מגישות

1. אביטל שוחט-קרן. שם משתמש: Avitalsho1. תעודת זהות: 301560322.
מייל: avital.shohat1@mail.huji.ac.il
2. נעמה גלזר-סלומינסקי. שם משתמש: Naamagl. תעודת זהות: 307880971.
מייל:naama.glazer@mail.huji.ac.il
3. נטע ברק. שם משתמש: netta.barak. תעודת זהות: 204635593.
מייל: netta.barak@mail.huji.ac.il

לינק לקוד ולקבצים:

https://drive.google.com/drive/folders/1iMJ-zqqwoOeK-9L6oeSrQFhgQINalbvE?usp=sharing

השאלה

האם ניתן לנבא את הפופולריות של שיר אך ורק באמצעות המאפיינים המוזיקליים שלו? ואם כן, איך? האם לשירים הפופולריים במדינות שונות יש מאפיינים מוזיקליים שונים? ומהם? בפרויקט זה ניסינו לגשת לשאלה הזו בדרכים שונות, ולהסתכל על המאפיינים המוזיקליים של שיר בפרספקטיבה חדשה.

ניסיון ראשון

כדי לגשת לשאלות שמעניינות אותנו, חיפשנו דאטה מתאים. רצינו שהדאטה יכיל הרבה מאוד שירים, את המאפיינים המוזיקליים שלהם, פרטים בסיסיים על האומן המבצע, וכמובן מדדים למידת הפופולריות של השיר במדינות שונות ובאופן כללי. בחרנו לשלב דאטה משלושה מקורות שונים:

1. Spotify – דגימה של כ-116,000 שירים שנאספו רנדומלית בנובמבר 2018 (השירים לאו דווקא יצאו בחודש זה, אבל הדגימה לא תכיל שירים שיצאו מאוחר יותר) והועלו לקאגל (https://www.kaggle.com/tomigelo/spotify-audio-features או SpotifyAudioFeaturesNov2018.csv המצורף). הדאטאסט מכיל את שם השיר, המזהה של השיר במערכות Spotify, שם האומן, הפופולריות של השיר, מאפיינים מוזיקליים בסיסיים כגון טמפו, משקל וסולם ומאפיינים מוזיקליים מורכבים יותר כגון אקוסטיות, דיבוריות, אנרגטיות, רקידות וכו'.
2. Google Trends – השימוש בחיפושי Youtube כפי שמוצגים ב-Google Trends נועד להצביע על הפופולריות היחסית של שיר במדינה מסוימת. השתמשנו ב-API בשם pytrnds שמבצע crawling ובאמצעותו הבאנו, לכל שיר ועבור כל מדינה בעולם, את הפופולריות היחסית מ-1 עד 100 של החיפוש המכיל את שם השיר ולאחריו שם האומן המבצע.
3. Youtube – השימוש בדאטה מ-Youtube נועד לייצר מדד לפופולריות הכללית של שיר, כזה שנוכל גם לשקלל ביחד עם פופולריות החיפוש בכל מדינה. בנינו בעצמנו זחלן שמבצע חיפוש המכיל את שם השיר ולאחריו שם האומן המבצע (בדומה לשאילתה ב-pytrends), ומחלץ מהקליפ הראשון שמוחזר את מספר הצפיות ותאריך העלייה לאתר, על מנת לייצר משני המדדים האלו מדד לפופולריות של השיר.

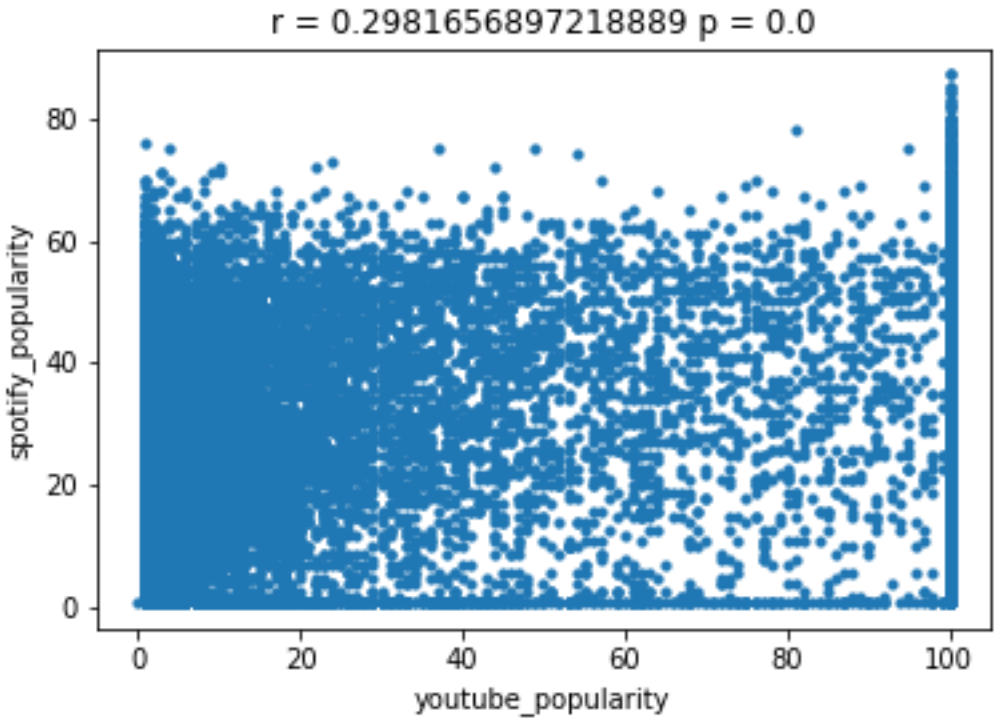
כדי לצמצם את הנזק שייגרם ע"י תקלות ולבצע את הזחילה בצורה יעילה יותר, פיצלנו את 116,000 השירים לקבצים קטנים יותר בני 100 שירים כל אחד וביצענו את פעולות הזחילה בנפרד (ראו crawlers.py). לאחר שאיחדנו אותם, ביצענו מספר פעולות כדי לנקות ולסדר את הדאטה (ראו random_songs_preprocessing.ipynb):

