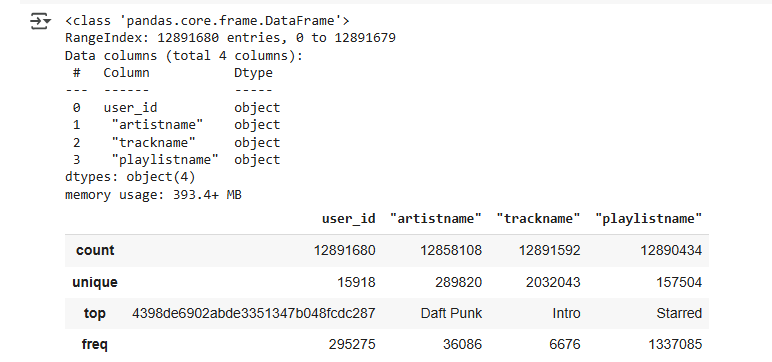


**ניתוח EDA במילים שלי:**

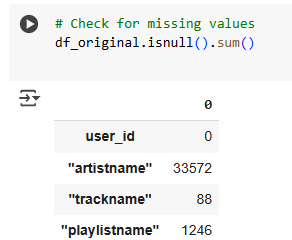
**הדאטה הגולמי ניראה כך-**



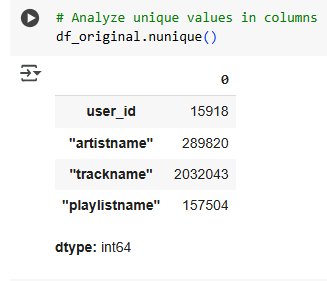
לאותו יוזר ישנן מספר שורות עם הבחירות שלו לגבי שיר שאהב , מי האמן שיצר אותו ומה הפלייליסט אליו משוייך השיר. בסה"כ 4 עמודות/פיצ'רים מתוכן עמודה אחת היא קוד היוזר.



הדאטה המקורי מכיל מעל 12 מיליון שורות ו 4 עמודות סטרינגיות בלבד.

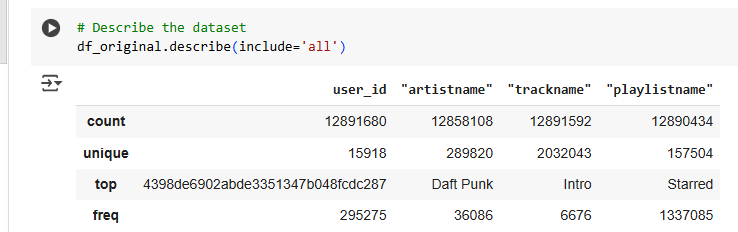


בדאטה היו 33572 שורות חסרות של אמנים, 88 שורות חסרות של שמות שירים ו 1246 שורות חסרות של שמות פלייליסטים.



הדאטה כלל בסה"כ 15918 יוזרים יוניקים , 289820 שמות אמנים, כ 2 מיליון שמות שירים (יותר שמות שירים מאשר אמנים, שזה בסה"כ הגיוני וזה אומר שקל יותר לאפיין את הדאטה לפי פיצ'ר האמנים כי יש להם יותר ייצוג בדאטה מאשר שמות השירים למשל).

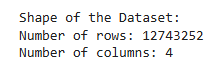
בדאטה יש בסה"כ 157504 שמות פלייליסטים.



ניתן לראות כאן ש ה top מראה שהפלייליסט הכי גדול הוא starred.

**שלב ניקוי הדאטה מכיל את השלבים הבאים:**

1. ניקוי ערכים חסרים df\_original.dropna()
2. המרת כל הסטרינגים מאותיות גדולות לאותיות קטנות ובכך לייצר אחידות בסטרינגים. col.str.lower()
3. הסרת כפילויות drop\_duplicates()
4. נורמליזציה/סטנדרטיזציה ע"י הסרת רווחים מיותרים, תווים נדירים או בעיות אחרות שקיימות בפורמט הטקסט.str.strip()
5. פילטריזציה של סטרינגים קצרים מדי , כי בסוף אנחנו רוצים שאורך הטוקנים שנייצר יהיה יחסית מאוזן וברור.

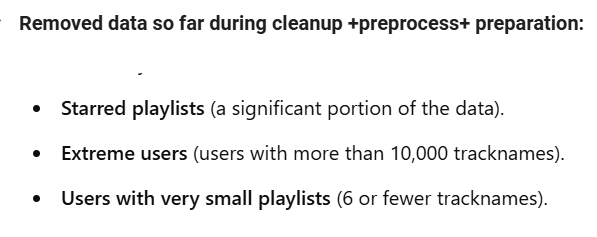


לאחר ניקוי הדאטה **הראשוני (בלבד)** רואים שמספר השורות הכללי בדאטה ירד בפחות מ 1%, כלומר רוב הדאטה העיקרי נשמר לאחר הניקיון.

**שלבי ניקיון נוספים שבאים בהמשך המחברת ולאחר אנליזות ומחקר מעמיק:**

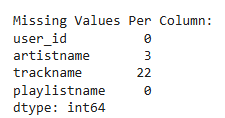
1. יש שלבי ניקוי (כמו הסרה של פלייליסט starred והסרה של יוזרים שהם משתמשים "כבדים" שמקבלים ייצוג בלתי סביר, או כאלה שבחרו רק בפלייליסט הקיצוני starred). בסוף מגיעים לבערך 10 מיליון שורות דאטה.
2. כמו כן יש בהמשך גם ניקוי והסרה של שירים עם סטרינג ריק ("רווח"), היו כ 186 כאלה. וניקוי שירים שהם NAN.
3. יש השלמה של 3 אמנים NAN עם placeholder ל 'unknown artist'.

**לסיכום לגבי הניקיון המלא לפני איזון משקלים-**

****

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ניתוח דאטה לאחר ניקיון ראשוני בלבד:**

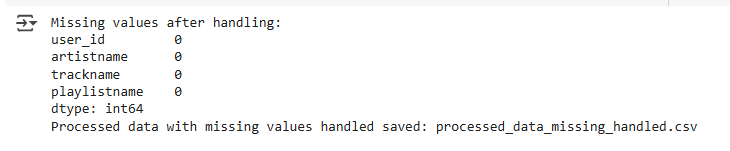


בדאטה הנקי ראינו שיש עדיין מספר ערכים חסרים NAN שיש לטפל בהם וגם הם טופלו:

בשמות האמנים הוחלט לעשות fillna לשים placeholder בשם 'unknown artist' -

את השירים ללא שמות הוחלט להסיר dropna.





לאחר מכן ראינו שהדאטה סוף סוף נקי ואפשר להמשיך ולהתמודד עם bias ו outlier שקיימים ב skewed data.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**באנליזה לגבי הפלייליסטים - (סטטיסטיקות)**

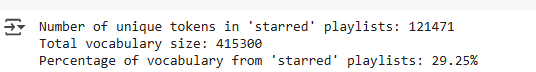
1. לאחר ניקוי הדאטה נותרו 137880 פלייליסטים יוניקים.
2. בממוצע מספר השירים בפלייליסט עומד על 92 אך זה כרגע לא מאוזן כי יש פלייליסט אחד שמושך את הנתון הזה למעלה בצורה קיצונית וזה ה starred שמכיל הרבה מאוד שירים.
3. סטיית התקן גדולה מאוד 3642 וזה אומר שהשונות בין הפלייליסטים גדולה מדי ואין איזון (יש outliers).
4. הפלייליסט הכי קטן מכיל 1 שיר אחד (אם מקודם ניקינו שירים עם שמות קצרים מדי אז כעת מתגלה מקרה קיצון נוסף של שיר בודד בפלייליסט וזה למעשה כלל לא פלייליסט בהגדרה).
5. ל 25% מהפלייליסטים יש פחות מ 11 שירים.
6. רוב הפלייליסטים (החציון בסטטיסטיקה) מכילים כ 21 שירים. מה שאומר ש50% מהפלייליסטים מכילים 21 שירים או פחות.
7. 75% מהפלייליסטים מכילים 52 שירים או פחות.
8. הפלייליסט הכי גדול - starred - מכיל **1,330,529 שירים !! BIG OUTLIER**

למעשה הפלייליסט הזה הוא אגרגטיבי ומכיל שירים מועדפים של כלל היוזרים (כך על פי סקירה קצרה באינטרנט).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

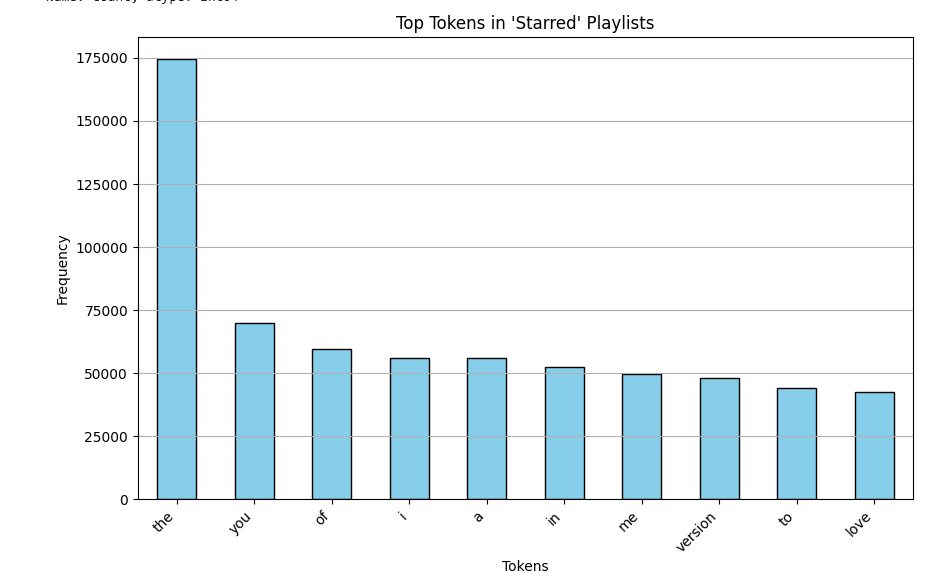
**אנליזה לגבי פלייליסט קיצוני starred (ניתוח באמצעות מילים, wordcloud):**

1. מספר המילים היוניקים בו הוא **417336**
2. אם נסתכל על ה"מילון" הכללי של הדאטה, אז נמצא כי 29.25% מהווים מילים רק מתוך פלייליסט starred בלבד, כלומר זה ייחודי רק ל starred !! זה סימן לרעש בדאטה.

