שחר יעקב 305255176, איה ברנשטיין 208521815

אפשר לגשת לקוד שלנו בכתובת: https://github.com/shaharjacob/hebrew-songs

רקע

מוזיקה היא מאגר עצום של מילים. בפרט, למילים של שירים יש משמעות, הן מבטאות תרבות ואופי של תקופות מסוימות. כמו כן, בעולם הטכנולוגי המוזיקה מקבלת מקום בקדמת התרבות מבחינת חשיפה ונגישות, מה שלדעתנו גורם לקשר עוד יותר ישיר לתרבות הרווחת.

מטרת הפרויקט

מטרתנו העיקרית בפרויקט היא אבחנה וזיהוי של מאפיינים מעשורים שונים. מה אפשר ללמוד על התרבות באותו עשור? מאפיינים יכולים להיות למשל: שפה דיבור (לדוגמה שימוש בסלנג), שירי מחאה, חרוזים, קונוטציה של השיר (חיובית או שלילית), מי האמן המבצע (גבר, אישה, להקה), אורך שיר וכו'. בנוסף, עניינה אותה השאלה האם אפשר למצוא הבדלים בין שירים שמוגדרים כלהיטים וכאלה שלא.

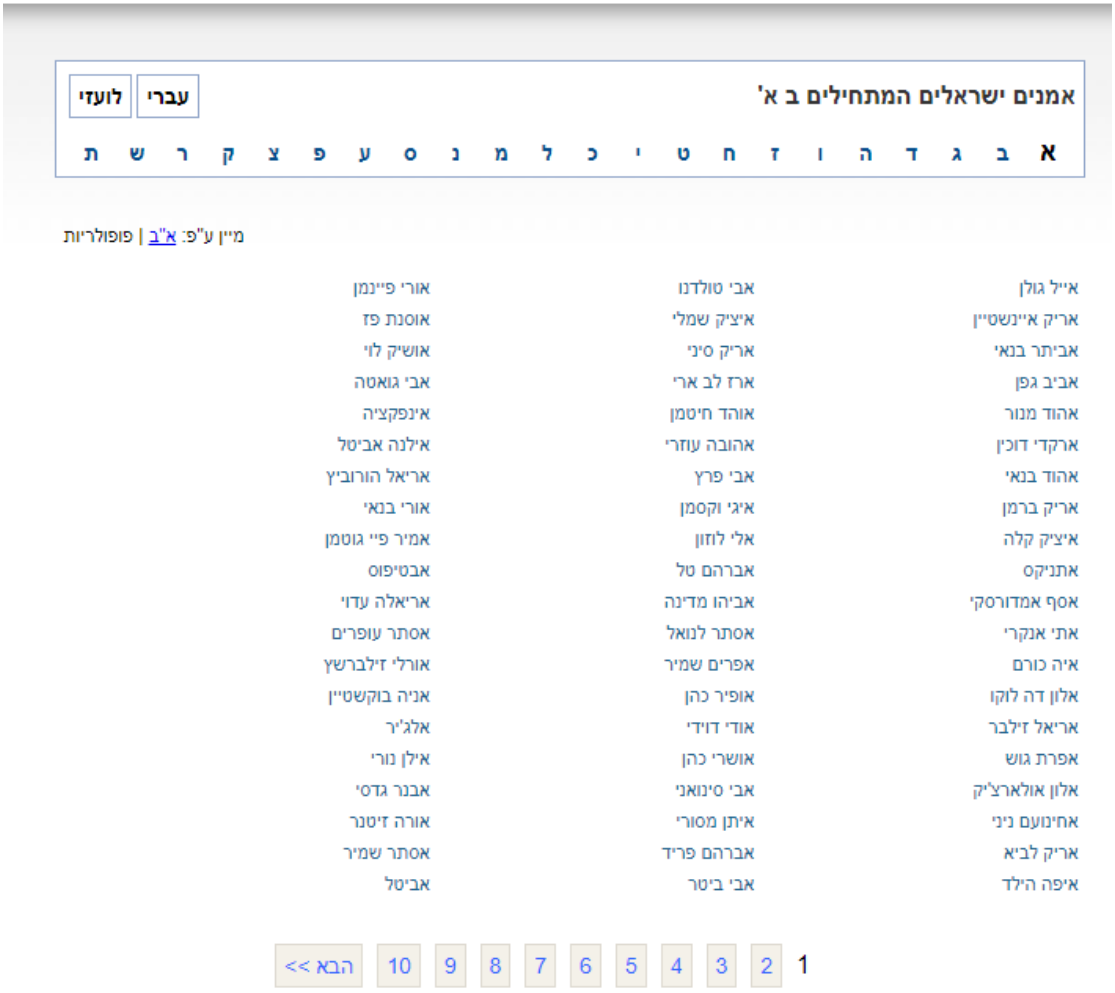
סקירה ספרותית

לא מצאנו פרויקטים דומים לפרויקט שלנו. מרבית התוכן שכן הצלחנו למצוא מבוסס בעיקר על ניתוח של השנה החולפת, טור או דעה אישית של כתב, ללא ניתוח נתונים אמיתי (NLP) כמו שאנחנו עשינו.

