**איתור אוטומטי של תאריכים בקורפוס החקיקה**

**פרויקט הדיגיטציה של החקיקה**

**מדעי הרוח הדיגיטליים - סמסטר א' תשפ"א, אוניברסיטת בן גוריון**

**מגישים: נדב שקד וניצן הוכמן**

**מטרת הפרויקט**

ברשותנו קבצי XML של חקיקה המתויגים בפורמט AKN.

מטרת הפרויקט היא לזהות ולתייג תאריכים בתוך מסמך החוק, תוך התייחסות לתפקיד התאריכים, לקשר ביניהם ומידע נוסף.

**תיאור הפרויקט במונחים של מדעי הרוח הדיגיטליים**

החוקים שמורים בתוך מבנה הxml שהינו חצי-מובנה – כלומר, המידע שמור בצורה בה הוא מחולק לתכונות, אך לא בצורה אחידה. חלק מהתיוגים יכולים שלא להופיע ועדיין המידע יהיה קריא לבני אדם.

בעזרת מבנה זה יכולנו לגשת לאזורים אותם אנו רוצים לנתח בחוק, וגם להוסיף את תיוג התאריך.

בפרויקט זה הרעיון הוא להעשיר את המידע שניתן לקבל מהדיגיטציה של קבצי החוק, ולכן אנו מתייגים את התאריכים תוך שאנו מוסיפים להם תכונות, וכך מוסיפים מידע נוסף (מטה דאטה) על התאריכים בחוק.

האתגר המרכזי בפרויקט היה זיהוי התאריכים בטקסט מכיוון שיש דרכים רבות לכתיבת תאריכים, וגם לזהות משמעויות של תאריכים בעזרת ביטויים מסוימים וההטיות המורפולוגיות שלהם הנמצאים לפני או אחרי כל תאריך שזוהה. כדי להתגבר על אתגר זה השתמשנו בביטויים רגולריים.

**פתרון הבעיה**

בתוך החוקים התאריכים יכולים להופיע בצורות שונות:

* בין המספרים יכולים להופיע סימנים שונים כגון - . , \ / ...
* סדר הופעת חלקי התאריך יכול להיות שונה – yyyy-dd-mm / dd-mm-yyyy / dd-mm / ...
* התאריך יכול להופיע בתור תאריך לועזי או בתור תאריך עברי
* התאריך יכול להיות מורכב ממילים ו/או מספרים – 21 לנובמבר , כ"ט בחשוון תשפ"א, הראשון לשני 2020 ...
* המילים יכולות להיות כתובות בדרכים שונות : חשון / חשוון, מרץ / מרס, תישע / תשע ...
* בין חלקי התאריך יכולים להופיע מילות קישור / סימנים פיסוק כגון : **ה**ראשון **ב-**נובמבר, **מה** 22 **ל**נובמבר **-** 1930...

בנוסף, לתאריכים בחוקים יכולים להיות משמעויות שונות, ולעיתים ניתן לזהות אותן על ידי מילות מפתח המופיעות בסביבת התאריך שנמצא, למשל: **החל מ** 4 לדצמבר 2011 **עד** 5 למרץ 2012 ...

תהליך העיבוד:

בנינו תהליך עיבוד אוטומטי בעזרת שפת Python, המקבלת כקלט קובץ XML של חקיקה המתויג בפורמט AKN, ומחזירה קובץ XML עם הרחבה עבור תיוג של תאריכים בחוק.

התהליך מחולק לשלושה חלקים עיקריים -

1. זיהוי התאריכים

תאריכים מופיעים בצורות שונות ( כפי שהראינו לעיל ), ולכן על מנת לזהות אותם בתוך מבנה החוק ה XML-י בצורה הטובה ביותר, השתמשנו בביטויים רגולריים שבנינו בעזרת Regex כדי לזהות כמה שיותר צורות שונות של תבניות.

