איתור אוטומטי של תאריכים בקורפוס החקיקה

מגישות: נועה תשובה וקים טלדן

מטרת הפרויקט

נושא על - תיוג סעיפי חקיקה או רכיבים בתוך סעיפים על פי התוכן שלהם

שאלה מרכזית - איך מתייגים את הטקסט באופן אוטומטי כאשר הכללים לא ודאיים?

רקע - חוקים מכילים תאריכים וזמנים המופיעים בפורמטים שונים.

למשל: תאריך עברי. לדוגמה: **א׳ בניסן התשס״ב**

תאריך לועזי. לדוגמה: **(14 במרס 2002)״.**

לתאריכים המופיעים בחוקים יכולות להיות משמעויות שונות:

- לעיתים הם מסמנים את מועד תחילת תוקפו של החוק.

לדוגמה: **תחילתו של חוק זה ביום כ״ז באייר התש״ס (1 ביוני 2000).**

- לעיתים הם מסמנים את מועד פקיעת תוקפו של החוק.

לדוגמה: חוק זה, למעט סעיף 36, יעמוד בתוקפו עד יום כ׳ בתמוז התשפ״א (30 ביוני 2021).

- לעיתים הם מגדירים את זמן תחולתו של החוק.

לדוגמה: בתקופה שמיום פרסומו של חוק זה ועד יום כ״ז בטבת התשפ״ב (31 בדצמבר 2021)

- לעיתים הם מגדירים שהחוק ישמש כהוראת שעה לתאריכים מסוימים.

לדוגמה: בתקופה שמיום פרסומו של חוק זה ועד יום כ״ג בטבת התשע״ט (31 בדצמבר 2018) (להלן – תקופת הוראת השעה) יקראו את חוק להסדרת הביטחון בגופים ציבוריים, התשנ״ח–1998[1] (להלן – החוק העיקרי) כך:

הדוגמה נלקחה מתוך: חוק להסדרת הביטחון בגופים ציבוריים(הוראת שעה), התשע"ו-2016

איתור אוטומטי של תאריכים בקורפוס החקיקה

תאריכים יכולים להופיע בחוק במספר וריאציות:

- תאריך מפורש – התאריך יהיה כתוב בצורה ברורה, יצוינו היום, החודש והשנה הרלוונטיים.

לדוגמה: ביום כ"ז באלול התשנ"ט (8 בספטמבר 1999).

- תאריך לא מפורש – התאריך יהיה כתוב בצורה שאיננה כוללת את התאריך המלא, אך ניתן להבין ממנה בדיוק באיזה תאריך מדובר.

לדוגמה: 91. (א) תחילתו של סעיף 4(א) לחוק יום חינוך ארוך, כנוסחו בסעיף 90 לחוק זה, ביום תחילתה של שנת הלימודים התשס"ד.

הדוגמה נלקחה מתוך: חוק המדיניות הכלילית לשנת הכספים 2004, התשס"ד-2004)

- תאריכים מותנים – תאריך התלוי בתאריך אחר. כלומר, תאריך שנקבע תקופת זמן מסוימת לאחר תחילת או סיום תאריך אחר (שיתכן שידוע לנו ויתכן שלא).

לדוגמה: תחילה אחרים, הכל כפי שתקבע, ובלבד שמועד תחילה שייקבע כאמור לא יהיה מאוחר מתום שלוש שנים ממועד התחילה לפי סעיף קטן (א).

הדוגמה נלקחה מתוך: חוק חופש המידע, תשנ"ח-1998)

יעמוד בתוקפו עד יום תחילתו של חוק הפיקוח על שירותים פיננסיים

המשימה –

איתור ותיוג של זמנים ושל תאריכים בתוך הטקסט. אם מדובר בתאריכים המופיעים יחד ומייצגים את אותו מועד –תיוג שלהם באופן שמקשר ביניהם.

לדוגמה: סעיף בו מופיע התאריך בו התקבל החוק בשני פורמטים – עברי ולועזי – זה לצד זה:

• נתקבל בכנסת ביום ח' בתמוז תשל"ו (6 ביולי 1976), הצעת החוק ודברי הסבר פורסמו בה"ח 1169, תשל"ח עמ' 176.
1 ס"ח תשכ"ג, עמ' 2.

הדוגמה נלקחה מתוך: חוק הבטיחות במקומות ציבוריים (תיקון), תשל"ו-1976)

הקלט – תיקייה עם קבצי XML של חקיקה מתויגת בפורמט AKN.

