گزارش پروژه داده کاوی

کسری آقابیکی

برای ران کردن کد فایل exe. وجود دارد ولی به دلیل امنیت ویندوز اجازه ران گرفتن نمی‌ده مگر اینکه آنتی ویروس غیر فعال بشه و ران از ادمینستر بگیریم. راه حل دیگه اینه که فیلم زیر رو ببینین که از مراحل اجراش گرفتم.

https://youtu.be/m1gj401NLT0

Kmeans

فایل مورد نظر را به انتخاب می‌کنیم، بعد تعداد خوشه ها رو انتخاب می‌کنیم و در نهایت یک seed است که عددی هست که برای شروع الگوریتم تولید اعداد تصادفی استفاده می‌شود. دلیل استفاده از seed این است که تولید اعداد تصادفی تا حدودی در دست ما باشد تا در صورت نیاز بتوانیم جواب ها را دنبال کنیم و یا حتی یک جواب را دوباره تولید و چک کنیم.

در پایان خروجی شامل کلاسترهای مورد نظر و اعضای آن است. همچنین خروجی شامل یک نوار هست که به صورت خیلی تقریبی درصد کلی پراکندگی داده ها در کلاستر ها را نشان می‌دهد.(یان نوار اصلا دقیق نیست چون هر کدام از I ها معادل یک درصد است. با تغییر عدد در خط زیر می‌توان تعداد I ها را زیاد کرد و دقت نمایش درصد بالا می‌رود)

```
int totalBarLength = 100; // Total number of I
```

Apriori

فایل مورد نظر را به انتخاب می‌کنیم، بعد تعداد دسته بندی ها از ما خواسته می‌شود. این دسته بندی با خوشه بندی دفعه قبل فرق دارد و برای تبدیل داده های پیوسته به گسسته استفاده می‌شود. برای مثال اگر من اعداد بین 0 تا 30 دارم که نشان دهنده دمای ها هست و می‌خواهم آن را به سه دسته سرد، معتدل و گرم تقسیم کنم، عدد 3 را وارد می‌کنم، اینطوری خود کد این بازه را به 3 بازه مساوی تقسیم می‌کند و دامنه هر بازه را کنار خودش می‌نویسه.(اگر داده های ما تماما گسسته هستند این عدد اهمیتی ندارد می‌تونیم اون رو 1 در نظر بگیریم چون روی هیچ داده ای اعمال نمی‌شه). این بازه های گسسته به ترتیب با 1، 2 ، 3 و ... نامگذاری می‌شوند.

بعد از اون نوبت minimum support هست که به صورت درصد بین 0 و 1 نوشته می‌شه. برای مثال 27 درصد برابر است با 0.27.

نتیجه شامل Frequent itemset ها در سطح های مختلف است. در هر سطح نام داده ها(اگر داده ها گسسته بوده باشند، یک عدد به عنوان نام و یک رنج که عدد نماینده آن است در کنار آن است) و در انتهای خط درصد حضور آن است.