פרויקט גמר במדעי הרוח הדיגיטליים

**יצירת מאגר מידע על שירים ישראליים**

**מגישים : שי מחנאי 308009356**

**שני צלח 205681562**

**תאריך הגשה: 30.7.18**

על הפרויקט:

מטרת הפרויקט היא יצירת מאגר מידע על שירים ישראליים שיצאו החל מתחילת המאה ה-20 ועד שנת 2012.

את הפרוייקט ביצענו במסגרת קורס "מדעי הרוח הדיגיטליים" בהנחיית ד"ר יעל נצר אוניברסיטת בן גוריון.

המידע המוצג בפרויקט, עבור כל שיר: שם השיר, מבצע, שנת ביצוע לועזית ועברית, מלחין, משורר, מעבד מוזיקלי, בעלי תפקיד שונים, מקום הוצאה ושם חברת התקליטים. בנוסף, לכל שיר שמרנו גם את מספר המזהה (ID) שלו מתוך המאגר של הספריה הלאומית, זאת כדי לשמור על קונוטציה בין המקור לפלט.

תוצאות הפרויקט מוצגות בפורמט json ובפורמט csv.

הרעיון לביצוע הפרויקט נבע מגילוי הצורך למאגר נוח וידידותי של השירים בעברית. אמנם קיים מאגר של אתר שירונט, שהינו נוח לשימוש ומסודר, אבל אינו כולל את שנת הוצאת השיר, מה שפוגע בראיונות מחקר רבים שניתן לבצע על השירים.

בעזרת המאגר שבנינו ניתן לבצע חיפושים שונים אודות השירים, תוך יכולות מיון וסינון, וכן לבצע כל עבודת מחקר אחרת על מאגר השירים. אנחנו מגיעים מעולם התכנות, ולכן ראינו לנכון ליצור מאגר נוח שכזה, שיכיל מידע הן בפורמט csv - נוח למשתמשי קצה, והן בפורמט json - נוח יותר לצרכי תכנות.

בסיס המידע של הפרויקט הינו מאגר השירים של הספרייה הלאומית – קובץ xml המכיל כ-30 אלף שירים . זהו קובץ שאינו נגיש ונוח לקריאה של אדם , במיוחד אדם שאינו מתכנת ומכיר את נהלי שפת xml - שפת תגים. המאגר כתוב בתקן marc, תקן לייצוג ולהעברת מידע ביבליוגרפי אודות פריטים שונים.

יש לשים לב כי גם עם סיום הפרויקט, נותרו שירים רבים ללא כיסוי מלא של הנתונים. הסיבה העיקרית לכך היא חלקיות המידע במאגר של הספריה הלאומית. סיבה נוספת ומשנית היא כמות ה"רעש" שהיה סביב חלק מהמידע, דבר שהפריע לנו לסנן אותו באופן אוטומטי.

כמו כן נציין כי רוב מוחלט של השירים ה"בעייתים" הללו הם תפילות וברכות ושירים ביידיש. קרי - שירים שהחיפוש שלהם פחות רלוונטי אצל המשתמש הפוטנציאלי הממוצע, להערכתנו.

קבצי הפרויקט:

1. קובץ קוד Main.py - קוד שיוצר json מהמאגר של הספריה הלאומית. ה - json מכיל "פריטים" מסוג שירים, בכל פריט שיר יש מידע כמו: שם השיר, מבצע, מלחין, שנת הוצאת השיר,שם הוצאה לאור וכדומה.
2. קובץ קוד jsonToCsv.py - הקוד יוצר קובץ csv מה json שיצרנו קודם. בקובץ csv יש שורה עבור כל שיר ובכל עמודה יש פרטים על שיר כפי שמופיע בקובץ json.
3. מאגר השירים של הספריה הלאומית - shir1.xml.gz
4. פלט:

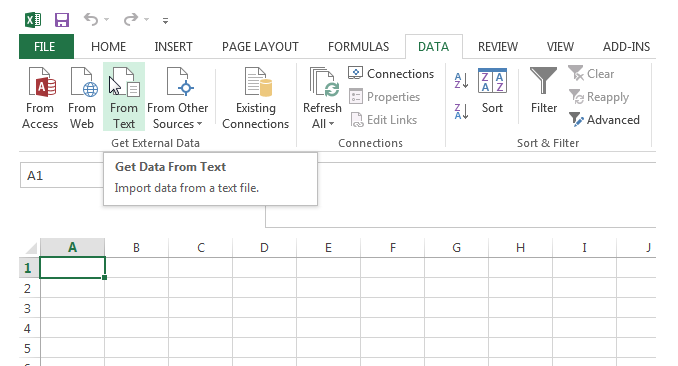
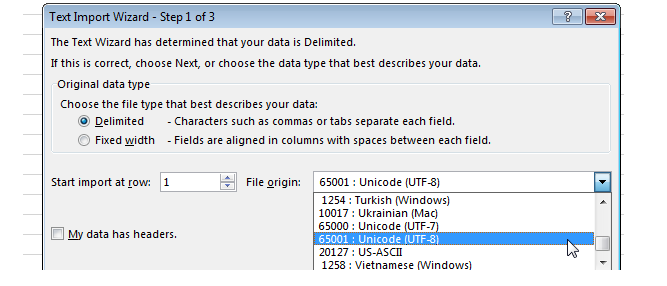
1.קובץ songs.json - הפלט של הקוד Main.py

2.קובץ songs.csv - הפלט של הקוד jsonToCsv.py

הוראות למשתמש:

1. הרצת הקוד Main.py על המאגר של הספריה הלאומית צריך להיות פורמט של קובץ ארכיון (.gz) הקוד מקבל כפרמטר את נתיב של המאגר באופן הבא : < python Main.py <path
2. הרצת הקוד jsonToCsv.py על מנת לקבל קובץ csv, הקוד מקבל את הפרמטרים הבאים : הנתיב של הקובץ json והנתיב של הקובץ csv .

python json\_to\_csv.py <json\_in\_file\_path> <csv\_out\_file\_path>

1. על מנת לפתוח את הקובץ song.csv בתוכנת excel יש לבצע את ההתאמות הבאות:
   1. פתח גיליון excel חדש
   2. לחץ על data -> from text
   3. בחר את קובץ songs.csv שנוצר כפלט של הקוד jsonToCsv
   4. קבע את הגדרות הקידוד להיות utf-8 באופן הבא:
   5. יש לקבוע את delimiter להיות פסיק (comma)

אתגרים, מגבלות ודילמות בביצוע הפרויקט:

1. קובץ מאגר המידע, קובץ xml של הספריה הלאומית, הינו קובץ עצום, כך שלא ניתן לבצע סריקה בצורה פשוטה. לכן, בסיוע המרצה בקורס, ד"ר יעל נצר, השתמשנו בספרייה xmltodict של python בשביל לבצע את הסריקה על קובץ tar.gz- כיווץ של הקובץ המקורי, כך שניתן היה לבצע עליו את הקוד במהירות רבה.
2. קושי נוסף הנובע מגודלו של הקובץ, היה הבנת המבנה שלו, שכן כל פריט הינו בעל תגים ומאפיינים שונים, ונגזרת מכך היא מבנה שונה של הפריטים. עובדה זו הקשתה על יצירת מבנה אחיד שיכלול את כל הפריטים.
3. בתוך תג אחד מסוים, יכול היה להתחבא מידע רלוונטי לקטגוריות שונות, כמו לדוגמא שנת ההוצאה של השיר, מיקום ההוצאה וחברת התקליטים.
4. אחד הקשיים הגדולים ביותר שהתמודדנו איתם היה מציאת שנת הוצאת השיר, הן באנגלית ובמיוחד בעברית. המידע היה חלק מטקסט נרחב יותר, כך שהיה עלינו לבנות ביטויים רגולריים מתאימים בשביל לזהות זאת.
5. התמודדנו עם הדילמה: האם ליצור מאגר מידע בעל נפח משמעותי, כזה המוציא יותר מידע ממאגר הספרייה הלאומית (לדוגמא - ביצוע מוקלט, מידע על האלבום אליו שייך השיר, שנת לידה של בעלי התפקידים וכו'), או ליצור מאגר מצומצם יותר, בעל הנתונים הרלוונטיים בעינינו עבור המשתמשים הפוטנציאליים בו.
6. המידע במאגר המקורי מפוזר על פני תוויות שונות, כך שלא ניתן לקבל באופן ברור ומהיר תוצאות רלוונטיות לחיפוש. לדוגמא: בשיר שעליו קיים מידע של מספר מבצעים או בעלי תפקיד שונים, המידע בקובץ הופיע תחת מספר תגים שונים, כאשר תחת כל תג היה מידע נוסף, שלא תמיד רלוונטי למשתמש שרוצה לקבל מידע מהיר על השיר. לכן במסגרת עבודתנו, קיבצנו את המידע על כל שיר תחת קטגוריות ברורות ומתאימות, כך שניתן לראות בבירור וללא מאמץ מהם בעלי התפקידים הנוגעים לשיר, או כל מידע רלוונטי אחר.
7. דילמה נוספת: האם לקבוע עבור בעל תפקיד מסויים את תפקידו, למרות שלא צויין כך במאגר: קיימים שירים בהם לא מצויין המבצע. למרות זאת, החלטנו לא לקבוע מבצע שהוא אחד מבעלי התפקידים שתפקידו אינו מוגדר במפורש. במידת הצורך, חיפוש שם המבצע ישלח את המחפש לבעל התפקיד הרלוונטי, ומשם לשיר עצמו.
8. התמודדנו עם תווים שונים (, ' " . / \ ? - וכו') והסרנו אותם כדי לשמור על אחידות וקלות החיפוש למשתמש. כמו כן, מאותה הסיבה, הורדנו בצורה אוטומטית הערות שונות שנכתבו בפריטים ולא היו אינפורמטיביים לצרכי המשתמשים במאגר.
9. עלתה השאלה - האם לגרוע ערכים שהאלגוריתם הניב עבורם תוצאות דלות (למשל - רק שם השיר). החלטנו שנשאיר למשתמש את האופציה לסנן פריטים כנ"ל ולהעלות בקובץ את כל התוצאות של ריצת האלגוריתם.

אפיון מאגר השירים ואבחנות:

* המבצע הפופלרי ביותר במאגר הוא : אריק איינטשטיין עם 324 שירים לאחר מכן חווה אלברשטיין עם 302 שירים
* המשורר הפופלרי ביותר במאגר הוא אהוד מנור שע"פ המאגר הלחין 416 שירים לאחר מכן זאב נחמה עם 162 שירים.
* המלחין הפופולרי ביותר הוא שלום חנוך עם 341 שירים ואחריו מתי כספי עם 298 שירים
* במאגר ישנם 5943 שירים ללא שנה לעוזית נתונה ו 3354 שירים ללא מבצע
* מספר השירים 29367
* מספר המבצעים במאגר - 3784