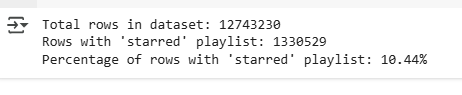


1. מתוך 121471 טוקנים (מילים) ייחודיים ל starred , יש 45.7% טוקנים שמופיעים רק פעם אחת (נדירים). עוד הוכחה כי התרומה של starred לרעש הכללי בדאטה היא גבוהה מאוד.
2. אם נסתכל על מודל WORD2VEC המודל לומד על קשרים בין טוקנים ולכן כמות גבוהה של טוקנים נדירים למעשה תוריד את איכות ה embeddings במודל והאיכות למידה שלו תהיה נמוכה. כלומר הפלייליסט הזה מהווה רעש עבור מודל מסוג זה.
3. ה bias שנוצר בעקבות כמות טוקנים נדירים גבוהה שמגיעה מ starred עלול ליצור דיפרופורציה בלמידה של המודל את הדאטה, באופן כזה שכביכול starred יקבל יותר ייצוג ממה שאמור לקבל בפועל.
4. אחוז הטוקנים השכיחים (הלא-נדירים) מ starred היא 98.90% מתוך סך השירים ב"מילון" הכללי של הדאטה, כלומר הפלייליסט מכיל כמעט את כל מה ששאר הפלייליסטים ביחד מכילים ולכן התרומה שלו בתור פלייליסט לכשעצמו היא לא ייחודית (לא מצביעה על העדפת יוזר ספציפית) ולכן אין בעיה להסיר את הפלייליסט הזה ולהיפטר מהרעש.





1. את הפלייליסט starred ניתן לראות בדאטה מופיע ב 10.44% מהשורות (על אף גודלו הייחודי).



1. ומסעיף 7 כשמעמיקים אז רואים שהפלייליסט starred למעשה צויין על ידי 587 יוזרים בלבד !! כלומר כ 3% מהיוזרים בסה"כ.



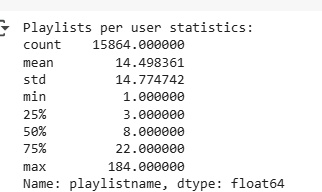
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

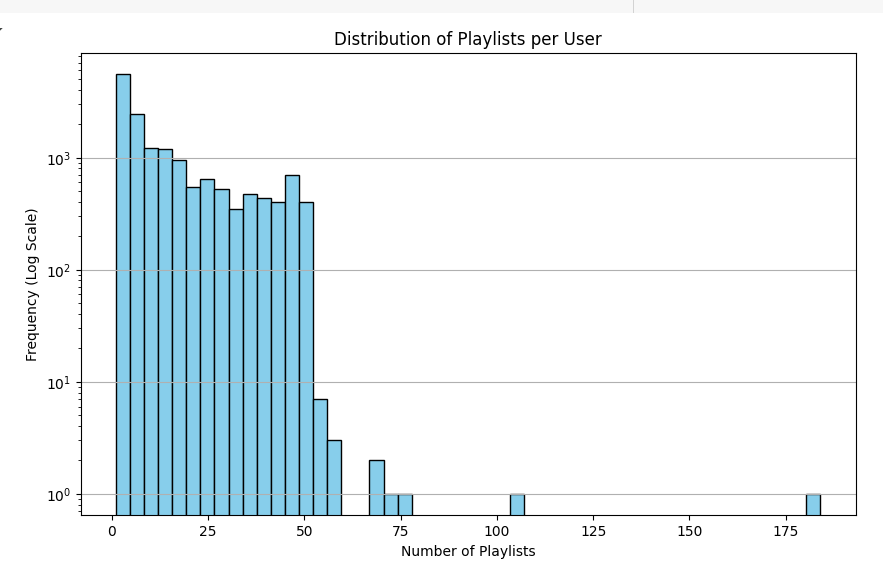
**אנליזה לגבי שמות שירים:**

1. השיר הארוך ביותר מכיל 51 מילים
2. 25% מהשירים מכילים מילה אחת
3. 50% מהשירים מכילים 3 מילים
4. 75% מהשירים מכילים 4 מילים
5. ברור ששירים המכילים מספר גבוה של מילים זה outlier נוסף שדורש טיפול. אבל אין הרבה כאלה. לרוב השירים יש בין 1-5 מילים. חשוב להדגיש כי כל outlier מהווה רעש בדאטה.

אנליזה לגבי היוזרים:

1. סטטיסטיקות -



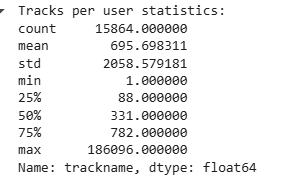


אפשר לראות שיש 15864 יוזרים יוניקים, הסטיית תקן 14 כלומר אין הטיה מסויימת ליוזר מסויים אבל עדיין ניבדוק מקרה קצה אפשרי שייתכן וניתן לטפל.

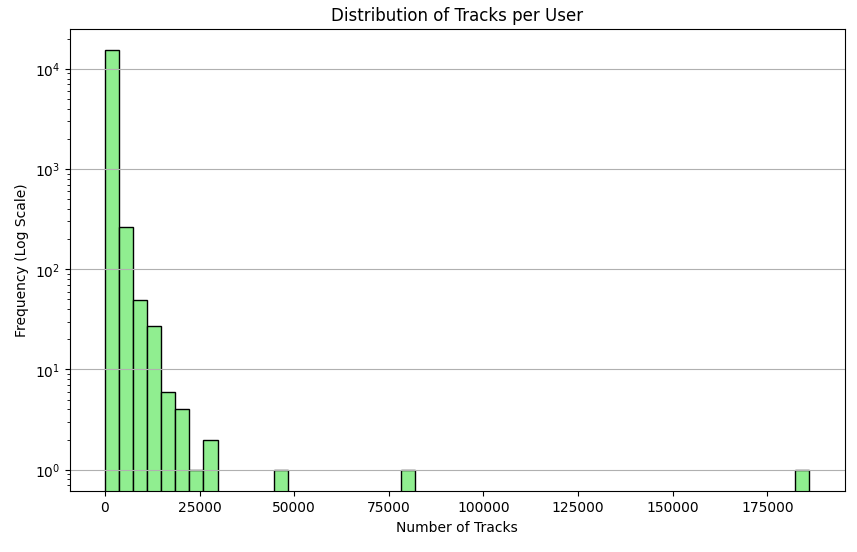
ניתן להסיק מהגרף שככל שמספר הפלייליסטים גדל כך מספר היוזרים קטן, כלומר לרוב היוזרים יש בין 0 ל 50 פלייליסטים, כאשר 13.88% מהיוזרים בחרו בפלייליסט אחד בודד (ו 587 יוזרים מתוכם בחרו אך ורק בפלייליסט בודד אחד שהוא starred).



1. אם נסתכל על מספר השירים ליוזר אז נגלה שונות יותר משמעותית (וזה הודות ל starred שמכיל מעל מיליון שירים).



1. מקודם ראינו ש 587 יוזרים בחרו רק בפלייליסט אחד שהוא starred . כעת ניתן לראות שמספר היוזרים שבחרו במספר פלייליסטים וביניהם גם starred, עומד על 5019 מתוך 15864 יוזרים יוניקים. כלומר 4432 יוזרים מראים את ההעדפות שלהם גם באמצעות פלייליסטים אחרים (וכבר ראינו שהמילון הכללי משותף בכ 99% בין starred ליתר הפלייליסטים).

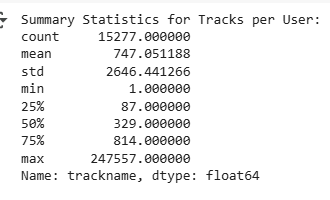


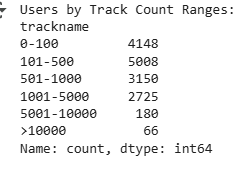
ככל שיש יותר שירים יוניקים כך יש פחות יוזרים, כלומר לרוב היוזרים יש בין 0-25000 שירים יוניקים, ולכן גם בכמות השירים צריך לטפל ולאזן. זה גם מהווה רעש לכשעצמו.

