گزارش کار:

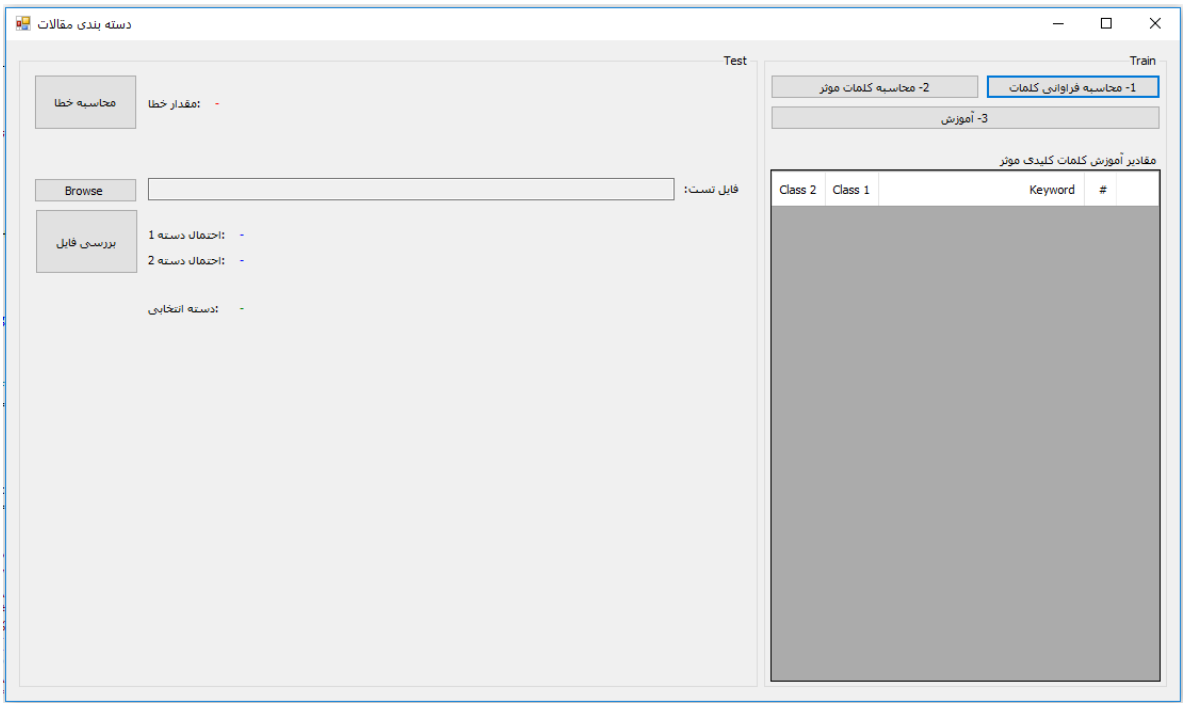
پروژه دسته بندی مقالات با استفاده از کلمات کلیدی موثر و طبقه بند بیزی

فرایند آموزش

در فرایند آموزش، برنامه می بایست با استفاده از مقالاتی که به عنوان مقالات آموزشی در نظر گرفته شده اند استفاده نموده تا به سطحی از دانش برسد که بتواند مقالات را دسته بندی نماید.

در این پروژه دو دسته مقالات و برای هر دسته 100 مقاله آموزشی موجود می باشد که در مجموع 200 مقاله آموزشی در دسترس می باشد.

فرایند آموزش شامل 3 مرحله می باشد که در ادامه مشاهده می کنید:



1- **محاسبه فراوانی کلمات:** در این مرحله برای تمامی مقالات آموزشی در هر دسته، فراوانی کلمات را بدست می آوریم و در فایلی به نام words_frequency.txt ذخیره می کنیم.

