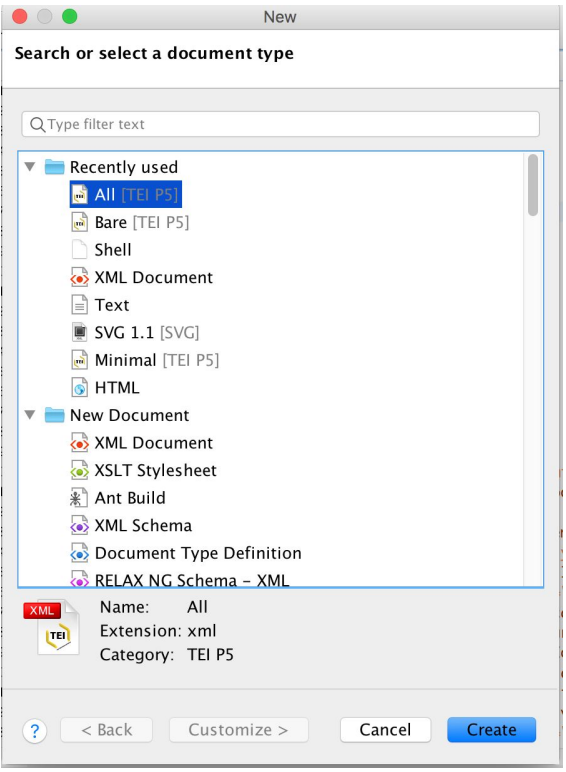
# תרגיל ראשון בקורס 'מדעי הרוח הדיגיטליים' סמסטר ב' 2017

מטרת התרגיל היא להכיר את העבודה עם קבצי ה xml/tei - תיוג של תוכן וצורה ומטה-דטה, ולהכיר כלים של עיבוד שפות טבעיות - תיוג חלקי דיבר.
העבודה תעשה בזוגות
תאריך ההגשה: 8.5.2017

## חלק א': תיוג TEI

- צריך לבחור ערך מה[לקסיקון](לקסיקון) (כתבו לי את מי בחרתם ואשלח את הטקסט)
- את הערך עליכם לכתוב כקובץ tei / xml

איך לעשות?
- להתקין oxygen (לא חובה, וכאמור יש שימוש חופשי לחודש, אח"כ צריך לשלם אז אם עובדים בזוג, תתקינו קודם אצל אחת ואחכ אצל השניה.
- לפתוח קובץ חדש, לבחור tei (כמו בתרשים משמאל)
- יפתח לכם קובץ שלדי של tei. אם אתם משתמשים בכלי אחר מאוקיסג'ן, יש לכם דוגמא בעמ' האחרון של התרגיל
- לקרוא את הפרק Common structure and elements באתר tei by examples




עכשיו -
1. צריך למלא את teiHeader - המטה-דטה על הערך.

יש ערכים שהם אופציונליים ויש ערכים שהם הכרחיים (למשל, title).
עליכם לחשוב - איזה אינפורמציה יש להכליל במטה-דטה?
אפשר לקרוא הסברים כאן למשל
- מטה-דטה: על מי הערך? מי המחבר? מה ההקשר הכללי יותר (ספר, הוצאה, וכו), מי המקודד?

2. בשלב השני עליכם לחשוב על הטקסט עצמו - מה המבנה של הטקסט? זהו את החלקים המשמעותיים (כותרת, פסקה, הפניות בסוף הערך וכו') ובחרו את תג המבנה שנראה לכם הכי מתאים.
כתבו את ההחלטות שלכם (בקובץ readme).

3. השלב השלישי - סימון זהויות בקובץ named entities - שמות מקומות, אנשים וארגונים.

https://prezi.com/8tqj70cwsnlo/named-entities-people-places-and-organisations/

## חלק שני - הרצת המתייג - חלקי דיבר

בחלק הזה של התרגיל תריצו על קובץ הטקסט את מתייג חלקי הדיבר של מני אדלר

הורידו את הקובץ
http://www.cs.bgu.ac.il/~nlpproj/lemlda.zip

יש בקובץ יותר ממה שצריך לתרגיל שלנו, אבל לא נורא
תעברו לספריה tagger ושם תריצו את הפקודה

```
java -Xmx1200m -XX:MaxPermSize=256m -cp
trove-2.0.2.jar:morphAnalyzer.jar:opennlp.jar:gnu.jar:chunker.jar:
splitsvm.jar:duck1.jar:tagger.jar vohmm.application.BasicTagger
inputfile outputfile -lemma -conll -ner
```

