنی بین کاهش بعد و از دست دادن اطلاعات ایجاد می کند.

از نقطه زانویی به بعد اطلاعاتی داریم که پیچیدگی را افزایش می دهند ولی ارزش افزوده قابل توجهی ندارند.

تصویر زیر یک نمونه نقطه زانویی در مقدار 3 را نشان می دهد:



سوال 3

خطای بازسازی شده در Truncated SVD به صورت نرم فروبینیوس تفاضل ماتریس اصلی و ماتریس کاهش یافته است.

سوال 4

Cosine similarity: یک معیار شباهت دو بردار در فضای برداری است این معیار شباهت بیان می کند که اگر دو بردار هم جهت باشند خروجی 1 است.

اگر دو بردار خلاف جهت هم باشند خروجی 1- است.

اگر دو بردار بر هم عمود باشند خروجی این معیار 0 است.

$$\text{Cosine similarity}(\mid X, Y \mid) = \frac{x \cdot y}{||x|| \ \ ||y||}$$

Euclidean Distance: یک معیار برای اندازه گیری فاصله ی بین دو نقطه است که از رابطه فیثاغورث محاسبه می شود. اگر این نقاط روی هم باشند این معیار 0 است و هرچه از هم فاصله بیشتری داشته باشند به بینهایت می رود اما در فضای محدود به قطر بین نقاط نزدیک می شود.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}.$$

سوال 5

کم کردن میانگین و تقسیم به انحراف معیار داده ها به دلیل کاهش تاثیر مقیاس های مختلف ویژگی ها و بهبود همگرایی الگوریتم ها لازم است. در ضمن اثرات نویز را کمتر می کند. برای مثال در بردار a:

$$a_{standard} = \frac{a - mean(a)}{standard\ deviation(a)}$$

سوال 6

شبه کد:

1-ابتدا برای ماتریس A_{m*n} یک رتبه k انتخاب میکنیم و یک ماتریس تصادفی P_{m*k} تشکیل میدهیم و سپس ماتریس A را در p ضرب میکنیم و به عنوان ماتریس اصلی قرار میدهیم.

$$Z_{m*k} = A * p$$

2- ماتریس Z فضای ستونی غالب ماتریس A را داراست. تجزیه QR ماتریس Z را انجام میدهیم.

3- ماتریس $Y = Q^T * A$ را تشکیل میدهیم و تجزیه SVD آن را انجام میدهیم.

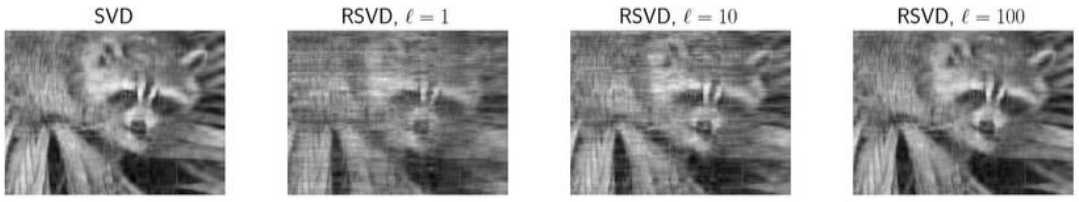
$$Y = U_{tiled} * \Sigma * V^T$$

4-$U = Q * U\_tiled$ را تشکیل می دهیم و در نهایت Randomized SVD را به صورت زیر گزارش میکنیم:

$$A = U * \Sigma * \mathrm{V^T}$$

کاربرد این الگوریتم در فشرده سازی عکس, کاهش بعد دیتاست ها در تحلیل های آماری و به طور کلی فشرده سازی است.

یک مثال از کاهش بعد تصویر را ملاحظه می کنیم(منبع)



سوال 7

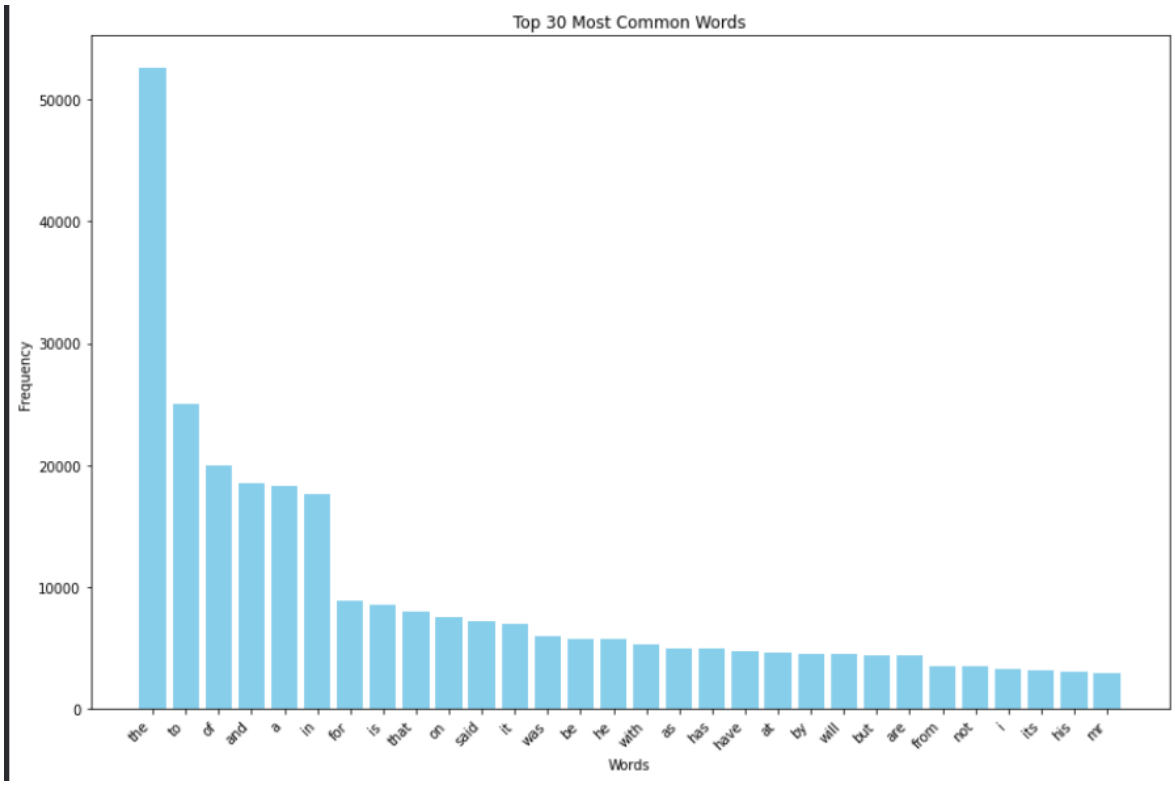
پاکسازی داده ها انجام شد.

