תרגיל בית 1

בתרגיל זה עליכם לעבד קורפוס של טקסטים בעברית. מטרת התרגיל היא להבין את הקשיים הכרוכים בעיבוד אוטומטי של טקסטים גדולים, ובפרט טקסטים שמקורם באינטרנט. הקלט הוא אוסף של קבצים מדברי ימי הכנסת, הקבצים הם בעברית עליכם לכתוב תוכנית שתעבד באופן אוטומטי את הקבצים, תשלוף את התכנים הרלוונטיים לנו ותעביר אותם .docx ובפורמט (לפורמט שנוח לעבוד איתו (חישובית.

שימו לב, הקורס בנוי כך שהפלט שתייצרו בתרגיל הזה ישמש אתכם גם לתרגילים הבאים, לכן הקפידו על עבודה יסודית ומדוייקת.

אחרי הפתיחה תקבלו כמה .(extract) ולפתוח אותו zipהיא לשנות את הסיומת של הקובץ ל doc הדרך לעבוד עם קבצי הקובץ הוא ,(word בתקייה document קובץ בשם) word/document.xml תיקיות וקבצים. בתרגיל אנחנו נשתמש בקובץ פורמט שהרבה יותר נוח לעבוד איתו. עליכם לזהות את מבנה הקובץ ואת התגים הרלוונטיים ,XML בפורמט.

עליכם לכתוב תוכנית שהפלט שלה הוא קובץ XML יחיד במבנה הבא:

<root>

<doc><name>היו"ר ש' וייס:</name>

<text>חברי הכנסת , נא לשבת .

אני פותח בזה את ישיבת הכנסת , יום רביעי , י' באב התשנ"ג , 28 ביולי 1993 .

...

...

</text>

</doc>

<doc><name>ש"חד) תופיק זיאד (:</name>

<text>

להצעות לסדר יש פרוצדורה , מגישים בכתב לפני הישיבה .

ראה תקנון .

תודה .

נא לשבת .

...

...

</text>

</doc>

<doc><name>(הליכוד דן תיכון):</name>

<text>נקיים את הדיון על-פי התקנון .

...

...

</text>

<doc/>

....

....

</root>

.<root/><root> כלומר, הקובץ מתחיל ונגמר בתגיות

וכל <name/><name> כאשר השם של הדובר ימצא בין התגיות <doc/><doc> לאחר מכן, הנתונים של כל דובר יהיו בתגיות <text/><text> הטקסט שנאמר ע"י אותו הדובר בכל הקובץ ימצא בתגיות

:הטקסט צריך לקיים את התנאים הבאים

.שליפת רק הטקסט המדובר: כל הכותרות, אמרות מנהליות, תוכן עניינים וכו' לאצריכים להכלל בטקסט 1.

צריך להופיע פעם אחת בקובץ הפלט, יחד עם <name/><name> כל דובר מופיע פעם אחת בלבד: השם בין התגיות 2. .הטקסטים המתאימים של הדובר לאורך כל קובץ הקלט

3. חלוקה של הטקסט למשפטים: עליכם לקבוע כיצד לזהות גבולות בין משפטים בעברית ולממש זאת על הטקסט הנתון. כל משפט מופיע בשורה נפרדת, ללא שורות ריקות.

4. טוקניזציה: עליכם לקבוע כיצד לחלק משפטים לטוקנים בעברית, ולממש זאת על הטקסט המחולק למשפטים שיצרתם בשלב הקודם. התוצאה של שלב זה הוא טקסט שבו כל משפט מופיע בשורה נפרדת, וכל טוקן מופרד ברווח בודד שימו לב שבאופן כללי, סימני הפיסוק צריכים להיות טוקנים נפרדים (אחרת, המילה *בית* והמילה *בית!* יהיו טוקנים משכניו. שונים, וזו אינה תוצאה רצויה). למרות האמור במשפט הקודם, הקדישו מחשבה לסימני פיסוק שונים, יתכנו מקרים יוצאי דופן אשר בהם דווקא תעדיפו לא להפריד בין סימן פיסוק למילה שאליה הוא מוצמד במקור. כמו כן, חשוב לציין כי אתם לא לא אמורים לבצע ניתוח מורפולוגי כלל, ולכן, לדוגמא, המילים *הבית* או *ובית* יופיעו ללא הפרדת התחילית מהמילה עצמה.

