به نام خدا

گزارش تمرین چهارم بازیابی پیشرفته اطلاعات

خوشه‌بندی مقالات علمی براساس چکیده

گروه ۳۰

پارسا حسینی

درسا مجدی

امیررضا باقری

چکیده روند پروژه

در این تمرین هدف آن است که مقالات را براساس چکیده، به تعدادی خوشه معنادار تقسیم کنیم و با امتحان کردن مقادیر مختلف k، تعداد خوشه‌های مناسب را مشخص کنیم. در مرحله اول، از دیتاست آماده شده در مرحله قبل استفاده کرده و آن را براساس زمینه علمی‌شان جدا کردیم. مقالات موجود در سه زمینه علمی روانشناسی (Psychology) و علم مواد (Materials Science) و ریاضی (Mathematics) جدا شده و برای مراحل بعدی مورد استفاده قرار گرفتند. سپس عملیات پیش‌پردازش بر روی متن چکیده مقالات اعمال می‌شود. سپس با استفاده از یک مدل transformer آماده بردارهای تعبیه (embedding) مقالات بدست آمد. در مرحله آخر الگوریتم kmeans بر روی بردارهای تعبیه اجرا شده و نتایج تجربی به صورت Confusion Matrix و نمایش دو بعدی مقالات و label متناظر آنها، نمایش داده شد. این الگوریتم برای مقادیر مختلف k امتحان شد و نتایج آنها براساس معیار Purity و امتیازهای RSS و silhouette_score مقایسه شد.

بخش ۱. آماده‌سازی داده‌ها

از دیتاست آماده شده در تمرین قبلی استفاده می‌کنیم و مقالاتی را که در لیست زمینه علمی‌شان یکی از علوم روانشناسی (Psychology) و علم مواد (Materials Science) و ریاضی (Mathematics) است را جدا کرده و در دیتافریم data ذخیره می‌کنیم.

```
[ ]  data = df[df['fieldsOfStudy'].str.contains('Psychology') |
                df['fieldsOfStudy'].str.contains('Materials Science') |
                df['fieldsOfStudy'].str.contains('Mathematics')].copy()
     data
```

	title	abstract	fieldsOfStudy
3	Molecule Attention Transformer	Designing a single neural network architecture...	['Computer Science', 'Physics', 'Mathematics']
15	Hierarchical Attention Transformer Architectur...	The attention mechanisms are playing a boostin...	['Computer Science', 'Mathematics']
21	Set Transformer: A Framework for Attention-bas...	Many machine learning tasks such as multiple i...	['Computer Science', 'Mathematics']
33	Transformer-Based Online CTC/Attention End-To-...	Recently, Transformer has gained success in au...	['Computer Science', 'Engineering', 'Mathemati...
49	Graph-Aware Transformer: Is Attention All Grap...	Graphs are the natural data structure to repre...	['Computer Science', 'Mathematics']
...	...	...	...
7906	Active Balancing of Li-Ion Battery Cells Using...	A circuit for balancing Li-ion battery cells i...	['Materials Science', 'Computer Science']
7919	Early social attention impairments in autism: ...	This study investigated social attention impai...	['Psychology', 'Medicine']
7925	The attention system of the human brain.	Illustration de trois fonctions principales qu...	['Psychology', 'Medicine', 'Geography']
7929	Predicates and predicate transformers for supe...	Discrete-event systems are studied, treating t...	['Computer Science', 'Mathematics']
7942	Understanding the ageing aspects of natural es...	Cellulose based insulation materials and miner...	['Materials Science']

1025 rows × 3 columns

اشتراک این سه دسته از مقالات با یکدیگر ∘ است و احتمالا بتوان به خوبی مقالات آنها را خوشه‌بندی کرد.

```
[ ]  sum(data['fieldsOfStudy'].str.contains('Psychology') &
                                df['fieldsOfStudy'].str.contains('Materials Science'))
```

```
0
```

```
[ ]  sum(data['fieldsOfStudy'].str.contains('Mathematics') &
                                df['fieldsOfStudy'].str.contains('Materials Science'))
```

```
0
```

```
[ ]  sum(data['fieldsOfStudy'].str.contains('Mathematics') &
                                df['fieldsOfStudy'].str.contains('Psychology'))
```

```
0
```

سپس به این مقالات برحسب اینکه کدام یک از زمینه‌های علمی اشاره شده را دارند، label می‌دهیم تا بتوانیم در مرحله بعدی از این label ها برای محاسبه امتیازهای خوشه‌بندی استفاده کنیم.

