# Word2Vec Professional Narration

## שקף ראשון: נתונים גולמיים (2–3 דקות)

הנתונים הגולמיים שלנו נאספו מתוך Kaggle והכילו שמות שירים, אמנים, וכותרות של רשימות השמעה. עם זאת, ההתפלגות לא הייתה אחידה – רשימות מסוימות כללו שירים נדירים, בעוד אחרות היו מבוססות על להיטים פופולריים. כדי להתמודד עם דלילות והטיות, ביצענו תהליך עיבוד שכלל:  
1. ניקוי נתונים – הסרת כפילויות ושורות פגומות.  
2. נרמול טקסט – המרה לאותיות קטנות.  
3. חילוץ טוקנים – פירוק שמות שירים למילים בעזרת `split`.

## שקף שלישי: מודל Word2Vec (8–9 דקות)

מערכת ההמלצות שלנו מבוססת על Word2Vec. המודל משתמש באלגוריתם Skip-Gram, שמטרתו לנבא אילו מילים (או טוקנים) סובבות מילה מסוימת. לדוגמה: אם שיר מסוים מופיע עם שירים אחרים באותה רשימה, Word2Vec לומד את הקשרים ביניהם.

## מהו Skip-Gram?

Skip-Gram הוא אלגוריתם שמטרתו לנבא את ההקשר של מילה מסוימת על סמך הופעתה במשפטים או רשימות. במודל שלנו:  
- 'מילה' היא שיר או אמן.  
- 'משפט' הוא רשימת השמעה.  
המטרה היא למפות שירים למרחב שבו שירים בעלי הקשרים דומים יופיעו קרובים זה לזה.

## פונקציית most\_similar() לעומת מפה דמיון ודמיון קוסינוס

- \*\*`most\_similar()`\*\*: פונקציה זו משתמשת בדמיון קוסינוס כדי לחשב קרבה בין שיר (וקטור) לשירים אחרים במודל. היא מחזירה רשימה של שירים דומים.  
- \*\*מפת דמיון (Similarity Map)\*\*: מפה זו נבנית על בסיס קרבה בין הווקטורים, ומדגישה את יחסי הקשרים בין כל השירים במודל.  
- \*\*דמיון קוסינוס\*\*: מדד זה משווה בין שני וקטורים ומחשב את הזווית ביניהם. ככל שהערך קרוב ל-1, השירים דומים יותר.

## Hard Negative Sampling

בתהליך האימון, לא כל דוגמה שלילית מועילה. Hard Negative Sampling בוחר דוגמאות שליליות קשות – שירים שנראים דומים על פי המודל, אך אינם רלוונטיים. לצורך כך, השתמשנו בגרסה מותאמת של Softmax, שמעניקה סיכויים גבוהים יותר לדוגמאות שליליות קשות, ושיפרה את הדיוק של המודל.

## מדוע בחרנו במערכת המלצות מבוססת אמנים עם התאמה מטושטשת?

ברשימות השמעה רבות יש אי-אחידות במידע. לדוגמה, שמות אמנים או שירים עלולים להיכתב בצורות שונות. התאמה מטושטשת (Fuzzy Matching) אפשרה לנו לזהות הקשרים בין אמנים, גם כשלא היה תיעוד מדויק של שמותיהם. מערכת מבוססת אמנים הייתה הגיונית כי אמנים יוצרים הקשרים ברורים יותר בין שירים, מה שמאפשר למודל להבין טעם מוזיקלי ברמה גבוהה.

## שקף אחרון: תובנות ואתגרים (4–5 דקות)

לאחר בניית המודל, המסקנות היו ברורות:  
1. Word2Vec זיהה קשרים סמויים בין שירים.  
2. שימוש בדמיון קוסינוס ומפות דמיון שיפר את דיוק ההמלצות.  
  
\*\*אתגרים עיקריים:\*\*  
1. \*\*הטיה ללהיטים פופולריים\*\* – נפתרה עם דגימה מאוזנת.  
2. \*\*יוצאי דופן\*\* – השתמשנו ב-Hard Negative Sampling.  
  
בסופו של דבר, המודל שלנו הצליח לספק המלצות מדויקות בעזרת שילוב של Word2Vec, התאמה מטושטשת וטכניקות מתקדמות אחרות.