سوال 8

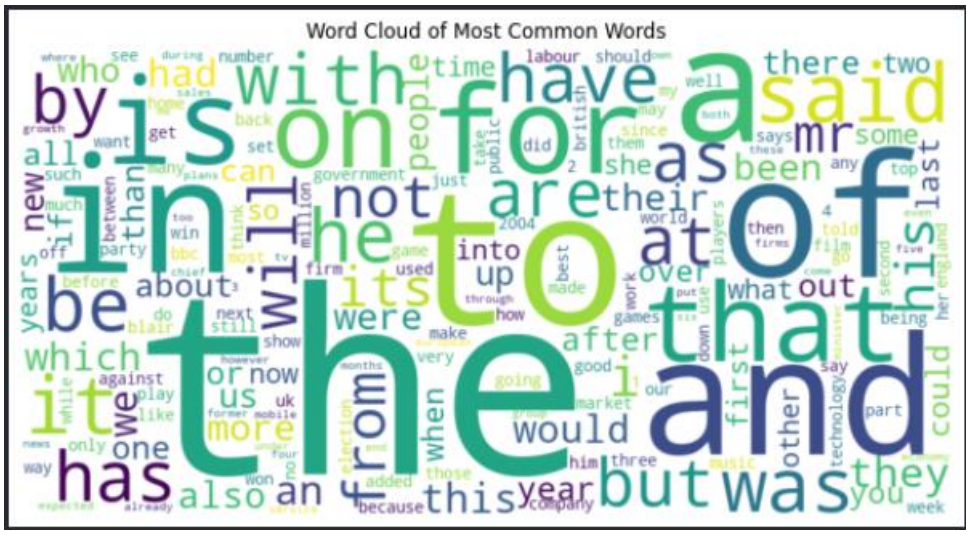
نمودار میله ای را گزارش میکنیم و مشاهده می شود که کلمه The پر تکرار ترین است.

دانستن تعداد تکرار کلمات وقتی مفید است که بتوان با استفاده از این معیار خبر را دسته بندی کرد.

در این جا پرتکرار بودن کلمه the اطلاعاتی به ما نمی دهد.



سوال 9



سایز کلمات با فراوانی آن ها رابطه دارد.

معمولا کلمات طولانی تر حذف می شوند.

سوال 10

52کلمه داریم و فراوانی این 52 کلمه را روی 2000 متن بررسی کزده ایم.

$$BagOfWordsMatrix_{2000*52}$$

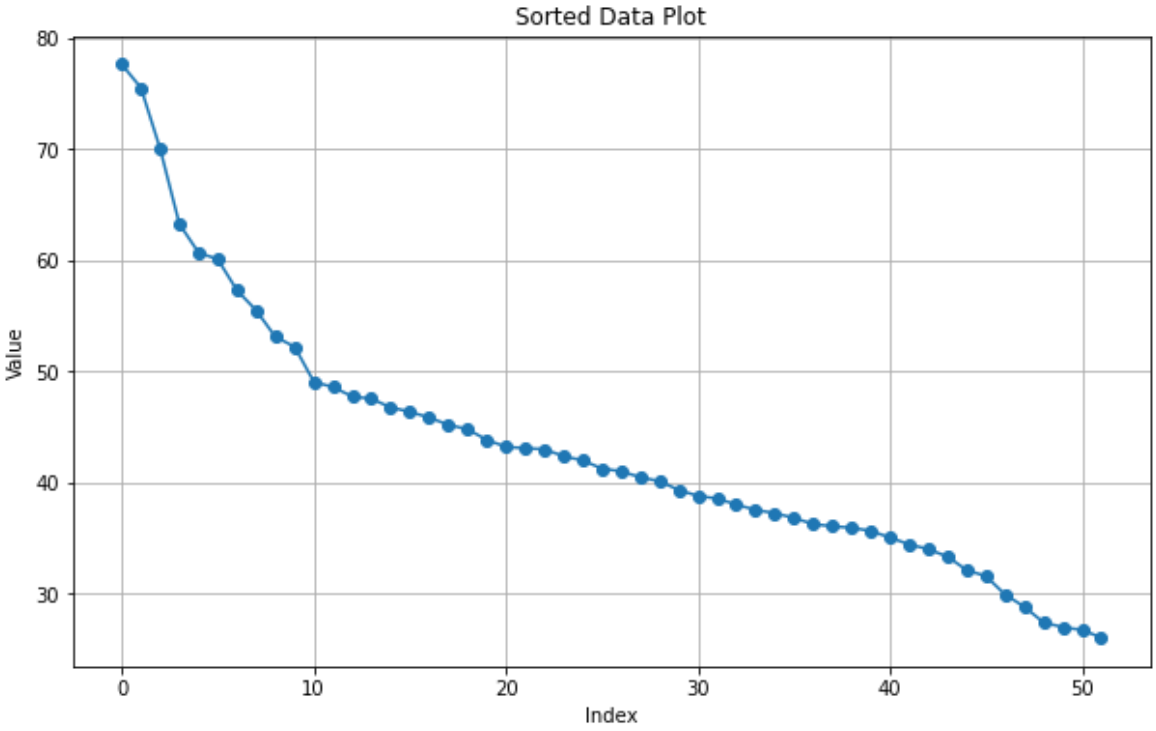
```
Number of rows in 'word_frequencies_train.csv': 2000
Number of columns in 'word_frequencies_train.csv': 52
```

سوال 11

ابتدا با کم کردن میانگین و تقسیم بر انحراف معیار ماتریس را استاندارد میکنیم سپس با تحزیه SVD ابعاد ماتریس ها را مشخص می کنیم.

```
Shape of U: (2000, 52)
Shape of S: (52,)
Shape of VT: (52, 52)
```

سوال 12



آستانه را 10 انتخاب میکنیم و خطای بازسازی را حدود 0.6 گزارش می کنیم.

```
Reconstruction error (MSE): 0.6173317912276224
```

سوال 13

طبق الگوریتم ارایه شده پیاده سازی را انجام میدهیم.

سوال 14

مشاهده میکنیم خطای بازسازی از حالت قبل بیشتر است که طبیعی است.

