پروژه نهایی - تحلیل و آنالیز دیتای یوتیوب (Youtube)

طراحان: مسعود مظلوم، بهناز ریوندی

دانشکده علوم ریاضی دانشگاه فردوسی مشهد

پاییز ۱۴۰۰

یوتیوب مدتی طولانی است که فعالیت خود را شروع کرده است – احتمالاً درباره آن زیاد شنیده اید. سایت اشتراک گذاری ویدیو به طور لایو و زنده در سال ۲۰۰۵ آغاز به کار کرد، و از آن زمان به بعد برای بارگذاری فیلم در وب، به رسانه ای مستقل تبدیل شده است. به این ترتیب نگاهی به اصول یوتیوب می تواند مفید باشد. یوتیوب یک سیستم عامل مبتنی بر ویدئو است که توسط دو نوع کاربر هدایت می شود:

- سازندگان ویدیو (افرادی که کانال دارند و فیلم ها را در آن کانال ها بارگذاری می کنند)
- بینندگان ویدیو (افرادی که فیلم ها را تماشا می کنند، با فیلم ها تعامل دارند و در کانال ها مشترک و عضو می شوند)

هدف از این پروژه تحلیل و آنالیز این شبکه است که در ادامه به بررسی آن می‌پردازیم. این پروژه برای شما در ۲ مرحله زیر، طراحی شده‌است:

- پیش‌پردازش داده‌ها
- آنالیز ، تحلیل و نتیجه‌گیری

توجه داشته باشید که برای انجام این پروژه، فقط مجاز به استفاده از کتابخانه‌های *matplotlib*, *pandas*, *numpy* و *seaborn* خواهید بود.

پیش‌پردازش

در اولین گام از پروژه نهایی، قصد داریم با پیش‌پردازش داده و پاسخ به تعدادی سوال ساده، ضمن دید پیدا کردن نسبت به داده، آن را برای مراحل بعدی آماده کنیم. مراحل پیش‌پردازش و سوالات ساده، جدا از همدیگر نیستند. یعنی ابتدا چند مرحله پیش‌پردازش داریم، سپس به تعدادی سوال تحلیلی جواب خواهیم داد و مجددا داده را پیش‌پردازش خواهیم کرد.

۱- ابتدا کتاب‌خانه‌های مورد نیاز خود را اضافه کنید.

۲- در این پروژه ما دو فایل (csv) در اختیار شما قرار می‌دهیم، فایل‌های $US\_youtube\_trending\_data.csv$ و $category\_ids.csv$. اولین قدم برای آشنایی با دیتا بررسی ویژگی‌های (feature) دیتاست مورد نظر است. (در مورد ستون‌های فایل دیتا تحقیق کنید و در نوت‌بوک ذکر کنید.)

۳- در ابتدا دو فایل csv را بخوانید و ستون‌هایی که اطلاعاتی ارزشمند به ما نمی‌دهند را حذف می‌کنیم. برای مثال ستون‌های:

$'video\_id', 'channelId', 'thumbnail\_link', 'comments\_disabled', 'ratings\_disabled'$.

۴- کتابخانه datetime برای کار با تاریخ و زمان در پایتون آماده شده‌است. این ماژول از ۵ نوع داده (type) پشتیبانی می‌کند. مواقع زیادی پیش می‌آید که رشته‌ای داشته باشیم و بخواهیم آن را به زمان تبدیل کنیم. در دیتاست ما دو ستون $publishedAt, trending\_date$ داریم و هدف در این مرحله تبدیل دیتای رشته به datetime است. (برای هر ویژگی دو ستون جدید به نام‌های $time\_trending$، $date\_trending$، $time\_published$، $date\_published$ تعریف کنید.)

۵- یکی از اطلاعاتی که برای ما می‌تواند مفید باشد، استخراج ماهی از سال است که ویدیو منتشر یا ترند شده است.(ستون‌های $month\_trending$، $month\_published$ را اضافه کنید و اعداد [8,9,10,11] را با نام ماه مناسب خود ['Aug', 'Sept',"Oct', 'Nov"] جایگزین کنید.