עבור (קטע) המשפט:
**מרדכי צ׳צ׳קס, שהיה מוסמך לרבנות ונטה לחסידות**
תקבלו פלט שנראה כך:

```
 10 מרדכי מרדכי מרדכי properName unspecified unspecified
unspecified unspecified unspecified
 11 צ'צ'קס צ'צ'קס צ'צ'קס properName unspecified unspecified
unspecified unspecified unspecified
 12 , , , punctuation unspecified unspecified unspecified
unspecified unspecified
 13 שהיה ש ש relativizer/subordinatingConjunction unspecified
unspecified unspecified unspecified unspecified 0
13 שהיה הוא היה copula masculine singular unspecified 3 past
 14 מוסמך מוסמך הוסמך verb masculine singular unspecified any
beinoni
 15 לרבנות ל ל preposition unspecified unspecified unspecified
unspecified unspecified 0
15 לרבנות ה ה definiteArticle unspecified unspecified unspecified
unspecified unspecified 0
15 לרבנות רבנות הרבנות noun feminine singular absolute unspecified
unspecified
 16 ונטה ו ו conjunction unspecified unspecified unspecified
unspecified unspecified 0
16 ונטה נטה נטה verb masculine singular unspecified 3 past
```

```
17 לחסידות ל ל preposition unspecified unspecified unspecified
unspecified unspecified 0
17 לחסידות ה ה definiteArticle unspecified unspecified unspecified
unspecified unspecified 0
```

הטור הראשון (משמאל) - מס׳ המילה במשפט - אם המספר מופיע בשתים או שלוש שורות עוקבות, זה בגלל שמדובר במילה שיש לה ניתוח מורפולוגי - למשל ״בבית״ ב-ה-בית

הטור השני - הtoken
הטור השלישי - ה לקסמה של המילה - האופן שבו היא נמצאת בלקסיקון, צורת הבסיס שלה
הטור הרביעי - המילה או המורפמה כפי שמופיעה בטקסט המקורי לאחר הניתוח המורפולוגי (ה הידיעה, ו החיבור)
ואח״כ - חלק הדיבר (לפי המדריך למתייגת https://www.cs.bgu.ac.il/~adlerm/tagging-guideline.pdf)
והתכונות שלו אם יש (זכר, נקבה, יחיד, רבים וכו׳)

תייגו את הערך בעזרת המתייג, ובדקו - איזה טעויות יש? איך תויגו השמות הפרטיים?שמות המקומות?
האם מה שסימנתם כשם פרטי זוהה ככזה?

## חלק שלישי

כתבו תכנית פייתון המקבלת את הטקסט שהוא ערך לקסיקון הספרות ומנסה לזהות שמות פרטיים, תאריכים ומקומות על סמך תוצאות המתייג, בעזרת ביטויים רגולרים וכדומה -- ומייצרת קובץ xml/tei (כפי שתייגתם בצורה ידנית).
בשלב הזה * המטה-דטה* יכול להיות כללי מאוד (הקובץ שנוצר צריך להיות tei תקין, אבל לא צריך לדאוג בצורה אוטומטית למלא את כל השדות, רק מה שברור לכם, כמו שם הערך)

עליכם לחשוב ולהעריך -- כמה טוב התכנית שלכם עובדת?
האם הכללים ׳תפורים׳ לטקסט הקצר שברשותכם?
כמה טוב יעבוד על טקסט אחר? על ערכים אחרים בלקסיקון הספרות? על טקסט כללי? בדקו!

ההגשה - קובץ zip ששמו שרשור שמות עם _ ביניהם, לשלוח לדואר yaeln@cs.bgu.ac.il עם הנושא ״תרגיל ראשון dh.bgu״

- חלק ראשון: קובץ tei של הערך + קובץ readme המתאר מה החלטתם לתייג, למה, ואם היו לכם לבטים בין תגים שונים.
- חלק שני: קובץ מתויג + ניתוח התוצאות, כמה טוב עבד המתייג, איפה היו טעויות, ואיך לגבי זיהוי שמות פרטיים, מקומות ותאריכים.
- חלק שלישי: תכנית ניתנת להרצה + דוגמא להרצה (קובץ הקלט, קובץ הפלט).

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model
href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng"
type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<?xml-model
href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng"
type="application/xml"
     schematypens="http://purl.oclc.org/dsdl/schematron"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>Title</title>
         </titleStmt>
         <publicationStmt>
            <p>Publication Information</p>
         </publicationStmt>
         <sourceDesc>
            <p>Information about the source</p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <body>
         <p>Some text here.</p>
      </body>
   </text>
</TEI>
```