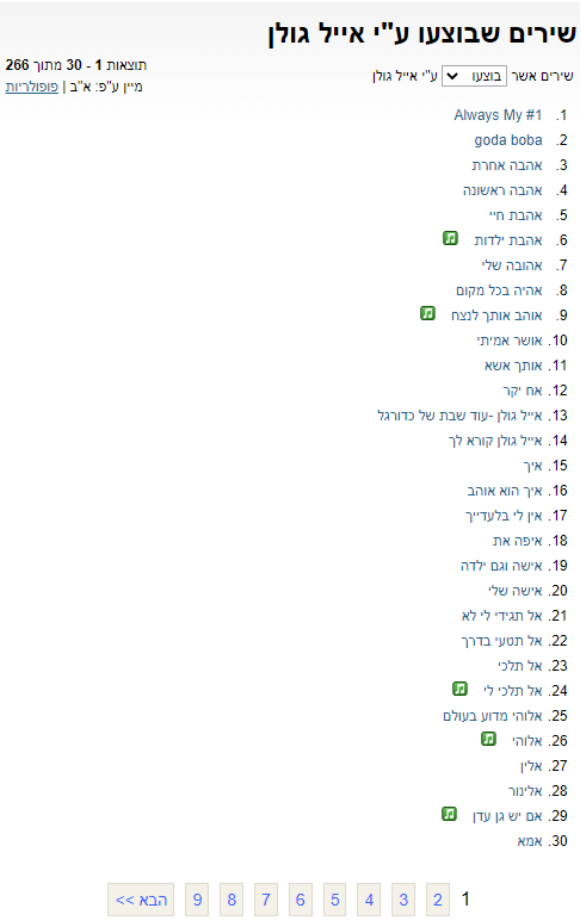
נתונים / Data

בתכנון המקורי (נרחיב על זה בסוף המסמך), מטרתנו העיקרית (בעצם שאלת מחקר) הייתה להצליח לנבא האם שיר הוא להיט או לא. הבעיה היא, שהיה לנו קשה מאוד להשיג דאטא נקי מרעשים (כמו שכבר נרחיב, אפילו את שנת הוצאת השיר לא היה פשוט למצוא). אז מבחינה אידיאלית היינו רוצים שהדאטה שלנו יכיל את שם השיר, מילות השיר, שנת הוצאה, ומדד נכון לכמה השיר היה להיט בתקופתו. מקשיים טכניים, לקחנו רק כמות קטנה של שירים אותם הגדרנו כ"להיטים" ואת היתר לא, כלומר בצורה בינארית. מעבר לבעייתיות בסיווג בינארי, מספר השירים שסיווגנו כלהיטים היה נמוך מדי בשביל להצליח ללמוד מכך משהו. מצד שני הגדלה של מספר השירים שיוגדרו כלהיטים גררה טשטוש גדול יותר בהבדל בין שירים שהם להיטים לכאלה שלא.

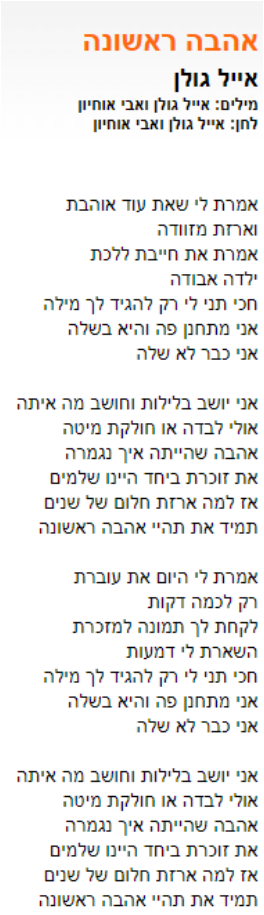
הדאטא שאספנו בנוי מקובץ tsv גדול, כאשר כל שורה מייצגת שיר. מבחינת עמודות, המידע שיש לנו הוא לגבי האמן שביצע את השיר, שם השיר, מילות השיר, תאריך התפוצה שלו, האם השיר הוא להיט (ביחס לאותה תקופה), האם האמן המבצע הוא גבר או אישה.
את המידע הבסיסי אודות שם השיר, האמן המבצע ומילות השיר, אספנו מכמה אתרים, כאשר המרכזי שביניהם הוא שירונט [1]. עשינו אוטומציה מלאה ואספנו בערך 30 אלף שירים (היינו צריכים קצת לנקות את הדאטה מאימוג'ים לדוגמה). ניגשנו בכל אות ל-50 האמנים המובילים (שירונט כבר מסדר אותם בצורה כזו), ואספנו את כל השירים של אותו אמן.



