# **דו"ח טכני**

**Collect Data**

1. ייבוא ספריות:

* Pandas
* googleapiclient.discovery
* tmdbv3api
* tqdm

1. הגדרת משאבי API ואיסוף הנתונים:

הקוד משתמש במפתחות כדי לאפשר גישה למידע מהשירותים הבאים:

* YouTube Data API – שימוש בשירות על מנת לחפש ולמשוך טריילרים של סרטים כדי לבצע תמלול ולהכניס את קבצי הטקסט לעמודה ב -Data Frame המכילה את מילות הטקסט בכל טריילר.
* TMDb API – השתמשנו בשירות על מנת להוסיף ל - Data Frame עמודה שמכילה את ה – Genres כדי שלכל סרט יהיה שיוך תחת איזה קטגוריה הוא נמצא.

1. הצגת התקדמות:

השתמשנו ב-tqdm כדי להציג פסי התקדמות עבור לולאות.

1. אחסון הנתונים:

הקוד שומר את המידע הרלוונטי לפי העמודות:

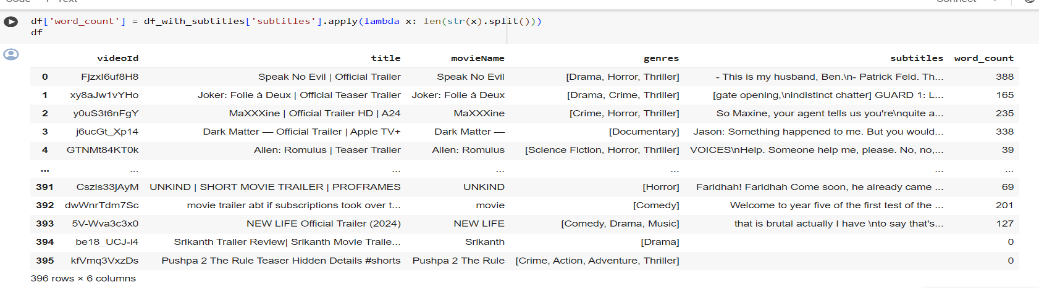
* Video ID
* Title
* Movie Name
* Genres
* Subtitle

1. פונקציה לחילוץ כתוביות מווידואי YouTube :
2. ייבוא הספריות הנדרשות לפעולת החילוץ:

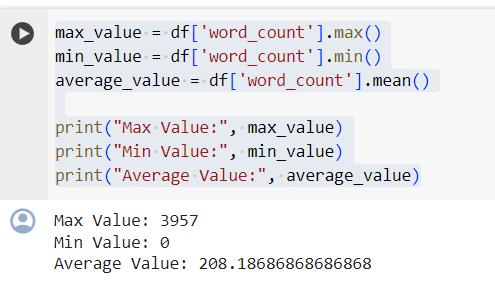
* Pandas
* pytube
* youtube\_transcript\_api
* tqdm

1. הגדרת פונקציה עם פרמטר df המייצג Data Frame שמכיל פרטי וידאו.
2. יצירת עמודה חדשה ב- Data Frame לשמירת הכתוביות.
3. לולאת עיבוד על כל השורות בData Frame על מנת לעבור על כל הטריילרים.
4. חיפוש וחילוץ הכתוביות לפי קוד הוידאו.
5. הוספת הכתוביות לשורה המתאימה בData Frame .
6. קוד זה מוסיף עמודה חדשה ל-Data Frame בשם “word\_count”.

הוא מחשב את ספירת המילים עבור כל ערך בעמודת “Subtitle” של ה-Data Frame על ידי פיצול הטקסט למילים ולאחר מכן ספירת מספר המילים בכל ערך באמצעות פונקציית היישום עם פונקציית lambda.

