بازیابی پیشرفته‌ی اطلاعات نیم‌سال اول ۱۳۹۹ـ۱۳۹۸

گزارش فاز دوم پروژه

مدرس: دکتر بیگی اعضای گروه: احسان سلطان‌آقایی، وحید بالازاده مرشت

مشارکت وحید بالازاده

درصد مشارکت: ۵۰ درصد.

کارهای انجام‌گرفته:

- بخش ۲: با استفاده از الگوریتم SVM و با پارامتر $C = 0.5$ که به عنوان بهترین پارامتر در قسمت validation انتخاب شده بود، دسته‌بندی موضوعی به سیستم فاز اول اضافه شد. کافی است در هنگام جست‌وجو شماره‌ی مربوط به موضوع موردنظر نیز وارد شود. هم‌چنین تمام مستندات فاز اول موضوع‌بندی شدند.
- بخش ۱ (پیاده‌سازی الگوریتمKNN): کد مربوط به این الگوریتم در فایل knn.py آمده‌است. این الگوریتم براساس فاصله‌ی اقلیدسی بردارهای مستندات کار می‌کند.
- مقایسه‌ی پارامترهای مختلف در الگوریتم‌های SVM, KNN با استفاده از validation روی ده درصد از داده‌های آموزش. کد مربوط به این بخش در فایل validation.py قرار دارد.

مشارکت احسان سلطان‌آقایی

درصد مشارکت: ۵۰ درصد.

کارهای انجام‌گرفته:

- بخش ۱ (پیاده‌سازی الگوریتم naive bayes): به دو روش پیاده‌سازی شده است. اولی به اسم کلاسیک بدون در نظر گرفتن tf-idf عمل می‌کند و طبق اسلاید ۱۵ درس پیاده‌سازی شده است. این روش دقت بالای ۹۰ درصد در داده‌های تست دارد. روش دوم به اسم gaussian برای هر مستند یک بردار tf-idf در نظر می‌گیرد و برای هر جفت کلمه و تگ به‌ازای همه مستندهایی که این کلمه در آن موجود است و متعلق به تگ مربوطه است tf-idf را در نظر گرفته و یک توزیع نرمال به آن فیت می‌کنیم. سپس در محاسبه posterior برای محاسبه likelihood با فرض مستقل بودن کلمه‌ها از هم برای هر کلمه از توزیع نرمال حساب‌شده در قسمت قبل استفاده می‌کنیم. اگر برای جفت کلمه و تگی هیچ داده‌ای موجود نبود ولی کلمه جزو دیکشنری بود، احتمال آن را مطابق روش کلاسیک حساب می‌کنیم تا دقت روش دسته‌بندی‌مان بهبود یابد. در نهایت روش دوم به دلیل داده آموزش کم برای فیت‌کردن توزیع نرمال دقیق از دقت کم‌تری برخوردار است و دقت بالاتر از ۶۰ درصد داراست.

- بخش ۱ (پیاده‌سازی الگوریتم SVM and Random Forest): این دو روش نیز به کمک کتاب‌خانه sklearn پیاده‌سازی شده‌اند. هر دوی آن‌ها دقت بالای ۸۰ درصد دارند که در قسمت ارزیابی اطلاعات بیشتر آمده است.
- پیش‌پردازش و پیاده‌سازی‌های کلی این فاز

یافتن بهترین پارامترها

الگوریتم SVM

شکل‌های زیر نتایج اجرای پارامترهای مختلف را روی نود درصد از داده‌های آموزش و تست آن‌ها روی ده درصد داده‌ی validation نشان می‌دهد. با توجه به نتایج بهترین پارامتر $C = 0.5$ انتخاب شد.

parameter: 0.5

	precision	recall	f1-score	support
1	0.91	0.88	0.90	218
2	0.94	0.98	0.96	221
3	0.85	0.84	0.84	222
4	0.87	0.87	0.87	239
accuracy			0.89	900
macro avg	0.89	0.89	0.89	900
weighted avg	0.89	0.89	0.89	900

$C = 0.5$ (آ)

parameter: 1

	precision	recall	f1-score	support
1	0.90	0.88	0.89	218
2	0.95	0.97	0.96	221
3	0.83	0.83	0.83	222
4	0.85	0.86	0.86	239
accuracy			0.88	900
macro avg	0.88	0.88	0.88	900
weighted avg	0.88	0.88	0.88	900

$C = 1$ (ب)

parameter: 1.5

	precision	recall	f1-score	support
1	0.89	0.88	0.88	218
2	0.95	0.97	0.96	221
3	0.82	0.82	0.82	222
4	0.85	0.86	0.86	239
accuracy			0.88	900
macro avg	0.88	0.88	0.88	900
weighted avg	0.88	0.88	0.88	900

$C = 1.5$ (ج)

parameter: 2

	precision	recall	f1-score	support
1	0.89	0.88	0.88	218
2	0.95	0.96	0.96	221
3	0.81	0.82	0.81	222
4	0.85	0.85	0.85	239
accuracy			0.88	900
macro avg	0.88	0.88	0.88	900
weighted avg	0.88	0.88	0.88	900

$C = 2$ (د)