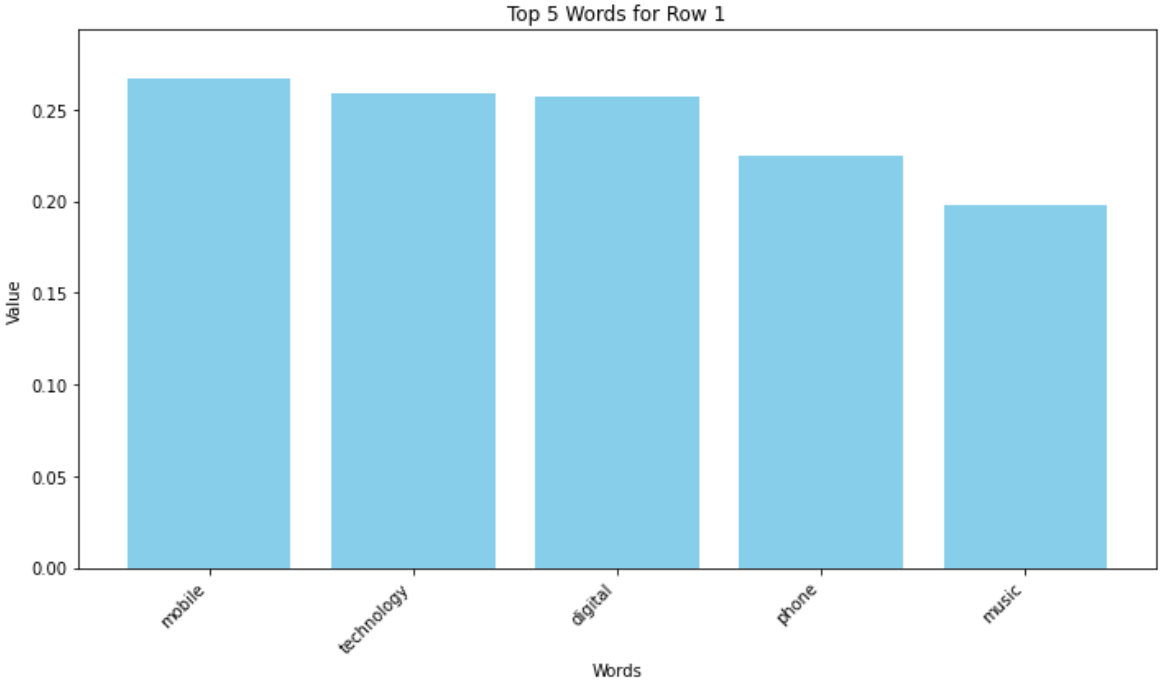
```
Reconstruction error for random SVD (MSE): 0.7657032675117402
```

سوال 15

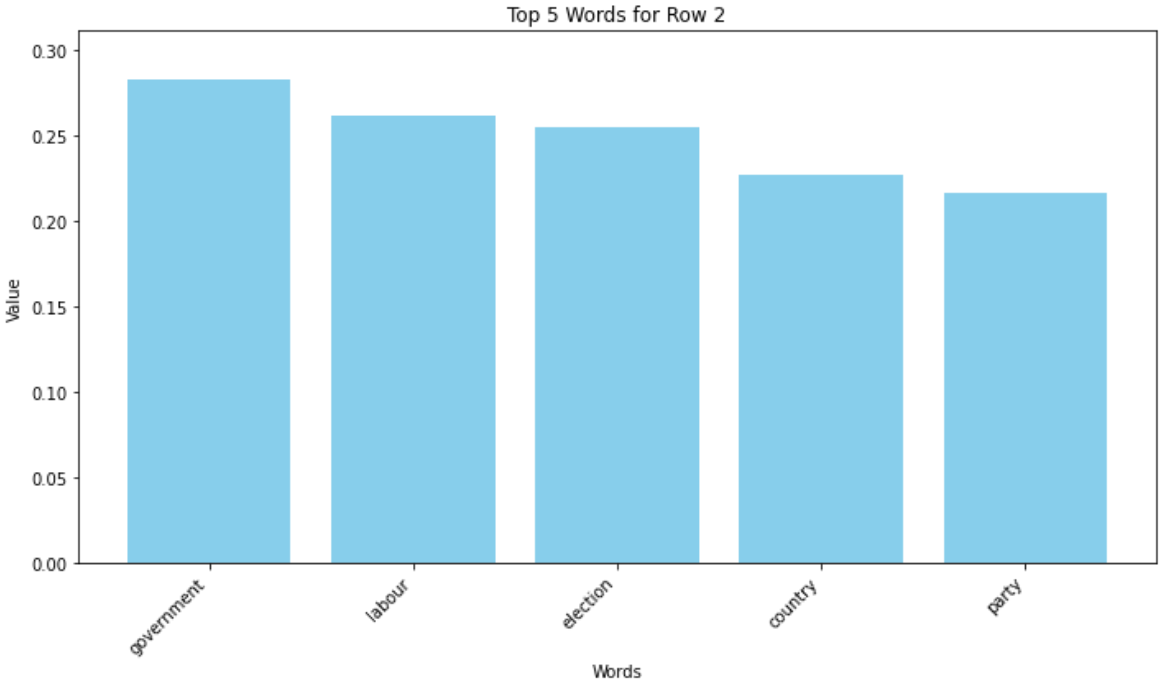
در ماتریس تجزیه شده Truncated SVD ماتریس V^T یک ماتریس با 10 سطر و 52 ستون است.

برای هر سطر 5 درایه ای که بیشترین مقدار را دارد را گزارش میکنیم و کلمات متناظر را پیدا میکنیم.

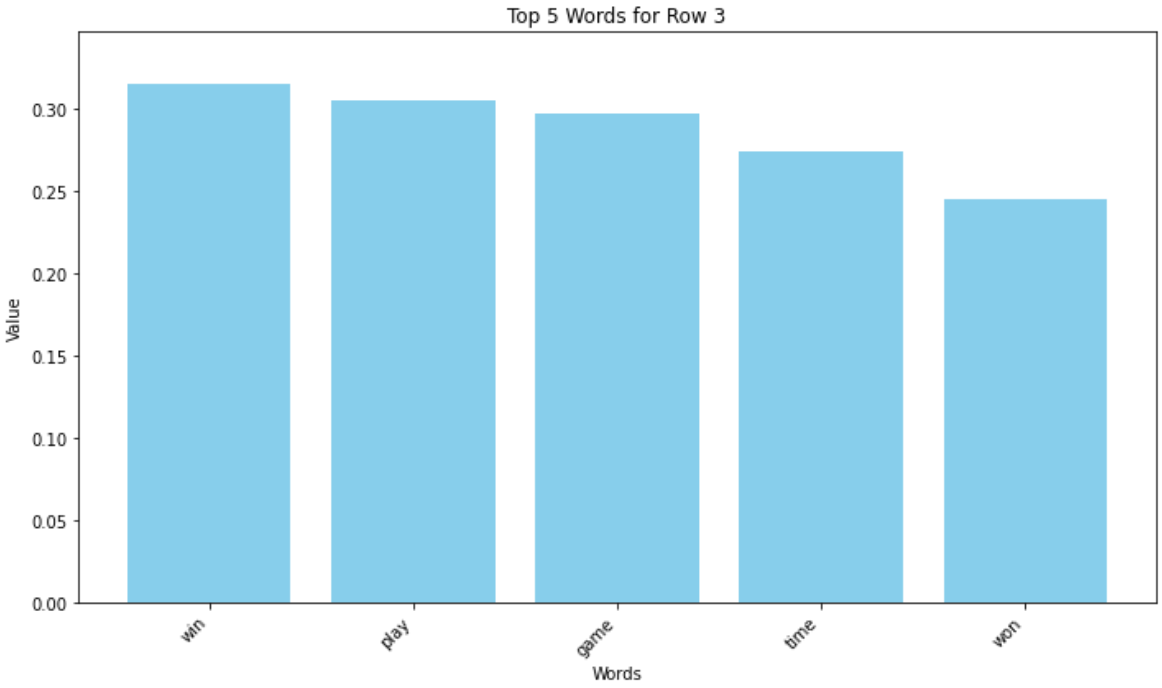
شدت این کلمات در این دسته موضوع بیشتر است.