****

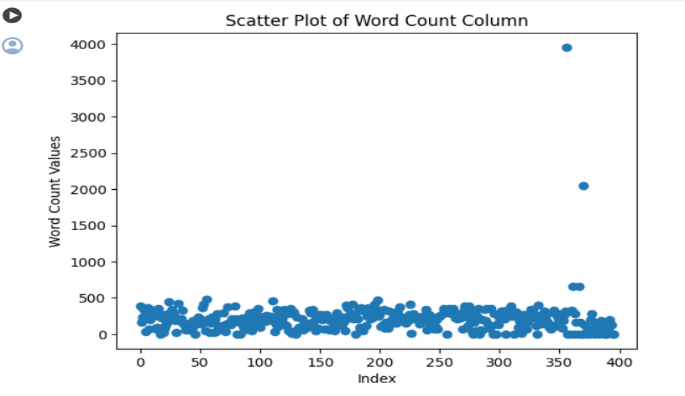
1. חישוב ערכי המקסימום, המינימום והממוצע של העמודה “word\_count” ב-Data Frame.



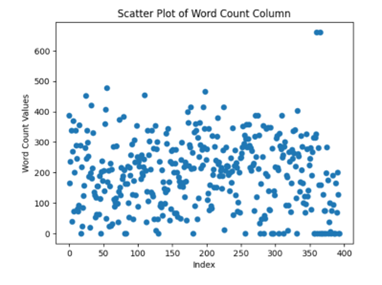
זיהוי ערכים חריגים ב- Data Frame כאשר יש שורות בדאטה שמכילות 0 מילים.

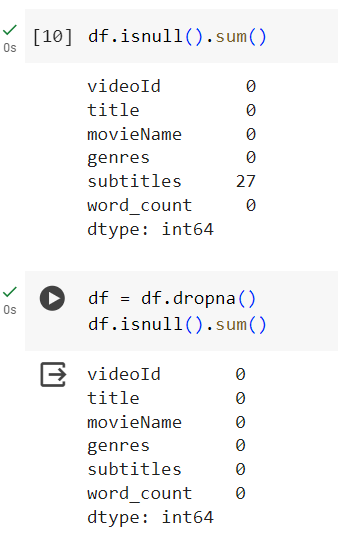
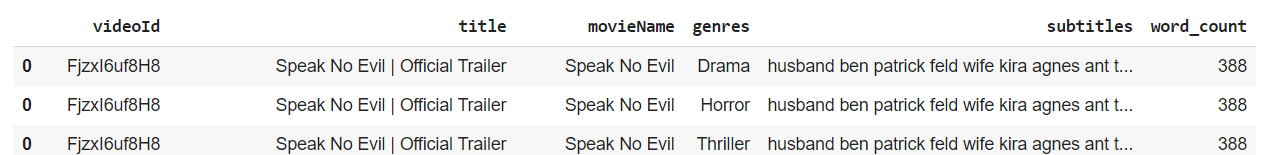
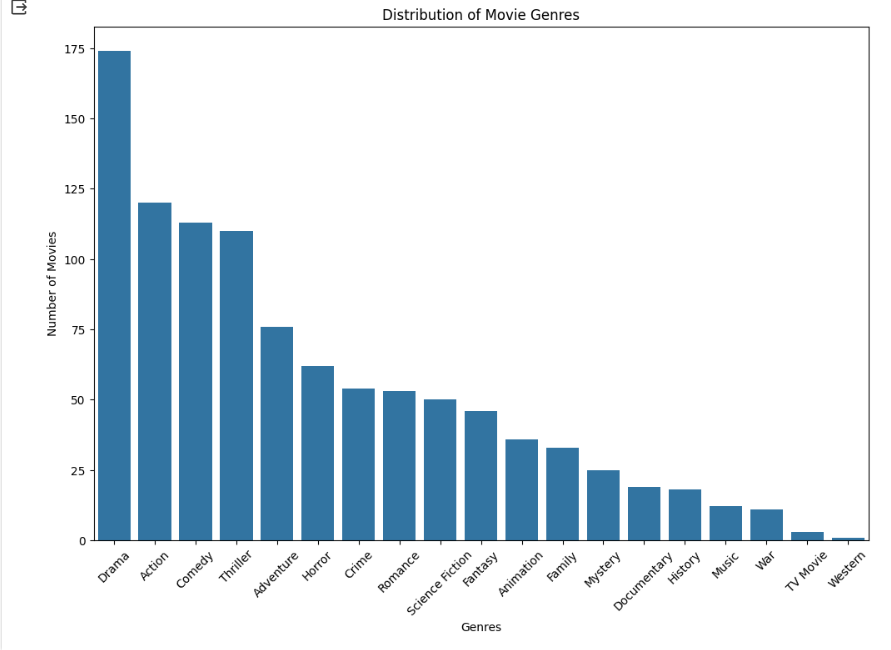
1. יצירת גרף פיזור של ערכי ספירת המילים המאוחסנים בעמודה “word\_count” של ה-Data Frame.

