מדעי הרוח הדיגיטליים – פרויקט גמר

שמות בשירה העברית

נטע כהן, אופיר כהן

על הפרויקט:

לשירים יש חלק מרכזי ביותר באופיין התרבותי של מדינות העולם בכלל, ומדינת ישראל בפרט. מטרת הפרויקט שלנו היא לחקור את נפוצות השמות השונים בזמר העברי, על מנת להבין מגמות ותהליכים תרבותיים אשר התרחשו בחברה הישראלית. המגמות שבחרנו לבדוק הן:

- האם לאורך הזמן ניתן להבחין בהפחתת השימוש בשמות תנכ"יים בשירים?
- האם גברים שרים בעיקר על נשים, ונשים בעיקר על גברים?
- מהם השמות הנפוצים ביותר בכל שנה?
- האם שמות של גברים נפוצים יותר משל נשים, או להפך?

על מנת להשיג מטרה זו, השתמשנו במגוון כלים שלמדנו בקורס – את מאגר השירים לקחנו מאתר "שירונט", ועליו הרצנו את מתייג השירים של ד"ר מני אדלר. בעזרת המתייג הוצאנו מכל שיר את השמות המופיעים בו, ובנינו שאילתה אל מאגר "Wikidata" שהכרנו בקורס על מנת לדעת מהו מין הזמר. את ממצאי הפרויקט בחרנו להציג בצורה ויזואלית בעזרת תרשימים, גרפים וצירי זמן, הממחישים את השינויים שהתרחשו לאורך השנים.

תיאור שלבי הפרויקט:

שלב 1: פרסור השירים

פרסור השירים מתוך אתר שירונט: https://shironet.mako.co.il/html/indexes/performers/

פלט:

1. תיקייה המכילה את כל מילות השירים.
2. קובץ json המכיל מידע נוסף על כל שיר – זמר/כותב, שנת יציאה, url (קישור לשירונט), כותרת (ממוספרת לכל זמר).

הקוד מצורף בקובץ ParseSong.py (נכתב ב-Python 2.7).

שלב 2: יציאת רשימות שמות

הרשימות נלקחו מהאתרים הבאים:

1. אתר הלמ"ס – דירוג השמות שניתנו לילדי שנת 2012-2015:
http://www.cbs.gov.il/reader/cw_usr_view_SHTML?ID=825
2. אתר מטרנה -
https://www.materna.co.il/%D7%94%D7%A8%D7%99%D7%95%D7%9F-%D7%95%D7%9C%D7%99%D7%93%D7%94/%D7%A9%D7%9E%D7%95%D7%AA-%D7%9C%D7%AA%D7%99%D7%A0%D7%95%D7%A7%D7%95%D7%AA
3. אתר שמות מהתנ"ך - http://tora.us.fm/tnk1/dmut/index.html

הקוד מצורף בקובץ Get_names.ipynb (נכתב ב-Python 3.6).

שלב 3: הכנת קבצי השירים להרצת המתייג

הסרת תווים (| ,\ ,/) ופיצול קבצים ארוכים (מעל 90 מילים) לקבצים קטנים.

הקוד מצורף בקובץ SplitLongFiles.py (נכתב ב-Python 3.6).

שלב 4: הרצת המתייג על השירים

שלב 5: תחקור הנתונים

פלט: 4 קבצי csv – עבור שמות גברים, נשים, שמות גברים מהתנ"ך ושמות נשים מהתנ"ך.

כל רשומה מכילה את השם (שנמצא בשיר), שם השיר, שם האמן, מין השם (שנמצא בשיר), שנת יציאה (של השיר), מינו של האמן, ו-url (משירונט).

עבור כל שיר, נבדק האם קיימת מילה שתויגה בישות I_PERS. במידה וכן, נבדק האם המילה קיימת באחת מרשימות השמות.

אם המילה קיימת הן באחת מרשימות השמות של הגברים והן באחת מרשימות השמות של הבנות, נבדק האם בסמוך למילה (2 מילים קדימה ו2 מילים אחורה) קיימת מילה שתויגה כמילה המשויכת למין זכר או נקבה, זאת על מנת לקבוע האם בשיר זה מדובר על שם של גבר או אישה. במידה והופיעה בסמוך למילה הן מילה שתויגה כמילה המשויכת למין זכר, והן מילה שתויגה כמילה

המשויכת למין נקבה, נקבע כי השם שייך ל2 המינים (השם הוכנס הן לקובץ השמות של הגברים והן לקובץ השמות של הנשים).

שנת יציאת השיר וה-url נלקחו מקובץ הjson שיצרנו בשלב 1.

מינו של האמן נלקח ע"ב ביצוע שאילתה לdbpedia.

שלב 6: חישוב סטטיסטיקות

חישוב סטטיסטיקות לצורך מענה על שאלות המחקר הבאות:

1. האם שמות מהתנ"ך נפוצים יותר בשירים ישנים לעומת שירים מודרניים?

עבור כל שנה נספרו כמה שמות מופיעים בשירים מכל קבוצה (גברים/נשים/תנ"ך..).

ניתן לראות את הפלט בקובץ - names_per_year.csv.

2. מהם חמשת השמות הנפוצים בכל שנה? מהם השמות הנפוצים ביותר?

עבור כל שנה נספר כמה פעמים מופיע כל שם.

ניתן לראות את הפלט בקובץ - top_names.json

3. האם שמות של נשים נפוצים יותר בשירים לעומת שמות של גברים?

עבור כל שנה נספרו כמה שמות מופיעים בשירים מכל קבוצה (גברים/נשים/תנ"ך..).

ניתן לראות את הפלט בקובץ - names_per_year.csv.

4. האם בשירים ששרים גברים המכילים שמות, מרבית השירים הם עם שמות של נשים?

האם בשירים ששרות נשים המכילים שמות, מרבית השירים הם עם שמות של גברים?

כך לדוגמה, אחוז שירי הגברים עם שמות בנות חושב לפי הנוסחה הבאה:

$$\frac{num_{sex_{female}} \cap num_{singer_{male}}}{num_{singer_{male}}}$$

כאשר: $num_{sex_{female}}$ – מספר השירים בהם יש שמות נשים.

$num_{singer_{male}}$ – מספר השירים שבוצעו ע"י גברים בהם מופיעים שמות.

הקוד מצורף בקובץ ComputingStatistics.py (נכתב ב-Python 3.6).

אתגרים ומגבלות:

1. למרבית השירים אין את נתון שנת היציאה.
2. חלק מהמילים שתויגו כשמות אינם כאלה. לצורך אימות נוסף ביצענו בדיקה האם השם קיים ברשימת השמות, דבר שהביא לשיפור קל בשמות שהתקבלו.
3. בחרנו לקחת כשם את הלקסמה של המילה כדי ששמות זהים לא יספרו פעמיים (לדוגמה: לאיתן, איתן, מאיתן). יתכן והדבר עלול להוביל לחילוץ שמות שגוי (כגון: רז מהמילה ברז).
4. בחישוב הסטטיסטיקות השונות – יתכן ואותו השם מאותו השיר נספר כמה פעמים, וזאת כיוון שבאתר "שירונט" השיר הופיע מספר פעמים – תחת המבצע, המלחין, הכותב וכו', כולם בעלי url שונה.