איור 1: האמנים המובילים באות א' לפי אתר שירונט

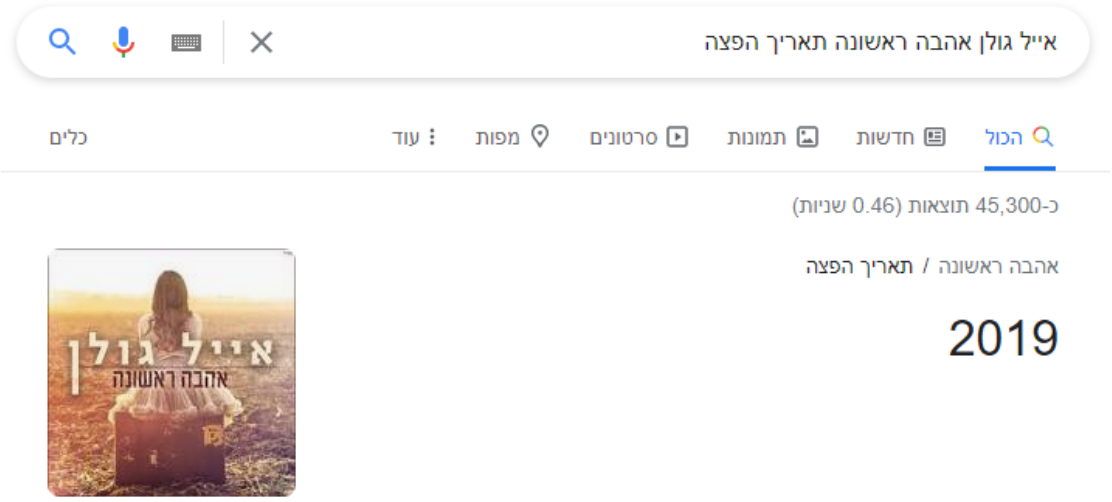


איור 2: השירים שבוצעו ע"י אייל גולן



איור 3: מילות השיר 'אהבה ראשונה'

הבעיה הייתה ששירונט לא סיפק לנו את תאריך ההפצה של השיר, לכן היינו צריכים לגשת ל-API של גוגל. סיננו את כל השירים שגוגל לא הצליח לספק לנו את תאריך ההפצה שלהם, ונשארנו עם קצת יותר מ-10 אלף שירים.



