گزارش انجام فاز دوم پروژه درس بازیابی پیشرفته اطلاعات

اعضای تیم:

سیدعلیرضا حسینی 96105701

سجاد شهابی 96105875

مهدی جوانمردی 96105683

پاییز ۹۹

ساختار کلی پروژه:

پروژه تشکیل شده از چندین مولفه است:

Transformer: این package وظیفه‌ی تبدیل داک‌ها به فضای tf-idf را برعهده دارد.

Classifiers: این package شامل 4 نوع دسته‌بند خواسته شده در قسمت اول این فاز پروژه را شامل می‌شود. علاوه بر این این package شامل کلاس ClassificationGenerator نیز است که وظیفه‌ی دسته‌بندی داکیومنت‌ها بر اساس یک دسته‌بند را برعهده دارد.

برای اجرای پروژه ابتدا لازم است pre-execution.py اجرا شود و سپس main.py در پوشه src اجرا شود. داده ها نیز باید در پوشه /data در مسیر root پروژه قرار بگیرند.

گزارش انجام قسمت اول (پیاده‌سازی دسته‌بندها):

همانطور که در قسمت قبل نیز گفته شد، این دسته‌بندها در پکیج Classifiers پیاده‌سازی شده‌اند.

علاوه بر این برای دسته‌بندهای KNN و SVM پارامترهای با توجه به داده‌های validation به صورت زیر انتخاب شده‌اند. برای اجرا validation از ۱۰ درصد داده‌های تمرین برای این‌کار و از بقیه برای تمرین استفاده خواهیم کرد و با توجه به نتایج زیر بهترین پارامتر را انتخاب می‌کنیم.

در دسته‌بند SVM برای هر کدام از پارامترهای زیر نتایج را بدست می‌آوریم:

parameters	Accuracy title/desc	F1-score class 1 title/desc	F1-score class -1 title/desc
1/2	0.568/0.607	0.596/0.618	0.536/0.596
1	0.600/0.627	0.592/0.617	0.608/0.636
3/2	0.589/0.627	0.577/0.630	0.6/0.627
2	0.584/0.627	0.578/0.638	0.590/0.615

با توجه به نتایج جدول صفحه‌ی قبل عدد 1 را برای پارامتر c در نظر می‌گیریم.

حال برای دسته‌بند KNN خواهیم داشت :

parameters	Accuracy title/desc	F1-score class 1 title/desc	F1-score class -1 title/desc
1	0.58/0.54	0.50/0.15	0.63/0.68
5	0.598/0.567	0.41/0.43	0.695/0.650
9	0.611/0.568	0.37/0.45	0.71/0.645

با توجه به نتایج جدول بالا عدد 9 را برای پارامتر k در نظر می‌گیریم.

گزارش انجام قسمت دوم (بهبود سیستم بازیابی اطلاعات فاز اول پروژه):

در این قسمت، موتور جستجوی آمده در فاز اول را بهبود داده‌ایم به گونه‌ای که در رابط کاربری در جست و جو به زبان انگلیسی قابلیتی اضافه شده‌است تا بتوانیم جست و جو را فقط در دسته داک‌هایی که از نظر دسته‌بند استفاده شده پربیننده هستند انجام دهیم که نتایج آن در پایین نمایش داده شده‌اند.

ابتدا جستجوی عادی را انجام دادیم که منجر به نتایج زیر شد.

```
1
Write 1 for title or 2 for description or 3 for both:
3
write your query:
invisible forces
Write 1 for most views or 2 for others:
2
your top 10 result is:
6,1921,2311,908,1043,1687,1309,1057,547,1956
Part 5:
1- Show Result of an Input Query Using tf-idf
```

سپس جستجو بر روی نتایج پربازدید را انجام دادیم. همانطور که مشخص است مستنداتی که خروجی داده است شامل همه‌ی موارد بالا نمی‌شود و موارد کم بازدید از نتایج خارج شده اند.

```
1
Write 1 for title or 2 for description or 3 for both:
3
write your query:
invisible forces
Write 1 for most views or 2 for others:
1
your top 10 result is:
6,1921,2311,2216,1963,559,2096,1334,312,2366
Part 5:
1- Show Result of an Input Query Using tf-idf
```

گزارش انجام قسمت سوم (ارزیابی نهایی):

روش Naive Bayes:

Confusion Matrix برای داده‌های آموزش به صورت زیر است:

Title	Predicted Class 1	Predicted Class -1
Actual Class 1	1090	64
Actual Class -1	150	991

Description	Predicted Class 1	Predicted Class -1
Actual Class 1	1153	1
Actual Class -1	55	1086

بنابراین معیارهای خواسته شده برای این داده‌های آموزش این دسته‌بند در زیر آمده‌اند.

	Title	Description

Accuracy	0.906753813	0.975599129
F1-score class 1	0.910609858	0.976291279
F1-score class -1	0.902550091	0.97486535
Recall class 1	0.879032258	0.954470199
Precision class 1	0.944540728	0.999133449
Recall class -1	0.939336493	0.999080037
Precision class -1	0.868536372	0.95179667

Confusion Matrix برای داده‌های آزمون به صورت زیر است:

Title	Predicted Class 1	Predicted Class -1
Actual Class 1	83	38
Actual Class -1	65	69

Description	Predicted Class 1	Predicted Class -1
Actual Class 1	102	19
Actual Class -1	82	52

بنابراین معایرهای خواسته شده برای این داده‌های آموزش این دسته‌بند در زیر آمده‌اند.