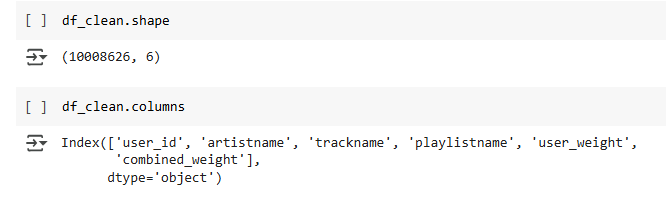
1. אפשר לראות שהסרה של פלייליסט starred לא מפחיתה בהרבה את כמות היוזרים שתורמים לדאטה המגוון:

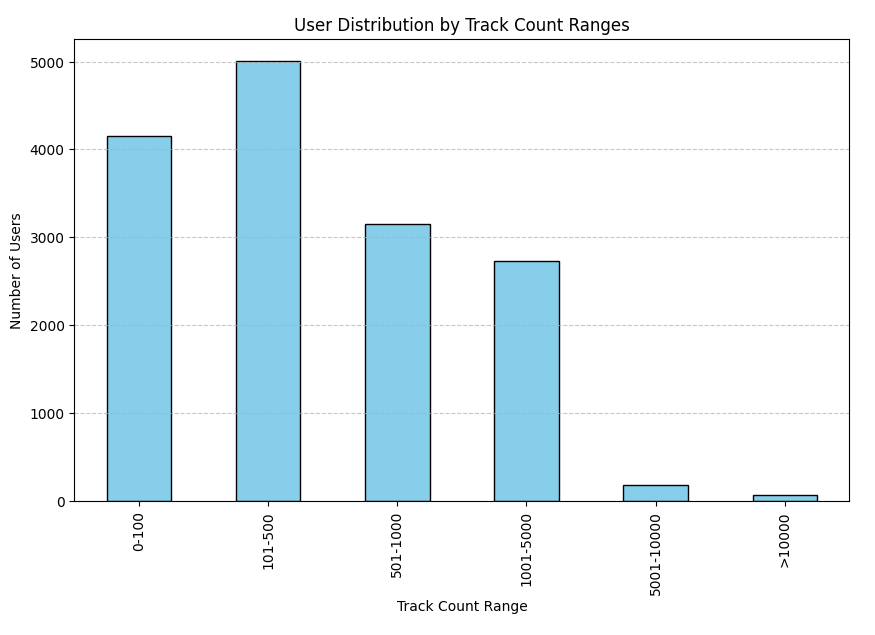


1. סטטיסטיקות של מספר שירים ליוזר לאחר הסרת פלייליסט starred:

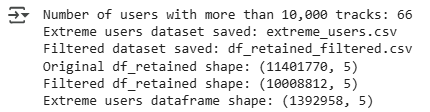






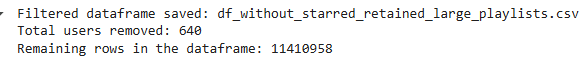


1. ניתן לראות בסטטיסטיקות שיש מקרה קיצון של 66 יוזרים להם יש המון שירים באופן בלתי סביר (יותר מ 10000) ולכן הייצוג שלהם לא מאוזן. מכיוון שמדובר ב0.42% מהיוזרים, מדובר במקרה קצה והוחלט להסירם (תחילה הוסרו יוזרים לפי סעיף 8 ולאחר מכן לפי סעיף 7). מדובר במשתמשים "כבדים" שמקבלים ייצוג בלתי סביר.



1. יש 640 יוזרים עם 1-6 שירים ומכיוון שהייצוג נמוך במצב כזה, התרומה שלהם לא גדולה ולכן הוחלט להסירם:





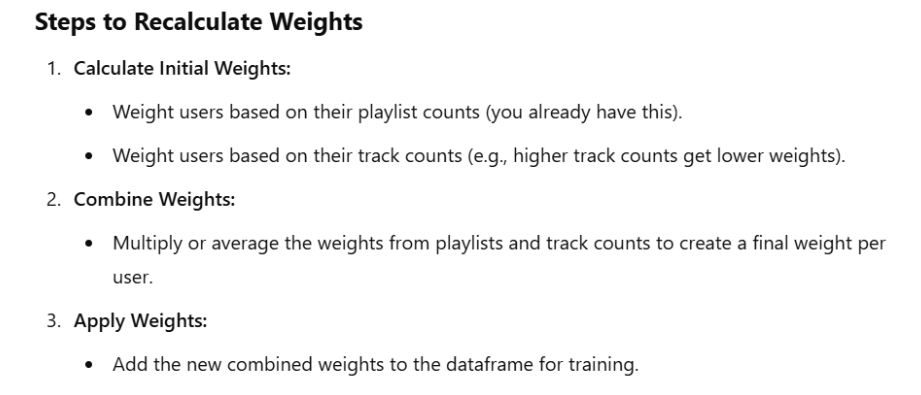
1. כעת נותרנו עם כ 10 מיליון שורות בדאטה ויוזרים עם 7 עד 9900 שירים.

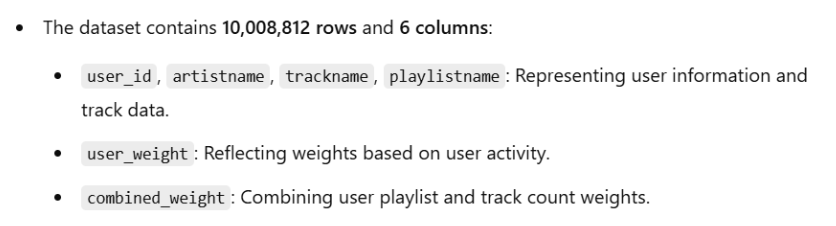


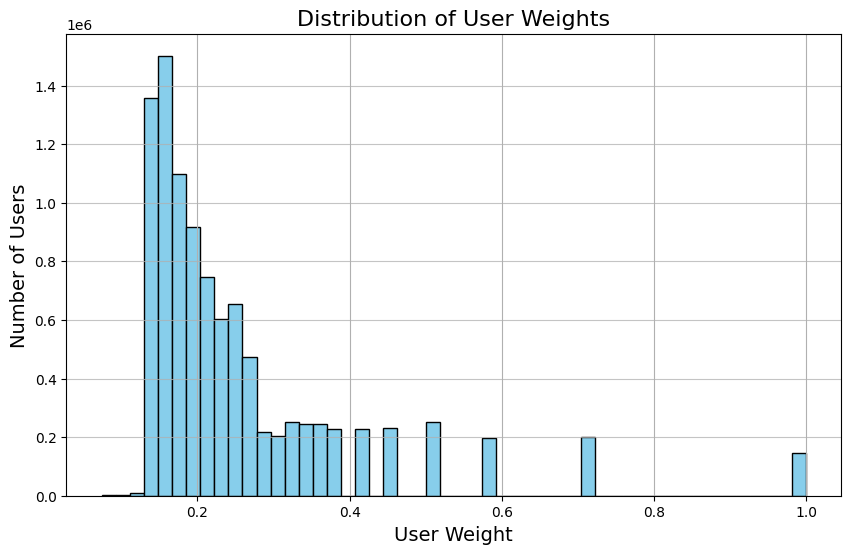
1. מכיוון ש word2vec רגיש לרעשים ולא יכול להתמודד עם רעש מאוד גדול אז צומצם הרעש ומקרי הקיצון ככל האפשר, תוך שמירה על דאטה שלם ומאוזן ככל האפשר. לקולאב יש משאבים מוגבלים בכל הקשור לעיבוד דאטה (מבחינת כוח חישובי CPU) , כמות שורות 10 מיליון היא על הקצה איתו יכול קולאב להתמודד ולכן צריך לעזור למודל ככל האפשר, בין אם זה איזון משקלים בדאטה (כי עדיין קיים חוסר איזון קל) על ידי יצירת עמודות נוספות לצורך זה, ובין אם מניפולציות של חלוקה לקבצי צ'אנקים קטנים איתם יוכל המודל להתמודד ביעילות ובמהירות יחסית.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**איזון דאטה על ידי משקלים:**

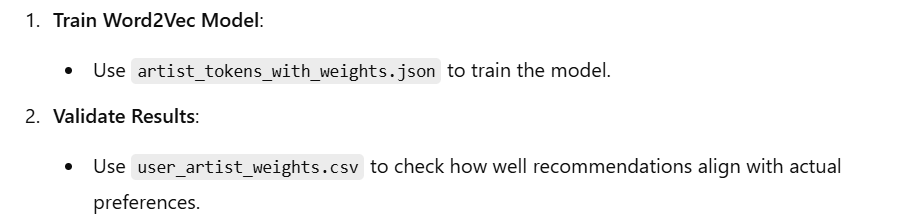






\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**תהליך יצירת הטוקנים:**



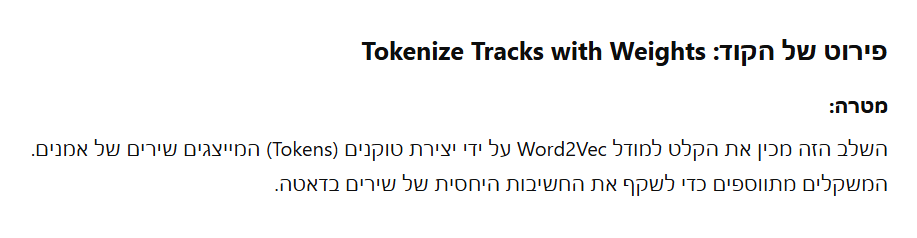
1. תחילה מייצרים מילון weighted\_user\_artist.dict תוך שימוש במשקלים שיצרנו לצורך איזון הדאטה ב2 עמודות נוספות שניקראות user\_weight ו combined\_weight: לדוגמה -