בנוסף, ישנם תאריכים בעלי משמעות אותם ניסינו לזהות על פי דפוס מסוים. עבור משמעויות מסוימות בחוקים יצרנו מילון שמכיל את כל מילות המפתח.

כדי למקסם ולהקל על תהליך בניית הביטויים הרגולריים השתמשנו במבנה המילון. עבור כל תת ביטוי אגרנו את האפשרויות השונות להופעתו בתוך המילון שבסופו של דבר הפכנו בעזרת פונקציות עזר שונות לביטוי רגולרי. הסיבה לשימוש במילון היא שניתן להוסיף מופעים נוספים של תתי ביטויים בקלות וכך להרחיב את המאגר של הביטויים הרגולריים שמזוהים על ידי האלגוריתם, זהו אפיק פשוט למשתמש להוסיף ולשפר את האלגוריתם.

1. עיבוד התאריכים

בחלק זה של הקוד רצינו לעבד את המידע שיש לנו על התאריכים שאספנו:

* זיהינו את החלקים השונים של התאריך - החלק שמייצג את היום, החלק שמייצג את החודש, והחלק המייצג את השנה.
* ייצגנו את התאריכים בצורה YYYY-mm-dd למען אחידות, וכפי שדרוש בפורמט AKN.
* עבור תאריכים עבריים שהאלגוריתם זיהה, השתמשנו ב API על מנת להמירם לתאריך לועזי.
* זיהוי חזרות של תאריך מסוים, עבור כל חזרה על תאריך אשר הופיע בקובץ קודם לכן הוספנו  
  התייחסות לתיוג של המופע הראשון של התאריך.
* זיהוי משמעות של תאריך, עבור תאריכים שמשמעותם התחלה של אירוע מסוים הוספנו סימן המעיד על כך, כך גם עבור תאריכים שמשמעותם סיום של אירוע מסוים.

1. תיוג התאריכים

עבור כל תאריך שהאלגוריתם זיהה, סימנו בעזרת התגית ה date בפורמט AKN, והוספנו את המידע שאספנו בתהליך העיבוד על התאריכים, בצורה הבאה:

* eId - כל מופע של תאריך סומן ב id ייחודי.
* date - כל מופע של תאריך מחזיק את התאריך בפורמט YYYY-mm-dd.
* alternativeTo - כל מופע של תאריך אשר כבר הופיע קודם לכן יחזיק את תכונה זו המכיל את ה id של התגית של המופע הראשון של התאריך.
* refersTo - כל מופע של תאריך בעל משמעות אותה זיהינו יחזיר את תכונה זו המכילה את הטיפוס של המשמעות.

לאחר יצירת התיוג הכנסנו אותו למקום המתאים בקובץ ה XML של החוק.

**הערכת התוצאות:**

להלן תוצאות מהפעלת האלגוריתם על קבצי חוק:

1.

<p>  
תקופת החירום נקבעה מיום **26.8.2020** עד יום **26.2.2021** או עד תום תקופת מצב החירום שהוכרז לפי  
</p>

ביטוי שאומר שהתאריך הוא תאריך סוף של אירוע מסוים

ביטוי שאומר שהתאריך הוא תאריך התחלה של אירוע מסוים

<ns0:p>  
תקופת החירום נקבעה מיום   
**<date eId="date\_1" date="2020-08-26" refersTo="#start">26.8.2020</date>**   
עד יום  
 **<date eId="date\_2" date="2021-02-26" refersTo="#end">26.2.2021</date>**   
או עד תום תקופת מצב החירום שהוכרז לפי  
</ns0:p>

2.