2- **محاسبه کلمات موثر:** در این مرحله می بایست با استفاده از مقدار فراوانی که برای کلمات در مرحله قبل بدست آمده است آنهایی که بیشترین تاثیر را برای انتخاب یک دسته دارند تعیین نمود که از فرمول زیر مقدار تاثیر هر کلمه در یک دسته بندی مشخص می شود:

effectiveCount(class) = Sum(class) – [Sum(Sum(OtherClass))]

effectiveCount(class): مقدار موثر کلمه کلیدی در کلاس

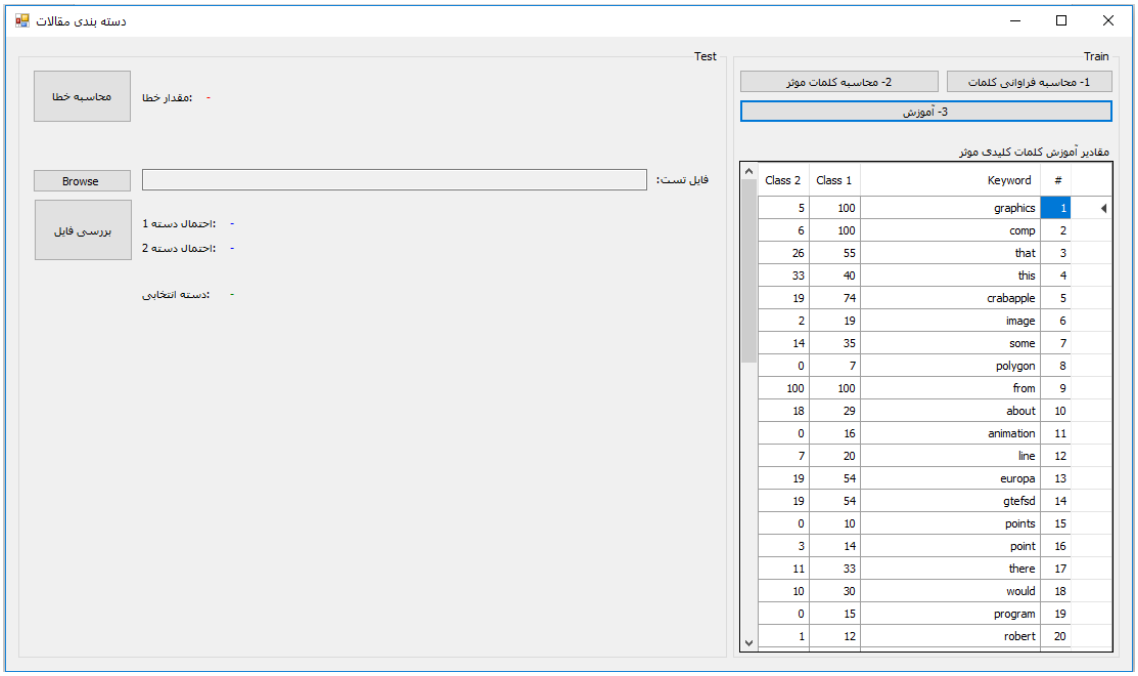
فروانی کلمه کلیدی در کلاس :Sum(class)

مجموع فراوانی کلمه کلیدی در دیگر کلاس ها :Sum(Sum(OtherClass))

پس از محاسبه مقدار موثر کلمات هر دسته از مقالات آنها را در فایلی به نام words_effective.txt ذخیره می کنیم.

همچنین تعداد 25 کلمه کلیدی که در هر کلاس دارای بیشترین مقدار موثر هستند را به عنوان خصوصیت های کلاس یا کلمات کلیدی نماینده آن کلاس در نظر می گیریم و در فایل class_attributes.txt ذخیره می کنیم.

3- **محاسبه مقادیر فراوانی کلمات مورد استفاده در فرمول بیز (آموزش):** در این مرحله برای هر یک از کلمات کلیدی نماینده کلاس که مجموعا 50 کلمه هستند (25 کلمه کلیدی به ازای هر دسته)، فراوانی در هر دسته یا کلاس را محاسبه می کنیم تا از این مقادیر در فرمول بیز برای مقالات تستی بتوان جهت دسته بندی استفاده نمود. برای محاسبه این مقدار در هر کلاس تعداد مقالاتی که دارای یک کلمه کلیدی هستند را می شماریم. در آخر برای تمامی کلمات کلیدی مقدار فراوانی آن در مقالات هر کلاس را خواهیم داشت.