1. הורדת רשומות לא תקינות או כאלה שבחרנו להתעלם מהן, למשל:
   a. שירים שלא נמצאו בחיפוש ב-Youtube .
   b. שירים שאורכם יותר מעשר דקות.

c. שירים שהחיפוש עבורם ביוטיוב העלה תוצאות חדשות יותר מנובמבר 2018 (כלומר, חדשות יותר מהדאטה סט המקורי).

2. נרמול כל הפיצ'רים המוזיקליים כך שיהיו בין 0 ל-1. במדדים מסוימים ראינו שההתפלגות היא בעלת "זנב ארוך" ולכן בחרנו לא לנרמל באופן סטנדרטי, אלא להתייחס לכל הערכים הגדולים מערך מסוים כ-1.
3. חישוב הפופולריות לפני יוטיוב – בחרנו לחשב את הפופולריות באופן הבא: חילקנו את מספר הצפיות במספר הימים שעברו מאז העלאת הסרטון לאתר, ואז נרמלנו את הציון למספר עגול בין 1 ל-100, בדומה לדירוג הפופולריות של ספוטיפיי.

לאחר שביצענו את הזחילה עבור כ-37,000 שירים, רצינו לבצע בדיקת שפיות בסיסית לאמינות של הדאטה שלנו. ציירנו Scatter Plot של הפופולריות לפי יוטיוב ולפי ספוטיפיי וחישבנו את הקורלציה ביניהם (ראו youtube_spotify_popularity_correlations.ipynb).



אז חשכו עינינו. ניתוח של המצב הוביל אותנו לכמה גורמים אפשריים לבעיה :

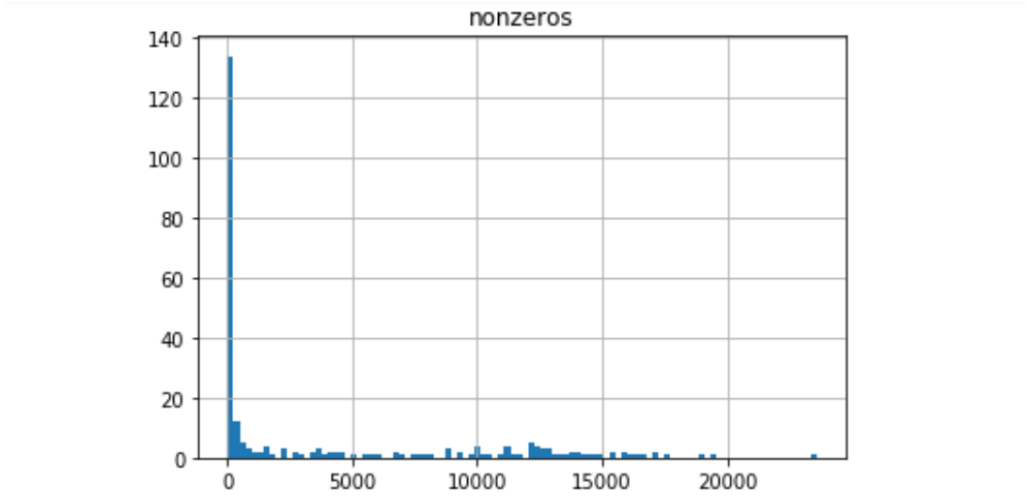
1. חוסר התאמה בין השיר שקיבלנו בספוטיפיי לשיר שמצאנו ביוטיוב – כלומר, חיפשנו שיר מסוים אבל החיפוש ביוטיוב החזיר לנו שיר שונה מזה שהתכוונו אליו.
2. חוסר התאמה בין הפופולריות של שירים מסוימים בין ספוטיפיי ליוטיוב. סביר, למשל, שקהלי היעד של יוטיוב וספוטיפיי הם בעלי מאפיינים שונים ולכן הפופולריות של שירים מסוימים תהיה שונה בשתי הפלטפורמות.
3. בעיות הנובעות מהאופן בו חישבנו את הפופולריות של שירים ביוטיוב. למשל, אולי מספר הצפיות אינו מעיד באופן מלא על הפופולריות של שיר, משום שזו מושפעת במקרים רבים ממערכת ההמלצה של יוטיוב. בנוסף, התפלגות הפופולריות של ספוטיפיי ויוטיוב הייתה מאוד שונה. בדקנו אפשרויות להפעלת פונקציות מונוטוניות עולות שונות על ציוני הפופולריות ביוטיוב כך שיהפכו את ההתפלגויות לדומות יותר מבלי לשנות את יחס הסדר. ניסינו מספר מניפולציות, אך אלה לא השפיעו על הקורלציה באופן מהותי, ולבסוף החלטנו להשאיר את הציונים כפי שהם.