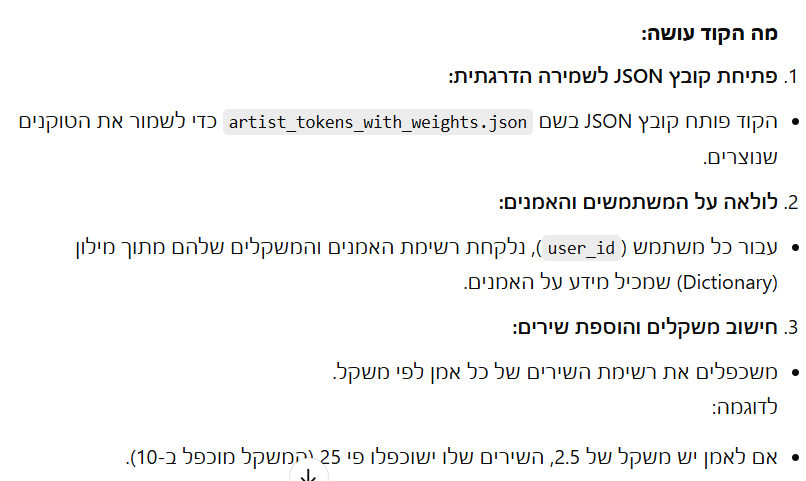
Example User (00055176fea33f6e027cd3302289378b): {'5 seconds of summer': {'tracks': ['disconnected', 'dont stop acoustic', 'english love affair', 'everything i didnt say', 'everything i didnt say live', 'if you dont know', 'just saying', 'long way home', 'long way home acoustic', 'the only reason'], 'weight': 5.4564354645876385}, 'abigail breslin': {'tracks': ['you suck'], 'weight': 0.5456435464587639}, 'against the current': {'tracks': ['another you another way', 'closer faster', 'gravity'], 'weight': 1.6369306393762917}, 'all time low': {'tracks': ['backseat serenade', 'break your little heart', 'dear maria count me in', 'jasey rae', 'merry christmas kiss my ass', 'somethings gotta give', 'the reckless and the brave', 'therapy'], 'weight': 4.365148371670111}, 'auryn': {'tracks': ['saturday im in love'], 'weight': 0.5456435464587639}, 'austin mahone': {'tracks': ['what about love'], 'weight': 0.5456435464587639}, 'avril lavigne': {'tracks': ['girlfriend', 'heres to never growing up'], 'weight': 1.0912870929175278}, 'bea miller': {'tracks': ['young blood'], 'weight': 0.5456435464587639}, 'becky g': {'tracks': ['shower'], 'weight': 0.5456435464587639}, 'blink182': {'tracks': ['after midnight', 'first date', 'ghost on the dance floor', 'happy holidays you bastard', 'i miss you'], 'weight': 2.7282177322938193}, 'bmike': {'tracks': ['baby dont cut'], 'weight': 0.5456435464587639}, 'bruno mars': {'tracks': ['treasure'], 'weight': 0.5456435464587639}, 'charli xcx': {'tracks': ['boom clap'], 'weight': 0.5456435464587639}, 'christina perri': {'tracks': ['human'], 'weight': 0.5456435464587639}, 'cimorelli': {'tracks': ['i got you', 'youre worth it'], 'weight': 1.0912870929175278}, 'demi lovato': {'tracks': ['give your heart a break', 'la la land', 'made in the usa', 'really dont care'], 'weight': 2.1825741858350556}, 'ed sheeran': {'tracks': ['all of the stars', 'dont', 'friends', 'photograph', 'tenerife sea', 'thinking out loud'], 'weight': 3.2738612787525834}, 'fall out boy': {'tracks': ['centuries', 'immortals from big hero 6', 'my songs know what you did in the dark light em up'], 'weight': 1.6369306393762917}, 'gerard way': {'tracks': ['brother'], 'weight': 0.5456435464587639}, 'green day': {'tracks': ['21 guns', 'american idiot', 'good riddance time of your life', 'kill the dj'], 'weight': 2.1825741858350556}, 'high school musical cast': {'tracks': ['breaking free'], 'weight': 0.5456435464587639}, 'imagine dragons': {'tracks': ['i bet my life'], 'weight': 0.5456435464587639}, 'jack jack': {'tracks': ['like that feat skate', 'tides'], 'weight': 1.0912870929175278}, 'jake miller': {'tracks': ['number one rule'], 'weight': 0.5456435464587639}, 'james arthur': {'tracks': ['impossible', 'youre nobody til somebody loves you'], 'weight': 1.0912870929175278}, 'jannik brunke': {'tracks': ['18', 'im a mess', 'shake it off', 'the hanging tree'], 'weight': 2.1825741858350556}, 'little mix': {'tracks': ['boy', 'change your life', 'little me', 'they just dont know you'], 'weight': 2.1825741858350556}, 'mark ronson': {'tracks': ['uptown funk'], 'weight': 0.5456435464587639}, 'maroon 5': {'tracks': ['animals', 'she will be loved acoustic'], 'weight': 1.0912870929175278}, 'meghan tonjes': {'tracks': ['blank space'], 'weight': 0.5456435464587639}, 'meghan trainor': {'tracks': ['all about that bass', 'lips are movin'], 'weight': 1.0912870929175278}, 'natalie imbruglia': {'tracks': ['torn'], 'weight': 0.5456435464587639}, 'nick jonas': {'tracks': ['jealous'], 'weight': 0.5456435464587639}, 'olly murs': {'tracks': ['wrapped up'], 'weight': 0.5456435464587639}, 'one direction': {'tracks': ['act my age', 'night changes', 'steal my girl', 'where do broken hearts go'], 'weight': 2.1825741858350556}, 'onerepublic': {'tracks': ['burning bridges', 'good life', 'i lived', 'something i need'], 'weight': 2.1825741858350556}, 'panic at the disco': {'tracks': ['northern downpour'], 'weight': 0.5456435464587639}, 'shawn mendes': {'tracks': ['a little too much', 'bring it back', 'imagination', 'life of the party', 'never be alone', 'one of those nights', 'show you', 'something big', 'stitches', 'the weight'], 'weight': 5.4564354645876385}, 'simple plan': {'tracks': ['gone too soon', 'perfect', 'welcome to my life', 'welcome to my life acoustic', 'you suck at love', 'your love is a lie single edit'], 'weight': 3.2738612787525834}, 'taylor swift': {'tracks': ['out of the woods'], 'weight': 0.5456435464587639}, 'the 1975': {'tracks': ['chocolate', 'girls', 'sex'], 'weight': 1.6369306393762917}, 'the cat empire': {'tracks': ['brighter than gold', 'still young'], 'weight': 1.0912870929175278}, 'the fray': {'tracks': ['how to save a life'], 'weight': 0.5456435464587639}, 'the vamps': {'tracks': ['high hopes', 'last night', 'move my way', 'on the floor', 'risk it all', 'wild heart'], 'weight': 3.2738612787525834}, 'the wanted': {'tracks': ['we own the night'], 'weight': 0.5456435464587639}}

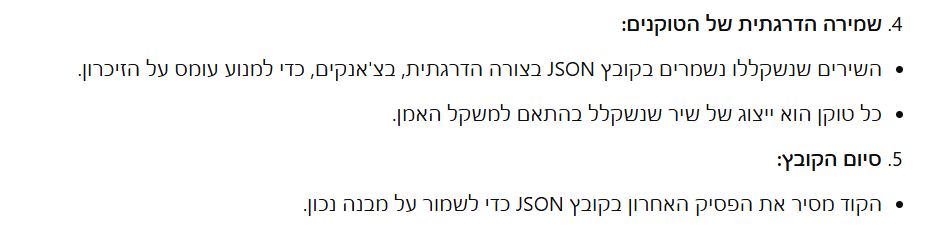
1. שלב יצירת הטוקנים ניראה כך:

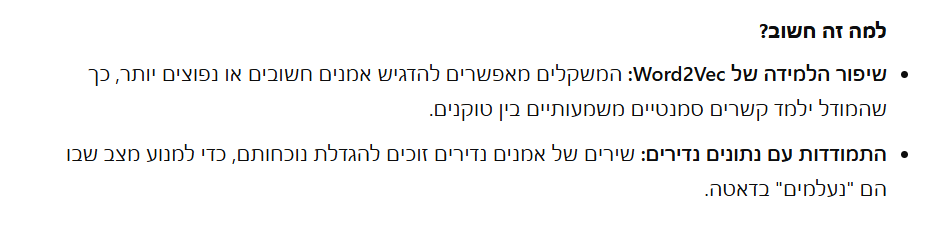


החלק שמוצג בתמונה מתאר שלב בתהליך יצירת הטוקנים למודל Word2Vec, תוך שימוש במשקלים על מנת להתאים את שכיחות השירים בדאטה בהתאם לחשיבותם.