הפלט –

- תוכנית שמקבלת קובץ XML מתויג בסטנדרט AKN, שהתאריכים בו לא מתויגים.

התוכנית תחזיר את הקובץ כאשר התאריכים מתויגים.

- תיקייה עם קבצי XML בהם התאריכים מסומנים בתגיות המבוקשות.

- בתוכנית שלנו, נריץ כל קובץ מהתיקייה בנפרד.

איתור אוטומטי של תאריכים בקורפוס החקיקה

תיאור הפרויקט במונחים של מדעי הרוח הדיגיטליים

בפרויקט קיבלנו כקלט תיקייה עם קבצי XML של חקיקה מתויגת בפורמט AKN, בכל שלב בתוכנית אנחנו עוברות על קובץ XML יחיד, שתוכנו הוא חוק מסוים.

הפורמט הזה הוא ניטרלי מבחינה טכנולוגית, שומר על המבנה המקורי של הטקסט ובו בעת מאפשר תיוג מורכב של חלקים רלוונטיים. (תיוג אונטולוגיות: מוסדות/מקומות/תאריכים. אנחנו התמקדנו בתאריכים).

זהו סטנדרט קבוע ופתוח – אפשרויות לתיוג עצמאי על ידי כל גורם מעוניין לאחר הפרסום(מגזר ציבורי/מחקר/תעשייה).

בתוכן של החוק אנחנו מחפשות תבניות של תאריכים. המטרה שלנו היא איתור ותיוג של זמנים ושל תאריכים בתוך הטקסט. אם מדובר בתאריכים המופיעים יחד ומייצגים את אותו מועד –תיוג שלהם באופן שמקשר ביניהם.

קבצי ה-XML שקיבלנו הם קבצים חצי מובנים. קובץ XML בנוי בחלקו מתיוגים המעידים על התוכן העוקב. למשל – כאשר רצינו לעבור על תוכן של חוק כלשהו שקיבלנו כקובץ XML, חיפשנו את המשפטים המופיעים אחרי תגיות <p> , האות p מייצגת את המילה paragraph (פסקה באנגלית) ואחריהן תגיע הפסקה. תיוגים כמו זה, עוזרים לנו לאתר בפשטות את התוכן הרלוונטי לנו.

על מנת לבצע את תיוג בתאריכים בקבצי ה-XML, עברנו על מספר חוקים וראינו כיצד תאריכים מיוצגים בהם והשתמשנו בביטויים רגולריים מתאימים על מנת לאתר אותם – צורות שונות של תאריכים עבריים ותאריכים לועזיים.

פתרון הבעיה

כפי שמצוין לעיל, קבצי הXML אלו קבצים חצי-מובנים, כלומר רק חלק מהמידע השמור בו מתויג. בקבצי ה-XML ששומרים את תוכן החוקים במדינת ישראל, לא קיים תיוג לתאריכים המופיעים בו. המטרה שלנו היא למצוא תאריכים בטקסט ולתייג אותם.

מכיוון שתאריכים יכולים להופיע בצורות שונות – למשל במילים, או במספרים, בתאריך עברי או בתאריך לועזי – יש מספר תצורות שצריך לזהות בכדי לתייג את כל התאריכים בצורה נכונה. עברנו על מספר רב של חוקים על מנת לראות צורות שונות בהן תאריכים יכולים להופיע, ולפיהם כתבנו את התוכנית שלנו.

בנוסף, שייכנו בין תאריכים לועזיים ותאריכים עבריים המייצגים תאריך זהה.

שלבי התוכנית findDates-

<u>קלט</u>: נתיב לקובץ XML המכיל תוכן של חוק.

<u>פלט</u>: קובץ XML בשם withDates המכיל את תוכן החוק עם תיוג התאריכים, כאשר לכל תאריך יש id ובמידה וישנם שני תאריכים בקובץ המצביעים על אותו היום בשנה, קישור ביניהם.

איתור אוטומטי של תאריכים בקורפוס החקיקה

<u>תהליך העיבוד:</u>

מכיוון שתאריכים בטקסט יכולים להופיע במספר צורות שונות, על מנת למצוא תאריכים בקובץ, השתמשנו בביטויים רגולריים.