כדי לרדת לשורש הבעיה, דגמנו כעשרים שירים מארבע קבוצות קיצוניות שונות של שירים לפי יחסי הפופולריות בין יוטיוב לספוטיפיי (youtube_popularity/Spotify_popularity), ואפיינו אותן איכותנית:

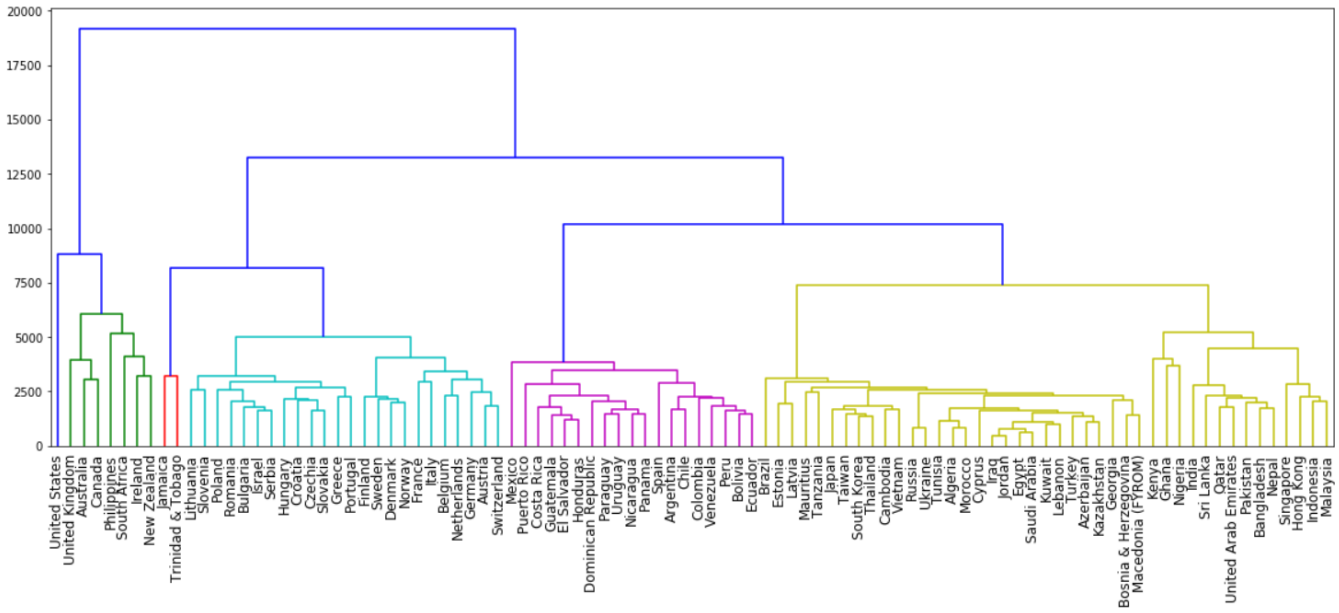
1. השירים שיחס הפופולריות שלהם הוא הגבוה ביותר (100 לפי יוטיוב ו-1 לפי ספוטיפיי) – ברוב המוחלט של שירים אלה הייתה טעות בזיהוי. השיר בספוטיפיי היה כל כך נידח, שהחיפוש ביוטיוב הוביל לשיר אחר, בדומה להשערה הראשונה שהעלנו.
2. שירים שיחס הפופולריות שלהם הוא הנמוך ביותר (1 לפי יוטיוב ו-50-60 לפי ספוטיפיי) – ברוב המוחלט של שירים אלה הזיהוי היה נכון, השיר שמצאנו ביוטיוב היה זהה לשיר בספוטיפיי.
אמנם הגדרנו את ציון הפופולריות לפי יוטיוב רק על פי תוצאת החיפוש הראשונה, אך תוצאת החיפוש הראשונה היא בד"כ התוצאה המתאימה ביותר בעלת מספר הצפיות הרב ביותר, ולכן ציון הפופולריות הנמוך נובע כנראה מסיבה אחרת. לדעתנו, ההבדלים במקרים הללו נובעים מהשוני בקהלי היעד ובמנגנוני ההמלצה של שני השירותים.
שירים שיחס הפופולריות שלהם הוא 1 – במחצית מהשירים החיפוש ביוטיוב הוביל לשיר אחר לחלוטין. עבור כרבע מהשירים השיר המבוקש נמצא ביוטיוב אך בביצוע של אומן אחר, החשבנו זאת כחוסר התאמה בחיפוש. ברבע האחרון החיפוש אכן תאם והפופלריות בספוטיפיי וביוטיוב, לפי המדדים שלנו, היתה זהה.
3. שירים שדגמנו באופן רנדומלי – הרוב המוחלט של השירים היו תואמים במידה זו או אחרת בין יוטיוב לספוטיפיי. לפעמים ההתאמה הייתה מלאה, לפעמים היה מדובר בגרסת אולפן מול גרסת הופעה חיה, לעתים היה מדובר בגרסה של הזמר המבוקש בשיתוף עם זמר אחר ולעתים בגרסת כיסוי של אמן שונה לגמרי (מבחינתנו מקרה זה לא נחשב כהתאמה מספקת). חלק מצומצם מהשירים היה טעות מוחלטת בזיהוי.