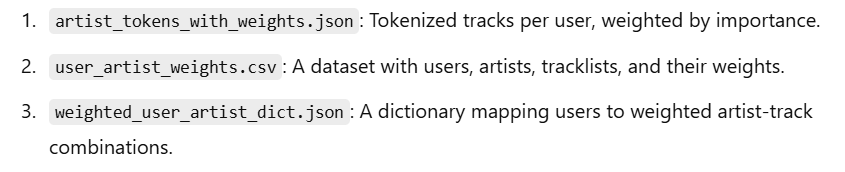


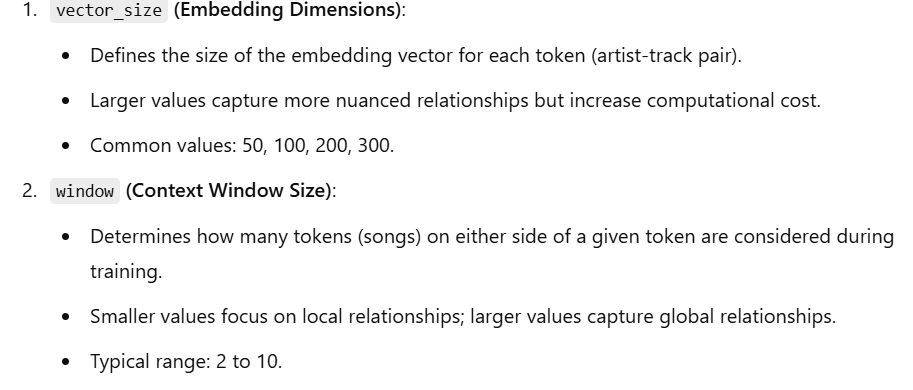


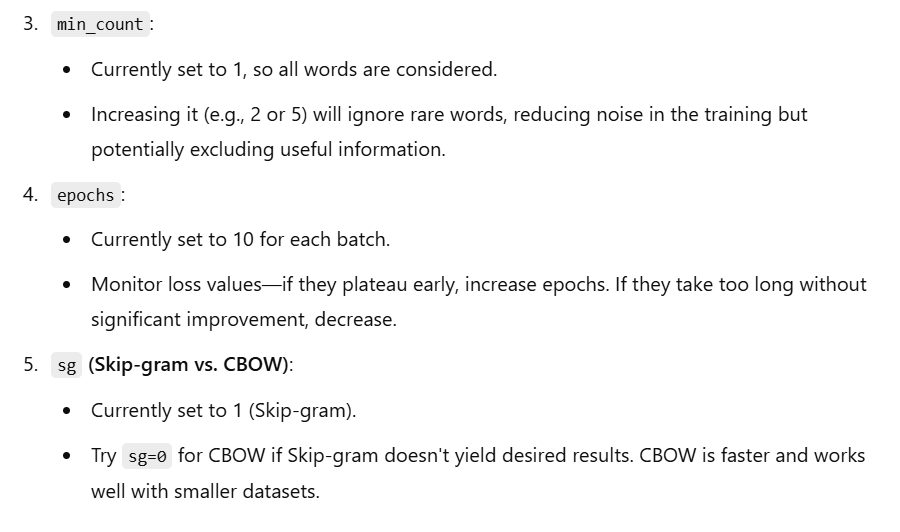




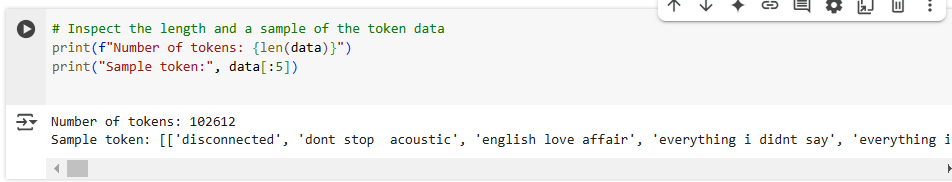








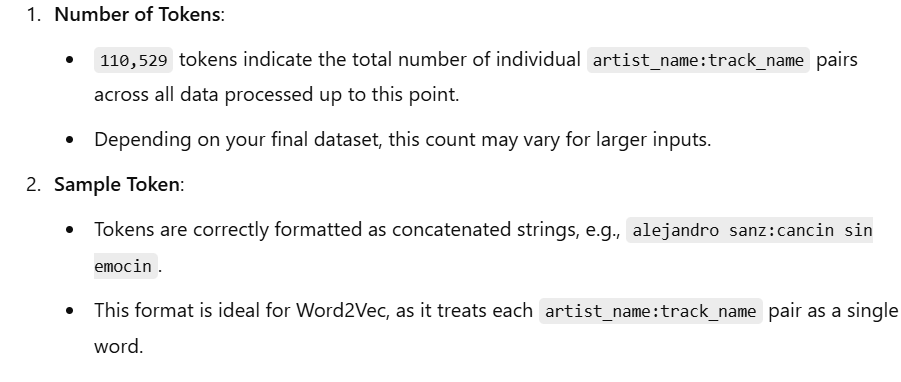
1. דוגמה לטוקנים שנוצרו: (סה"כ 102612 טוקנים) רואים כאן ליסטים של שמות שירים, אבל חסר האמנים ולכן יש לבצע מודיפיקציה.



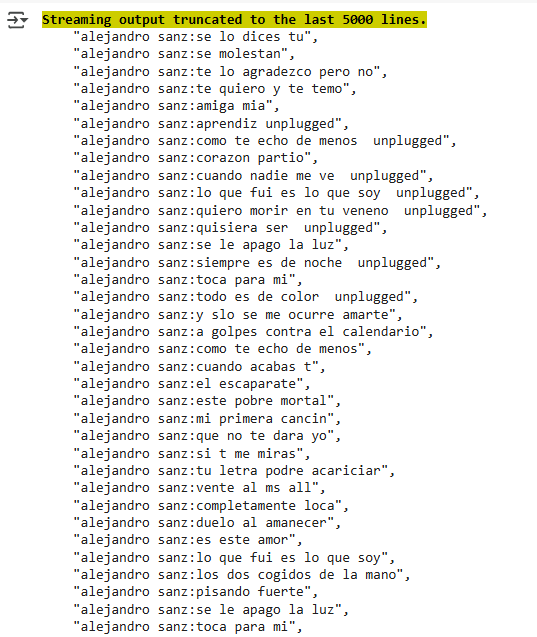
1. יצירת concatenated strings של אמן:שיר על מנת שהטוקנים יכילו את כל המידע הרלוונטי:



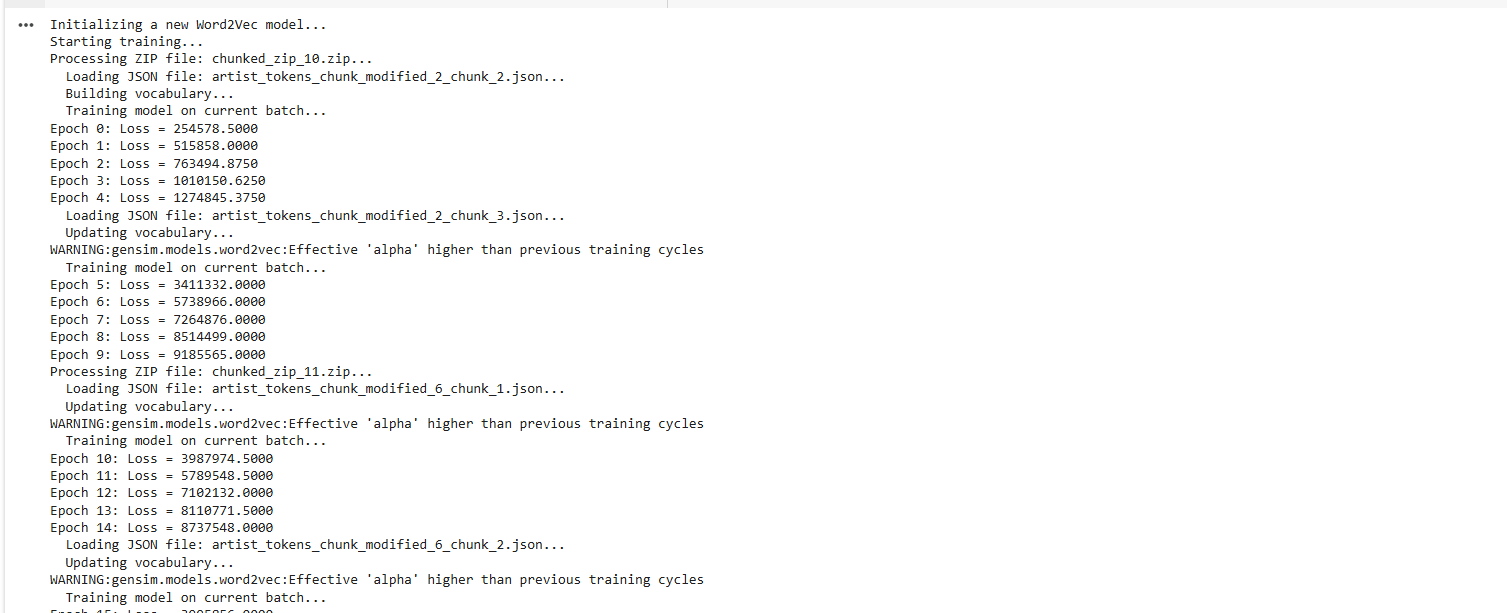
1. דוגמה לטוקנים הרלוונטים לאחר מודיפיקציה:

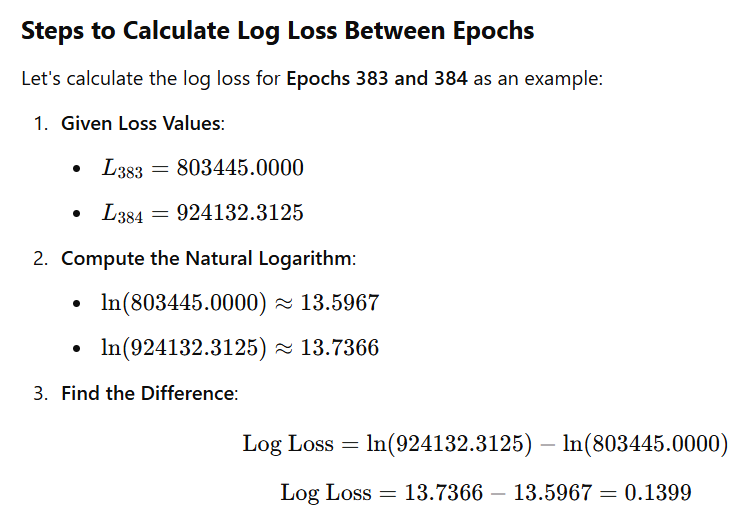


1. כדי להתמודד עם הכוח החישובי והמגבלות באופן כללי של word2vec בסביבת קולאב, חילקתי את הטוקנים ל 101 קבצי זיפ של ג'ייסון (כל קובץ 100MB) וכל 5 קבצים בתוך קובץ זיפ של 500MB (בהתחלה חילקתי ל 500MB אבל המודל לא רץ מהר ולכן היה צורך להקטין את גודל הצ'אנקים). סה"כ 9 ג'יגה !!! קשה להעלות דאטה כזה מנופח לקולאב ישירות לדירקטורי ואפילו ספרית gdown יש לה מגבלות כך שאי אפשר לייצר לינק ישיר לקבצים. ולכן נאלצתי להשתמש ב mount drive ישירות מהדרייב האישי ומשם למשוך את הקבצים.
2. חילקתי את הקבצים ל SEED , למעשה עשיתי ספליט אבל עם seed שמור בדרייב כדי שכל פעם ירוץ אותו seed של train במהלך האימון , כאשר 20% מהטוקנים נישמרו לצורך test או ולידציה בהמשך.
3. זו דוגמה לאחד מקבצי המיני ג'ייסון (100MB) שמכיל טוקנים לאימון המודל:



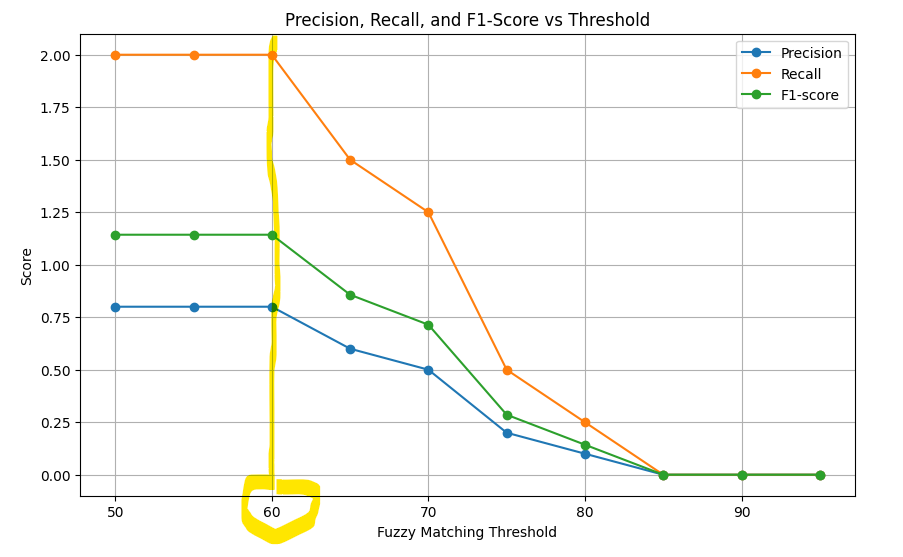
**שלב האימון מודל WORD2VEC -** דוגמה: כל פעם נפתח batch והמודל מתאמן עליו עד להתכנסות סך הקבצים לאימון.



לצורך האימון משתמשים בוריאציה של softmax והיא negative sampling או hirarcheal sampling

**שלב הולידציה (קבעתי שטסט יהיה על 1000 יוזרים בלבד אחרת זמן הרצה לטסט גדול מדי כי הוא רץ ומשווה וקטורים וזה דורש כוח חישובי גדול מדי וזמן הרצה של לפחות 15 שעות על כל קובץ הטסט):**





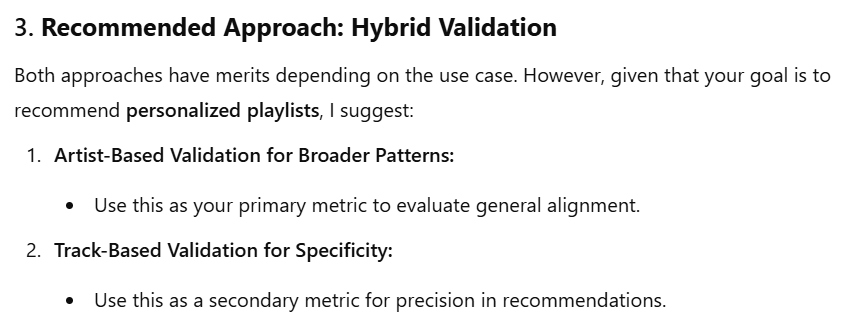
ככל שה threshold גבוה יותר אז רמת הדימיון שנלקחת היא יותר גבוהה אבל אז המטריקות ייצאו לא טוב אם החלק השני ב אמן:שיר (אין עדיין אימון על בסיס השירים) לא זוהה ב 100% ולכן מחפשים נקודה בה המטריקות יחסית מאוזנות כלומר ה f1 score הוא הכי גבוה וזו הנקודת מרפק וכך בחרתי threshold=60 שזה מבטיח שנקבל דימיון חלקי עד כמה שניתן ועדיין המטריקות יצליחו להיות טובות כאשר הן מבוססות אמן.

ניתן לראות בקוד האימון ש fuzzy נקלח עם threshold=60 במהלך האימון כי זו נקודת המרפק בה ניתן לשמור על איזון בין precision ל recall ולהבטיח שיהיה לנו כמה שפחות false positive (המלצות שממליצים ליוזר אבל בפועל הוא לא אוהב). כדי לוודא מחקרית בהמשך אם fuzzy הוא התשובה, אנחנו נצטרך לקבל פידבק ישיר מיוזרים על ההמלצות שהמודל מבצע. אבל זו דרך התמודדות קלה יחסית שמאפשרת לקבל מטריקות מדוייקות לשלב הראשון באימון על הקשרים בדאטה לפי האמנים והעדפות היוזרים. דרך לוודא שההמלצות הן לא false positive לפחות כשלב טרום מחקרי, הוא לראות שאכן הקוסינוס דימיון קרוב ל 1 ואכן כך יצא בתוצאה הסופית כשנלקח אינפוט של יוזר ויוצא לו אאוטפוט של מערכת המלצות מבוססת אמן.

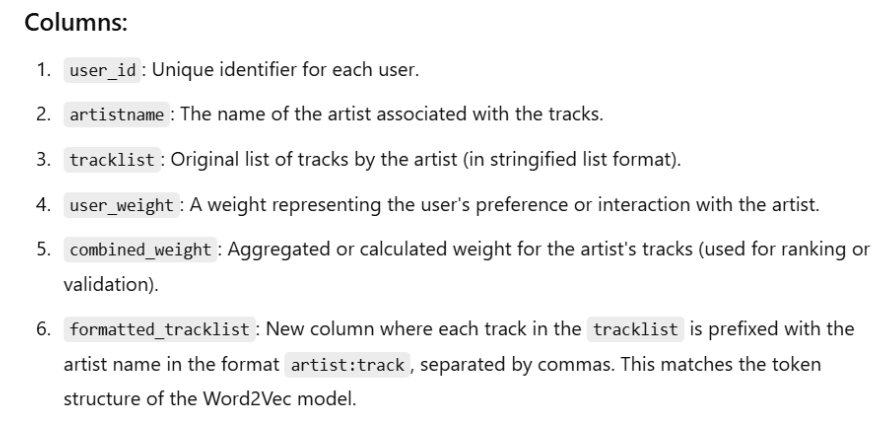
**כך ניראה קובץ ground truth לולידציה:**

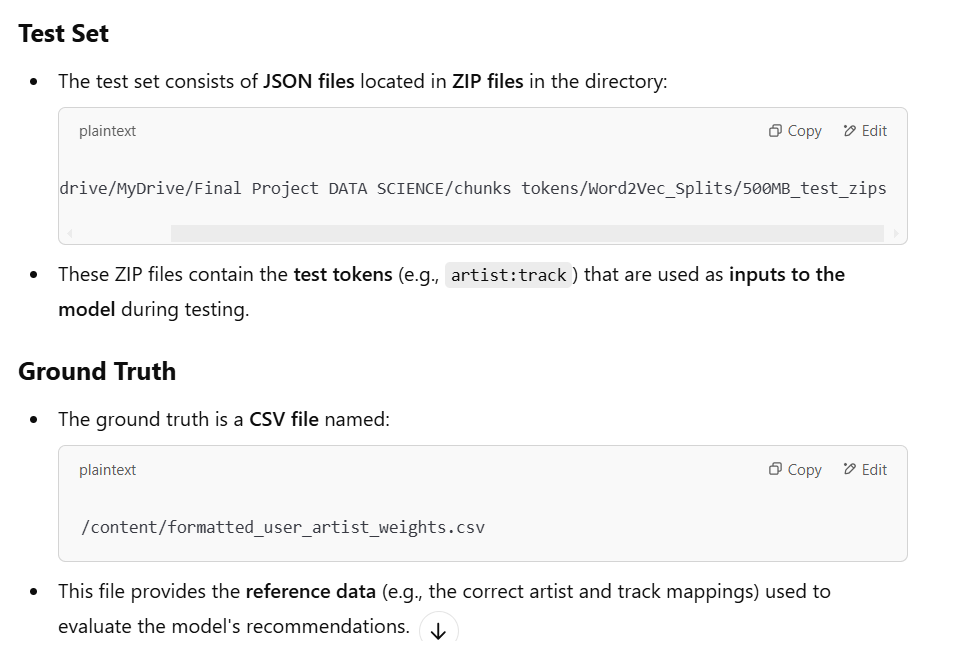


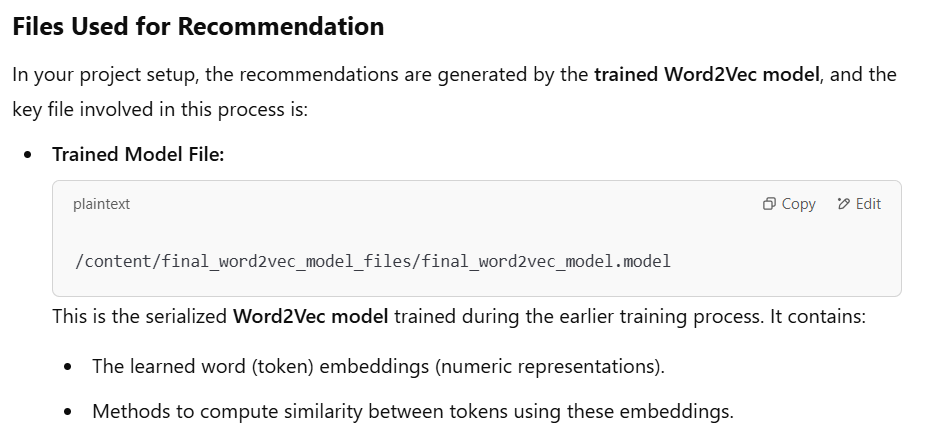
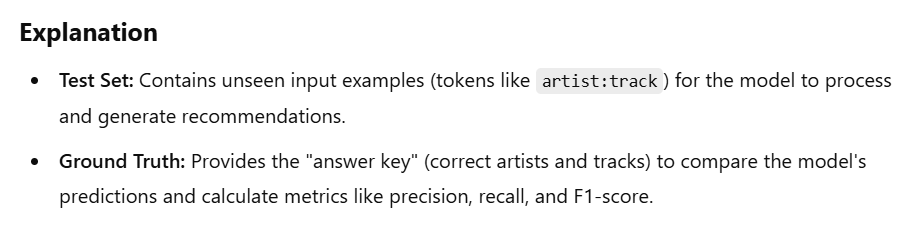
**השלב הסופי של מערכת המלצות ייראה כך- מערכת היברידית של קשרים בין אמנים וקשרים בין שירים:**