איור 4: תאריך ההפצה של השיר 'אהבה ראשונה' של אייל גולן כפי שמתקבל בגוגל

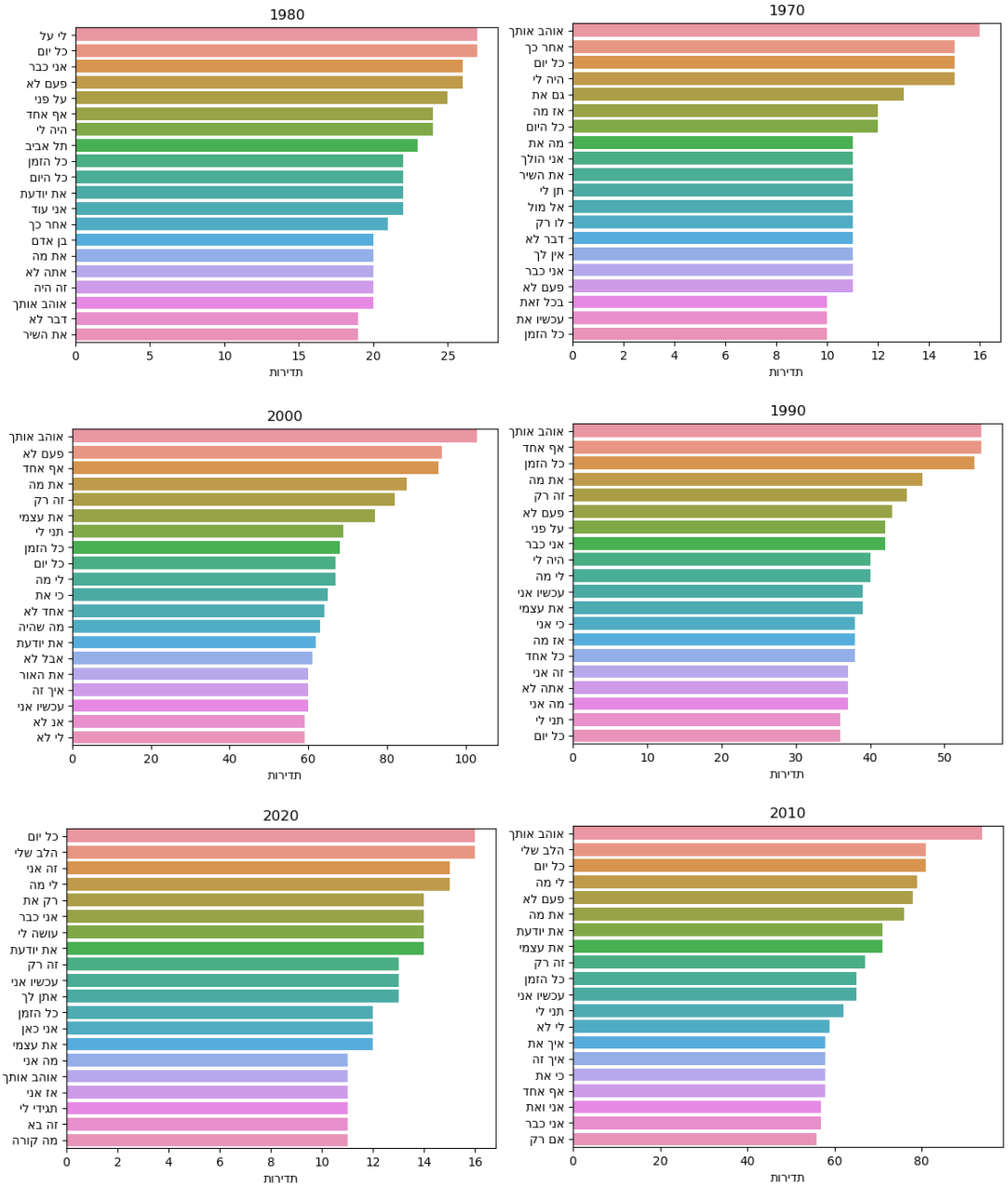
בנוסף, רצינו לדעת האם שיר נחשב ללהיט באותה תקופה (עשור) בה יצא. לשם כך השתמשנו במצעדי פזמונים של ויקיפדיה [2]. בכל שנה לקחנו את ה-5 שירים המובילים, כך שבכל עשור היו לנו 50 שירים שהוגדרו כלהיטים. עשינו אוטומציה שעבדה חלקית ונדרשה עבודה ידנית לקבלת תוצר סופי. בנוסף, עשינו ניסיון להיעזר ב-YouTube ולאפיין האם הוא להיט או לא לפי מספר הצפיות. הבעיה הייתה שהיה רעש גדול מדי (קשה לאבחן מהי הגרסה הרשמית של השיר, וכאשר ממיינים לפי מספר צפיות, לא תמיד מקבלים את השיר הרצוי). בנוסף קשה לנרמל שירים ישנים לעומת חדשים מבחינת כמות צפיות. לכן החלטנו לא להכניס את זה ל-data שלנו.

Features

מילים נפוצות

הפיצ'ר הראשון שנגשנו אליו על מנת לנסות לראות מגמות בעשורים השונים היה מציאת המילים הנפוצות. בכל עשור מצאנו את המילים הנפוצות ביותר, כולל n_grams באורכים שונים. כאשר יצרנו stopwords מותאמים לשירים ישראלים לפי דעתנו.

דוגמה ל-n_grams של עשורים בשנים 1970-2020 למילים באורך 2 :

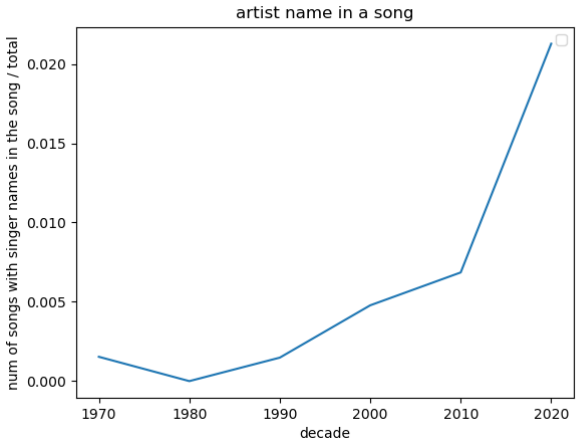


איור 5 : צירופי המילים באורך 2 הנפוצים ביותר לאורך העשורים

יצרנו פונקציה שבודקת בכל עשור מה מייחד אותו לעומת עשורים אחרים, סוג של tf-idf.
למשל עבור מילים באורך 2 קיבלנו :
1970 : אין לך, דבר לא, לו רק, אל מול, תן לי, את השיר, אני הולך
1980 : בן אדם, אני עוד, תל אביב, לי על
1990 : כל אחד, כי אני
2000 : אבל לא, מה שהיה, אחד לא, כי את
2010 : איך זה, איך את, לי לא
2020 : אני כאן, אתן לך, עושה לי, רק את, זה אני