<p>  
הממשלה, לפי הצעת שר האוצר, תגיש לכנסת עד יום **י"ב בחשוון התשע"ח (1 בנובמבר 2017),** דוח לעניין קיום או אי־קיום הפרש הוצאות או הפרש גירעון (בסעיף זה – דוח לעניין קיום הפרשים); ועדת הכספים של הכנסת תקיים דיון בדוח לא יאוחר מהמועד האמור; הוראות לעניין הגשת הדוח ייקבעו בחוק.  
</p>

חזרה על התאריך הקודם אך מומר לתאריכו הלועזי

ביטוי שאומר שהתאריך הוא תאריך התחלה של אירוע מסוים

<ns0:p>  
הממשלה, לפי הצעת שר האוצר, תגיש לכנסת עד יום  
 **<date eId="date\_1" date="2017-11-01" refersTo="#end">י"ב בחשוון התשע"ח</date>**  
 **(<date eId="date\_2" date="2017-11-01" alternativeTo="date\_1">1 בנובמבר 2017</date>),**   
דוח לעניין קיום או אי־קיום הפרש הוצאות או הפרש גירעון (בסעיף זה – דוח לעניין קיום הפרשים); ועדת הכספים של הכנסת תקיים דיון בדוח לא יאוחר מהמועד האמור; הוראות לעניין הגשת הדוח ייקבעו בחוק.  
</ns0:p>

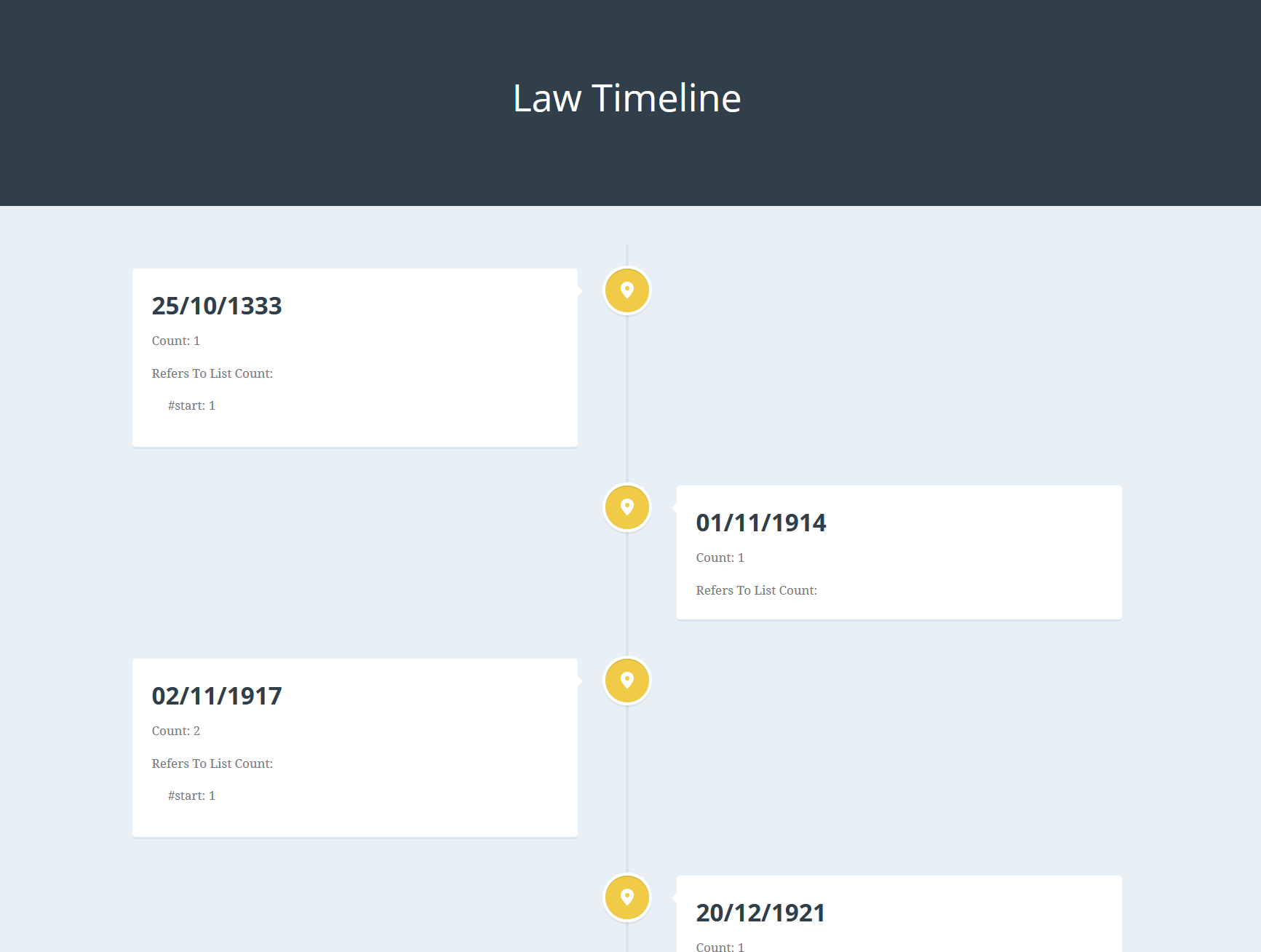
לפי בדיקותינו, זיהוי התאריכים עצמם נעשה יחסית בהצלחה, יחד עם זיהוי תאריכים זהים (עבריים ולועזיים). הסיבה המרכזית לכך שאין דוגמאות לתאריך שלא זוהה היא שעבור כל כישלון כזה עדכנו את האלגוריתם כך שיזהה את התאריך. במקרים רבים היה מדובר בהוספה פשוטה למילון של תת ביטוי.  
נציין כי זיהוי משמעות התאריכים יכול להיות מקיף יותר. לתאריכים בחוקים ישנם תפקידים שונים – החלת חוק, תיאור תקופה מהעבר או תיאור תקופה מהעתיד, וכו'. זיהוי תפקידים אלו דורש עבודה נוספת. עם זאת, הבסיס לעבודה זו קיים.