	Title	Description
Accuracy	0.596078431	0.603921569
F1-score class 1	0.617100372	0.668852459
F1-score class -1	0.572614108	0.507317073
Recall class 1	0.560810811	0.554347826
Precision class 1	0.685950413	0.842975207
Recall class -1	0.644859813	0.732394366
Precision class -1	0.514925373	0.388059701

روش k-NN:

Confusion Matrix برای داده‌های آموزش به صورت زیر است:

Title	Predicted Class 1	Predicted Class -1
Actual Class 1	471	683
Actual Class -1	94	1047

Description	Predicted Class 1	Predicted Class -1
Actual Class 1	778	376
Actual Class -1	339	802

بنابراین معایرهای خواسته شده برای این داده‌های آموزش این دسته‌بند در زیر آمده‌اند.

	Title	Description
Accuracy	0.66	0.69
F1-score class 1	0.54	0.685
F1-score class -1	0.729	0.691
Recall class 1	0.833	0.696
Precision class 1	0.408	0.674
Recall class -1	0.605	0.680
Precision class -1	0.917	0.702

Confusion Matrix برای داده‌های آزمون به صورت زیر است:

Title	Predicted Class 1	Predicted Class -1
Actual Class 1	28	93

Actual Class -1	22	112

Description	Predicted Class 1	Predicted Class -1
Actual Class 1	56	65
Actual Class -1	61	73

بنابراین معایرهای خواسته شده برای این داده‌های آموزش این دسته‌بند در زیر آمده‌اند.

	Title	Description
Accuracy	0.549	0.506
F1-score class 1	0.327	0.471
F1-score class -1	0.661	0.537
Recall class 1	0.56	0.479
Precision class 1	0.231	0.462
Recall class -1	0.546	0.529
Precision class -1	0.836	0.545

روش SVM:

Confusion Matrix برای داده‌های آموزش به صورت زیر است:

Title	Predicted Class 1	Predicted Class -1
Actual Class 1	1097	57
Actual Class -1	30	1111

Description	Predicted Class 1	Predicted Class -1
Actual Class 1	1096	58
Actual Class -1	15	1126

بنابراین معایرهای خواسته شده برای این داده‌های آموزش این دسته‌بند در زیر آمده‌اند.

	Title	Description
Accuracy	0.962	0.968
F1-score class 1	0.961	0.967
F1-score class -1	0.962	0.968
Recall class 1	0.973	0.986
Precision class 1	0.950	0.949
Recall class -1	0.951	0.951
Precision class -1	0.973	0.986

Confusion Matrix برای داده‌های آزمون به صورت زیر است:

Title	Predicted Class 1	Predicted Class -1
Actual Class 1	72	49
Actual Class -1	44	90

Description	Predicted Class 1	Predicted Class -1
Actual Class 1	63	58
Actual Class -1	34	100

بنابراین معایرهای خواسته شده برای این داده‌های آموزش این دسته‌بند در زیر آمده‌اند.

	Title	Description
Accuracy	0.635294	0.639215
F1-score class 1	0.607594	0.577981
F1-score class -1	0.659340	0.684931
Recall class 1	0.620689	0.649484
Precision class 1	0.595041	0.520661

Recall class -1	0.647482	0.632911
Precision class -1	0.671641	0.746268

روش Random Forest:

Confusion Matrix برای داده‌های آموزش به صورت زیر است:

Title	Predicted Class 1	Predicted Class -1
Actual Class 1	785	369
Actual Class -1	372	769

Description	Predicted Class 1	Predicted Class -1
Actual Class 1	971	183
Actual Class -1	213	928

بنابراین معایرهای خواسته شده برای این داده‌های آموزش این دسته‌بند در زیر آمده‌اند.

	Title	Description
Accuracy	0.677124183	0.82745098
F1-score class 1	0.679359585	0.830624465

F1-score class -1	0.674857394	0.824156306
Recall class 1	0.678478825	0.820101351
Precision class 1	0.680242634	0.841421144
Recall class -1	0.675746924	0.835283528
Precision class -1	0.673970202	0.813321648

Confusion Matrix برای داده‌های آزمون به صورت زیر است:

Title	Predicted Class 1	Predicted Class -1
Actual Class 1	75	46
Actual Class -1	53	81

Description	Predicted Class 1	Predicted Class -1
Actual Class 1	73	48
Actual Class -1	50	84

بنابراین معایرهای خواسته شده برای این داده‌های آموزش این دسته‌بند در زیر آمده‌اند.

	Title	Description

Accuracy	0.611764706	0.615686275
F1-score class 1	0.602409639	0.598360656
F1-score class -1	0.620689655	0.631578947
Recall class 1	0.5859375	0.593495935
Precision class 1	0.619834711	0.603305785
Recall class -1	0.637795276	0.636363636
Precision class -1	0.604477612	0.626865672