۶- در این گام، شما باید ستونی جدیدی به نام lag اضافه کنید که این ویژگی(feature) مدت زمان بین پست شدن و ترند شدن هر ویدیو را مشخص می‌کند.($\#Calculate\ lag\ time\ between\ posting\ and\ trending$)

۷- با مشاهده‌ی دیتاست متوجه می‌شوید که ستونی به نام $categoryId$ وجود دارد، شما باید از فایل csv دیگری که در اختیارتون قرار گرفته است، این آی‌دی را به نام کتگوری مپ کنید. ($\#Covert\ category\ IDs\ to\ category\ names$)

۸- برای راحتی کار با دیتاست نام‌های ویژگی‌های (feature) زیر را به شکل زیر تغییر دهید.

$$\{'channelTitle': 'channel',\ 'categoryId': 'category',\ 'view\_count': 'views',\ 'comment\_count': 'comments'\}$$

۹- دیتاهای تکراری را حذف کنید. $(drop\_duplicates(subset = 'title',\ keep = 'first'))$

آنالیز ، تحلیل و نتیجه‌گیری

۱۰- در این قسمت به تحلیل دیتاست پس از پیش‌پردازش می‌پردازیم، دیتا را بر اساس تعداد like ها به صورت نزولی مرتب کنید. ($\#Sort\ by\ 'like',\ most\ to\ least$)

۱۱-نشان دهید در هر روز از تاریخ چه تعداد ویدیوای منتشر شده است.

($\#See\ how\ many\ videos\ were\ published\ each\ day\ in\ the\ dataset$)

۱۲- نتیجه‌ای که از بالا بدست آوردید را plot کنید. بعد از رسم نمودار آیا بین روزهای هفته و منتشر شدن ویدیو ارتباطی وجود دارد؟(نتیجه خود را به طور کامل شرح دهید.)

۱۳- در این بخش به بررسی چهار ماه سال و تعداد کل $published,\ trending,\ views$ بپردازید.

	published	trending	views
Aug			
Sept			
Oct			
Nov			

۱۴- در قسمت ۶، ستونی جدید به نام lag به دیتاست مون اضافه کردیم، در این بخش با استفاده از نمودار هیستوگرام به بررسی آن می‌پردازیم.$plt.hist(lag\_data, density = True)$. (نتیجه و برداشت خود را از نمودار رسم شده به طور کامل شرح دهید.)

۱۵- در این قسمت رابط کاربری بین کاربر و دیتاستمون طراحی می‌کنیم به گونه‌ای که براساس درخواست کاربر تعداد فیلم‌های منتشر شده در ماه مشخص شده را نشان دهد. سوالاتی که باید از کاربر پرسیده شود به صورت زیر است.

$Month\ for\ most\ liked\ videos\ (Aug,\ Sept,\ Oct,\ Nov)?$

$How\ many\ videos\ to\ see?$

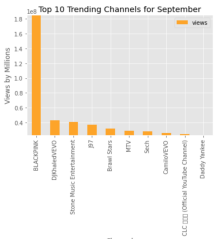
۱۶- در بخش‌های بعدی به بررسی ماه و ترند شدن ویدیو می‌پردازیم. در مرحله‌ی اول دیتاست را براساس ماه ترند شده، مرتب کنید.
دیتا فرمی که مربوط به این بخش است:

channels = youtube_data[['channel','views','month_trending']]

- در این بخش ده ویدیو برتر مربوط به ماه‌های (Aug, Sep, Oct, Nov) را نمایش دهید (برتر بودن براساس views ای که داشته است مشخص می‌شود.)
- دیتا و نتایجی که به دست اوردید را پلات کنید. (۴ نمودار)

برای مثال:

channel	views	month_trending
BLACKPINK	184778248	Sept
DJKhaledVEVO	43394819	Sept
Stone Music Entertainment	41213361	Sept
J97	37422074	Sept
Brawl Stars	32114735	Sept

