עיבוד שפות טבעיות סמסטר א' תשפ"א

תרגיל בית מספר 1: **קורפוסים**

מועד הגשת התרגיל: 14.11.20 בשעה 23:59

מטרת התרגיל היא להמחיש את האתגרים הכרוכים באיסוף, ארגון ועיבוד אוטומטי של טקסט בהיקף נרחב, ובמיוחד כזה שמקורו באינטרנט.

בתרגיל זה נתנסה בהקמת קורפוס חדש לגמרי. הקורפוס יורכב מטקסט של ערכי ויקיפדיה בתצורה שתוסבר בהמשך.

בתרגיל זה תשתמשו ב-API של ויקיפדיה, המאפשר להתחבר ולאסוף מידע באופן אוטומטי (כלומר, בתוכנה). ל-API הבסיסי נכתבו לא מעט עטיפות (wrappers) שמקלות מאוד על הגישה והעבודה מול ויקיפדיה. לדוגמה: https://pypi.org/project/wikipedia/ . כדאי לקרוא את התיעוד ולעבור על הדוגמאות שממחישות באופן בהיר את דרך העבודה.

הקורפוס שתיצרו בתרגיל זה מבוסס על אוסף של ערכי ויקיפדיה בשפות שונות. ב-API ניתן להגדיר בקלות את השפה שבה רוצים לעבוד– כפי שמפורט בדוגמאות שתראו online. רשימת קודי השפות הנדרשות יפורטו בקובץ שיתקבל כקלט לתכנית שתכתבו. הקורפוס הסופי יורכב מקבצי טקסט המבוססים על ערכי ויקיפדיה, קובץ לכל ערך בכל אחת מהשפות שאיתן תעבדו. רשימת הערכים שעליה יתבסס הקורפוס תפורט בקובץ שיתקבל כקלט לתכנית שתכתבו. קבצי הקלט יהיו קבצי טקסט פשוטים בהם כל ערך/שפה יופיעו בשורה נפרדת. דוגמאות לקבצי קלט יופיעו באתר בהמשך להוראות התרגיל. הקבצים בקורפוס שייווצר יהיו קבצי טקסט (סיומת txt.) ושמותיהם יורכבו מתחילית של קוד השפה קו תחתון מפריד ושם הערך. כך לדוגמה עבור רשימה שכוללת את הערכים ***Apple, Banana, Watermelon*** בשפות ***en*** (אנגלית), ***es*** (ספרדית) ***fr*** (צרפתית) – הקורפוס שייווצר יכלול את 9 הקבצים הבאים:

en_apple.txt
en_banana.txt
en_watermelon.txt
es_apple.txt
es_banana.txt
es_watermelon.txt
fr_apple.txt
fr_banana.txt
fr_watermelon.txt

הטקסט צריך לקיים את התנאים הבאים:

1. **נקיון**: הטקסט צריך להכיל את הטקסט שמופיע בעמוד הויקיפדיה הרלוונטי, ורק אותו (ללא תמונות, תגי html, וסימנים שאינם חלק מהשפה).
2. **חלוקה למשפטים**: עליכם לקבוע כיצד לזהות גבולות בין המשפטים ולממש זאת על הטקסט הנתון. בפלט, כל משפט צריך להוות שורה נפרדת, **ללא שורות ריקות**. שימו לב לכותרות ולכותרות המשנה שמופיעות בגוף הדף וחשבו איך לארגן אותן.
3. **טוקניזציה**: עליכם לקבוע כיצד לחלק משפטים לטוקנים, 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 משכניו. שימו לב שבאופן כללי, סימני הפיסוק צריכים להיות טוקנים נפרדים (אחרת, המילה ?today והמילה today יהוו טוקנים 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 אמורים לבצע ניתוח מורפולוגי כלל, ולכן, לדוגמא, המילים *lived* או *live* יופיעו כמות שהן, ללא הפרדת הסופית *ed* מהמילה עצמה.

עליכם לכתוב תכנית שהפלט שלה הוא קבצי טקסט – קובץ לכל ערך ויקיפדיה בכל שפות המטרה שהוגדרו בקובץ הקלט, כפי שהוסבר למעלה. את קבצי הקלט יש לכתוב בעזרת ***קידוד utf-8*** לטובת תאימות בין מערכות הפעלה שונות. דוגמאות לקבצי קלט אפשריים עבור התכנית נמצאים באתר בהמשך להוראות התרגיל.

יש להגיש שני קבצים (**נפרדים** ולא ארוזים ב zip):

1. דו"ח ובו תיאור של ההחלטות שקיבלתם בתהליך פיתוח הקוד – איך חילקתם למשפטים ואיזו מדיניות נקטתם בשלב הטוקניזציה. בדו"ח מאוד כדאי לכלול פירוט של ההתלבטויות שלכם והבעיות שנתקלתם בהן. למשל, אם קיבלתם החלטה מסוימת כדי לבצע חלוקה למשפטים, והחלטה זו גורמת לחלוקה נכונה במקרים מסוימים, אך לחלוקה שגויה במקרים אחרים, פרטו זאת כאן והראו דוגמאות. שימו לב שאם בתוצאות יהיו מקרים שגויים ללא כל הסבר, תאבדו 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), תקבלו את רוב אם לא את מלוא הנקודות. מאוד מומלץ להתבונן בפלטים שלכם ולעדכן את המימוש לפי הצורך, וכן ולתת דוגמאות מתוך הנסיון שלכם בדו"ח. אנא הגישו את הדו"ח בפורמט PDF, וציינו שם ומספר תעודת זהות. שם הקובץ צריך להיות hw1Report.pdf.
2. קובץ קוד אחד, בשפת התכנות פייתון, גרסה 3.7 ומעלה. שימו לב שגרסה זו אינה תואמת בהכרח גרסאות קודמות של פייתון. לא יתקבלו הגשות בשפות תכנות אחרות או בגרסאות קודמות של פייתון. למען הסר ספק - אין להיעזר בספריה NLTK בפתרון תרגיל זה. לקובץ הקוד עליכם לקרוא **hw1.py**, ועליו להיות מופעל משורת הפקודה. הקוד יופעל עם הפקודה הבאה:

```
python hw1.py <value_list> <language_list> <output_dir_path>
```

באשר:

- **<value_list>** הוא הנתיב ושם קובץ הקלט שמכיל את ערכי הויקיפדיה שעליהם יתבסס הקורפוס.
- **<language_list>** הוא הנתיב ושם קובץ הקלט שמכיל את קודי השפות שאיתן תעבדו.
- **<output_dir_path>** הוא הנתיב המלא לתיקייה שאליה ייכתבו קבצי הטקסט המעובדים.

דוגמה להרצה משורת הפקודה:

```
(base) C:\Users\liatn\Documents\Liat\TA_NLP\2020A\HW1>python hw1.py  C:\Users\liatn\Documents\Liat\TA_NLP\2020A\HW1\langs.txt
C:\Users\liatn\Documents\Liat\TA_NLP\2020A\HW1\values.txt  C:\Users\liatn\Documents\Liat\TA_NLP\2020A\Corpus\
```

ניתן להניח שיש הרשאות כתיבה לתיקיית הפלט וכן שקבצי הקלט תקינים ומכילים ערכים חוקיים.
שימו לב לשגיאות שעלולות להיווצר ודאגו להוסיף טיפול הולם למניעת קריסה של התכנית.

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

יש להקפיד על עבודה עצמאית. צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד.
תאריך הגשה: 14.11.2020, עד השעה 23:59.
שאלות על התרגיל אפשר לשאול בפורום תרגילי בית.

בהצלחה!