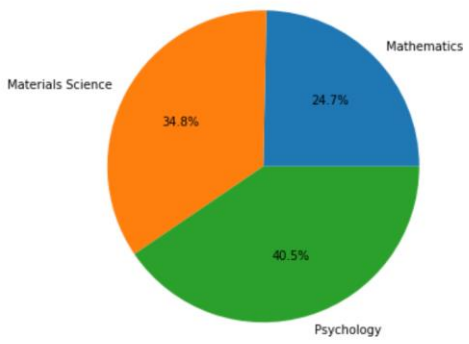
```
[ ]  def f(fields):
        if 'Materials Science' in fields:
            return 'Materials Science'
        if 'Psychology' in fields:
            return 'Psychology'
        return 'Mathematics'

    data['label'] = data['fieldsOfStudy'].apply(f)
    data.head()
```

	title	abstract	fieldsOfStudy	label
3	Molecule Attention Transformer	Designing a single neural network architecture...	['Computer Science', 'Physics', 'Mathematics']	Mathematics
15	Hierarchical Attention Transformer Architectur...	The attention mechanisms are playing a boostin...	['Computer Science', 'Mathematics']	Mathematics
21	Set Transformer: A Framework for Attention-bas...	Many machine learning tasks such as multiple i...	['Computer Science', 'Mathematics']	Mathematics
33	Transformer-Based Online CTC/Attention End-To-...	Recently, Transformer has gained success in au...	['Computer Science', 'Engineering', 'Mathemati...	Mathematics
49	Graph-Aware Transformer: Is Attention All Grap...	Graphs are the natural data structure to repre...	['Computer Science', 'Mathematics']	Mathematics

نسبت تعداد مقالات هر دسته به صورت یک pie chart در شکل بعدی نمایش داده شده است.

```
[ ]  d = Counter(data['label'])
     count = list(d.values())
     labels = list(d.keys())
     print(count)
     print(labels)
     plt.figure(figsize=(6, 6))
     plt.pie(count, labels=labels, autopct="%.1f%%")
     plt.show()
```

```
[253, 357, 415]
['Mathematics', 'Materials Science', 'Psychology']
```



در مرحله آخر آماده‌سازی داده‌ها، پیش‌پردازش بر روی متن چکیده اعمال شده و داده آماده اجرای خوشه‌بندی می‌شود.

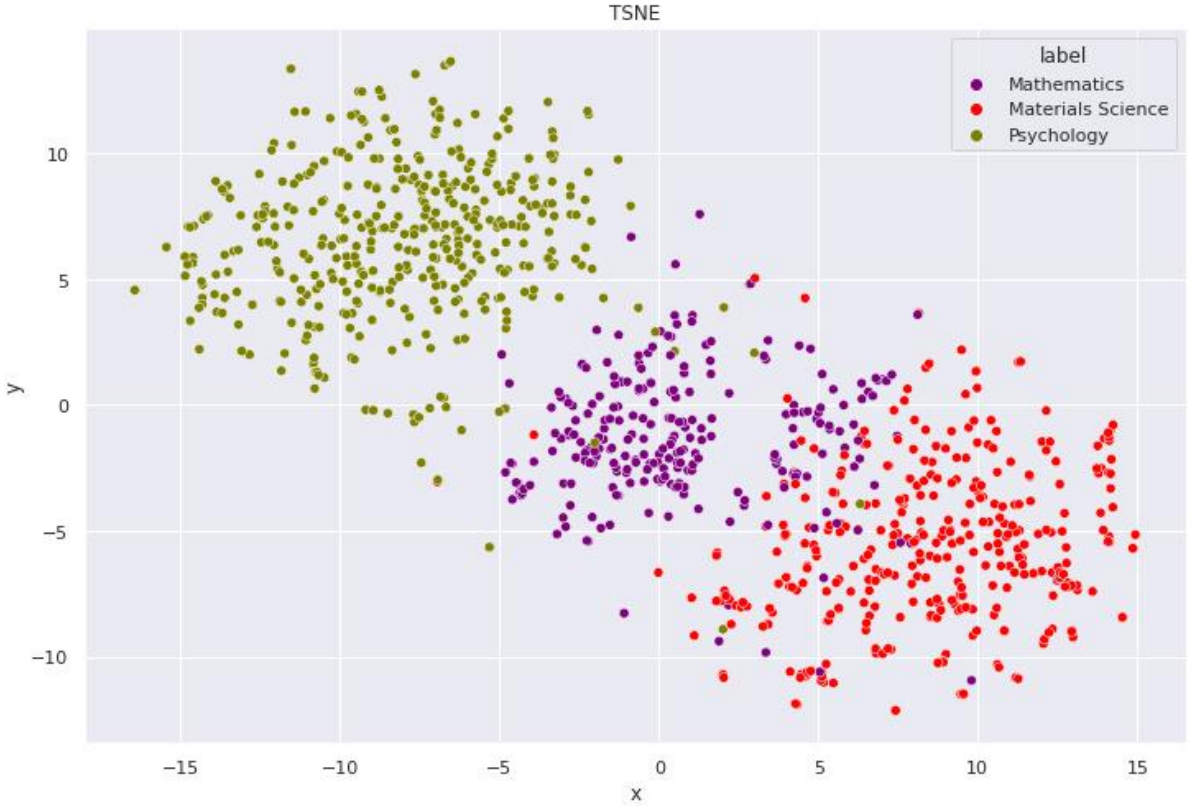
```
[ ] # Preprocess corpus
    tqdm.pandas()
    data['clean_abstract'] = data['abstract'].progress_apply(preprocessor.run)
    data.head()
```

```
100%|██████████| 1025/1025 [00:33<00:00, 30.58it/s]
```