עשינו דבר דומה גם עבור n_grams באורכים 3+ וקיבלנו :
1970 : לי שום דבר, אני רואה אותו, אין לך מה, על אם הדרך, בו מכל צד, מה את חושבת, זה רק חלום
1980 : מה קורה לי, לא חשוב מה, דבר לא השתנה, כל זה היה
1990 : את לא לבד, כל כך קשה, לא יודע אם, אני כל כך
2000 : אני רוצה להיות, אני רוצה אותך
2010 : לא יודע איך, אני לא יכול, כל מה שיש
2020 : אשיר לך שיר, אתן לך את הכל, לומר לך תודה, מה את עושה, רוצה לומר לך, מה שבא לי

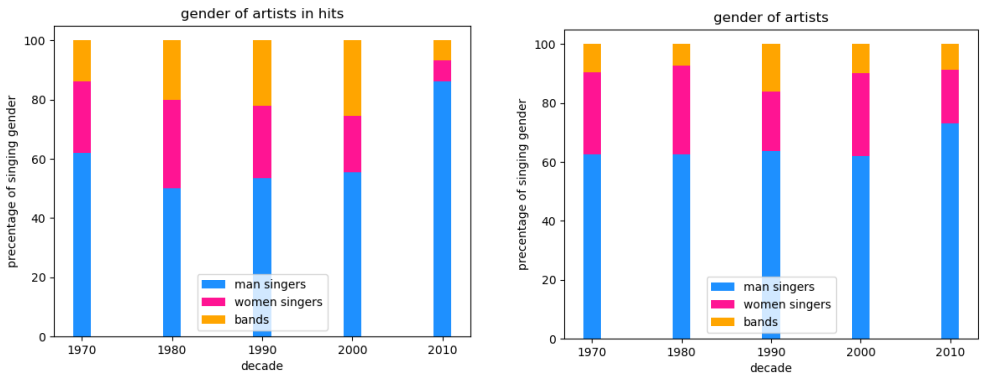
כשעברנו על הדאטא שמנו לב לתופעה מעניינת, שבשנים האחרונים זמרים נוטים להשתמש בשם של עצמם במילות השיר. בדקנו את זה והתוצאות מעניינות :



איור 6 : מדד לכמה פעמים אמן משתמש בשם שלו במילות השיר לאורך העשורים

מין / Gender

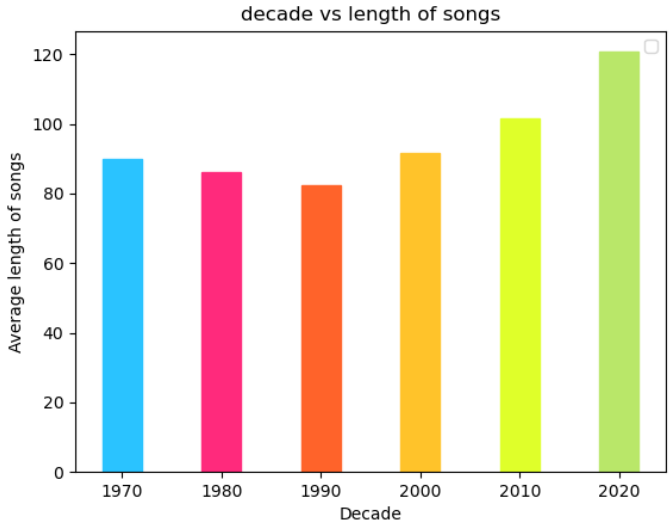
רצינו לבדוק מגמות לאורך השנים גם בתחום המגדרי. אנחנו שמים לב לתופעה מעניינת שכבר שאנחנו מתקדמים בשנים, נשים פחות בולטות בלהיטים. כמו כן להקות באופן כללי פחות בולטות בלהיטים.



איור 7 : אמן המבצע (גבר/אישה/להקה) לאורך העשורים

אורך שיר

דבר נוסף שבדקנו היה אורך השיר לפי עשורים. הופתענו לגלות שדווקא ככל שמתקדמים השנים, כך השירים נהיים ארוכים יותר. לדעתנו זה נובע מכך שהשפה נהיית פחות ופחות גבוהה ולכן יש צורך ביותר מילים כדי להביע כוונה. הכוונה לא רק בשפה גבוהה מבחינת המילים עצמן, אלא מבחינת עומק ואבסטרקטיות של הטקסט. כשהטקסט עמוק, נדרשים לדעתנו פחות מילים בכדי לתאר אותו.