- עליכם להגיש שני קבצים

א. דו"ח ובו תיאור של מה שעשיתם בעת הפיתוח, כלומר כיצד החלטתם איך לחלק את הטקסטים למשפטים ואיך לעשות טוקניזציה. זה המקום לפרט התלבטויות שהיו לכם במהלך העבודה. למשל, אם קיבלתם החלטה מסוימת כדי לבצע טוקניזציה, והחלטה זו גורמת לחלוקה נכונה לטוקנים במקרים מסוימים, אך גורמת לחלוקה שגויה במקרים אחרים, פרטו זאת כאן. שימו לב - כדאי לכם לכלול כאן תיאור מפורט של ההתלבטויות וההחלטות שלכם, כיוון שאם בתוצאות הטוקניזציה יהיו מקרים שגויים, ללא כל הסבר, יורדו על כך נקודות (מתוך הנחה שלא חשבתם על המקרה הזה וגם לא שמתם לב לטעות בפלט שלכם), בעוד שאם תכללו הסבר משכנע מדוע לא טיפלתם במקרה מסוים (או טיפלתם בו דווקא בצורה מסוימת), לא .יורדו על כך נקודות

אל תשכחו לציין את שמכם, עם מספר סטודנט. בנוסף, אנא ציינו על איזו מערכת .PDF הדו"ח צריך להיות בפורמט .hw1report.pdf הפעלה הרצתם את הקוד שלכם. לקובץ זה עליכם לקרוא

ב. קובץ קוד אחד, בשפת התכנות פייתון, גרסא 3.5 ומעלה. שימו לב - המטלות ייבדקו בגרסא 3.5, אשר אין לה תאימות בהכרח עם גרסאות קודמות של פייתון, לכן מומלץ לכתוב את הקוד בגרסה זו. לא יתקבלו הגשות בשפות תכנות אחרות או בפתרון תרגיל זה. לקובץ הקוד עליכם NLTK בגרסאות קודמות של פייתון. למען הסר ספק - **אין** להיעזר בספריה .ועליו להיות מופעל משורת הפקודה ,hw1.py לקרוא

:הקוד יופעל עם הפקודה הבאה

python hw1.py <dir_name_input> <file_name_output>

כאשר <dir_name_input> הוא שם התקייה שנוצרה אחרי פתיחה של קובץ הZIP

ו <file_name_output> הוא קובץ ה XML שפורט לעיל - הפלט.

לדוגמה: python 13_ptm_235386 output.xml

הגבלות:

מותר להשתמש בספריה xml.etree.ElementTree ובכל ספריה סטנדרטית של פייתון.

אסור להשתמש בספריה NLTK.

מצורפים שני קבצים לדוגמה, אתם מוזמנים להשתמש בהם לפיתוח. התרגיל שלכם יבדק על קבצים אחרים באותו הפורמט.

הוראות מיוחדות:

הכוללת את מודול הפייתון ומספר רב של חבילות בסיסיות יותר אשר ישמשו anaconda מומלץ בחום להתקין את חבילת אותכם לאורך כל הקורס ויחסכו מכם התקנה של חבילות אחרות במהלך התרגילים. להורדה והתקנה מי שבוחר לעבוד בצורה אחרת, מתבקש להתקין אותה בעצמו. https://www.continuum.io/downloads -

אחד ובו שני הקבצים המפורטים לעיל - קוד פייתון ודו"ח zip יש להגיש קובץ.

יש להקפיד על עבודה עצמאית. צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד.

תאריך הגשה: 11.11.2018, עד השעה 23:59.

שאלות על התרגיל אפשר לשאול בפורום תרגילי בית.