الگوریتم KNN

شکل‌های زیر نتایج اجرای پارامترهای مختلف را روی نود درصد از داده‌های آموزش و تست آن‌ها روی ده درصد داده‌ی validation نشان می‌دهد. با توجه به این‌که تفاوت خاصی بین نتایج $K = 5$ و $K = 9$ نیست، ما $K = 5$ را انتخاب می‌کنیم.

parameter: 1

	precision	recall	f1-score	support
1	0.84	0.89	0.86	218
2	0.91	0.90	0.91	236
3	0.82	0.81	0.81	232
4	0.80	0.78	0.79	214
accuracy			0.84	900
macro avg	0.84	0.84	0.84	900
weighted avg	0.84	0.84	0.84	900

$K = 1$ (آ)

parameter: 5

	precision	recall	f1-score	support
1	0.90	0.89	0.90	218
2	0.92	0.93	0.92	236
3	0.82	0.85	0.83	232
4	0.84	0.81	0.83	214
accuracy			0.87	900
macro avg	0.87	0.87	0.87	900
weighted avg	0.87	0.87	0.87	900

$K = 5$ (ب)

parameter: 9

	precision	recall	f1-score	support
1	0.89	0.89	0.89	218
2	0.91	0.93	0.92	236
3	0.83	0.86	0.84	232
4	0.85	0.80	0.83	214
accuracy			0.87	900
macro avg	0.87	0.87	0.87	900
weighted avg	0.87	0.87	0.87	900

$K = 9$ (ج)

ارزیابی نهایی

در زیر معیارهای خواسته‌شده برای هر یک از چهار الگوریتم آمده‌است. دقت کنید که پارامتر مربوط به SVM و KNN براساس validation به ترتیب $C = 0.5$ و $K = 5$ در نظر گرفته‌شده است.

	precision	recall	f1-score	support
1	0.96	0.94	0.95	2250
2	0.97	0.99	0.98	2250
3	0.92	0.92	0.92	2250
4	0.93	0.92	0.92	2250
accuracy			0.94	9000
macro avg	0.94	0.94	0.94	9000
weighted avg	0.94	0.94	0.94	9000

Classic Naive Bayes (آ)

	precision	recall	f1-score	support
1	0.94	0.92	0.93	2250
2	0.96	0.97	0.97	2250
3	0.90	0.91	0.90	2250
4	0.91	0.90	0.91	2250
accuracy			0.93	9000
macro avg	0.93	0.93	0.93	9000
weighted avg	0.93	0.93	0.93	9000

k-NN (ب)

	precision	recall	f1-score	support
1	0.99	0.99	0.99	2250
2	1.00	1.00	1.00	2250
3	0.99	0.98	0.99	2250
4	0.98	0.99	0.99	2250
accuracy			0.99	9000
macro avg	0.99	0.99	0.99	9000
weighted avg	0.99	0.99	0.99	9000

SVM (ج)

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2250
2	1.00	1.00	1.00	2250
3	1.00	1.00	1.00	2250
4	1.00	1.00	1.00	2250
accuracy			1.00	9000
macro avg	1.00	1.00	1.00	9000
weighted avg	1.00	1.00	1.00	9000

Random Forest (د)

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2250
2	1.00	1.00	1.00	2250
3	0.99	0.99	0.99	2250
4	0.99	0.99	0.99	2250
accuracy			0.99	9000
macro avg	0.99	0.99	0.99	9000
weighted avg	0.99	0.99	0.99	9000

Gaussian Naive Bayes (ه)

شکل ۳: نتایج روی داده‌های آموزش

	precision	recall	f1-score	support
1	0.90	0.85	0.87	250
2	0.92	0.93	0.93	250
3	0.78	0.82	0.80	250
4	0.82	0.82	0.82	250
accuracy			0.85	1000
macro avg	0.86	0.85	0.86	1000
weighted avg	0.86	0.85	0.86	1000

k-NN (ب)

	precision	recall	f1-score	support
1	0.93	0.88	0.90	250
2	0.92	0.96	0.94	250
3	0.83	0.83	0.83	250
4	0.84	0.84	0.84	250
accuracy			0.88	1000
macro avg	0.88	0.88	0.88	1000
weighted avg	0.88	0.88	0.88	1000

Classic Naive Bayes (آ)

	precision	recall	f1-score	support
1	0.85	0.82	0.83	250
2	0.86	0.95	0.90	250
3	0.80	0.80	0.80	250
4	0.84	0.78	0.81	250
accuracy			0.84	1000
macro avg	0.84	0.84	0.84	1000
weighted avg	0.84	0.84	0.84	1000

Random Forest (د)

	precision	recall	f1-score	support
1	0.91	0.88	0.89	250
2	0.92	0.97	0.95	250
3	0.84	0.85	0.84	250
4	0.86	0.84	0.85	250
accuracy			0.88	1000
macro avg	0.88	0.88	0.88	1000
weighted avg	0.88	0.88	0.88	1000

SVM (ج)

	precision	recall	f1-score	support
1	0.66	0.66	0.66	250
2	0.81	0.69	0.74	250
3	0.58	0.59	0.59	250
4	0.60	0.68	0.64	250
accuracy			0.66	1000
macro avg	0.66	0.66	0.66	1000
weighted avg	0.66	0.66	0.66	1000

Gaussian Naive Bayes (ه)

شکل ۴: نتایج روی داده‌های آزمون