- חיפשנו ביטויים המתאימים לצורה של תאריך עברי, ושמרנו אותם במערך של מערכים.
  כאשר כל מערך בגודל 3, כך שבמקום ה-0 ישנו היום, במקום ה-1 החודש, ובמקום ה-2 השנה.
  על מנת שנוכל לזהות בין תאריכים זהים בעלי צורות שונות המרנו אותם לצורה אחידה של מספרים באמצעות שימוש בתוסף HebrewDate.
  כך למשל התאריך כ"ג בכסלו התשפ"א הומר ל "23-09-5781".
- חיפשנו ביטויים המתאימים לצורה של תאריך לועזי, ושמרנו אותם במערך של מערכים.
  כאשר כל מערך בגודל 3, כך שבמקום ה-0 ישנו היום, במקום ה-1 החודש, ובמקום ה-2 השנה.
  על מנת שנוכל לזהות בין תאריכים זהים בעלי צורות שונות המרנו אותם לצורה אחידה של מספרים באמצעות שימוש בתוסף GregorianDate.
  כך למשל התאריך 9 בדצמבר 2020 הומר ל "09-12-2020".
- על מנת למצוא לאחר מכן לכל תאריך את הID שלו, שמרנו כל תאריך במפה datesWithId, אשר הkey שלה הוא התאריך המומר לצורה אחידה והערך הוא הID.
  על מנת למצוא תאריכים לועזיים המסמנים את אותו היום של תאריכים עבריים, השתמשנו ב2 מפות נוספות:
  1. hebToGreg – תחילה, כל תאריך עברי המרנו לצורתו הלועזית באמצעות שימוש בפונקציה to_greg. לאחר מכן שמרנו בkey תאריך זה ובערך את הID.
  2. gregToHeb - תחילה, כל תאריך לועזי המרנו לצורתו העברית באמצעות שימוש בפונקציה to_heb. לאחר מכן שמרנו בkey תאריך זה ובערך את הID.
- כעת, עברנו פעם נוספת על קובץ ה XML וזיהנו את התאריכים מחדש.
  עבור כל תאריך שמצאנו ביצענו את הפעולות הבאות:
  1. בדקנו האם הוא תאריך עברי או לועזי, והמרנו באותו בהתאם לצורה האחידה.
  2. חיפשנו האם הוא קיים באחת מ2 המפות (בהתאם לאם הוא לועזי או עברי) במידה וכן שמרנו את ערך הID של התאריך המצביע לאותו היום בפורמט אחר.
  3. אם קיים לו תאריך אלטרנטיבי הוספנו תיוג במקום המתאים בטקסט באופן הבא:

```
ET.Element('date', attrib={'date eId': Id, 'alternativeTo': alterDate, 'date': numFormat})
```

date eId – הID של התאריך הנוכחי.
alternativeTo – הID של התאריך המצביע לאותו היום בפורמט אחר.
date – התאריך בפורמט הזהה. הסיבה שבחרנו לשים את התאריך בפורמט הזהה ולא בפורמט שבו הוא מופיע בטקסט היא על מנת לשמור על אחידות התיוגים.

אם לא קיים לו תאריך אלטרנטיבי הוספנו תיוג במקום המתאים בטקסט באופן הבא:

```
ET.Element('date', attrib={'date eId': Id, 'date': numFormat})
```

date eId – הID של התאריך הנוכחי.
date – התאריך בפורמט הזהה.

איתור אוטומטי של תאריכים בקורפוס החקיקה

דוגמאות:

תאריך לועזי -

```
<date date eId="2" alternativeTo="1" date="2020-12-09">2020 בדצמבר 9</date>)
```

תאריך עברי –

```
</date>).<date date eId="1" alternativeTo="2" date="5781-09-23">א"תשפ בכסלו כ"ג</date>
```

* הערה: תאריכים (של שנים) שמהווים חלק משם של חוק – לא תייגנו. (כנדרש בהגדרת המשימה). למשל: "בחוק השידור הציבורי הישראלי, תשע"ד-2014" איננו מתויג.

כיצד הפרויקט שלנו יכול לסייע?

- תיוג תאריכים –

תיוג התאריכים יוכל לתרום במספר דרכים:

1. איתור כל התאריכים המשויכים לחוק מסוים בדרך נגישה וברורה.
2. שיוך תאריך עברי לתאריך לועזי התואם לו ולהפך.
3. מתן אפשרות למעקב אחר כניסה לתוקף של חוקים ופקיעת תוקף של חוקים.

- איתור טעויות:

כאשר אנחנו מתייגים את התאריכים, אנחנו מקשרים בין כל תאריך עברי ותאריך לועזי שמייצגים את אותו התאריך. הקוד שלנו מתרגם בין תאריכים עבריים ללועזיים (ולהפך) על מנת להשוות ביניהם.

בעזרת ההשוואה, מתגלות טעויות בתרגום התאריכים השמורים בחוקים.