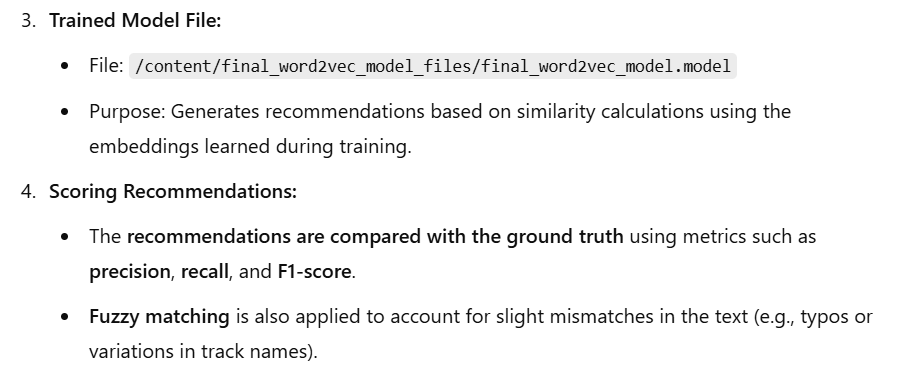
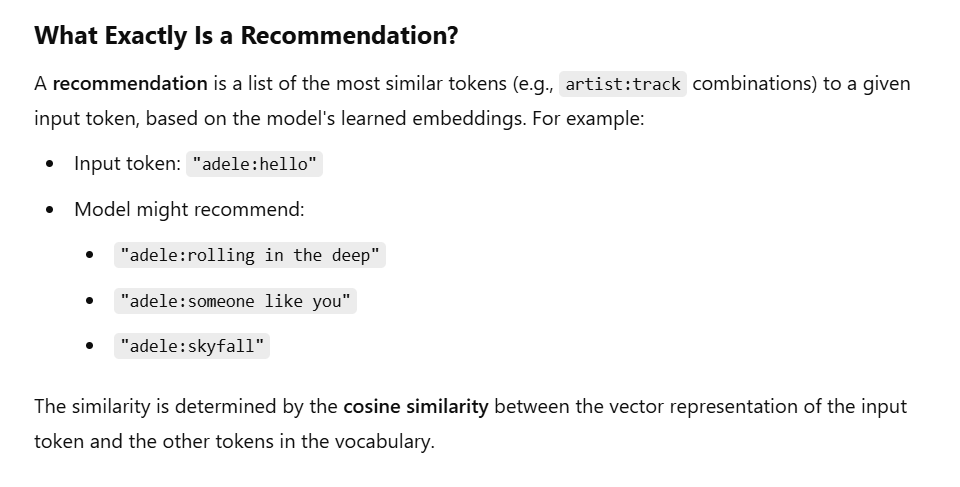


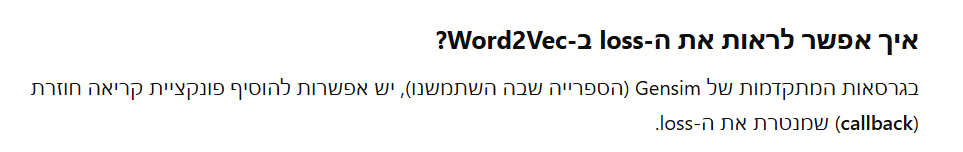
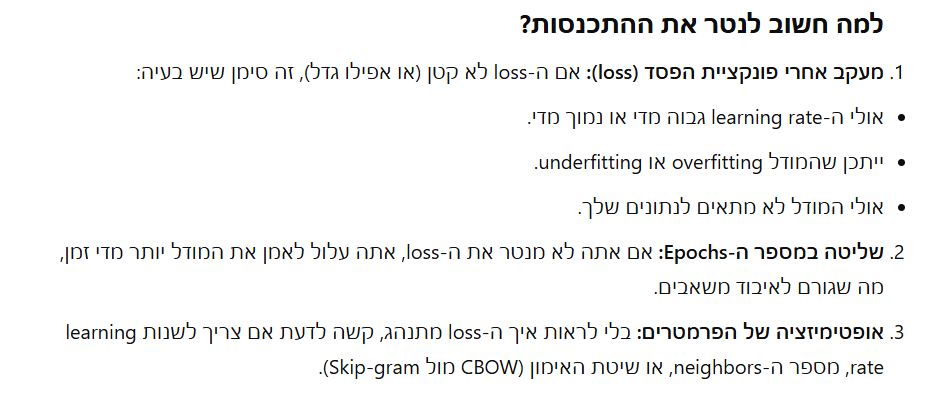
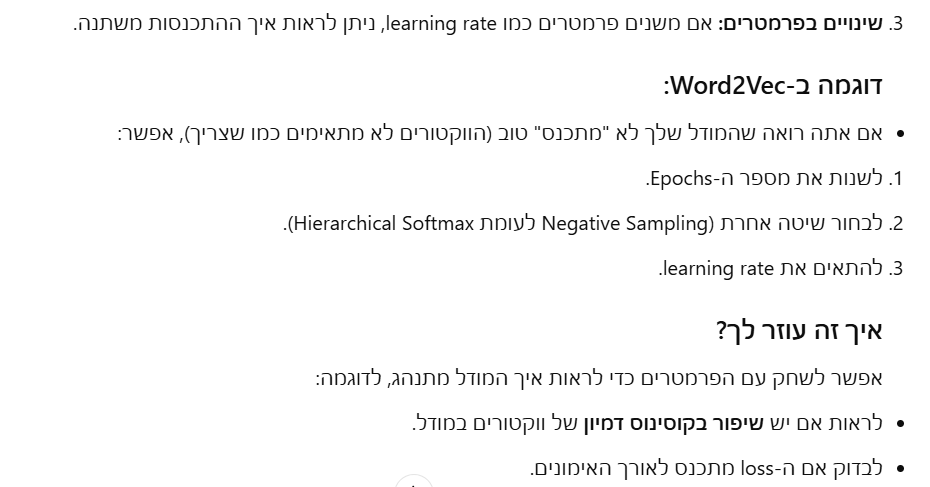
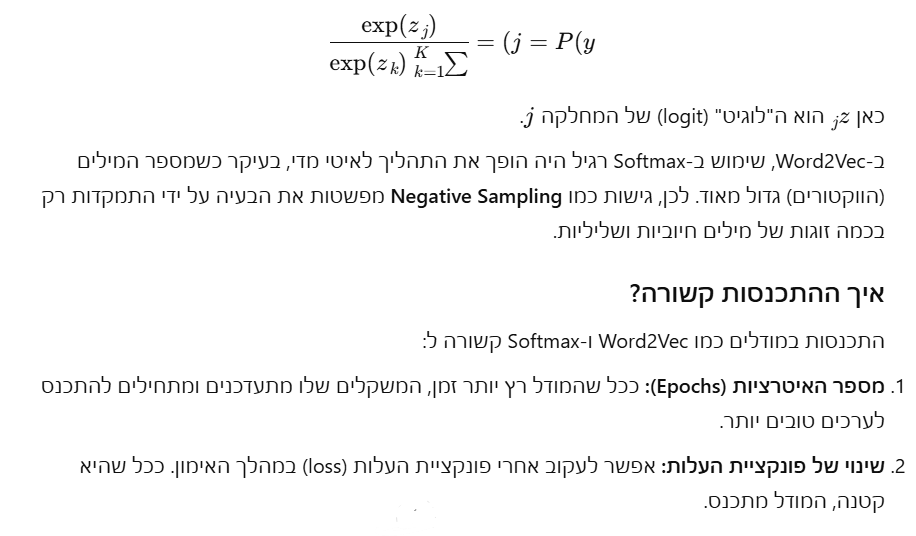
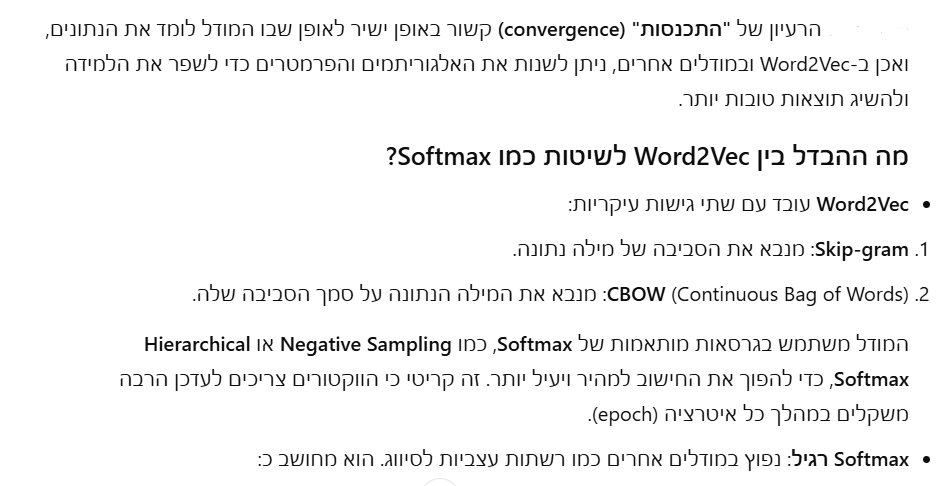
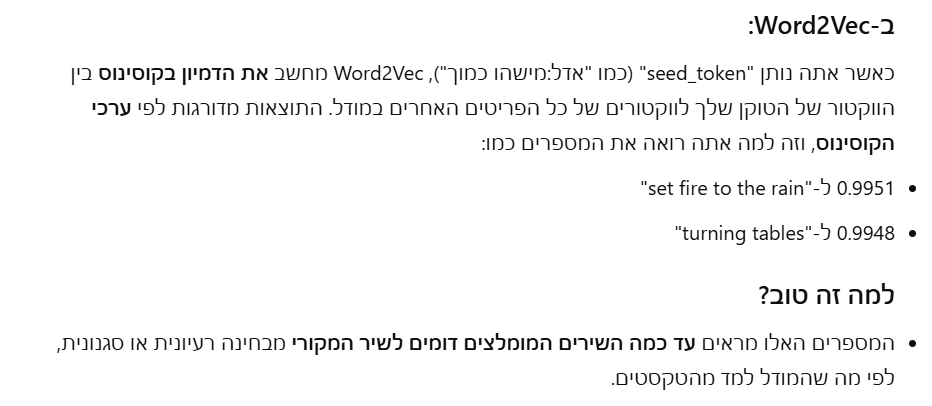
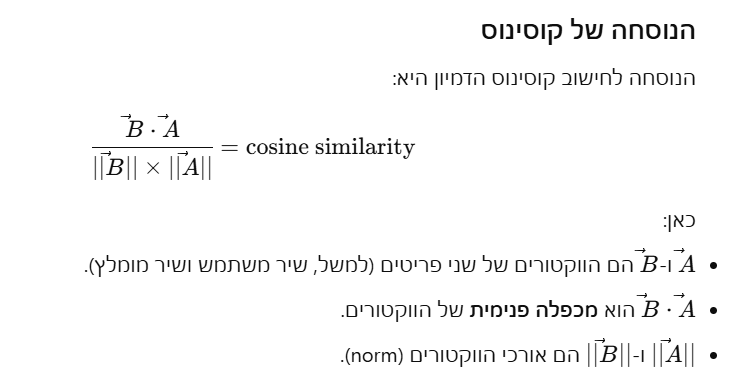
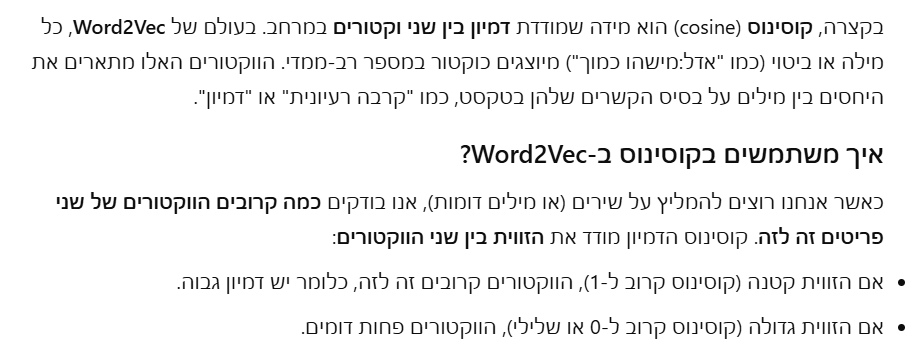
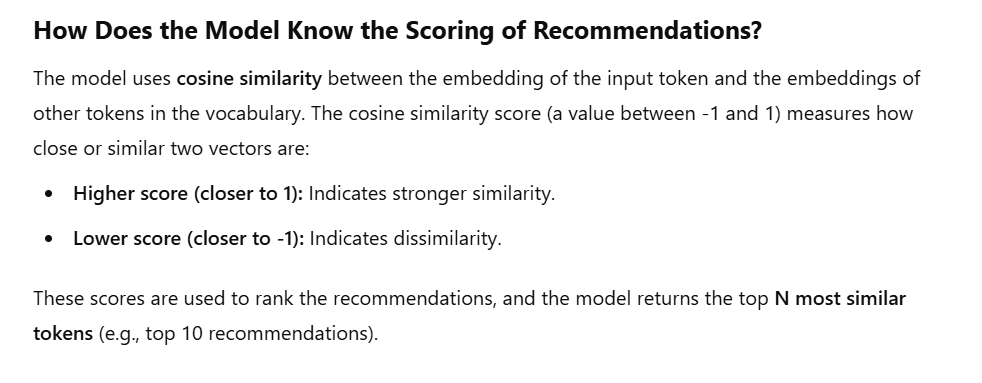
**כדי שהמודל יוכל לתקשר עם ה ground truth היה צורך להוסיף עמודה נוספת שממירה את עמודות אמן ושיר לעמודה משותפת אחת:**