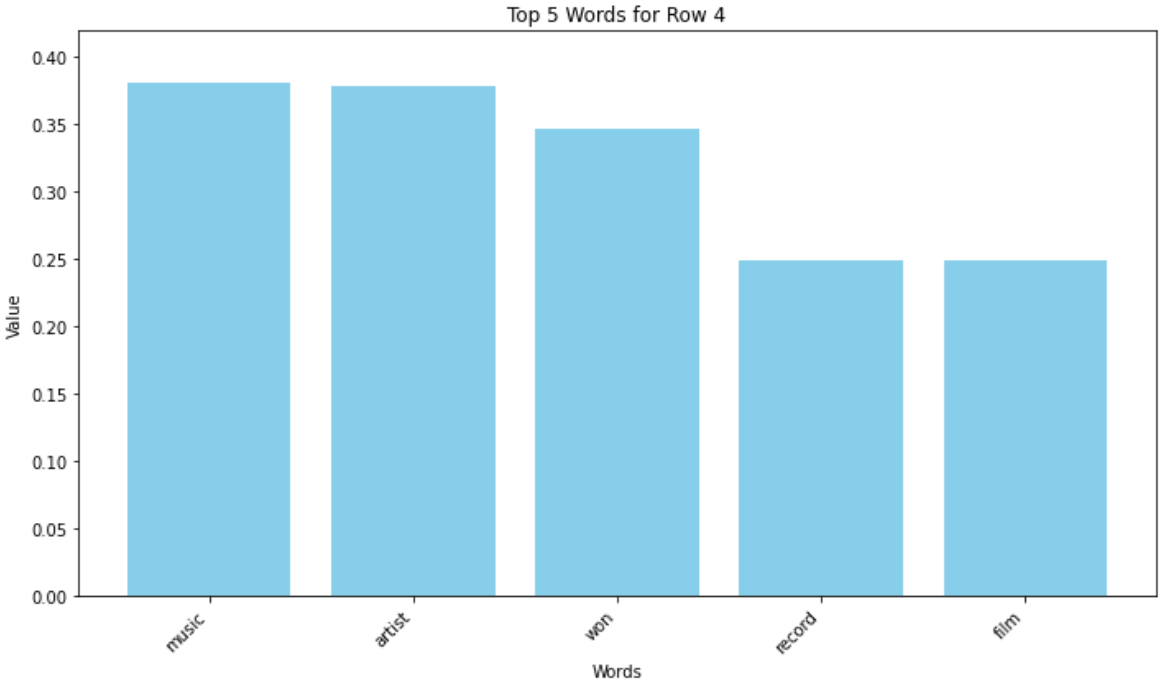
پیش بینی: تکنولوژی



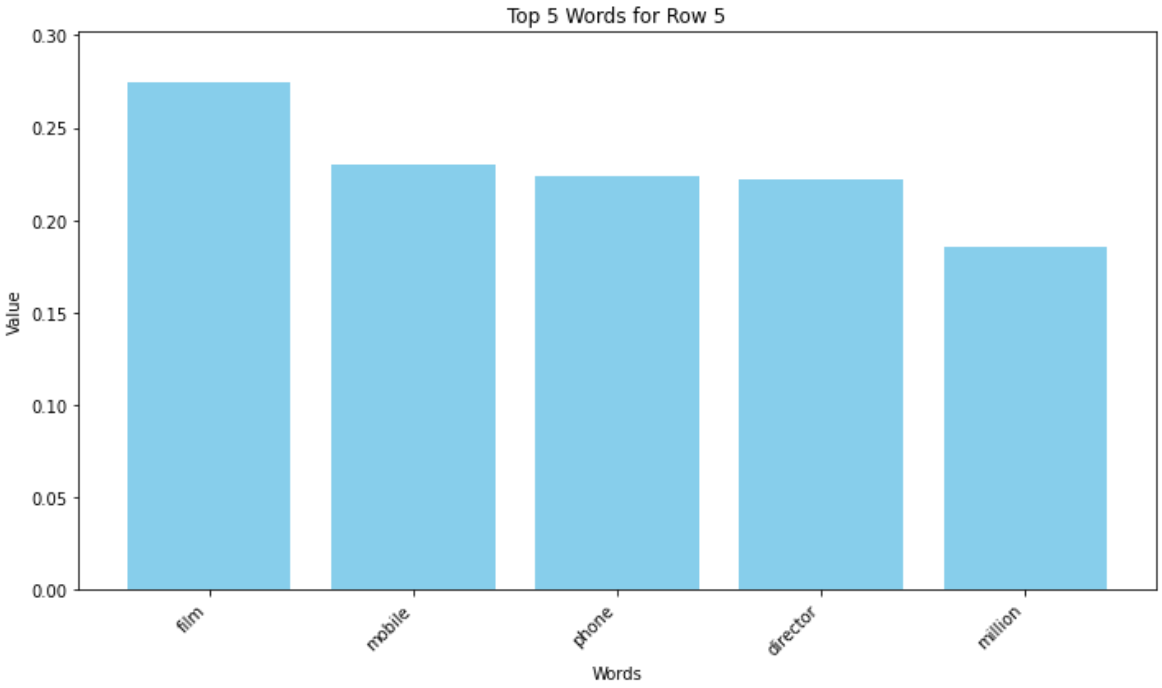
پیش بینی: سیاسی



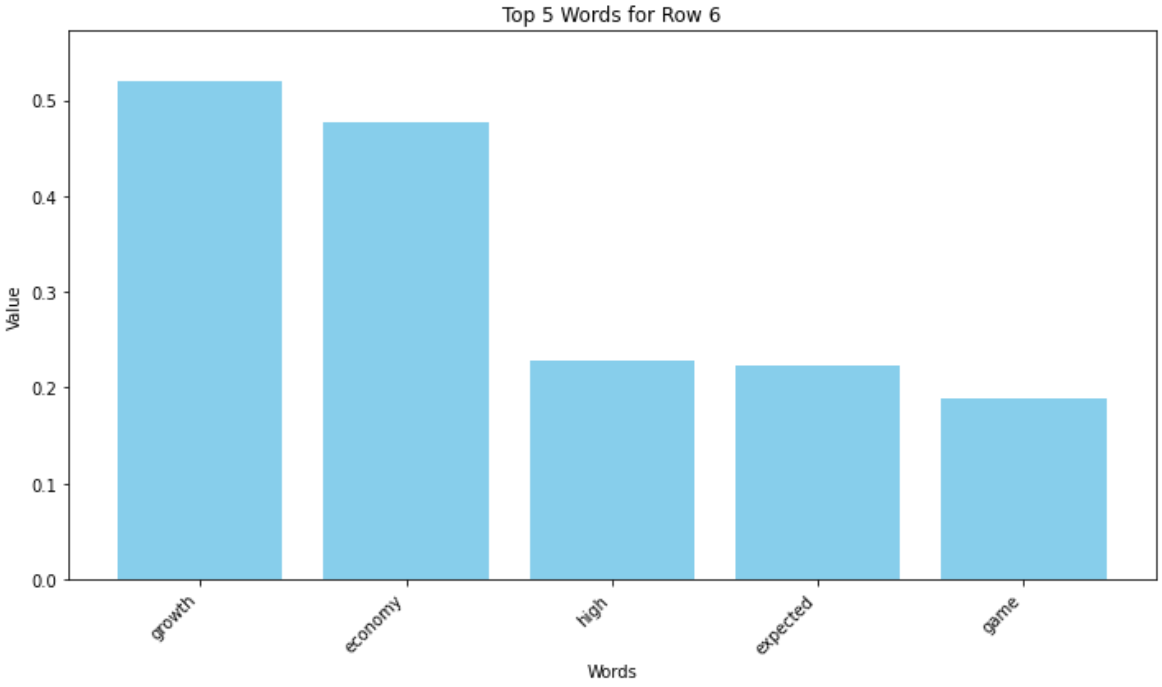
پیش بینی: سرگرمی



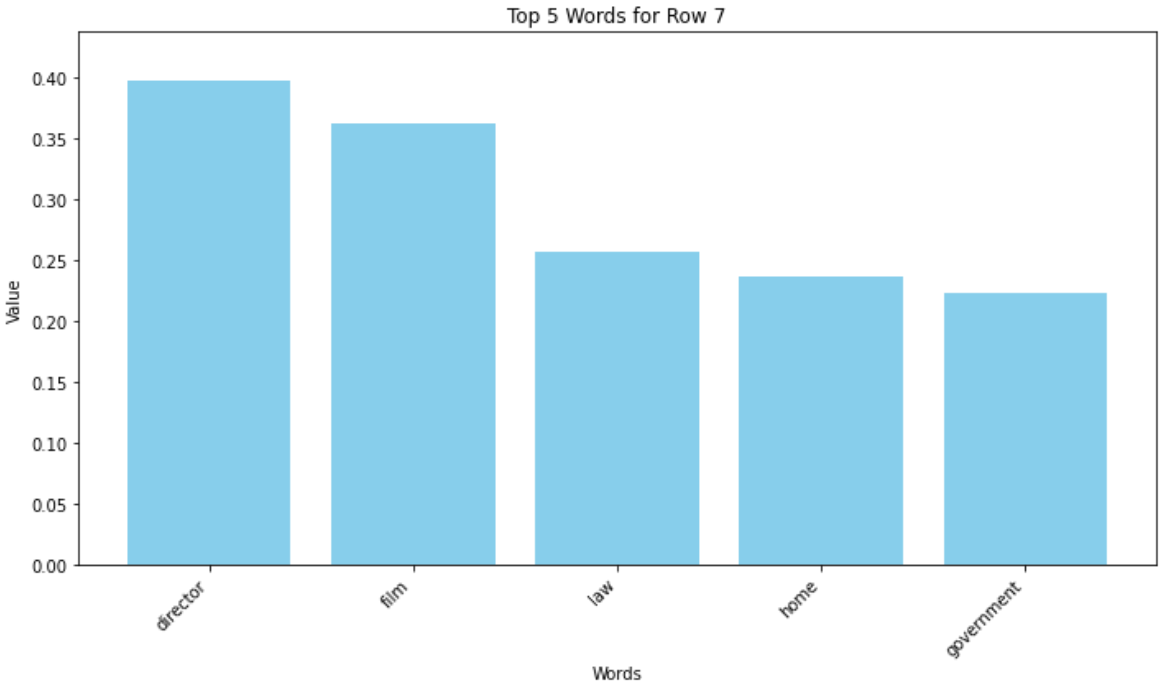
پیش بینی: موسیقی



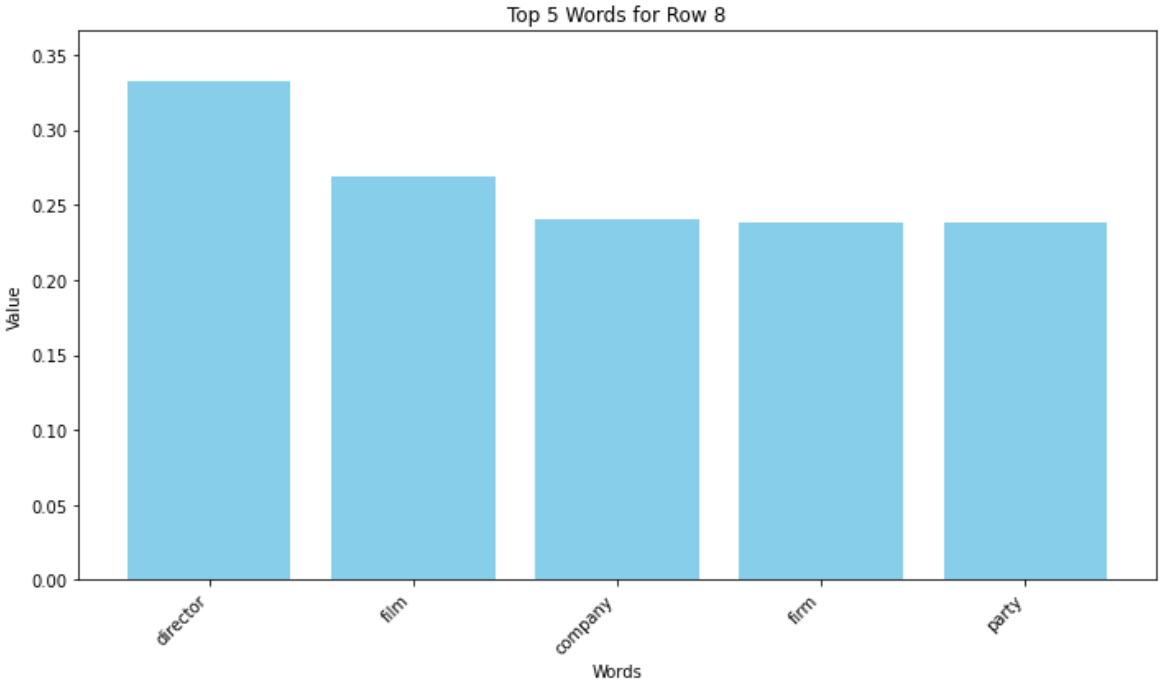
پیش بینی: سینما



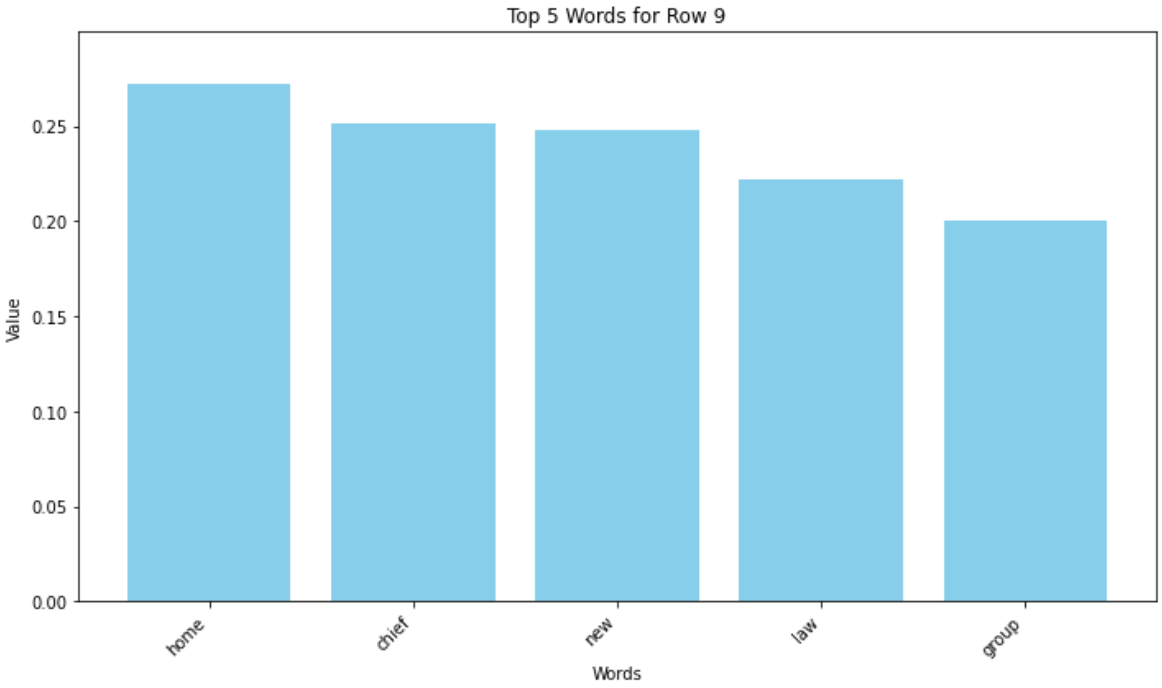
پیش بینی: اقتصاد