	title	abstract	label	clean_abstract
3	Molecule Attention Transformer	Designing a single neural network architecture...	Mathematics	design single neural network architecture perf...
15	Hierarchical Attention Transformer Architectur...	The attention mechanisms are playing a boostin...	Mathematics	attention mechanism play boosting role advance...
21	Set Transformer: A Framework for Attention-bas...	Many machine learning tasks such as multiple i...	Mathematics	machine learn task multiple instance learning ...
33	Transformer-Based Online CTC/Attention End-To-...	Recently, Transformer has gained success in au...	Mathematics	recently transformer gain success automatic sp...
49	Graph-Aware Transformer: Is Attention All Grap...	Graphs are the natural data structure to repre...	Mathematics	graph natural datum structure represent relati...

بخش ۲. خوشه‌بندی

در این بخش ابتدا با استفاده از معماری ترنسفورمر مشابه تمرین گذشته داده‌های متنی که نیازمند خوشه‌بندی بودند را به بردار تبدیل کردیم. از آنجایی که بعد از این تبدیل بُعد داده‌ها زیاد است (۳۸۴) برای نمایش آنها روی نمودار scatter هم با استفاده از PCA وهم با استفاده از TSNE آنها را به داده‌های دوبعدی تبدیل کرده و با کمک labelهای اولیه‌شان نمایش می‌دهیم.



قسمت اصلی خوشه‌بندی داده‌ها با استفاده از تابع run_kmeans انجام می‌شود. Kmeans و توابع مربوط به آن در کتابخانه‌ی sklearn وجود دارند و در این تمرین از آنها استفاده می‌شود. با استفاده از تابع آماده‌ی Kmeans و پارامتر n_clusters داده‌ها با تعداد خوشه‌ی تعیین شده خوشه‌بندی می‌شوند و لیبل می‌خورند. داده‌ها این‌بار با لیبل‌های اختصاص داده‌شده و دوباره با PCA و TSNE نمایش داده می‌شوند. Confusion matrix لیبل‌های نسبت داده‌شده و لیبل‌های اولیه با کمک تابع آماده‌ی contingency_matrix کشیده می‌شود. برای ارزیابی عملکرد خوشه‌بندی از معیارهای Purity، Silhouette و RSS استفاده شده است.

Purity که با تابع purity_score محاسبه شده است از فرمول زیر بدست می‌آید:

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N}\sum_{k}\max_{j}|\omega_k \cap c_j|$$

معیار silhouette که با تابع آماده‌ی silhouette_score از sklearn محاسبه شده که مقادیر بین [-1,1] را می‌گیرد و هرچقدر به ۱ نزدیکتر باشد بهتر است.