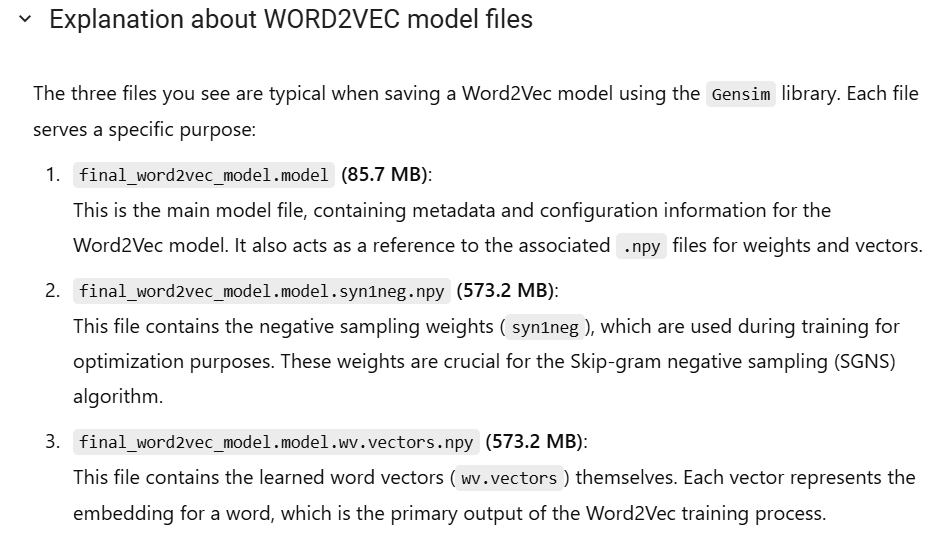


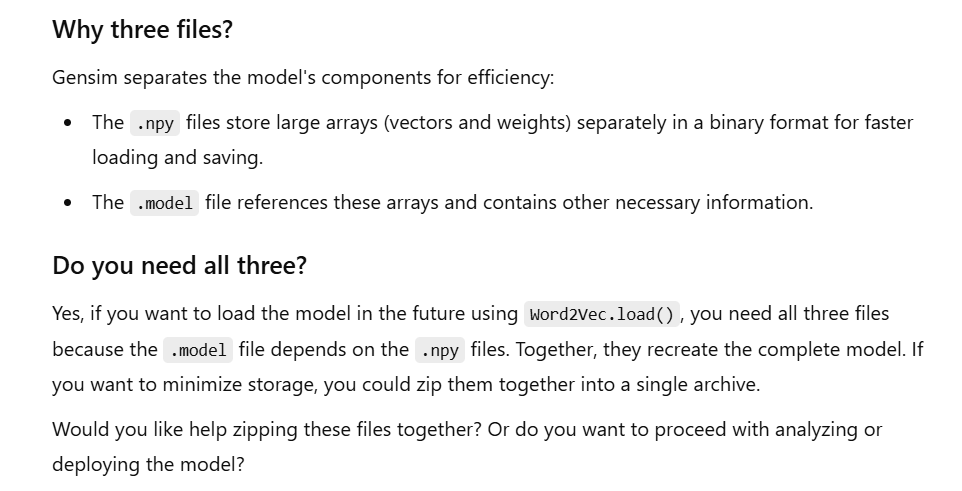




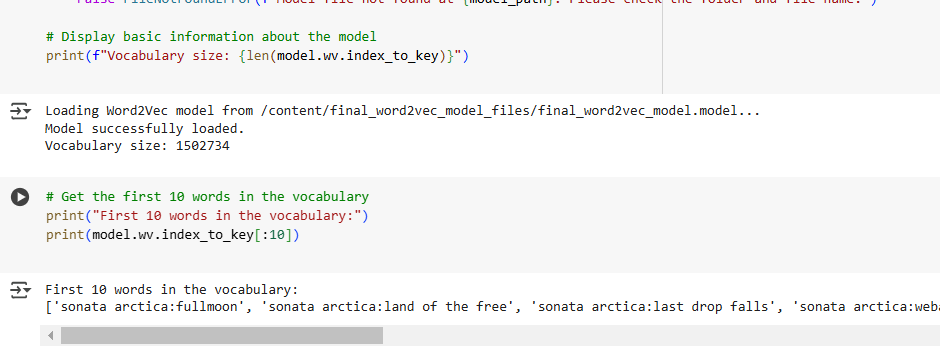


**הסבר לגבי קבצי המודל (נוצרים 3 כאלה):**





**כך המודל ניראה מבפנים:**



**שלב האינפוט בו היוזר נותן אינפוט ומקבל מהמודל אאוטפוט של המלצות, זה השלב אליו רצינו להגיע (לפחות לאחר אימון לגבי קשרים בין אמנים):**