איור 8 : אורך השיר לאורך העשורים

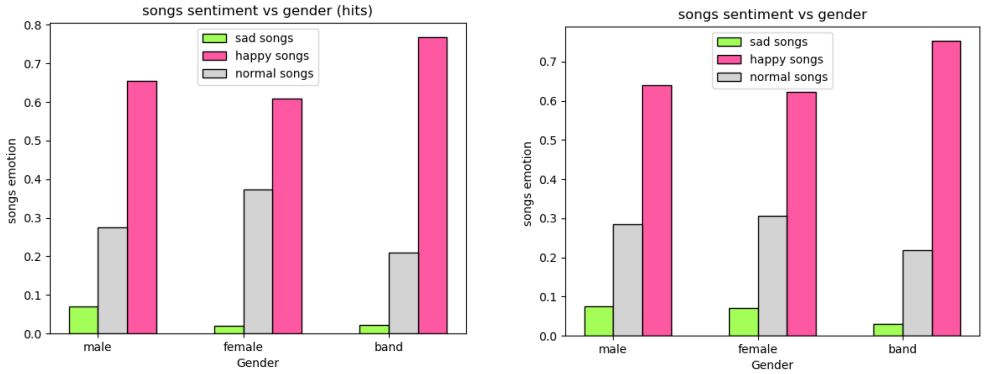
קונוטציה

זה אולי הפיצ'ר המרכזי בפרויקט שלנו שהשקענו בו את מירב הזמן. עניין אותנו מאוד לדעת להעריך האם שיר הוא בעל קונוטציה חיובית או שלילית. כלומר שיר שמח ואופטימי זה שיר עם קונוטציה חיובית, ואילו שיר פרידה או שיר שמדבר על מוות הוא בעל קונוטציה שלילית. מעבר ליכולת לסווג שירים ככאלה, שזו משימה מעניינת בפני עצמה, רצינו גם לבדוק קונוטציה כללית של עשור מסוים ולהשוות לעשורים אחרים. כמו-כן, התעניינו גם בשנים בהם היו מלחמות.

אז איך עשינו את זה?

תחילה לקחנו מודל AlephBERT מ-HuggingFace, ועשינו לו Fine-tuning על ציוצים בטוויטר שקשורים לרובי ריבלין, כאשר הציוצים האלה מתויגים (קונוטציה שלילית או חיובית). את הקוד לאימון אפשר למצוא תחת AlephBERT.py ואילו את הדאטה תחת data/train_sentiment.csv. את האימון עשינו ב-Google Colab כדי להשיג כוח חישובי (GPU).

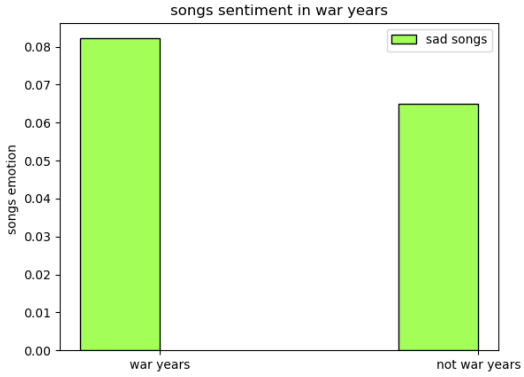
לאחר מכן, פירקנו את השיר לפי שורות וסכמנו את מספר השורות עם הקונוטציה השלילית/חיובית, וכך החלטנו האם שיר הוא בעל קונוטציה חיובית, שלילית או ניטרלי.



איור 9: מימין כמה אחוז מהשירים הם בעלי קונוטציה שלילית/חיובית/ניטרלית. משמאל: אותו הדבר רק עבור להיטים

אפשר לראות שלא ראינו הבדל ניכר לעין בין להיטים לכאלה שלא, מלבד נשים, שכאשר מוציאות להיטים זה בדרך כלל לא יהיה שיר עם קונוטציה שלילית.

דבר נוסף שרצינו לראות זה השפעה של שנים בהם היו מלחמות על הקונוטציה בשירים. ואכן זה השתקף בתוצאות:



איור 10: שנות מלחמה, כמה מהשירים מוגדרים בעלי קונוטציה שלילית.

כלומר בשנים בהם היו מלחמות או מבצעים גדולים (1973, 1982, 2006, 2008, 2012, 2014, 2021), ניתן לראות שיש יותר שירים עם קונוטציה שלילית. חשוב לציין שלקחנו גם שנה אחת קדימה לכל שנה כי לפעמים לוקח זמן עד שהרגש מחלחל לשירים.

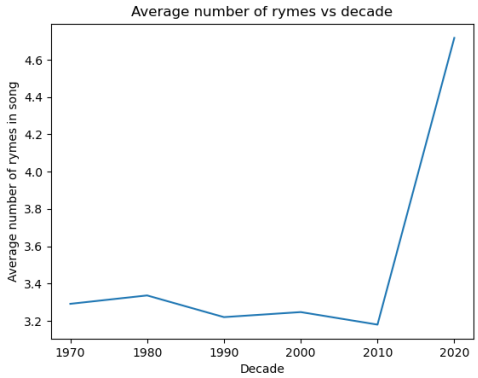
בנוסף, עשינו פונקציה שנותנת לנו מדד (ערך מספרי רציף ולא בינארי) לשאלה האם שיר הוא בעל קונוטציה שלילית או חיובית. ועשינו את זה בצורה דומה לסיווג הבינארי עליו הסברנו בהתחלה, רק ששקללנו את התוצאה ביחס לאורך השיר. התוצאות היו אכן כמצופה. למשל, בין השירים עם הקונוטציה הכי שלילית הוא 'מה נעשה' של הדג נחש. שזה הגיוני מאחר ומדובר בשיר מחאה.
שיר נוסף בלע קונוטציה שלילית, של נורית גלרון:

"אלוהיך נתן לך למות
בהרגשת אשמה
כתוב למרומים:
הוא עזב אותך."