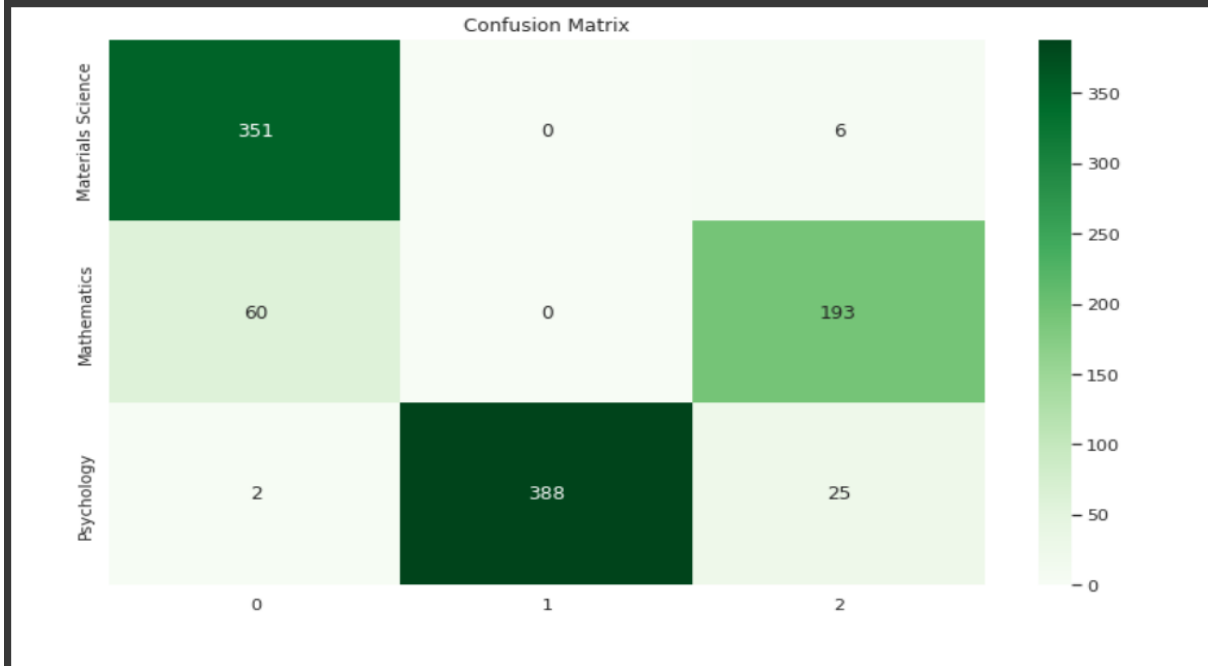
معیار RSS هم که با تابع calc_rss محاسبه شده از فرمول زیر بدست می‌آید:

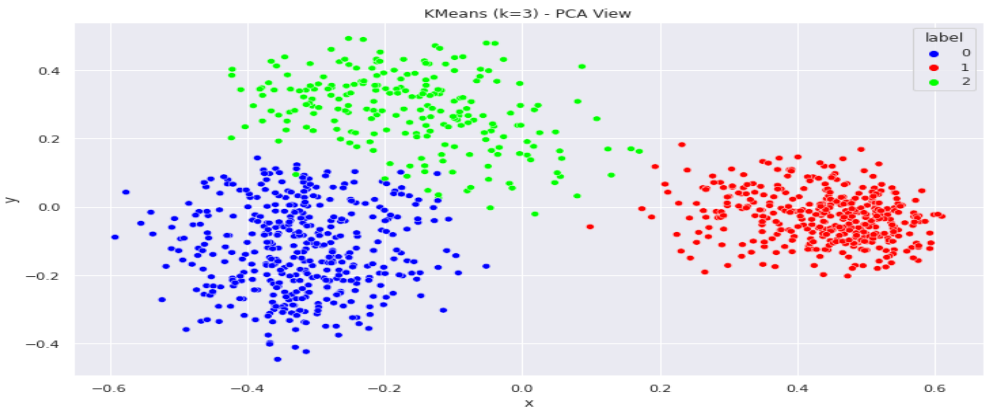
$$\text{RSS} = \sum_{k=1}^{K}\text{RSS}_k \qquad \text{RSS}_k = \sum_{\vec{x}\in\omega_k}|\vec{x}-\vec{\mu}(\omega_k)|^2 \qquad \vec{\mu}(\omega) = \frac{1}{|\omega|}\sum_{\vec{x}\in\omega}\vec{x}$$

