דו"ח תרגיל 1 – עיבוד שפות טבעיות

מגיש: שגיא גוילי, ת.ז. 203638804

בתרגיל התבקשנו לכתוב תוכנית בפייתון אשר מקבלת: קובץ ערכים, קובץ שפות ונתיב של תיקיית יעד אליה היא תייצר קורפוס חדש אשר מורכב מקבצי טקסט אשר יכילו את התוכן של כל ערך בקלט כפי שהוא מופיע ב-Wikipedia, בכל שפה בקלט.

קבצי הטקסט יווצרו בעזרת חלוקה לטוקנים כך שיהיו נקיים מתגיות HTML, תמונות, וסימנים שאינם קשורים לשפה, מחולקים למשפטים ללא שורות ריקות בניהם.

אפרט את השלבים השונים שעברתי במהלך יצירת התוכנית והשיקולים שהייתי צריך לעשות בדרך

שלב הניקיון שליפת הנתונים בעזרת ספריית Wikipedia:

בכדי לקבל את תוכנו של הערך כפי שהוא מופיע ב-Wikipedia בכל שפה שמופיעה בקלט, רצתי בלולאה על קובץ השפות- בכל איטרציה הגדרתי את השפה בה ארצה לקבל את הערך על ידי הפונקציה set_lang בספרייה Wikipedia, לאחר מכן השמתי לתוך משתנה את דף הערך על ידי שימוש בפונקציה page (גם היא מתוך הספרייה Wikipedia), לאחר מכן שמרתי את התוכן על ידי הפונקציה content (גם היא מתוך הספרייה Wikipedia).

בעיה ראשונה ששמתי לב אליה בשלב זה היא קבלת ערכים שונים מהמצופה בספרייה Wikipedia, למשל עבור ADHD קיבלתי את הערך של Addition (שהוא פירש כפעולת חיבור ולכן קיבלתי את הדף של פעולת חיבור בויקיפדיה) הדבר ככל הנראה נובע מתקלה בזיהוי ראשי תיבות בספרייה; למשל כאשר ניסיתי את לקבל את הדף עבור הערך PC נזרקה לי שגיאת Disambiguation שכן הספרייה זיהתה את הערך PR משום מה ולו יש פירושים רבים שהוא לא ידע לעכל (בשונה מ-Addition שהוא במקרה כן ידע "לעכל" אבל פשוט נתן פלט לא נכון) הפתרון שלי לבעיה היה השימוש בפונקציה search של הספרייה Wikipedia, הפונקציה מחזירה לי את רשימת הערכים המתאימים לערך שהיא מקבלת, במקרה של ADHD הערך הראשון שהיא החזירה הוא Attention deficit hyperactivity disorder שהתאים לצורך שלי ואכן פתר את הבעיה. חשוב לציין שישנם ערכים שונים שלאו דווקא הצלחתי לתקן בעזרת search (למשל עבור PC הערך הראשון שקיבלתי היה עדיין PC ולכן זה לא עזר לי בכלום, אבל במרבית הדוגמאות ש-page לא ידעה "לעכל" search כן פתרה לי את הבעיה בערך הראשון ברשימה שהיא החזירה- למשל HTPC, TBD ועוד..

הבעיה השנייה שנתקלתי בה היא הקושי של הספרייה למצוא ערכים בשפות השונות, הבעיות בלטו בערכים הבאים:

- ADHD בשפה הצרפתית מצא לי את הערך Adam ובשפה הספרדית את הערך Adán ששניהם מבטאים את אותו הערך- אדם כלומר האדם הראשון במקום את הפרעת בעיות הקשב והריכוז אותה בעצם חיפשתי.
  ניסיתי לפתור את הבעיה בדומה לפתרון של הבעיה הראשונה- לחפש את הערך Attention deficit hyperactivity disorder במקום ADHD ואכן כאשר שיניתי את השפה והשתמשתי ב-search הערכים שקיבלתי היו מדוייקים יותר וקיבלתי את התוצאות הרצויות.

- Banana בשפה הצרפתית מחזיר לי את הערך של סדרת המנגה Banana Fish, גם כאשר אני משתמש בשיטת השימוש ב-Search אני עדיין מקבל את אותה התוצאה. ההשערה שלי שהסדרה יותר פופולרית בצרפת או מדינות אשר משתמשות בשפה הצרפתית יותר מאשר הפרי עצמו ולכן אלה התוצאות הרצויות יותר עבור הערך הזה בשפה הצרפתית.
  פתרון לבעיה זו לא מצאתי שכן השיטה היחידה היא תרגום המילה ושימוש במילה Banane

- Hurricane בשפה הצרפתית והספרדית גם החזיר פלטים פחות רצויים- כדוגמת קבוצות כדורגל ורוגבי (Hurricanes, Kiel Baltic Hurricanes) , ככל הנראה כי הם יותר נפוצים מאשר

הוריקנים אמיתיים, ובמקרה הטוב יותר שמות של הוריקנים ספציפיים (Huracán Katrina), גם פה הדרך היחידה לפתרון שמצאתי הוא תרגום המילה לשפות הללו ואז חיפושן.

