**סיכום**

העבודה שלי התחלקה למס’ שלבים:

**בחירת הנושא וכתיבת רקע היסטורי עליו** – כבר בשיעור הראשון כשנודע לי שנושא הפרויקט הוא על מסמכים בארכיון המדינה, חשבתי לבדוק את הנושא של פרשת ילדי תימן. זאת משום שכאמור היה לי ידע קודם בנושא (בעקבות עבודה שעשיתי בעבר) וידעתי שיש ערפל כבד סביב הפרשה הזו, לכן ידעתי שלנתח מידע של רשימת ילדים ולהפוך אותה לנגישה וחכמה יותר יהיה נושא מעניין ומאתגר.

**המרת מסמך pdf לקובץ טקסט** – כאן נכנס השימוש בטכנולוגיית ocr להמרת קבצי pdf. בשלב הזה נחשפתי למס’ תוכנות למשל abbyy finereader ועוד מספר אתרים, התוצאה הטובה ביותר התקבלה ע"י אתר שנקרא newocr, אך תרגום מסמך בעברית שמטושטש בחלקו אינו פשוט והייתי צריך לתקן את התוצאה של הocr – חלק מהתיקונים היו בקוד (למשל הפרדה בין כל משבצת בטבלה על ידי | או המרת גרשיים " למה שכתוב בשורה מעל) וחלק מהתיקונים היו ידניים (למשל שמות שלא נקראו טוב).

**המרת המידע לקבצי tsv,json,xml** – בשלב הזה התחלתי מלהמיר את הקובץ טקסט שהתקבל בשלב (2) לקובץ tsv באמצעות קוד פשוט בpython. לאחר מכן הכל כבר היה הרבה יותר פשוט, יכולתי לעבוד על הקובץ בopenRefine ולעשות שם שינויים פשוטים (למשל להמיר את כל עמודת הגילאים לפי יחידות מידה של שנים). הצעד הבא היה להמיר את הקובץ tsv לקבצי json וxml, לצורך כך השתמשתי במס’ אתרים שעושים זאת (בסוף המסמך צירפתי את כל התוכנות והאתרים בהם נעזרתי), בנוסף ניתן לייצג את המידע הנ"ל על ידי תצוגת tree (למשל הצגת קובץ json בצורת tree).

**ניתוח המידע ע"י גרפים, קשרי משפחה ורשימת נפטרים** – לאחר שקיבלנו ייצוגים חכמים למידע היה הרבה יותר פשוט לנתח את המידע, באמצעות תוכנית בpython עם ספריית matplotlib יצרתי מספר גרפי עמודות, למשל גרף שמראה מאיפה הגיעו אותם נעדרים (אכן לא כולם מתימן, יש מגוון מדינות שמהם היו נעדרים למרות שכמובן הרוב מתימן), גרף של גילאים ועוד... בנוסף באמצעות מעבר על שמות המשפחה של הנעדרים יצרתי רשימה של נעדרים בעלי אותם שמות משפחה, יתרה מזאת באתר הוספתי קישורים לכל אחד שיש עליו מידע לאתרים כמו "עמותת עמרם" , מסמכים ספציפיים בארכיון המדינה ועוד. בנוסף, יצרתי רשימה של ילדים אשר נכתב ברשימה שהם נפטרו (כתוב אצלם "נפל").

**יצירת אתר להצגת העבודה** – לאחר כל השלבים הנ"ל, יצרתי אתר אשר בו יכולתי להציג בצורה מסודרת ויעילה את ממצאי העבודה והשלבים בה היא נעשתה. בנוסף, נפתח repository בgithub שם מרוכזים כל ממצאי העבודה.

**רפלקציה אישית**

בתחילת הסמסטר לא הייתי בטוח כל כך מה אפשר לעשות עם מסמך (רשימה במקרה שלנו) pdf ואיך אפשר לעבוד עם זה. במהלך הסמסטר, נחשפתי להמון טכנולוגיות ותוכנות שלא הכרתי – למשל טכנולוגיית המרת pdf ocr, תוכנת openRefine ועוד הרבה אתרים ותוכנות שנעזרתי בהם (מצורף למטה). בנוסף, בשל היכרותי עם הנושא שילבתי בעבודה גם רקע כללי וידע על הפרשה עצמה, מה שיצר שילוב מעניין של עבודה עם טכנולוגיה וכלים תכנותיים עם עבודה היסטורית על פרשה מסתורית. בנוסף, יצא לי לעבוד עם ייצוגי קבצים בפעם הראשונה xml,json,tsv וניתוחם בpython עם ספריות שלא הכרתי ולמדתי עליהן במהלך הפרויקט. לסיכום, אני מרוצה מאוד שבחרתי לעשות את המיני פרויקט במדעי הרוח הדיגיטליים, ומרגיש שלמדתי המון תוך כדי עבודה מעניינת הן בצד הטכנולוגי והן ובצד ההיסטורי.

**תוכנות ואתרים בהם נעזרתי:**

openRefine, newocr, abbyy finereader, convertio, onlinexmltools, codebeautify, csv library, matplotlib library, wix, convertsimple