سیستم هوشمند ارزیابی کیفیت کالاها با استفاده از نظر مشتریان

مقدمه:

این پروژه مربوط به نظرات مشتریان نسبت به محصولات شرکت دیجی‌کالا می‌باشد. هدف ما ارزیابی کیفیت محصولاتی است که در دیجی‌کالا در حال فروش می‌باشند که در سه قسمت به تفسیر آن می‌پردازیم. (به دلیل حجم بالای محاسبات آزمایش‌ها بر روی ۱۰۰۰ تا از داده‌ها انجام شده است.)

قسمت اول:

در این قسمت به دسته‌بندی کالاها براساس کاربرد آن‌ها می‌پردازیم. برای انجام این کار از الگوریتم‌های خوشه‌بندی(clustering) استفاده می‌کنیم. جهت استفاده از الگوریتم‌های خوشه‌بندی بایستی دیتاست مربوط را آماده کنیم. برای این کار کلمات زائد مربوط به ستون product_title را حذف کرده و بقیه کلمات پر تکرار را به عنوان متغیرهای دیتاست در نظر گرفته و در نهایت با استفاده از الگوریتم K_medoids به دسته‌بندی محصولات می‌پردازیم. دلیل استفاده از الگوریتم k_medoids ، وجود داده‌ها به صورت categorical می‌باشد. (کدهای مربوطه در فایل part1 می‌باشد.)

قسمت دوم:

در این قسمت به دسته‌بندی نظرات مشتریان به دو دسته راضی و ناراضی می‌پردازیم. جهت انجام این کار پس از حذف کلمات زائد، کلماتی که معانی رضایتمندی دارند را در لیستی به نام "satisfaction" قرار می‌دهیم و در صورت استفاده آن‌ها در هر سطر عدد **یک** را در دیتاست به خود اختصاص می‌دهند و کلماتی که دارای معانی منفی هستند را در لیست دیگری به نام "not satisfaction" قرار داده و در صورت استفاده در هر سطر، عدد **دو** را به خود اختصاص می‌دهند و بقیه کلمات با مقدار **صفر** در دیتاست ظاهر می‌شوند. الگوریتم "K_means" را جهت دسته بندی نظرات مشتریان اجرا می‌کنیم.

در پایان اجرای الگوریتم مقدار silhouette_score برابر **٪۶۷** می‌شود. و مقدار هیچ کدام از silhouette_sample منفی نخواهد شد؛ که نشان می‌دهد الگوریتم دسته‌بندی به درستی انجام گرفته است. (کدهای مربوط به این قسمت در فایل part2 می‌باشد.)

قسمت سوم (۱_۳):

در این قسمت به بررسی بیشترین علّت رضایتمندی مشتریان از تلفن‌های همراه می‌پردازیم. برای انجام این کار ابتدا از میان کلمات ستون advantages ، کلمات کاربردی را انتخاب کرده و تعداد تکرار هر یک از آن‌ها را در هر سطر با استفاده از بسته CountVectorizer شمارش کردیم و در انتها با استفاده از الگوریتم k_mean علل رضایت مندی از تلفن‌های همراه را به ۵ دسته تقسیم بندی کردیم که دسته اول مربوط به **کفیت بالای صفحه نمایش و دوربین** با بیشترین تکرار به عنوان اصلی‌ترین علّت رضایتمندی انتخاب می‌شود.

sum(c.iloc[:,2]==0) ⟶ 26

sum(c.iloc[:,2]==1) ⟶ 17

sum(c.iloc[:,2]==2) ⟶ 16

sum(c.iloc[:,2]==3) ⟶ 20

sum(c.iloc[:,2]==4) ⟶ 9

در پایان الگوریتم، مقدار average_silhouette برابر ۶۸٪ شد و مقدار هیچ یک از silhouette_samples ، منفی نشده است؛ که نشان می‌دهد الگوریتم k_mean به درستی انجام گرفته است. لازم به ذکر است با استفاده از تمامی کلمات ، این مراحل دوباره اجرا شده است و نتیجه حاصل همانند نتیجه فوق می‌باشد. (کدهای مربوط به این قسمت در فایل part 3.1 ذخیره شده است.)

قسمت سوم (۲_۳) :

در این قسمت همانند قسمت قبلی عمل می‌کنیم با این تفاوت که به دلیل تکرار کم تمامی کلمات در ستون disadvantages ، تمامی کلماتی که بیشتر یا مساوی ۲ بار تکرار در کل دیتاست داشته‌اند را به عنوان متغیرهای مسئله می‌پذیریم سپس تعداد دفعات تکرار آن‌ها در هر سطر را با استفاده از بسته CountVectorizer شمارش کرده و با استفاده از الگوریتم k_means بشترین علل نارضایتی را به ۵ دسته تقسیم بندی کرده‌ایم. بیشترین تکرار مربوط به دسته ۴ ام یعنی **باتری ضعیف و قابلیت‌های کم** به عنوان اصلی‌ترین علّت نارضایتی از تلفن‌های همراه انتخاب می‌شود.

در پایان الگوریتم، مقدار average_silhouette برابر ۷۱٪ شد و مقدار هیچ یک از silhouette_samples ، منفی نشده است؛ که نشان می‌دهد الگوریتم k_mean به درستی انجام گرفته است.(کدهای مربوط به این قسمت در فایل part3.2 ذخیره شده است.)

sum(c.iloc[:,2]==0) ⟶ 18

sum(c.iloc[:,2]==1) ⟶ 9

sum(c.iloc[:,2]==2) ⟶ 13

sum(c.iloc[:,2]==3) ⟶ 23

sum(c.iloc[:,2]==4) ⟶ 15

دلیل شمارش کلمات در قسمت سوم با استفاده از بسته CountVectorizer این است که با بررسی دیتاست متوجّه آن می‌شویم که تعدادی از کلمات در هر سطر ، حداقل ۲ بار تکرار شده‌اند که می‌تواند کمک زیادی برای پیدا کردن علّت رضایتمندی یا نارضایتی مشتریان بکند. با توجه به این که برخی از علّت‌ها در هر دسته مشترک بوده است و شاهد خطاهایی در الگوریتم دسته بندی خواهیم بود مبنای انتخاب علّت اصلی در دو قسمت قبلی ، تکرار زیاد آن علّت خاص در دسته‌ای که بیشترین تکرار داشته است ، می‌باشد.

تهیّه و تنظیم:

حمید ابراهیمی