בשורה התחתונה, נראה היה כי חוסר ההתאמה בין ציוני הפופולריות ביוטיוב וספוטיפיי נובע משילוב בין ההשערות שהעלינו. בחנו אפשרויות שונות לניקוי הדאטה, למשל התעלמות משירים עם יחס פופולריות נמוך מאוד אך מחיקת שירים אלו לא שיפרה משמעותית את המצב. הצענו גם כמה הצעות שיכלו לסייע לנו בניקוי הדאטה, כגון חילוץ אורך השיר כפי ושם האומן כפי שנמצאו ביוטיוב (אז יכולנו למחוק שירים שבהם הייתה חוסר התאמה בין יוטיוב לספוטיפיי), אך פסלנו אותן כי הן דרשו זחילה נוספת שהייתה מורכבת במסגרת הזמן שנותר לנו (הזחילה על יוטיוב דרשה זהירות ונאלצנו לבצע אותה באיטיות ובמקביל על מנת שלא ניחסם ע"י האתר).

עד לשלב זה, לא השתמשנו עדיין בנתונים שחילצנו מ-Google trends לגבי פופולריות החיפוש של השירים השונים ביוטיוב במדינות שונות. החלטנו לעשות Agglomerative Clustering של המדינות, כאשר הפיצ'רים היו הפופולריות של כל אחד מהשירים במדינה (ראו countries_clustering_by_songs_popularity - random_songs_data.ipynb). בחינה של הדאטה הראתה שעבור כ-150 מתוך 250 מדינות יש מספר נמוך מאוד של שירים שעברו את סף הפופולריות המינימלי עבורו Google Trends מספק מידע.



חלק ניכר מהאזורים שאליהן מתייחס Google Trends הם מדינות קטנות מאוד או אזורים קטנים שאינם מדינות. לכן החלטנו לסנן את המדינות שאליהן נתייחס במחקר, ולקחת רק מדינות שבהן לפחות 1000 (מתוך כ-37,000) חיפושים עברו את סף הפופולריות המינימלי. נותרנו עם כ-100 מדינות ויצרנו דנדוגרם שיתאר אותן.



הנטייה של קלאסטרינג מהסוג הזה הוא ל-rich is getting richer ולכן בחרנו את הפרמטרים שיצמצמו את התופעה עד כמה שניתן. נוסף על כך, נעזרנו בגרף ובחרנו לחלק את המדינות לארבעה קלאסטרים, משום שזהו מספר הקלאסטרים הגדול ביותר שבו לא קיבלנו קלאסטר שמכיל מדינה אחת בלבד (את החלוקה המפורטת לקלאסטרים ניתן לראות במחברת). ניתן לראות שהקלאסטרים מאופיינים במידה רבה ע"י השפה המדוברת (ייתכן וזו נובעת פשוט מהשפה בה נכתב שם השיר בחיפוש), למשל קלאסטר המכיל בעיקר מדינות דוברות אנגלית, וקלאסטר המורכב ממרבית מדינות דרום אמריקה וספרד. נוסף על

כך, ניתן לומר שלקלאסטרים יש מאפיינים תרבותיים דומים – אם נסתכל למשל על הקלאסטר שמכיל את מרבית מדינות אירופה. הקלאסטר הרביעי מכיל מדינות רבות שלא ניתן במבט חטוף להבין מה המכנה המשותף ביניהן. בסך הכל, הממצאים אינם מפתיעים מאוד, למרות שישנן הפתעות – למה הפיליפינים נמצאת בקלאסטר של המדינות דוברות האנגלית?