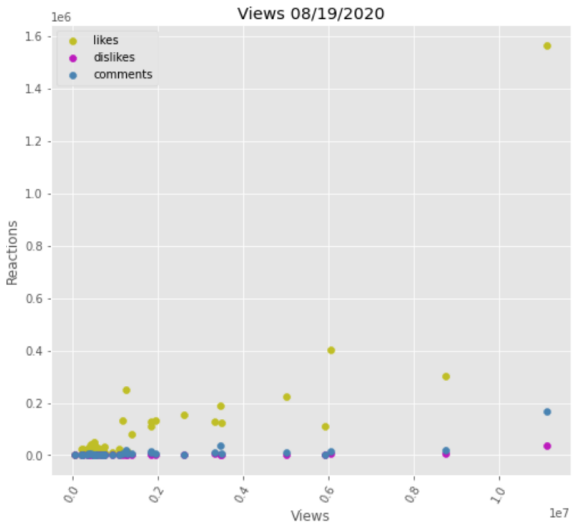


۱۷- در این بخش به بررسی نظر کاربران به ویدیوهای منتشر شده براساس دسته بندی (category) آن می‌پردازیم. براساس دسته بندی دیتا را گروه بندی کنید. تعداد $dislikes$ و $likes$ را برای هر دسته بندی نشان دهید. ویژگی(feature)جدیدی به نام $total\_opinions$ به دیتا فریم جدیدی که تعریف کردید اضافه کنید و مجموع $dislikes$ و $likes$ را در این ستون قرار دهید. علاوه بر این ستون‌هایی به نام‌های $\%dislike$، $\%like$ اضافه کنید که درصد این دو ویژگی را بدهد.

۱۸- با استفاده از نمودار میله‌ای، نشان دهید که در هر دسته بندی (category)، چند تعداد ویدیو منتشر شده است.

۱۹- در این بخش قصد داریم ۴ تاریخ، با روز یکسان اما ماه‌های مختلف انتخاب کنیم و تمامی ری‌اکشن‌های بینندگان ویدیو را نسبت به ویدیوهای منتشر شده، با استفاده از نمودار scatter نمایش دهیم. (چهار روز انتخابی '08/15/2020'، '09/15/2020' ، '10/15/2020', '11/01/2020' و likes و dislikes و comments را بر روی نمودارها مشخص کنید.)

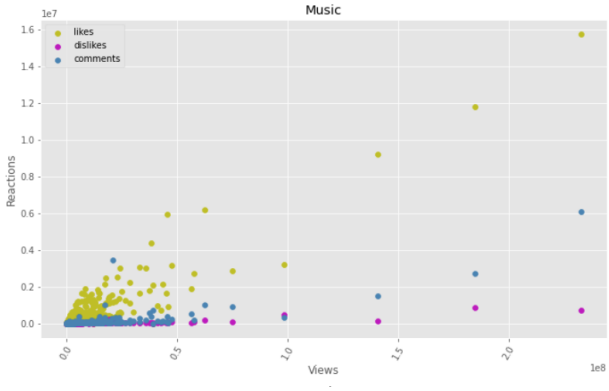
برای مثال نمودار زیر برای تاریخ 08/19/2020 است که رسم شده است. (نتیجه و برداشت خود را از نمودار رسم شده به طور کامل شرح دهید.)



۲۰- در این بخش ۱۰۰ ویدیو برتر را ابتدا جدا کنید سپس با رسم نمودار ویژگی‌ها scatter, *likes*، *dislikes*، *comments* را بررسی کنید .

۲۱- در این قسمت قصد داریم تحلیل‌هایی که برای قسمت بالا انجام دادیم را تکرار کنیم اما با تفاوت اینکه نظر بینندگان ویدیو را نسبت به هر دسته بندی (category) بررسی کنیم. (دسته بندی‌هایی (category) که بررسی می‌کنید شامل لیست زیر باشد: 'Gaming'، "News & Politics")

برای مثال نمودار زیر در دسته بندی Music است که رسم شده است.



۲۲- در این گام به بررسی کلید واژه‌های خاص در ستون title می‌پردازیم. اولین کلید واژه‌ای که سرچ می‌کنیم "trump" است. قصد داریم در این چهار ماه نظرات و ری‌اکشن‌ها رو بررسی کنیم. از نظر شما چه نموداری با توجه به دیتای کمی که در اختیار داریم بهتر است. (دقت کنید قبل از بررسی [df=df[df['category']=='News & Politics])

نمودار ابر کلمه (word cloud): نمودار word cloud با شکستن متون به کلمات اجازه می دهد که بیننده مشاهده کند که چه کلماتی بیشتر استفاده شده اند و چه کلماتی کمتر. هرچه اندازه یک کلمه در ابر بیشتر باشد ، بیشتر از آن استفاده شده است .
روشی برای مصور سازی اطلاعات که با نمایش کلمات با سایز های مختلف (اندازه هر کلمه بر اساس تکرار / فرکانس) در متن مشخصی نشان می دهد. سپس تمام کلمات در یک خوشه یا ابر از کلمات مرتب می شوند. در این نمودار می توان از متا دیتا ها برای بصری سازی اطلاعات نیز بهره برد.

۲۳- در این بخش قصد داریم با بررسی دیتای title، ابر کلماتی بسازیم و نشان دهیم چه کلماتی فرکانسی و تعداد بیشتری داشته است.