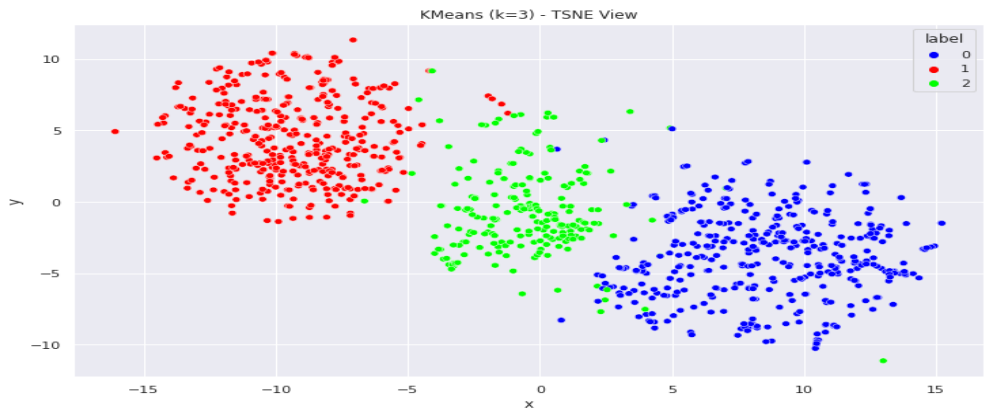
نتیجه‌ی این خوشه بندی را به ازای k = 3,4,7 در زیر مشاهده می‌کنید:

K = 3

```
Purity: 0.9092682926829269
silhouette_score = 0.14156734943389893
RSS = 538.1109245296685
```

