סמסטר ב', 2021

נושאים במדעי הרוח הדיגיטליים

פרסומים אחרונים בארכיון

יגאל אדרעי ויעל גיסלר

הצעת הפרוייקט:

בשנים האחרונות החל ארכיון המדינה לעבור תהליך דיגיטיזציה, על מנת להתאים את עצמו לעולם שבו הציבור רגיל לצרוך מידע באפיקים דיגיטליים וחברתיים.

מאחר ויש בארכיון מאות מיליוני עמודים (שהולכים ומתרבים) זוהי משימה שאורכת זמן רב.

כחלק מתהליך זה, התיקים החדשים שנסרקים מועלים לרשימת "פורסם לאחרונה" המתעדכנת אחת לחודש, ומרכזת את התיקים לפי החודש בו נסרקו.

התיקים ברשימה זו אינם ממויינים ולכן החיפוש ברשימה אינו מזמין ואינו יעיל, ואינו מאפשר מעקב אחר תיקים בנושאים מסויימים.

בפרוייקט זה כתבנו כלי המקבל קישור לעמוד של חודש מסויים, ומלחץ את המטא דאטה של התיקים ברשימה של חודש זה.

על ידי מימוש הפרוייקט, אנו מקווים להציע חווית משתמש טובה יותר ולאפשר למשתמשים מזדמנים באתר למצוא תיקים חדשים בתחומי העניין שלהם בקלות.

רקע היסטורי:

המידע המובא בחלק זה נלקח מדו"ח הגנז על נושא החשיפה הארכיונית, אשר פורסם בחמישה עשר בינואר 2018.

בשנת 2011 ארכיון המדינה החל לבצע רפורמה מקיפה בדרכי העבודה שלו, בהובלת גנז המדינה דאז יעקב לזוביק.

כתוצאה מרפורמה זו גבר השימוש בחומרי הארכיון ואיתו חל גידור במספר המשתמשים בו.

עד שנת 2018 נסרקו למעלה ממאה מיליון עמודים לארכיון, ומספר זה הולך וגדל מדי יום.

כל התיעוד הדיגיטלי של הארכיון נמצא במערכת התפעולית של הארכיון, מערכת רימון. ניתן להסתכל על מערכת רימון כאילו היא הארכיון עצמו, כיוון שכל התיעוד הדיגיטלי נמצא בה, יחד עם כל הסריקות. במערכת זו נמצא גם כל המטה-דאטה, כלומר כל המידע אודות החומר הארכיוני: תאריכים, תיאורים ועשרות סוגי מידע נוספים.

חשוב להבין כי תיק בארכיון המדינה מכיל בממוצע כ80-100 עמודים. עד לפני מספר לא רב של שנים, הסתפק הארכיון בתיוג התיקים באמצעות כותרת שנוצרה עבורו במשרד המוצא, שיכולה הייתה להיות אינדיקטיבית או חסרת משמעות כלל (למשל "שונות"), ובדרך כלל אינה הכילה מידע מידע מדוייק אודות התיק או הנושאים הנדונים בו. ללא תאורים עקביים, לא ניתן לבצע חיפוש אחיד על פני כל הארכיון, ובעצם לא ניתן למצוא את מה שמחפשים באופן מלא גם לאחר חיפוש ממושך.

מערכת רימון פתרה את הבעיה הזו. היא איפשרה לעובדי הארכיון לייצר "ישויות ידע": תיאורים קצרים אודות אישים, אירועים, ארגונים ועוד, ולהשתמש התאורים בכל התיקים העוסקים בהם.

במילים פשוטות, מערכת רימון איפשרה 'לתייג' את המידע ולבצע חיפוש בארכיון על פי נושאים.

מהלך הפרוייקט והכלים שהיו בשימוש:

העבודה על הפרוייקט נחלקה לשני חלקים עיקריים: חילוץ המידע מרשימת הפרסומים האחרונים ואיחסונו במאגר נתונים, ויצירת ממשק משתמש עבור המידע.

הכלים העיקריים בהם השתמשנו:

1. פייתון
   a. beautifulSoup
   b. selenium
   c. requests
   d. pymongo
2. ג'אווה סקריפט
   a. React
   b. mongoose

<u>חילוץ המידע</u>

לשם חילוץ המידע כתבנו תכנית פייתון המקבל כתובת url של הפרסומים האחרונים בחודש מסויים, ובעזרת selenium ו-beautifulSoup מחלץ את הקישורים של כל התיקים שפורסמו באותו חודש.

לאחר מכן, לאחר צפיה במקור הדף של התיקים, הצלחנו למצוא בשרת הארכיון קובץ מקור מסוג json ממנו נבנה עמוד התיק. השתמשנו בספריה requests על מנת לקבל את התוכן של קובץ זה. מהקובץ הזה חילצנו את המטה-דאטה של התיקים.

הרצנו את תכנית זו על כל החודשים ברשימת הפרסומים האחרונים.

<u>איחסון המידע</u>

על מנת לאחסן את המידע עבור כל תיק, בנינו מאגר נתונים ב-mongoDB ובו בנינו אוסף עבור כל חודש.

העלינו את התיקים למאגר הנתונים באמצעות הספריה pymongo.