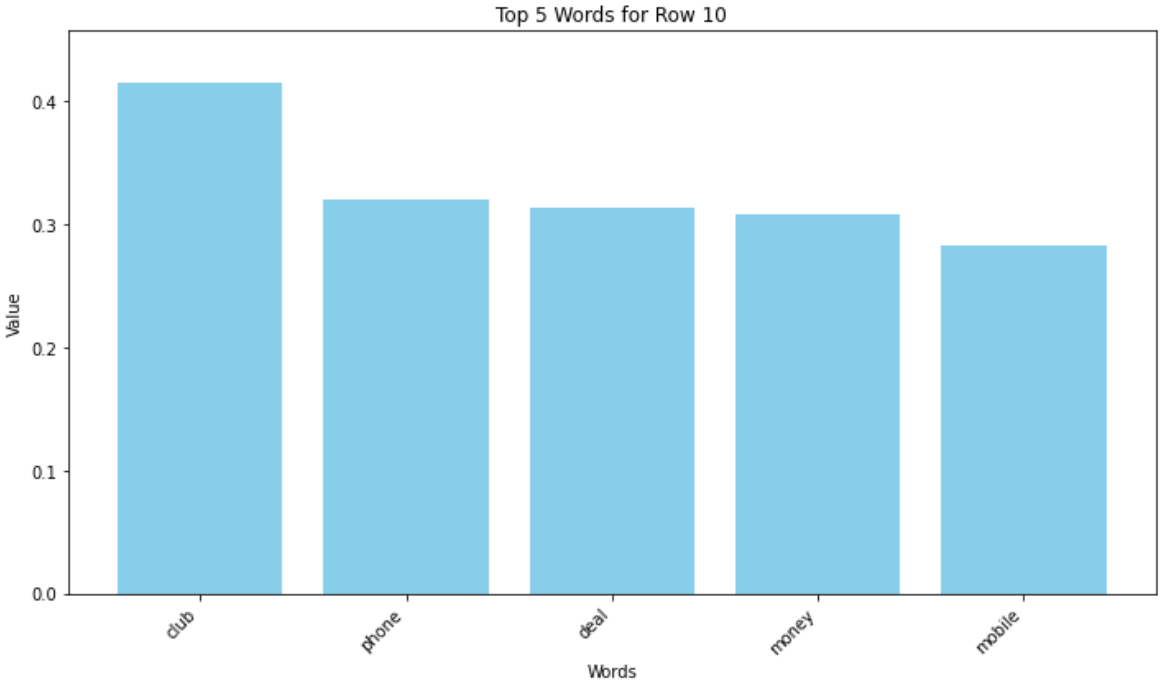
پیش بینی: سینما



پیش بینی: سینما



پیش بینی: اجتماعی



پیش بینی: سرگرمی

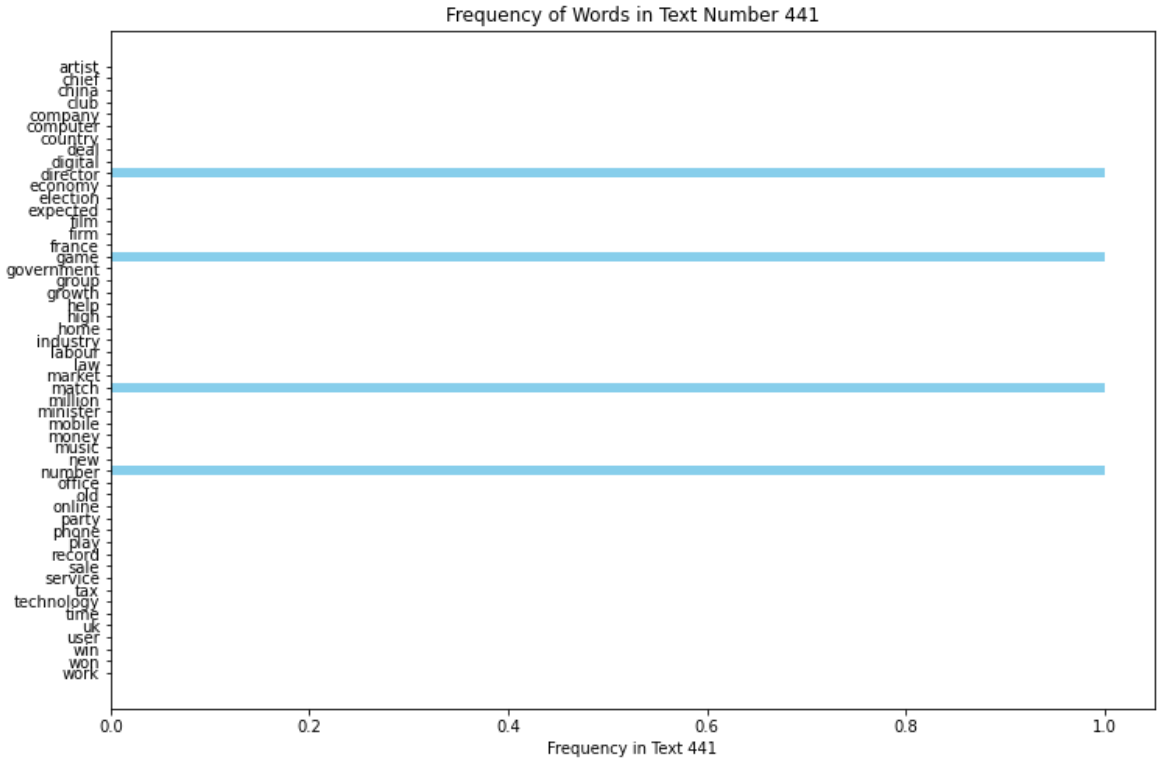
سوال 16

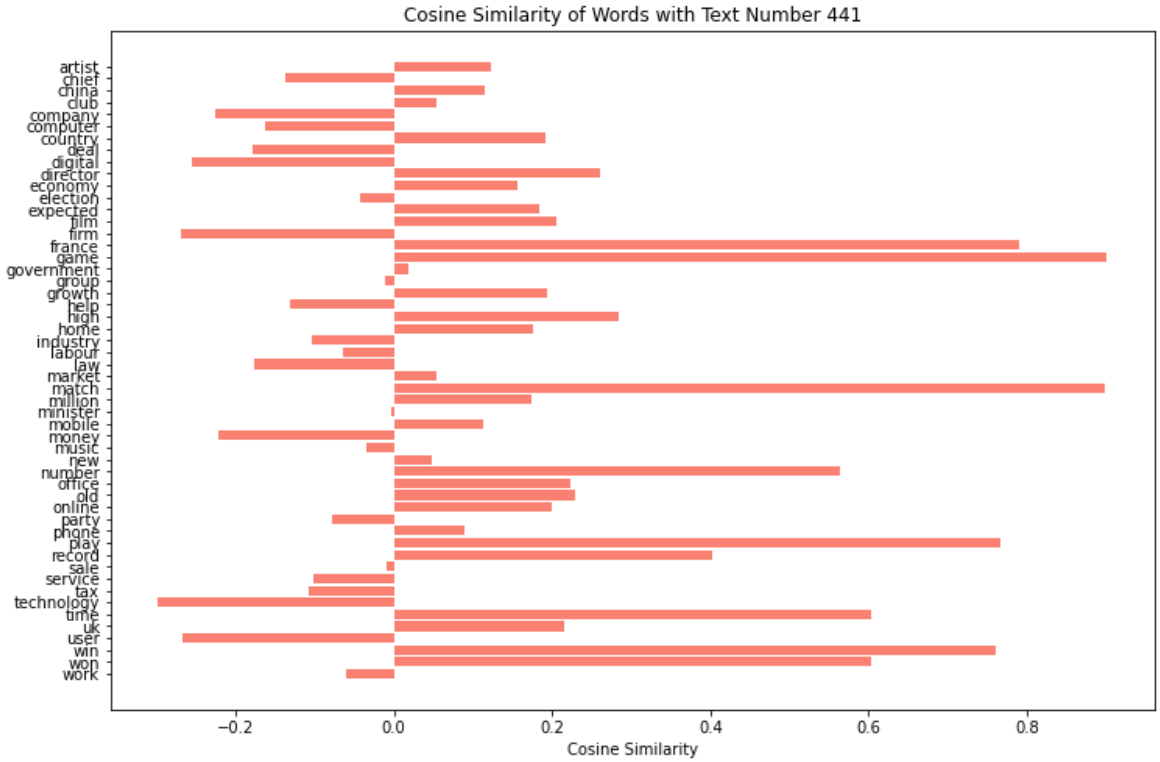
```
Cosine Similarity between 'mobile' and 'technology': 0.7329992051395805
Euclidean Distance between 'mobile' and 'technology': 0.38406623843972704
Cosine Similarity between 'director' and 'film': 0.9841561160696878
Euclidean Distance between 'director' and 'film': 0.1116357254243781
Cosine Similarity between 'win' and 'won': 0.7294240372616616
Euclidean Distance between 'win' and 'won': 0.32994985287496625
Cosine Similarity between 'play' and 'game': 0.9735899762085679
Euclidean Distance between 'play' and 'game': 0.11860283462314454
Cosine Similarity between 'play' and 'law': -0.29697197032723954
Euclidean Distance between 'play' and 'law': 0.7530550004677794
Cosine Similarity between 'government' and 'music': 0.03798486638983322
Euclidean Distance between 'government' and 'music': 0.7234775500771116
```

Cosine similarity بیان می کند بیشترین شباهت را director & film دارند و کمترین شباهت را government & music دارند.