לעומת זאת, בין השירים עם הקונוטציה החיובית אפשר למצוא את 'אלילת האגם' של אהובה עוזרי, 'דוניה דוניה' של יהורם גאון וגם 'אני קיים' של החברים של נטשה.

חרוזים

בנינו פונקציה שבודקת כמה חרוזים יש בשיר. לאחר קריאה ספרותית של כיצד מוגדר חרוז ולאחר כמה ניסיונות, החלטנו בסוף ללכת עם הגרסה הכי נאיבית שמצאנו שדווקא לדעתנו עבדה הכי טוב. אנחנו בודקים את שתי האותיות האחרונות של המשפט ומשווים לשתי אותיות האחרונות במשפט שקדם לו. אנחנו שמים לב שבעשור שהחל מ-2010 יש קפיצה משמעותית בשימוש בחרוזים:



איור 11: מספר ממוצע של חרוזים בשיר לאורך השנים

לאחר נרמול של מספר החרוזים בהתאם לאורך השיר, אפשר לדבר על "השיר הכי חרוזי". אצלנו, השיר 'על ראש החרמון' של חווה אלברשטיין, נמצא במקום הראשון:

"על ראש החרמון
היה היה ארמון
ועל ראש הארמון
מגדל עם פעמון

פעם ילד קטון
עלה על החרמון
מצא את הארמון
וצלצל בפעמון

דין דין..."

חיזוי שנת השיר

בנינו מודל naïve base שמבוסס על מילות השיר, שמטרתו לחזות מהי שנת ההוצאה לאור של השיר. לדוגמה, עבור 6 עשורים אופציונליים (1970, 1980, 1990, 2000, 2010, 2020), המודל מצליח במשימה בדיוק של קרוב ל-40%, כאשר ניחוש רנדומלי הוא 16.667%.

התוצאות ברורות יותר כאשר מציגים את ה- confusion matrix, שלמעשה העמודות מייצגות את הערך הנכון ואילו השורות מייצגות את הערך שהתקבל מהמודל. כך שבצורה אידיאלית היינו מצפים לראות ריכוז גדול באלכסון הראשי.

	1970	1980	1990	2000	2010	2020
1970	3	10	25	64	44	0
1980	4	29	21	101	67	0
1990	5	24	84	174	140	0
2000	4	25	40	288	257	0
2010	6	28	33	159	285	2
2020	0	2	2	21	33	3
Accuracy: 38.56%						

אומנם לא קיבלנו תוצאה אידיאלית בה הכל נמצא על האלכסון הראשי, אבל כן אפשר לראות שרוב המסה מתרכזת סביב המרכז, כאשר רוב השגיאות הן עבור עשורים עוקבים שזה הגיוני.

חיזוי האמן המבצע

מודל נוסף שבנינו שמבוסס על עץ החלטה (decision tree), מקבל רשימה של אמנים ופיצ'רים עליהם אנחנו רוצים שילמד, כאשר המטרה היא לחזות מי האמן המבצע בהינתן שיר. למשל בדוגמה הבאה, נתנו למודל שלנו ללמוד על האמנים: אייל גולן, שרית חדד, עומר אדם, כוורת, נועה קירל והדג נחש. הפיצ'רים עליהם אפשרנו לו ללמוד הם: אורך השיר, האם השיר להיט או לא, ומה הקונוטציה של השיר.

```
1.  |--- lyrics length <= 190.50
2.  |   |--- lyrics length <= 109.50
3.  |   |   |--- lyrics length <= 69.50
4.  |   |   |   |--- lyrics length <= 46.50
5.  |   |   |   |   |--- prediction: שרית חדד
6.  |   |   |   |--- lyrics length >  46.50
7.  |   |   |   |   |--- prediction: שרית חדד
8.  |   |   |--- lyrics length >  69.50
9.  |   |   |   |--- song sentiment <= 1.50
10. |   |   |   |   |--- prediction: אייל גולן
11. |   |   |   |--- song sentiment >  1.50
12. |   |   |   |   |--- prediction: כוורת
13. |   |--- lyrics length >  109.50
14. |   |   |--- lyrics length <= 119.50
15. |   |   |   |--- lyrics length <= 114.50
16. |   |   |   |   |--- prediction: אייל גולן
17. |   |   |   |--- lyrics length >  114.50
18. |   |   |   |   |--- prediction: אייל גולן
19. |   |   |--- lyrics length >  119.50
20. |   |   |   |--- hit <= 0.50
21. |   |   |   |   |--- prediction: עומר אדם
22. |   |   |   |--- hit >  0.50
23. |   |   |   |   |--- prediction: כוורת
24. |--- lyrics length >  190.50
25. |   |--- lyrics length <= 250.50
26. |   |   |--- lyrics length <= 239.50
27. |   |   |   |--- lyrics length <= 227.50
28. |   |   |   |   |--- prediction: נועה קירל
29. |   |   |   |--- lyrics length >  227.50
30. |   |   |   |   |--- prediction: נועה קירל
31. |   |   |--- lyrics length >  239.50
32. |   |   |   |--- hit <= 0.50
33. |   |   |   |   |--- prediction: עומר אדם
34. |   |   |   |--- hit >  0.50
35. |   |   |   |   |--- prediction: הדג נחש
36. |   |--- lyrics length >  250.50
37. |   |   |--- lyrics length <= 295.50
38. |   |   |   |--- prediction: הדג נחש
39. |   |   |--- lyrics length >  295.50
40. |   |   |   |--- lyrics length <= 313.00
41. |   |   |   |   |--- prediction: נועה קירל
42. |   |   |   |--- lyrics length >  313.00
43. |   |   |   |   |--- prediction: הדג נחש
```