**הרחבה – הצגה ויזואלית של תאריכים**

הוספנו לפרויקט את האפשרות להציג את התאריכים שתויגו במהלך המעבר על החוק בצורה ויזואלית דרך ממשק web שנראה כך:



בנוסף, אפשרנו להציג את התאריכים יחד עם התכונות שלהם:



**סיכום**

המשימה שלנו הייתה זיהוי ותיוג של תאריכים בתוך מסמך חוק ברובד התוכן והמשמעות. על מנת לקיים אותה, עבדנו עם כלים חדשים כגון ביטויים רגולריים – אשר עזרנו לנו בזיהוי התאריכים ומילות הקישור, וכלים לפרסור המבנה הxml-י – אשר עזרו לנו לעבור על המבנה בצורה מסודרת כדי למצוא את התאריכים, וגם להוסיף למבנה את התיוג החדש. למדנו על סטנדרט הAKN, על תהליך המחשבה בתיוג מסמכי xml – מה המידע שניתן להוסיף, איזה מידע הוא רלוונטי וכו'.

במהלך עבודתנו נכחנו לדעת כי יש לתאריכים משמעויות רבות בחוקים, כפי שציינו למעלה. חשבנו על דרכים להביא את המשמעויות האלה לקדמה וליחס להם מקום בתיוג התאריכים, אך הדבר דורש עבודה מרובה ולכן בפרויקט זה אין מענה לכל המשמעויות, אך אנו מאמינים שהדרך לזיהוי משמעויות נוספות סלולה.

העבודה הייתה מאתגרת ומהנה, ואנו מרוצים מפרי עבודתנו ומרגישים שלמדנו המון בדרך.

**מקורות**

<https://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html>   
Akoma Ntoso Version 1.0. Part 1: XML Vocabulary

<https://docs.python.org/3/library/re.html>   
re.py documentation

<https://www.hebcal.com/home/195/jewish-calendar-rest-api>   
Jewish calendar REST API

<https://www.baeldung.com/java-date-regular-expressions>   
RegEx for matching Date Pattern