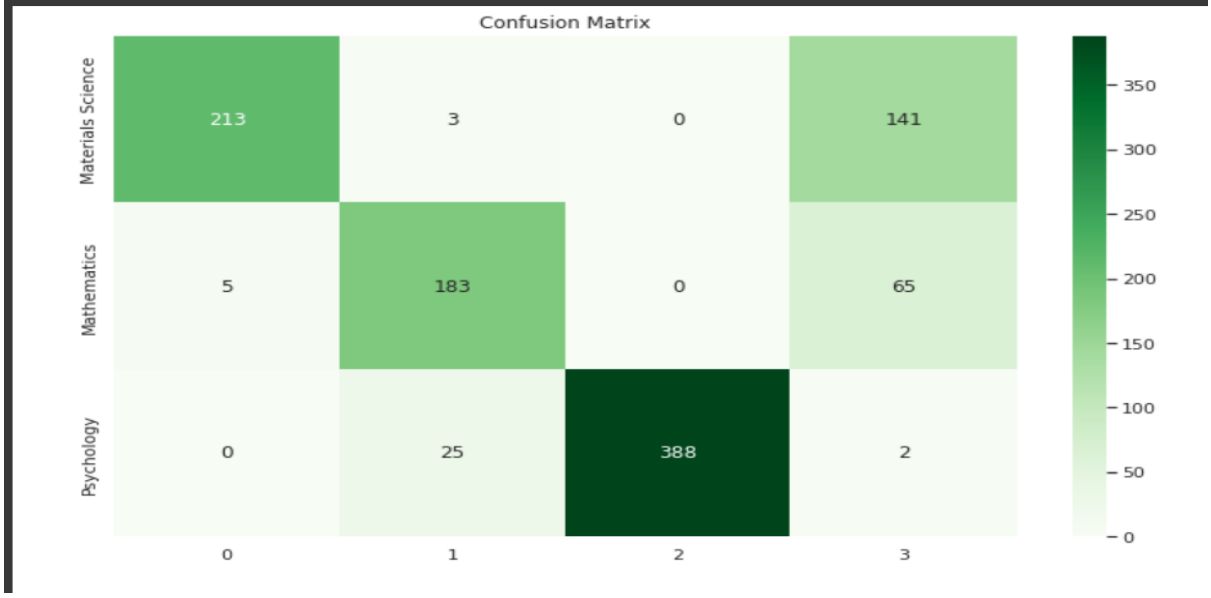


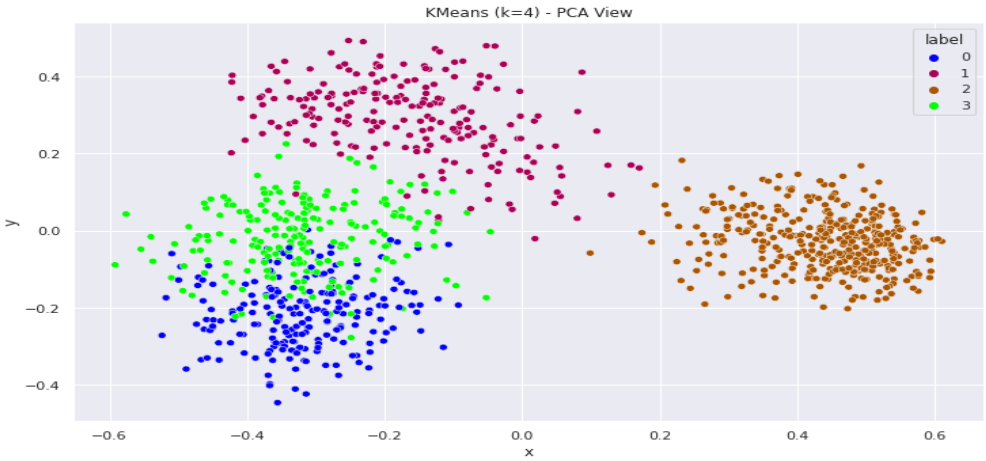


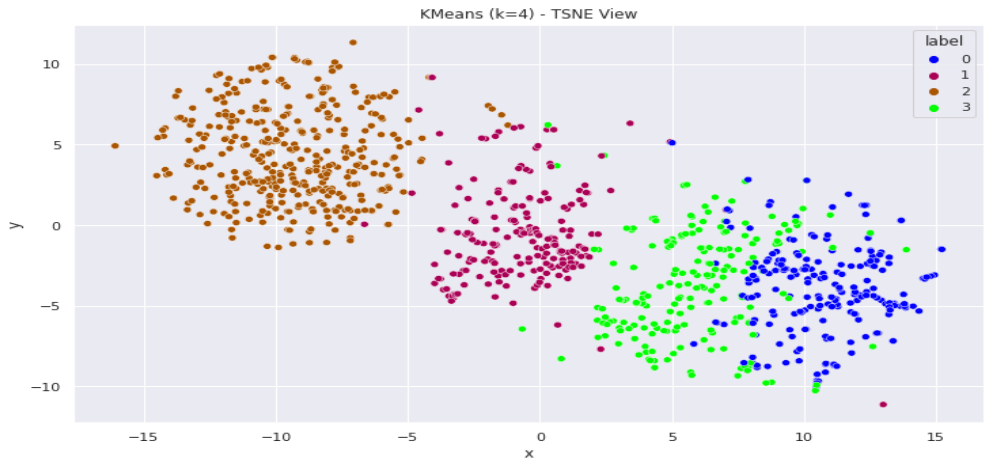


K = 4

```
Purity: 0.9024390243902439
silhouette_score = 0.12977702915668488
RSS = 526.402283429702
```