על אף שראינו שהדאטה שלנו אינו נקי לחלוטין ובעקבות ההצלחה היחסית של הקלאסטרינג, החלטנו לבדוק האם ניתן לבנות באמצעותו מודל רגרסיה, ולחזות את הפופולריות של שיר במדינה מסוימת לפי המאפיינים המוזיקליים שלו. בחנו את המודל על מדינות שונות (בחרנו מדינות מקלאסטרים שונים) ועל חישוב הפופולריות הכולל לפי יוטיוב, תוך הבנה שאין התאמה בין הפופולריות לפי יוטיוב לזו לפי ספוטיפיי, וישנה התאמה טובה יותר בין ההאזנות ביוטיוב והחיפושים ביוטיוב כפי שקיבלנו מגוגל טרנדס. בחנו שימוש ברגרסיה ליניארית ולאחר מכן ב-Random Forest עם פרמטרים שונים. בחרנו במודל הרגרסיה הליניארית כמודל בסיסי ופשוט על מנת שיהווה בייסליין שנוכל להשוות אליו. לאחר מכן בחרנו ב-Random Forest משום שזה מודל שמאפשר לבטא התנהגות מורכבת יותר תוך שימור ה-explainability. בשני המקרים המודל היה גרוע מאוד בלחזות פופולריות של שיר - coefficient of determination R^2 קרוב ל-0, ולעתים אף קטן מ-0, כלומר גרוע יותר משל מודל שמחזיר תמיד את הפופולריות הממוצעת (ראו random_songs_regressions.ipynb).

את הביצועים הגרועים של המודלים ניתן היה להסביר בשתי דרכים עיקריות:

1. ניקיון ועקביות הדאטה – חוסר ההתאמה הנפוץ מדי בין השירים מספוטיפיי שנמצאו בדאטסט הבסיסי לאלו שמצאנו ביוטיוב וגוגל טרנדס, וחוסר ההתאמה המחשידה לפעמים בין הפופולריות בספוטיפיי לזו ביוטיוב.
2. חוסר היכולת של מאפיינים מוזיקליים בלבד להסביר הצלחה או אי הצלחה של שיר. ייתכן כמובן שחוסר היכולת הזו נובעת מאופן חילוץ המאפיינים המוזיקליים ע"י ספוטיפיי, ושיום אחד יצליחו מדעני המידע לחלץ פיצ'רים מוזיקליים אחרים שלפיהם נצליח לנבא הצלחה של שיר. לעת עתה, אנחנו סומכות על האיכות של המאפיינים המוזיקליים כפי שחולצו ע"י ספוטיפיי.

על מנת לנסות ולשלול את האפשרות הראשונה, החלטנו לבצע בדיקות דומות על דאטה סט נקי ועקבי יותר.