ציר ה-x מייצג את המדדים של ה-Data Frame, בעוד שציר ה-y מייצג את ערכי ספירת המילים.



1. הורדת ערכים חריגים ב - Data Frame



1. ייבוא ספריות נוספות הנחוצות והורדת משאבי NLTK הנדרשים עבור משימות עיבוד מקדים של טקסט כגון Token, stopword removal ו – lemmatization .
2.  הורדת ערכים חסרים
3. פיצלנו את ערכי העמודה הז'אנרים מערכים של רשימות לשורות מרובות.
4. הדפסת גרף שמציג את מספר הסרטים הכי נפוצים לפי ז'אנר.
5. A screenshot of a computer

   Description automatically generatedהגדרת פונקציה על עמודת “subtitles” שמורידה את המילים שמוגדרות במאגר, וקיצור המילים לפי שורשים.

A word cloud with colorful text

Description automatically generated

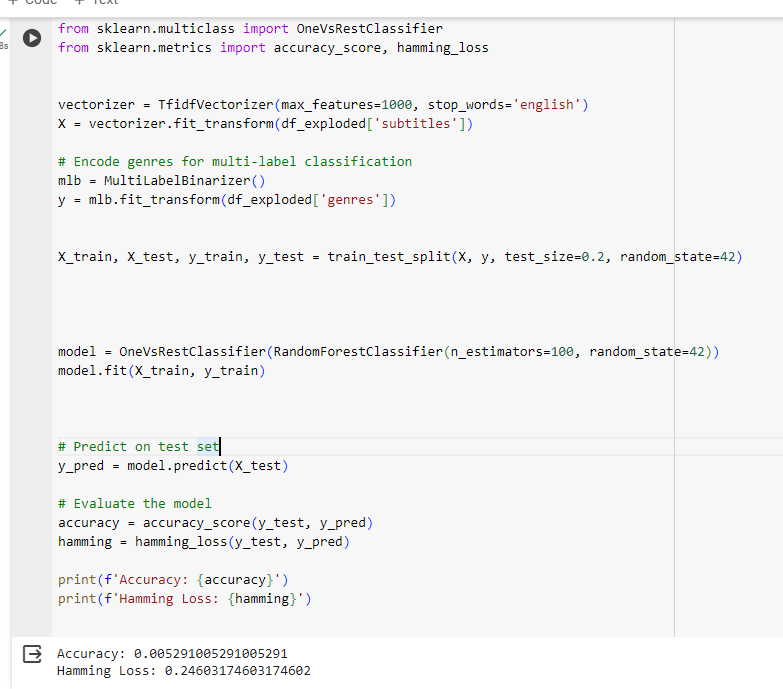
1. יצרת ענני מילים לכל זאנר:

A close up of words

Description automatically generated

A close up of words

Description automatically generated

1. 

נסינו לפתח מודל אשר יחזה את הדיוק בין חיזוי ז'אנר סרט על פי מילים, אבל לא הצלחנו להגיע למודל עם ציון גבוה.

גם לאחר שנסינו לצמצמם את ה"עולם" של המילים.

מסקנתי שקשה מאוד לחזות ז'אנר על פי מילים כי כל מילה בעולם יכולה לחזור באין סוף סיטואציות וז'אנרים בעולם.