דוגמה לטעות שמצאנו בעזרת הפרויקט:

בחוק "חוק מעמד ותיקי מלחמת העולם השניה, התש"ס-2000", מופיעה הפסקה הבאה:

> "מעצמות הברית" – המדינות שחתמו על הצהרת האומות המאוחדות ביום ב' בטבת התש"ב (1 בינואר 1942) או שהצטרפו אליה בתקופת מלחמת העולם השניה, וכן שאר המדינות אשר לחמו באותה תקופה נגד גרמניה ובני בריתה, שעה שלחמו נגדן;

נשים לב לחלק המסומן בקו – "ביום ב' בטבת התש"ב (1 בינואר 1942)".

איתור אוטומטי של תאריכים בקורפוס החקיקה

הפלט שהתקבל בהרצת התכנית שלנו כלל את התיוג הבא:

```
/ns0:content>
s0:point>
0:point eId="point_1__point_מעצמות_הברית">
ns0:content>
 <ns0:p>
   <ns0:def><date eId="5" date="5702-10-02">ב' בטבת התש"ב</date> (<date eId="6" date="1942-01-01">1942 בינואר 1</date>;
/ns0:content>
s0:point>
0:point eId="point_1__point_שלטון_עוין">
ns0:content>
```

בעוד שהתיוג המתבקש (כאשר מדובר בתאריך לועזי ותאריך עברי המייצגים את אותו התאריך)

נראה כך:

```
ת<date date eId="9" alternativeTo="10" date="5760-02-27">ט"כ"ז באייר התש</date>.(2000 ביוני 1) <date date eId="10" alternativeTo="9" date="2000-06-01">2000 ביוני 1</date>).
```

כלומר, התוכנית זיהתה שלא מדובר בייצוג אלטרנטיבי לתאריך זהה.

לכן ביצענו בדיקה -

הבדיקה העלתה שהתאריך העברי הרשום בחוק - "ב' בטבת התש"ב" - מייצג את התאריך הלועזי הבא:

המרת תאריך עברי ללועזי

יום ב חודש טבת שנה תש"ב

שנים נוספות

לפני 12 בלילה

חשב תאריך

ב טבת ה'תש"ב = יום שני 22/12/1941

בעוד שהתאריך הלועזי הרשום בחוק – "1 בינואר 1942" – מייצג את התאריך העברי הבא:

המרת תאריך לועזי לעברי

יום 1 חודש ינואר שנה 1942

אחרי השקיעה

חשב תאריך

יום חמישי 1/1/1942 = י"ב טבת ה'תש"ב

כך זיהינו כי התוכנית אכן זיהתה כראוי, ושני התאריכים הכתובים כתאריך זהה, למעשה מייצגים תאריכים שונים, ומדובר בטעות שיש צורך לתקן אותה.

בעזרת שימוש בקוד שכתבנו, תהיה אפשרות לאתר ביעילות טעויות כגון זו המפורטת לעיל, לתקן את החוקים בהתאם ובכך לשפר את רמת הדיוק בתיעודם.

איתור אוטומטי של תאריכים בקורפוס החקיקה

הערכת תוצאות

הדרך בה נמדוד את טיב התוצאות היא באמצעות הרצה של הקוד על מספר קבצי XML, ובדיקה שאכן תיוג התאריכים תויג כראוי.

הרצנו את הקוד על מספר קבצי XML וברובם המוחלט התיוג בוצע כראוי. בקבצים בהם ביצוע התיוג נתקל בבעיות, איתרנו את הסיבות לכך, בדקנו כיצד ניתן לפתור את זה, וסידרנו את הקוד בהתאם.

דוגמא לקובץ XML בו התיוג לא בוצע כראוי:

בתי המשפט האזרחיים ישפטו בהתאם לחוק העותומני שהיה נוהג בארץ ישראל ביום 1 בנובמבר, 1914, ובהתאם לאותם ה
במועצה ולפקודות ולתקנות הנוהגים בארץ־ישראל בתאריך דבר מלך זה או אשר ינהגו או יוחקו בעתיד; ובהתחשב עם החוקי

כפי שניתן לראות בחלק המסומן בקו, בין החודש לשנה יש פסיק.

כאשר כתבנו את הקוד, לא חשבנו על מקרה כזה ולכן לא תמכנו בצורת הכתיבה הזו.

לאחר הרצה על קוד זה גילינו כי לא בוצע תיוג בתאריכים אלו, לכן הוספנו באמצעות ביטוי רגולרי את האפשרות לפסיק לאחר החודש בתאריכים.