<u>ניסיון שני</u>

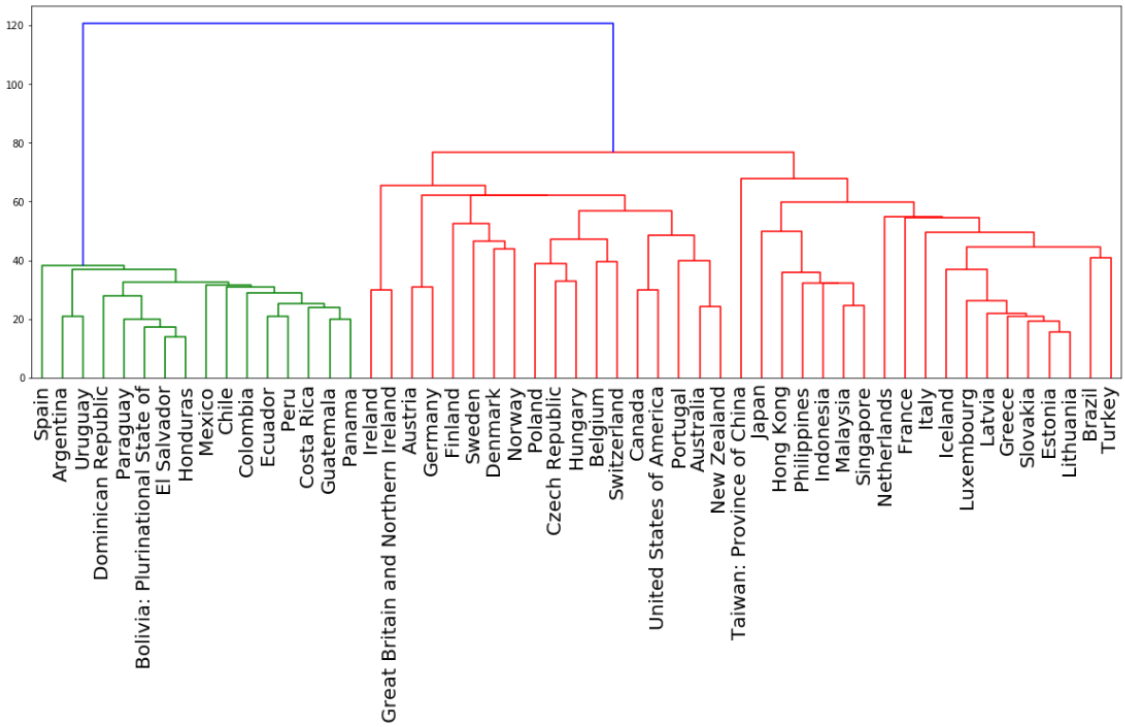
על מנת לנטרל את הבעיות שנבעו מהשילוב של דאטה בין ספוטיפיי ליוטיוב, ומהעבודה עם דאטה סט שחלק ניכר מהשירים בו הם לא פופולריים כלל עד כדי נידחים ביותר, החלטנו לעבוד עם דאטה סט המתאר את 200 השירים הפופולריים ביותר בכל יום בשנת 2017 ב-50 מדינות שונות (https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking או הקובץ spotipy_countries.csv המצורף). אם מתעלמים מהחזרות, הדאטה הכיל כ-20,000 שירים שונים של כ-7000 אמנים שונים.

העיבוד הראשוני של הדאטה כלל מספר פעולות (ראו top_hits_preprocessing.ipynb):

1. חישוב הפופולריות המשוקללת של כל שיר (מדובר כמובן בפופולריות עבור שנת 2017, ולא בפופולריות כללית). כשעשינו זאת, התחבטנו בשאלה הכמעט פילוסופית "איזה שיר נחשב יותר פופולרי? שיר שהאזינו לו 300 פעמים ביום אחד או שיר שהאזינו לו פעם אחת ביום במשך 300 ימים?". החלטנו שלא להכריע בסוגייה, ולהתייחס לשירים הללו כפופולריים באותה מידה. משום כך, הגדרנו את הפופולריות של שיר מסוים במדינה מסוימת לפי מספר ההשמעות הכולל שלו ב-2017, ונרמלנו את הפופולריות עבור כל מדינה, כך שקיבלנו ציונים בין 1 ל-100.
2. פנייה ל-API של Spotify על מנת לחלץ את המאפיינים המוזיקליים של השירים מתוך המזהים של השירים.

3. פנייה ל- API של ספוטיפיי על מנת לחלץ משם האמן את מזהה האמן, ובאמצעותו את הפופולריות של האמן.

לאחר שסיימנו את העיבוד הראשוני של הדאטה, ניגשנו לעשות קלאסטרינג של המדינות לפי השירים הפופולריים בהם (ראו countries_clustering_by_songs_popularity - top_hits_data.ipynb). לכל מדינה, שיר שהיה ברשימת השירים הפופולריים במדינה זו קיבל ציון בין 0 ל-1 בהתאם למספר ההשמעות שלו, ושיר שלא היה ברשימת השירים הפופולריים במדינה זו (אך היה בדאטה סט משום שהיה פופולרי במדינות אחרות) קיבל ציון 1-. בדומה לעבודה על הדאטסט הקודם, ייצרנו דנדוגרם ובעזרתו בחרנו את מספר הקלאסטרים ב- Agglomerative Clustering.



רצינו מצד אחד לייצר הרבה קלאסטרים כדי שכל קלאסטר יהיה יותר מדויק, ומצד שני להמנע מקלאסטרים קטנים מאוד. לבסוף בחרנו ליצור 7 קלאסטרים . החלוקה של המדינות לקלאסטרים יוצאת שונה מאוד מזו שלפי הדאטה סט של השירים הרנדומליים. מלבד קלאסטר מובהק של מדינות דוברות ספרדית, אין הרבה דמיון בין שתי החלוקות. ואולם, גם כאן ניתן להסביר לפחות חלק מהקלאסטרינג במונחים של שפה ותרבות משותפות (למשל אירלנד ואנגליה שחולקות קלאסטר, או דנמרק, פינלנד, נורבגיה ושוודיה שחולקות קלאסטר).