אפשר לראות בתוצאות שלמשל אם אורך השיר גדול מ-70 מילים (שורה 8) אך קטן מ-110 מילים (שורה 2), ואם מדד הקונוטציה החיובית גדול מ-1.5 (שורה 11) אז כנראה שמדובר בשיר של כוורת. בצורה דומה אם אורך השיר גדול מ-239 מילים (שורה 31) אך קטן מ-251 מילים (שורה 25), אז אם מדובר בלהיט (שורה 34) סיכוי גדול שמדובר בשיר של הדג נחש.

מסקנות

אחת המסקנות העיקריות שלנו היא שאין מקום מרוכז ומסודר שמחזיק בדאטה לגבי שירים במוזיקה הישראלית. כלומר אתרים כמו שירונט אומנם מחזיקים בדאטה גדול של אמנים והשירים שלהם, אבל חסר מידע לגבי שנת ההוצאה של השיר ובעיקר האם השיר הזה הצליח או איזשהו מדד הצלחה של השיר. כמו שציינו בתחילת המסמך, מטרת הפרויקט המקורית הייתה לנבא האם שיר הוא להיט או לא, אבל המחסור במידע בנוגע להצלחתו של שיר גרמה לנו לשנות כיוון ולהתעסק יותר בדאטה שכן יש לנו.

מבחינת תוצאות, אנחנו כן מסיקים שבצורה חד משמעית ניתן לראות שינוי מגמתי בתוכן ובסגנון השירים. לאורך השנים השירים נהיים ארוכים יותר (בצורה מפתיעה, אבל זה כנראה נובע משירי ראפ ומחאה), יש יותר שימוש בחרוזים, אמנים נוטים להשתמש בשמות שלהם במילות השיר, וכן המילים עצמן משתנות. אין לנו ספק שבהינתן דאטא גדול יותר מזה שהיה לנו, ובהינתן דאטה שמכיל מדד הצלחה של השיר, אפשר לקבל תוצאות מאוד מעניינות.

שימושים ושיתופי פעולה

לדעתנו השימוש העיקרי בכלי הנוכחי הוא בעיקר לצרכי מחקר והבנה של התרבות הקיימת.
במידה ונצליח לממש בצורה טובה את השאלה המקורית של האם שיר הוא להיט – זה יכול לשמש את האמנים עצמם.

הצעות להמשך

אנחנו חושבים שהצעת המשך לפרויקט הזה, יכולה להיות מדד (שהוא לא בינארי), לגבי כמה השיר הוא להיט. בגלל הבעייתיות בשירים ישנים, אולי אפשר להתרכז רק בשירים מהמילניום הנוכחי. אנחנו חושבים שכלי שגם משתמש ב-YouTube וגם משתמש בצורה יותר פרטנית במצעדי שירים כמו גלגל״צ (ולא רק לקיחת 5 שירים הכי מצליחים באותה שנה), יכולים להביא לשינוי שאנחנו ״נתקענו״ בגללו. כלומר ייתכן שהתמקדות בשירים רק מהמילניום הנוכחי הייתה פותרת את הבעיה שלנו, כי אז היינו יכולים ממש להיכנס לגלגל״צ ולבדוק כמה פעמים הושמע שיר ברדיו (מדד מעולה להצלחה של שיר) – כמובן שזה לא אפשרי כשאנחנו מדברים על עשורים רבים אחורה.

References / הפניות

[1] - https://shironet.mako.co.il
[2] - מצעד הפזמונים העברי השנתי - ויקיפדיה