באופן זה טיפלנו במקרי הקצה, ובכך שיפרנו את רמת הדיוק של התוצאות המתקבלות מהרצת התוכנית.

להערכתנו, על אף שעברנו על חוקים רבים ובקפדנות, ככל הנראה ישנן תצורות נוספות של כתיבת תאריכים שייתכן שפספסנו ומכך שהקוד שלנו אינו מזהה אותן. בניסיון להגיע לתוצאות מדויקות ככל שניתן, ביצענו בדיקות רבות, בכדי לכלול את רובן המוחלט של צורות הכתיבה השונות של התאריכים בחוקים.

סיכום

בחודשים האחרונים לקחנו חלק בקורס "מדעי הרוח הדיגיטליים". אל הקורס הגענו מבלי שנחשפנו לנושא הזה בעבר כלל, אך עם הרבה עניין וסקרנות. במהלך הקורס נחשפנו לנושאים חשובים ומעניינים, וכעת ניתנה לנו אפשרות ליצור פרויקט בעצמנו ולנסות להשפיע.

הנושא שבחרנו, מטרתו לסייע במשימת תיוג סעיפי חקיקה או רכיבים בתוך סעיפים על פי התוכן שלהם. וביותר ספציפיות, התמקדנו באיתור אוטומטי של תאריכים בקורפוס החקיקה.

האתגר שלנו היה להבין כיצד ניתן לתייג את הטקסט באופן אוטומטי כאשר הכללים אינם ודאיים.

את תהליך ההתמודדות עם האתגר עשינו בעבודת צוות, נדרש מאיתנו מעבר על החוקים בכדי לנסות לאתר צורות בהן נכתבים החוקים בדרך כלל, להבין את הכללים, ואת כל המסקנות שלנו לתרגם לקוד שידע "להתמודד" עם מגוון תצורות.

במהלך העבודה למדנו הרבה על חשיבות השימוש בטכנולוגיה, כיצד זה יכול לסייע.

חשיבות הדיוק בתיעוד של דברים בעלי חשיבות גבוהה – כמו חוקי מדינת ישראל – הינה רבה מאוד. כל צעד שנוכל לעשות על מנת לשפר ולייעל את ביצוע משימה זו הוא חשוב ומהווה תרומה חשובה לשימור ידע בזמנים אלו ולדורות הבאים.

איתור אוטומטי של תאריכים בקורפוס החקיקה

רק כאשר התחלנו לעבוד על הפרויקט הבנו כמה הנושא רחב ומקיף, זו משימה שדורשת זיהוי של סוגי כתיבה שונים, תרגום תאריכים, שיוך תאריך לצורות הכתיבה הרבות השונות הקיימות לו ועוד.

עברנו על חוקים רבים, בכדי ללמוד על צורות הכתיבה של התאריכים בחוקים, על דרכי כתיבת החוקים, ובכדי למצוא דוגמאות שיעזרו לנו להבין ולהעביר את המידע הלאה בצורה ברורה ומובנת.

העבודה על הפרויקט הייתה מהנה והיוותה חוויה לימודית מעניינת ומאתגרת, שונה במהותה מהדברים בהם התנסנו עד היום. החשיבות של התוצר והידיעה שהוא יכול להשפיע נתנו לנו מוטיבציה משמעותית שליוותה אותנו לאורך כל זמן העבודה על הפרויקט. על אף שתמיד קיימת התחושה שיש אפשרות לשפר את המחקר והעבודה עוד, אנחנו מרוצות מאוד ממה שהצלחנו ליצור ומלאות תקווה שהפרויקט שלנו ישמש לעזרה בשיפור התיעוד במאגר החקיקה הלאומי ולכל הנוגעים בדבר.

מקורות

אתר אינטרנט להמרת תאריכים (תאריך עברי לתאריך לועזי ולהפך) בו נעזרנו לצורך בדיקת טעויות שמצאנו בחוקים:

לוח שנה ותאריך עברי לועזי • MyLush.net

אתר מאגר החקיקה הלאומי, המרכז את החוקים, ממנו לקחנו את הדוגמאות המצורפות לאורך הקובץ ועליו התבססנו בניתוח צורות כתיבת התאריכים:

מאגר החקיקה הלאומי

את הקוד כתבנו בפייתון (גרסה 3.9.2)

ספריות של פייתון בהן נעזרנו:

- ElementTree –

`xml.etree.ElementTree` — The ElementTree XML API

- Regex Expression –

Regular Expression

- Pyluach –

pyluach 1.2.1