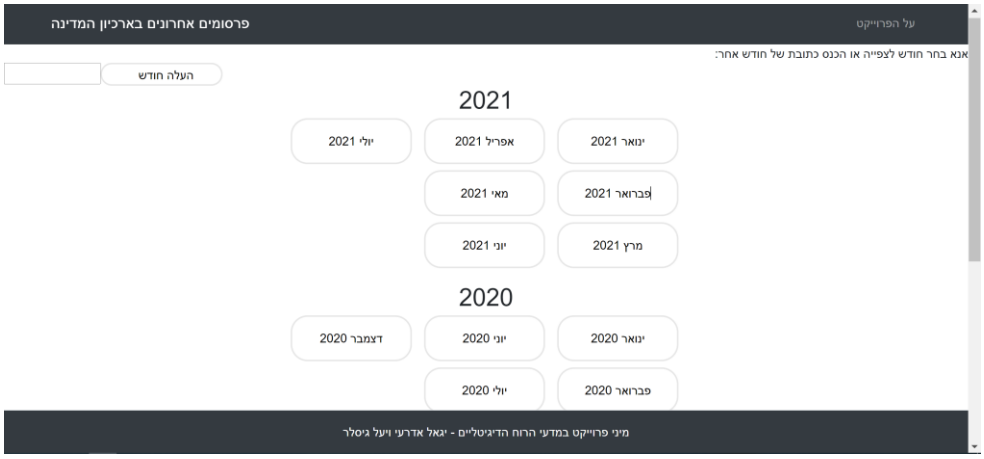
<u>הצגת המידע</u>

על מנת לייצר ממשק משתמש נוח, בנינו אפליקציית ווב באמצעות ריאקט וג'אווה סקריפט.

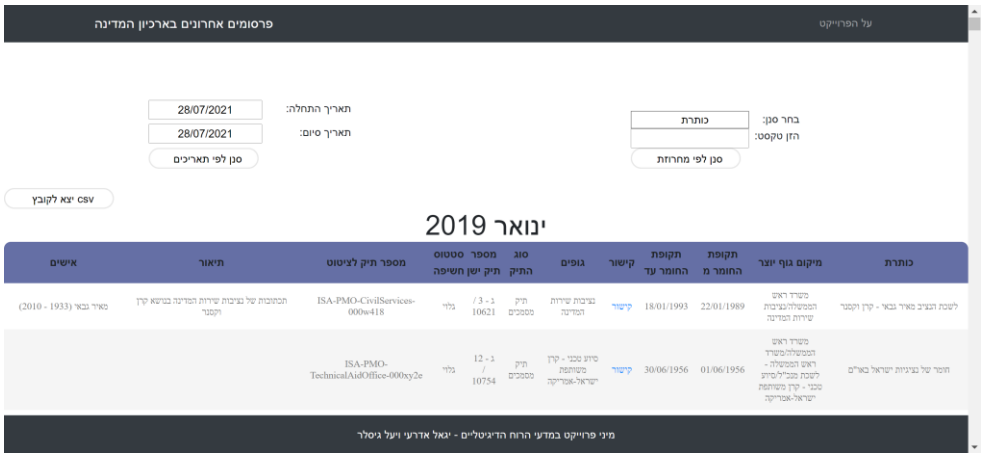
השתמשנו בספרייה mongoose על מנת לצפות בתוכן מאגר הנתונים שבנינו.

האפליקציה בנויה מ3 קומפוננטות:

1. קומפוננטה ראשית, בה ניתן לבחור חודש לצפיה או להעלות קישור לחודש חדש שאינו ברשימה, הקורא לקוד הפייתון בלחיצה על "העלה חודש".



2. קומפוננטה המאפשרת לצפות בתיקים של חודש מסויים, לפלטר אותם ולייצא אותם לקובץ CSV.



3. קומפוננטה בה ניתן לצפות בהסבר קצר על הפרוייקט ולפנות לקוד המקור של התכנית.

רפלקציה ומסקנות:

פרוייקט זה הוא למעשה הפרוייקט הראשון בתואר בוא לא קיבלנו פריימוורק או תכנית עבודה ובנינו אותו בעצמינו לגמרי מתחילתו ועד סופו.

לכן, מעבר לכלים הטכנים אותם רכשנו ולמדנו במהלך העבודה על הפרוייקט, למדנו גם על תכנון וחלוקת עבודה וניהול זמנים.

מבחינה טכנית, עבור שנינו זו הייתה התנסות דיי ראשונית בעבודה עם כלי אוטומציה (סלניום) ובפיתוח אפליקציות ווב. נדרשנו לחקור הרבה בעצמינו וללמוד להשתמש בכלים ובספריות חיצוניות באופן עצמאי.

כחלק מתהליך הלמידה על השימוש בכלים אלו, שיפרנו ושידרגנו את התכנית שלנו במהלך העבודה על הפרוייקט. בתחילת הפרוייקט, חילוץ המידע מתיקים שפורסמו בחודש אחד (סדר גודל של 2000-3000 תיקים) ארך בערך כשלוש שעות, וכן לא עבד במלואו. הצלחנו להוריד את זמן הריצה לסדר גודל של 20 דק ולייעל את התהליך באופן משמעותי.

לסיכום, העבודה על הפרוייקט איפשרה לנו לנהל פרוייקט בשלמותו ולהתנסות בכלים חדשים. העבודה עם כמויות גדולות של מידע דרשה מאיתנו סבלנות וקפדנות, ודחפה אותנו לייעל ולשפר את הקוד שלנו גם כשהגענו כבר למוצר עובד, אך לא טוב מספיק.

בנוסף במסגרת העבודה על הפרוייקט נחשפנו באקראיות לתיקים שונים ומגוונים בארכיון, שהרחיבו את אופקינו בנושאים שונים, שלא היינו נחשפים אליהם בפלטפורמה אחרת במהלך התואר.