**סיכום הפרויקט:**

**על הפרויקט:**

הפרויקט הינו פרויקט במסגרת הקורס "מדעי הרוח הדיגטליים" בהנחיית ד"ר יעל נצר.

במסגרת הפרויקט ניתחתי אוסף של ספרי ילדים מאוניברסיטת חיפה.

הבחירה באוסף זה להיות הפרויקט שלי נבע מהאהבה שלי לספרי ילדים והרצון ללמוד יותר על ז'אנר ספרים זה.

במהלך העבודה למדתי מהו קובץ marc xml, כיצד בנוי הmarc xml ומה פירוש השדות השונים, השתמשתי בכלים שנלמדו בכיתה ובזכות זה למדתי המון על הנושא ובנוסף הרחבתי את הידע שלי בתכנות בפייתון, שימוש בספריות חדשות שלא ידעתי קיומן בפייתון וכיצד ניתן לעבוד עם קובץ xml.

**מטרת הפרויקט:**

מטרת הפרויקט הייתה לנסות לקבץ את הספרים באוסף לקטגוריות אב חדשות שיצרתי ולסווג אותם לפי עשורים (תאריכי הוצאה לפועל).

בפרויקט זה ניסיתי להראות את הקשר של קטגוריה מסוימת לעשור ההוצאה של הספר, כלומר האם קטגוריה מסויימת הייתה פופולרית יותר בעשור מסוים לעומת עשור אחר.

בנוסף, רציתי לבדוק איזה קטגוריות היו פופולריות בכל עשור.

**תהליך הפרויקט:**

לפני שניגשתי לקובץ marc xml שקיבלתי הייתי צריכה ללמוד על השדות שלו ולהבין אותם על מנת שאוכל לעבד את הקובץ בצורה המיטבית ביותר.

בעזרת הכלי open refine, שלמדנו עליו בכיתה, יכולתי לראות את כל השדות שיש בקובץ.

בעזרת האתר של Library of Congress למדתי מה כל שדה מייצג, לדוגמא השדה 100 מייצג את שם הסופר.

בעזרת שני המקורות הללו גיליתי שהשדות שעליהם אעבוד יהיו שדה 260 (שמייצג שנת הוצאה של הספר) ושדה 650 (שמייצג את הקטגוריות שמשויכות לספר) ויכולתי לעבד את הקובץ בצורה הטובה ביותר.

כמו שנכתב לעיל הקובץ שקיבלתי מאוניברסיטת חיפה היה קובץ marc xml ועשיתי לו עיבוד ראשוני בעזרת הכלי open refine.

1. בעזרת open refine ביצעתי איחוד של קטגוריות אשר זהות במהותן אך שונות בצורה כלשהי בכתיבתן, לדוגמא איחוד של הקטגוריות שמים ושמיים.
2. לאחר מכן ייצאתי מספר קבצים:
   1. קובץ שמכיל את כל השדות של כל רשומה.
   2. קובץ שמכיל רק מספר רשומה ואת הקטגוריות המשויכות לה.
   3. קובץ שמכיל רק את מספר רשומה ואת שנת ההוצאה שלו.
3. תחילה מצאתי את כל הקטגוריות בעברית שיש לספרים באוסף. זה נועד על מנת שאוכל לסווג את הקטגוריות לקטגוריות אב החדשות שיצרתי (נעשה בתוכנית נפרדת משאר התוכנית).
4. לאחר מכן סיווגתי כל רשומה לכמה פרמטרים:
   1. כל הקטגוריות שלה בעברית, על מנת שאוכל להצמיד אחר כך את הקטגוריות החדשות שנוצרו.
   2. שנת ההוצאה של הספר.
5. אחר כך פרסרתי את הקובץ המלא (עם כל השדות) למילון חדש שמכיל את המידע הנדרש.
6. ביצעתי הצלבה בין מספר הרשומה לבין הקטגוריות החדשות שצריך להוסיף לכל רשומה ובין מספר הרשומה לבין עשור ההוצאה של הספר והוספתי את המידע הזה למילון משלב 6 (ניתן לייצא את המילון הזה לקובץ csv או קובץ json על מנת שאפשר יהיה להשתמש במידע בהמשך).
7. יצרתי מילון חדש שמכיל רק את השדות הרלוונטיים לייצוג הגרפי (מספר הרשומה, הקטגוריות החדשות שנוספו ועשור ההוצאה).
8. יצרתי 2 מילונים חדשים שבעזרתם נבנו הגרפים בשלב הבא:
   1. המילון הראשון הוא מילון שהמפתח שלו הוא הקטגוריות והערך של כל קטגוריה הוא מילון נוסף שהמפתח שלו הוא עשור ההוצאה והערך שלו הוא כמות הספרים הרלוונטיים.
   2. המילון השני הוא מילון שהמפתח שלו הוא העשורים והערך של כל עשור הוא מילון נוסף שהמפתח שלו הוא קטגוריה חדשה והערך שלו הוא כמות הספרים הרלוונטיים.
9. לבסוף השתמשתי בספרייה בפייתון בשם matplotlib.pyplot על מנת לייצר את הגרפים בצורה הנוחה והיעילה ביותר.
10. פתחתי עמוד בgit בו ניתן למצוא את כל המידע הרלוונטי לפרויקט.

לתוכנית הראשית ניתן 4 קבצים (parse\_xml.py):

1. קובץ שמכיל רק את המספר רשומה של הספר והקטגוריות שלו: recordID\_categories.txt.
2. מיפוי של קטגוריות מקוריות לקטגוריות אב החדשות שיצרתי: old\_tags\_to\_new\_tags.txt.
3. הקובץ המלא עם כל השדות של הרשומות: childrens\_collection.txt.
4. קובץ שמכיל רק את המספר רשומה של הספר ואת שנת ההוצאה שלו: recordID\_published\_year.txt.

לתוכנית הנפרדת (שלב 3, parse\_tags.py) ניתן את הקובץ recordID\_categories.txt.

**מסקנה מהפרויקט:**

בעזרת המידע החדש שנוסף לקובץ המקורי ניתן יהיה לבצע הסקות מושכלות יותר ביחס לעשורי ההוצאה של הספרים ו/או לקטגוריות החדשות שנוספו.

הגרפים שיצרתי מאפשרים לראות את המידע החדש בצורה ויזואלית נוחה בהרבה מקובץ json או קובץ csv ומאפשרת להמשיך לנסות להסיק הסקות ותובנות מהמידע המוצג.