فرایند تست

روش بررسی مقاله تستی: برای بررسی مقاله تستی و تعیین کلاس آن مراحل زیر انجام می شود:

1- تعیین کلمات کلیدی مقاله تستی
2- انتخاب کلمه کلیدی های معرف هر کلاس از میان کلمات کلیدی مقاله تستی
3- انتخاب نسبت فراوانی کلمه کلیدهای انتخاب شده به ازاي هر کلاس که در مرحله آموزش برای آن بدست آمده است.
4- محاسبه احتمال وجود مقاله در يك کلاس با استفاده از احتمالات بدست آمده براي کلمات کليدي آن در همان کلاس

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

$\hat{y} = C_k$ for some k as follows:

5- مقايسه احتمالات بدست آمده وجود مقاله به ازاي هر كلاس و انتخاب بزرگترين احتمال به عنوان كلاس انتخاب شده براي مقاله تستي.

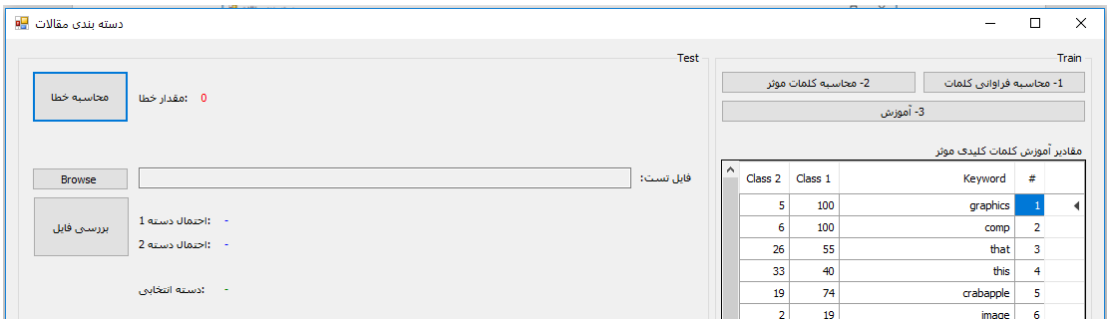
$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

در فرايند تست دو عمليات پياده سازی شده است كه در زير مشاهده می كنيد:

1- **محاسبه خطا:** كلاس انتخاب شده براي مقاله هاي تستي كه 10 مقاله براي دسته اول و 10 مقاله براي دسته دوم مي باشد كه در مجموع 20 مقاله تستي موجود است را با كلاس معرفي شده براي آنها مقايسه مي كنيم و در نهايت از فرمول زير براي نمايش نسبت خطا استفاده مي كنيم. مقدار بدست آمده براي خطاي محاسبه شده عددي بين 0 و 1 مي باشد كه هر چه كوچكتر باشد يعني سيستم داراي خطاي كمتري است. همانطور كه مشاهده مي كنيد براي برنامه پياده سازي شده عدد 0 بدست آمده كه نشان مي دهد برنامه تمامي مقالات تستي را در گروه درست طبقه بندي كرده است.

تعداد طبقه بندي نادرست

تعداد كل مقالات تستي



2- **بررسي يك فايل انتخابي توسط كاربر:** كاربر كه فايل مقاله را انتخاب مي كند و سيستم با روش طبقه بندي بيز احتمال وجود مقاله در هر دو دسته به روشي كه پيشتر توضيح داده شد محاسبه مي كند و دسته اي را كه احتمال بيشتري دارد را به عنوان كلاس انتخابي پيشنهاد مي دهد. كه در شكل زير مشاهده مي كنيد. لازم به ذكر است به علت مساوي بودن تعداد مقالات آموزشي و در نتيجه مساوي بودن احتمال وجود مقالات در هر كلاس كه برابر 0.5 مي باشد در محاسبه احتمال صرف نظر شده است.