ההבדל בקלאסטרינג הוכיח שהדאטה סט השונה אכן מספק תובנות שונות. לכן החלטנו לנסות לאמן מודלי רגרסיה על הדאטה החדש (ראו מחברת top_hits_regressions.ipynb). בדומה למה שעשינו עם הדאטה הישן, בחנו שימוש ברגרסיה ליניארית וב-Random Forest. אך בדומה לנסיונות הקודמים שלנו, המודלים נחלו כישלון חרוץ, וה- coefficient of determination R^2 היה לרוב קטן מ-0.2.

בניגוד לעבודה עם הדאטה סט הישן, הפעם לא ניתן לחשוד בחוסר עקביות פנימית, או בחוסר אמינות של הדאטה. הוא אמנם לא מדגם מייצג של כל השירים בעולם, משום שהוא מכיל רק שירים שהיו פופולריים מאוד במדינה כלשהי בשנת 2017, אך הוא עדיין גדול ומגוון יחסית, משום שחלק ניכר מהשירים בו היו פופולריים רק במדינות מסוימות ולא באחרות.

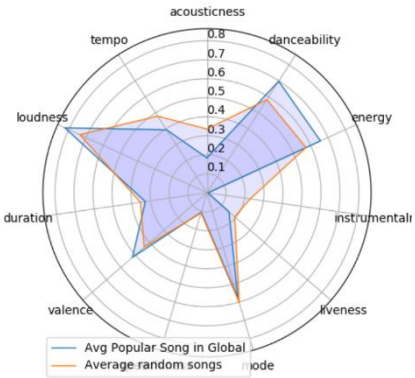
<u>סיכום - מחשבות מסלול מחדש</u>

חוסר ההצלחה הבולט של המודלים החזיר אותנו לשאלה היסודית – "האם ניתן לנבא את הפופולריות של שיר רק לפי המאפיינים המוזיקליים שלו?". ניתחנו את השאלה הזו והצענו כמה נימוקים איכותניים שמסבירים מדוע התשובה היא "לא":

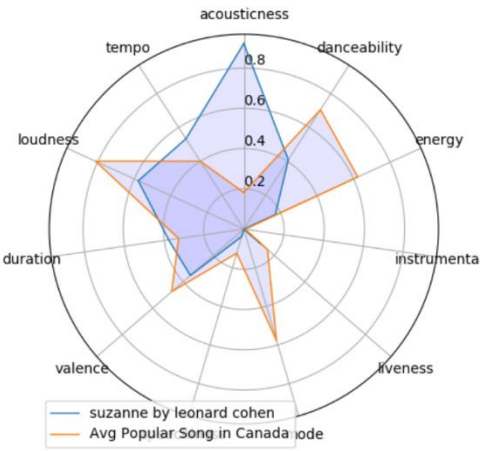
1. פופולריות האמן המבצע – אם אמן מפורסם ואמן אלמוני היו מפרסמים את אותו שיר באותו זמן, האם השיר היה זוכה לאותה פופולריות?
2. השפעה של פלטפורמות אחרות על הפופולריות של שיר – למשל שיר שהושמע בסרט/ פרסומת/ אירוויזיון/ אירוע ספורט וכד' בוודאי יהיה פופולרי יותר מאשר יכול להיות במידה ולא היה מושמע בהן.
3. שיווק ומיתוג השיר – התקציב שהוקצה לקליפ שלו, התקציב שהושקע בפרסום שלו (למשל בפלטפורמות כמו ספוטיפיי).
4. תוכן השיר – הנושא בו עוסק השיר, האם השיר פונה לקהל יעד רחב או לקבוצה מסוימת באוכלוסייה (כמו שירי דת, שירי ילדים, שירי כדורגל).

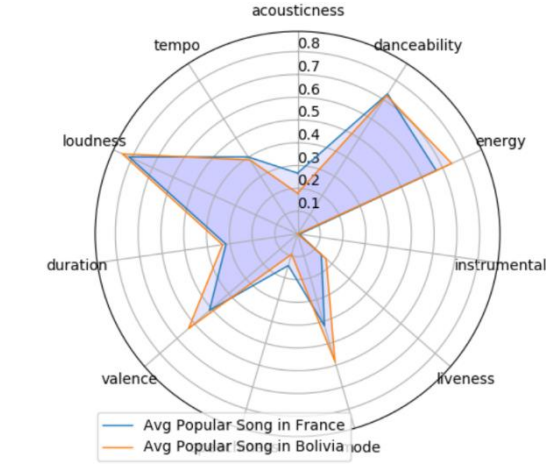
כלומר, אנחנו טוענות שניתן למצוא שירים עם מאפיינים מוזיקליים דומים עד זהים, שהפופולריות שלהם שונה באופן משמעותי. החלטנו לחזור לדאטה סט הראשוני שהשתמשנו בו, המכיל 116,000 שירים רנדומליים ואת המאפיינים המוזיקליים שלהם. בחרנו את השירים הפופולריים ביותר וחיפשנו את השירים הדומים להם ביותר לפי מרחק אוקלידי של וקטור המאפיינים המוזיקליים. ניתן לראות כי השירים הפופולריים ביותר דומים מאוד לשירים שאינם פופולריים כלל – השיר הדומה ביותר לשיר עם פופולריות 100 הוא בעל פופולריות 34, השיר הדומה ביותר לשיר עם פופולריות 95 הוא 23 וכך הלאה (ראו פירוט ב- songs_similarity.ipynb).