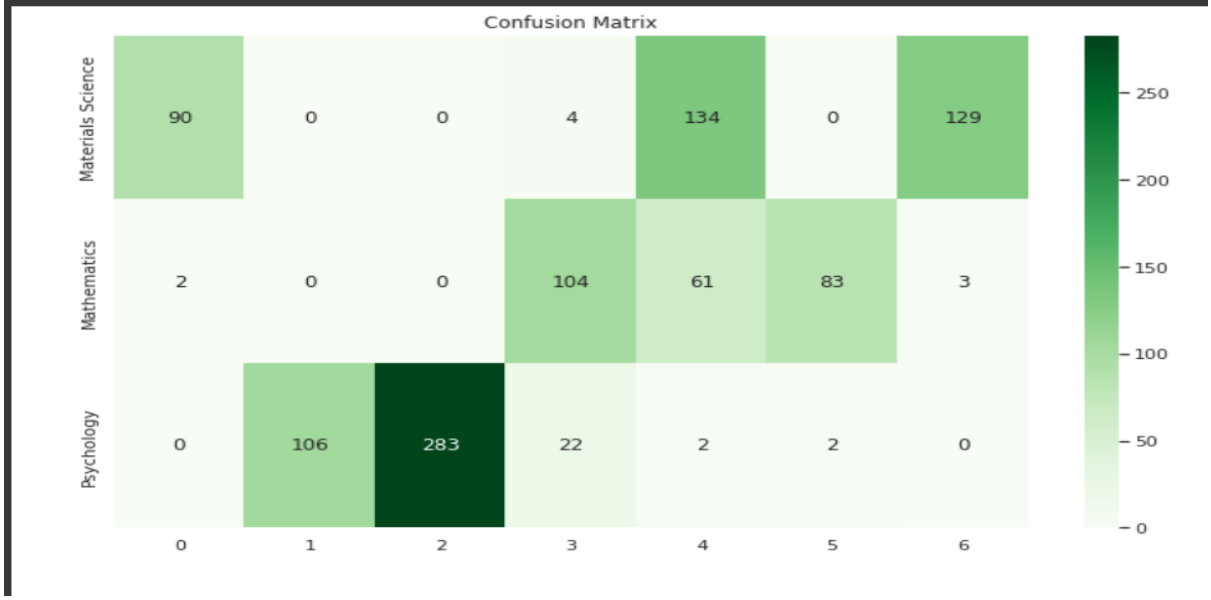


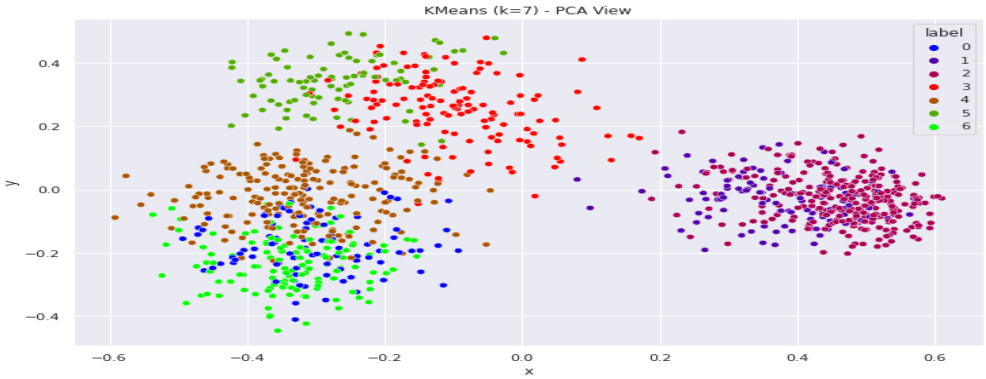


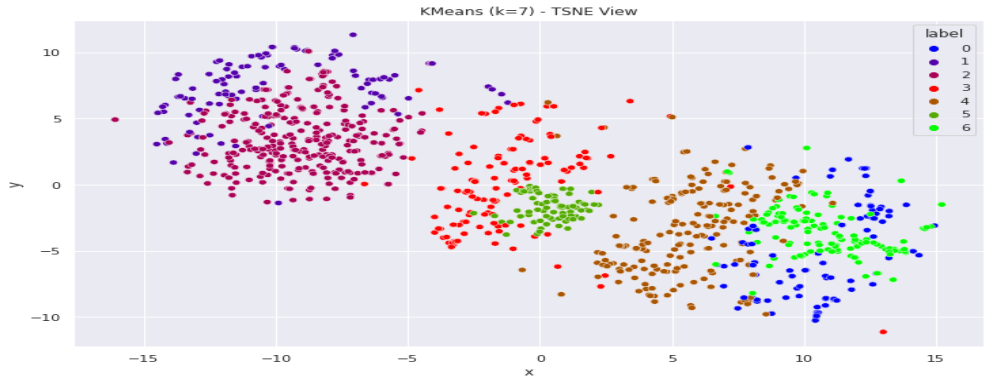


K = 7

```
Purity: 0.9063414634146342
silhouette_score = 0.059119388461112976
RSS = 501.91431903913826
```







مشاهده می‌شود معیار RSS با افزایش خوشه‌ها طبق توقع زیاد شده است. هرچند معیار silhouette در k=3 بیشینه است که نشان می‌دهد سه دسته برای خوشه‌بندی این داده مناسب است که این با شهود اینکه داده به صورت اولیه هم سه دسته بوده همخوان است.