תרגיל 1

קורפוסים

מבוא

בהרצאה האחרונה למדתן על קורפוסים: מאגר לשוני רחב, לרוב מעובד לשימוש נוח, המכיל מידע שימושי המאפשר לבצע ניתוחים לשוניים סטטיסטיים. בעידן כיום מציאת מאגר טקסטים אינה משימה מסובכת במיוחד, אך טיפול בטקסט על-מנת שיתאים לעבודה עמו היא כבר משימה מאתגרת הרבה יותר.

בתרגיל זה נכיר קורפוס מפורסם, The British National Corpus (BNC), ונלמד לעבוד איתו. כמו כן, נתנסה בבניית קורפוס מטקסט נתון. באופן כללי, נעבוד עם ה-BNC במהלך הקורס.

לתרגיל מצורף קובץ Python ובו template לתרגיל. עליכן לכתוב את הקוד לתרגיל בקובץ זה, ולהגישו יחד עם דו"ח המתעד את עבודתכן. הוראות ההגשה המפורטות נמצאות בסוף התרגיל. נמליץ להתקין את חבילת anaconda הכוללת מודל Python ומספר רב של חבילות בסיסיות שישמשו אתכן במהלך הקורס ויחסכו התקנה של חבילות אחרות.

חלק 1: היכרות עם קבצי XML

הקורפוס איתו נעבוד מגיע בקבצי XML. לתרגיל מצורפים כמה מהקבצים להתרשמות, אך תוכלו למצוא את הקורפוס המלא באתר הרשמי. בחלק זה עליכן לבחון את הקבצים המצורפים (ניתן לעשות זאת גם דרך ממשק אינטרנטי לצפייה בקבצים מסוג זה), ולהתרשם ממבנה הקורפוס. צרפו לדו"ח את התרשמותכן:

- איך מבנה הקורפוס משרת את המשתמשות בו?
- מה הופך את המבנה שלו ליעיל?
- מה היתרונות והחסרונות של ה-data בו (מבחינת התוכן)?

חלק 2: הכנת מבני נתונים ייעודיים עבור הקורפוס

במהלך הקורס נשתמש לא מעט בקורפוסים. למשימות שנבצע לא נצטרך כמות כה גדולה של טקסט, אך לרוב המשימות נצטרך להשתמש במאגר יחסית גדול. לאחר שהתרשמתן מה-BNC, ניצור 3 מבני נתונים ייעודיים (איתם נעבוד במהלך הקורס):

1. מבנה נתונים של token – לפחות 5 שדות
2. מבנה נתונים של משפט – לפחות 3 שדות
3. מבנה נתונים של הקורפוס כולו – אין מינימום למספר השדות

היעזרו בקבצי ה-XML על-מנת להחליט לגבי השדות למבנים אלו. לדוגמא, לכל token בקורפוס יש תגית המציינת את חלק הדיבר (part of speech, מסומן כ-POS), לכן נרצה להוסיף שדה כנ"ל למבנה הנתונים שלנו. מלבד מספר השדות המינימאלי יש לכן חופש מלא במימוש מבנים אלו, אך יש לנמק על החלטותיכן בדו"ח.

חלק 3: בניית קורפוס מטקסט

בחלק זה של התרגיל תתנסו בבניית קורפוס בהינתן טקסט. נשתמש בטקסט מעמודי ויקיפדיה (Wikipedia) הנתון בקבצי טקסט. קבצי טקסט לדוגמא מצורפים לתרגיל.

על-מנת לבצע זאת, עקבו אחר השלבים הבאים:

1. <u>טיפול בטקסט</u>: הטקסט מגיע בצורתו הגולמית, כפי שמופיע באתר. בחנו את מבנה הטקסט (כותרות, קישורים וכדומה) והחליטו אילו חלקים ממנו רלוונטיים לקורפוס, ואילו יכולים להיות רלוונטיים לשדות של האובייקטים השונים בקורפוס, לפי השדות שקבעתן בחלק 2.
2. <u>חלוקה למשפטים</u>: כפי שניתן להניח מחלק 2, בקורס שלנו יש חשיבות רבה למשפטים עצמם. חלקו את הטקסט למשפטים בהסתמך על כללי אצבע, והשאירו מקום למקרי קצה.
3. <u>טוקניזציה</u>: היחידות האטומיות איתן נעבוד הן tokens. בצעו (ללא שימוש ב-tokenizers ממומשים) טוקניזציה למשפטים.
4. <u>הכנסה לאובייקטים</u>: את כל ה-tokens הכניסו לאובייקטים הייעודיים, כמו גם את המשפטים. ניתן כמובן להשאיר שדות ריקים (וודאו כי אלו מאותחלים כראוי), אך אם ניתן למלא שדות – עשו זאת.
5. <u>יצירת מתודה מבצעת</u>: הוסיפו לאובייקט Corpus מתודה האחראית לממש את הנ"ל. בסופו של דבר, נרצה שביצירת אובייקט מסוג זה תהייה מתודה המקבלת קובץ טקסט כנ"ל ו"מכניסה" אותו לקורפוס (ז"א, את האובייקטים הייעודיים).
6. <u>מתודת יצירת קובץ טקסט</u>: הוסיפו מתודה נוספת ל-Corpus המדפיסה את הקורפוס לקובץ טקסט נתון באופן הבא: בין כל שני tokens יש רווח אחד, ובין כל שני משפטים ישנה ירידת שורה. ה

ניתן כמובן לממש את השלבים הללו בכל סדר שנראה לכן לנכון. הקפידו לפרט בדו"ח אודות כל החלטה שלקחתן במהלך העבודה (למשל, איך התייחסתן לכותרות? איך הדפסתן אותן לקובץ הטקסט במתודה של שלב 6?). שלדי המתודות שיש לממש נמצאים בקובץ ה-template, אך כמובן שניתן ליצור מתודות עזר נוספות. כמו כן, בתכנית הראשית הסברים לקלט והפלט של התכנית. **עקבו אחר ההוראות שם**.

אופן הגשה

1. לתרגיל מצורף קובץ Python בשם **<surname>_<first_name>_ex1.py**. עליכן לממש את התכנית (בגרסה 3.5 ומעלה) ולשנות את שם הקובץ בהתאם (לדוגמא: zeira_yuli_ex1.py). **אין להשתמש בספריות nltk, spacy או transformers**. על הקובץ לרוץ תחת הפקודה:
python <name_ex1.py> <xml_dir> <wiki_dir> <output_file>
(כמצוין בקובץ עצמו). **לקובץ הפלט יש לכתוב בקידוד utf-8**. הקפידו שזמן הריצה של התכנית לא יעלה על 5 דקות.
2. קובץ PDF בשם **<surname>_<first_name>_report1.pdf** (לדוגמא: zeira_yuli_report1.pdf) ובו דו"ח המפרט על הקוד שכתבתן ועל ההחלטות שקיבלתן במהלך העבודה. על הדו"ח להיות באורך שני עמודים לכל היותר.

*** יש להגיש את שני הקבצים הללו בנפרד, ולא כקובץ zip (ודומיו) ***

יש להקפיד על עבודה עצמית, צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד.
ניתן לשאול שאלות על התרגיל בפורום הייעודי לכך במודל.
יש להגיש את התרגיל עד לתאריך 19.03.22 בשעה 23:59.

בהצלחה!