همچنین Euclidean distance بیان میکند play & game بیشترین فاصله و mobile & technology کمترین فاصله معنایی را از هم دارند.

سوال 17





مشاهده می شود در کلمات پرتکرار معیار شباهت هم زیاد است.

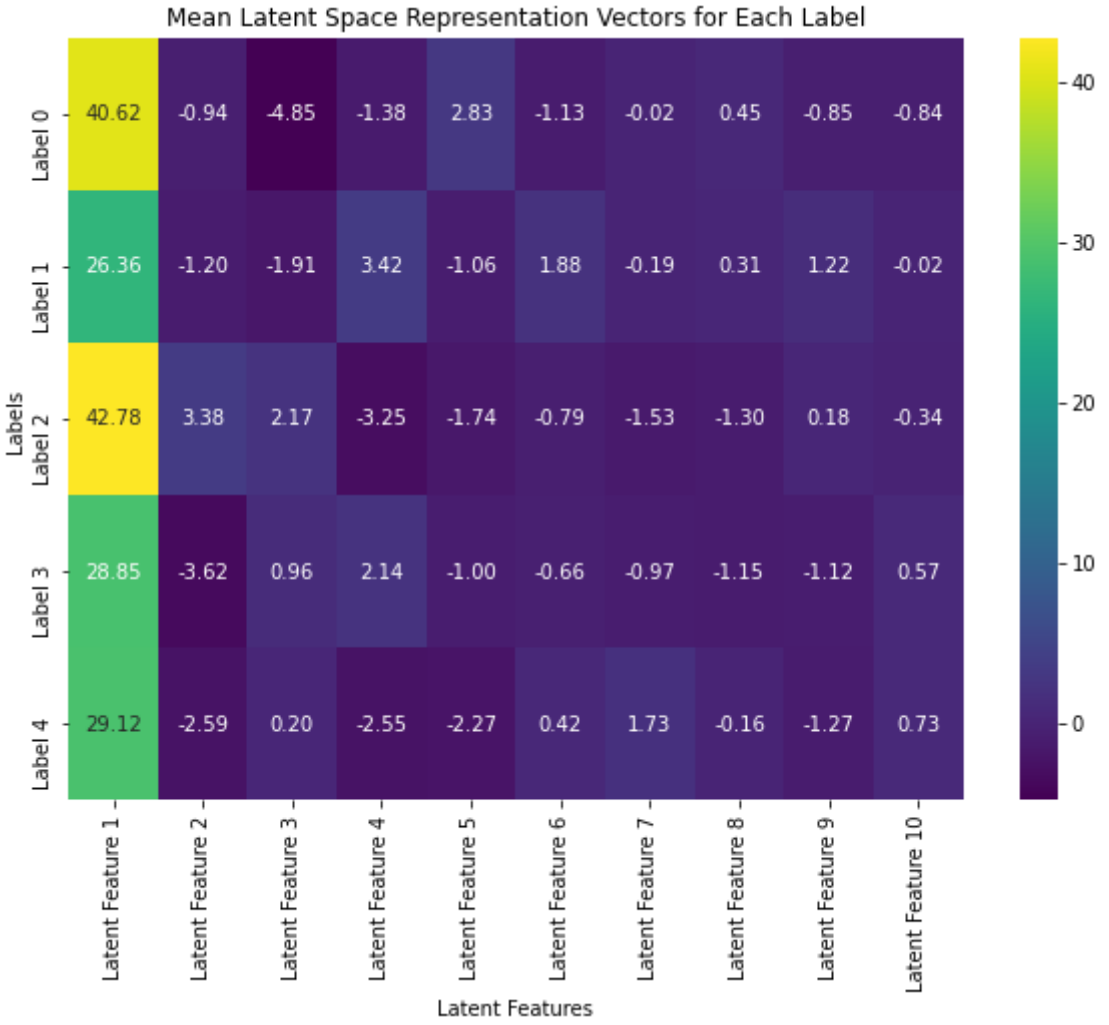
سوال 18

استفاده از فضای نهان می تواند ارتباط بقیه کلمات با technology را بررسی کند سپس با توجه به بقیه کلمات هم مفهوم با technology متن را تشخیص بدهد.

در جستجو در فضای نهان، می‌توان به داده‌های بیشتری دسترسی پیدا کرد، چرا که این جستجو به دنبال مفاهیم و ارتباطات جدید در داده‌های متنی می‌گردد که در جستجوی کوله کلمات قابل توجه نیستند.

همچنین این روش می‌تواند الگوها و روابطی را در داده‌های متنی شناسایی کند که به‌صورت مستقیم در متن ظاهر نشده‌اند.

سوال 19



با بررسی این بردار میتوان از SVM برای پیدا مردن مرز برچسب گذاری استفاده کرد.

سوال 20

با استفاده از ماشین بردار پشتیبان مرز تصمیمی گیری برای label داده ها را در فضای نهان مشخص می کنیم:

روی داده های train ماشین بردار پشتیبان را آموزش میدهیم و نتایج را گزارش می کنیم:

```
Accuracy of SVM Classifier: 0.75
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.73      0.75       381
           1       0.83      0.89      0.86       465
           2       0.73      0.74      0.73       357
           3       0.72      0.61      0.66       342
           4       0.71      0.76      0.73       455

    accuracy                           0.75      2000
   macro avg       0.75      0.74      0.75      2000
weighted avg       0.75      0.75      0.75      2000

Confusion Matrix:
[[279  20  33  14  35]
 [ 12 413   3  29   8]
 [ 14   9 263  15  56]
 [ 24  53  14 209  42]
 [ 34   5  48  24 344]]
```

دقت این روش 75 درصد شده است که قابل قبول است.

حال روی 225 متنی که از ایندا به عتوان داده test جدا کردیم اجرا میکنیم و دقت را میسنجیم:

```
Accuracy of SVM Classifier on Test Data: 0.73
Classification Report on Test Data:
              precision    recall  f1-score   support

           0       0.65      0.78      0.71        36
           1       0.85      0.85      0.85        46
           2       0.77      0.68      0.72        44
           3       0.71      0.61      0.66        44
           4       0.69      0.75      0.72        55

    accuracy                           0.73       225
   macro avg       0.73      0.73      0.73       225
weighted avg       0.74      0.73      0.73       225

Confusion Matrix for Test Data:
[[28  1  0  0  7]
 [ 2 39  0  4  1]
 [ 6  1 30  4  3]
 [ 3  4  3 27  7]
 [ 4  1  6  3 41]]
```

دقت حدود 73 درصد شده که قابل قبول است.