אז האם ניתן לומר שלמאפיינים המוזיקליים של שיר אין שום קשר לפופולריות שלו? כדי לבחון את הטענה הזו, הסתכלנו על השיר הפופולרי הממוצע (ממושקל לפי פופולריות) משנת 2017 והשווינו אותו לממוצע השירים מהמדגם הרנדומלי. מהגרף הסקנו שכן קיים קשר מסוים בין המאפיינים המוזיקליים לפופולריות של השיר. ייתכן כי המאפיינים המוזיקליים יכולים להוות תנאי הכרחי אך לא מספיק להצלחתו של השיר, בשילוב עם תנאים אחרים, חוץ מוזיקליים.



בחרנו בגרף ראדאר משום שהוא מאפשר לבחון את כלל המאפיינים המוזיקליים בבת אחת, וכך, למשל, להתבונן בשיר מסוים ביחס לשיר הפופולרי הממוצע במדינה כלשהי, להשוות שיר ל"שיר הממוצע", או להשוות בין השיר הפופולרי הממוצע במדינות שונות. השימוש בגרף הראדאר יכול להמחיש בצורה ויזואלית הבדלים בהעדפה המוזיקלית בין מדינות שונות – ולהראות, למשל, שבצרפת יש העדפה לשירים אקוסטיים יותר מבבוליביה, ובבוליביה ישנה העדפה לשירים אנרגטיים יותר מבצרפת. באופן דומה, הגרף מלמד אותנו שכדי שהשיר suzanne של לאונרד כהן יהיה פופולרי בקנדה של היום עדיף להפוך אותו לקצת יותר רועש, אנרגטי ורקיד וקצת פחות אקוסטי





כדי לחקור שאלות מסוגים דומים, יצרנו כלי שמאפשר למשתמש להסתכל על השיר הפופולרי הממוצע במדינה מסוימת (ממוצע ממושקל של השירים הפופולריים לשנת 2017), על השיר הפופולרי הממוצע בעולם (ממוצע ממושקל של השירים הפופולריים לשנת 2017), על השיר הממוצע בספוטיפיי, על שיר ספציפי לבחירת המשתמש (באמצעותו ייצרנו את שלושת הגרפים המובאים מעלה) ועל השילובים ביניהם. כדי להריץ אותו, היכנסו לתיקייה GUI והריצו את קובץ הפייתון gui.py. מוזמנים כמובן לצפות בדמו בקובץ Gui trailer.mp4, אבל אנחנו ממליצות בחום לשחק בו בעצמכם! הוא סיפק לנו שעות של הנאה.

<u>רעיונות לפרויקטי המשך</u>

1. להמשיך ולרדוף אחרי המתכון להצלחה של שיר:
   a. לבחון האם וכיצד משפיעה הופעה של שיר בסרט/פרסומת/תוכנית ריאליטי וכו' על הפופולריות שלו, והאם היא עוזרת לחזות את הפופולריות שלו בעתיד או רק משקפת את הפופולריות בעבר?
   b. לבחון האם המילים של השיר משפיעות על הפופולריות שלו? האם למשל שירי אהבה מצליחים יותר משירים אחרים בעלי מאפיינים דומים?
   c. האם פרסום ממומן של שיר בספוטיפיי למשל, אכן משפיע על הפופולריות שלו?
   d. האם גלגל"צ אכן קובעת בעצמה את הפלייליסט הישראלי? האם השיר הישראלי הפופולרי אכן משקף את בחירות העורכים ברשתות הרדיו הפופולריות או שלציבור הישראלי יש טעם מוזיקלי משלו?
2. להמשיך לבחון את המאפיינים המוזיקליים:
   a. לחקור את האבולוציה של מוזיקאים לאורך זמן - האם המוזיקה שלהם הופכת לאקוסטית יותר? או אולי לרקידה יותר? האם התהליך דומה אצל אמנים שונים?
   b. לחקור את האבולוציה של הטעם המוזיקלי במדינות שונות – האם הטעם המוזיקלי המועדף הופך לדומה יותר ויותר?
   c. האם קיימים שירים פופולריים מאוד שרחוקים מהשיר הפופולרי הממוצע? אם כן, מה הופך אותם לפופולריים?