היו כמובן עוד ערכים עם תופעות דומות, אך דאגתי שבכל מקרה גם אם ה-Search עזר ו-page בכל מקרה החזיר שגיאה (כי לא נמצא שום ערך עבור value כלשהוא) אזי נקבל פשוט דף ריק עבור הערך והתוכנית לא תיפול)

הבעיה השלישית שנתקלתי בה הייתה תגי HTML שונים וסימנים לא מובנים או פחות קשורים לשפה, הדבר בעיקר קרה בערכים בהם היו משוואות וכל מיני חישובים מתמטיים, למשל התגית displaystyle חזרה על עצמה בכל משוואה מתמטית (לדוגמא בערך Hurricane), הפתרון שלי היה להשתמש בביטויים רגולריים על מנת להוציא את כל התגיות הללו. תופעות נוספות של ביטויים מתמטיים היו רווחים גדולים בין אות לאות ואף שורות ריקות, הפתרון שלי היה בשלב בטוקניזציה לשמור על הביטויים הללו ובעצם כל משתנה או סימן מתמטי יהווה "token" בפני עצמו, העדפתי זאת על מחיקתם שכן חלק מהמשוואות קשורות לתוכן ובלעדיהם חלק מהדברים לא יובנו (מה גם שלפעמים הן היו חלק ממשפט ולא עמדו בפני עצמן)

דבר נוסף שהתקבל כתוצאה מכך היא מחיקת השורות הריקות בטקסט (כחלק מהדרישה בשלב הבא)

<u>שלב החלוקה למשפטים:</u>

כפי שנכתב בשלב הקודם, קיבלתי את ערך ה-value באמצעות הפונקציה content, התוכן שהתקבל היה די מסודר ברוב המקרים- מסודר בשורות ורווחים, לרוב בצורה ברורה וקריאה.

הדבר הראשון שעשיתי היה לנקות כל מיני ביטויים לא הגיוניים שמצאתי בטקסטים השונים, לדוגמא: "()", ",,", אות קטנה-נקודה-אות גדולה שהיו מחוברות אבל בעצם ברוב המקרים מדובר בהתחלה של משפט חדש אחרי הנקודה ועוד נוספים (כמובן שלא הכל ניתן לאתר אבל אלו מקרים ששמתי לב אליהם לכן החלטתי כן לטפל בהם)

לאחר מכן ביצעתי את החלוקה הראשונית של התוכן, עשיתי זאת על ידי איתור כל "\n", כלומר כל ירידת שורה, והשמת כל מה שבין לבין בתא משל עצמו במערך התוכן שהרכבתי. התוצאה הייתה חלוקה של כל פסקה וכותרת בתא משלה בתוך מערך התוכן של הערך (כאשר היו ביטויים מתמטיים גם הרבה ערכים מתמטיים בהרבה תאים שונים) שזו הייתה תוצאה לא רעה בשלב זה, אך בנוסף התקבלו הרבה תאים ריקים.

לאחר ניפוי תאים ריקים אלו, עברתי לשלב הטוקניזציה.

<u>שלב הטוקניזציה:</u>

בשלב זה העבודה עברה לרמת התא במערך התוכן, כלומר יכולה להיות לנו פסקה מאוד ארוכה עם הרבה סימני פיסוק (בייחוד ? . ! שבסבירות רבה מציינים התחלה של משפט, כלומר צורך בירידת שורה) או כותרת או ביטויים קצרים מאוד באורך של תו אחד לרוב בכל שורה (כמו ביטויים מתמטיים).

אציין שאינני שולט יותר מדי בכלל השפות השונות, לכן רוב החוקים על פיהם עבדתי הם חוקי השפה האנגלית ולכן הטוקניזציה אמנם תתאים לרוב השפות, אבל בעיקר לשפה האנגלית- לכן צפויות אי-טוקניזציות שונות בכל מיני שפות אחרות.

אפרט כעת את השינויים שעשיתי בטקסט בשלב זה:

- הדבר הראשון אותו עשיתי היה להוסיף רווח בין הנקודה בסוף השורה שיש בתא הנוכחי לאות שלפניה (אם אכן יש נקודה בסוף השורה) שכן בסבירות גבוהה מאוד היא תציין סוף משפט ולא כחלק מראשי תיבות כלשהם או כל שימוש אחר ולכן יזוהה כ-token בפני עצמו.

- **<u>צירופי נקודות (dots)</u>**- כלומר הצירופים כמו ".." וגם "..." שצצו לא פעם בכמה ערכים (וזה הרי שונה מהופעה של נקודה בודדת), לצירופים אלה הוספתי לצירופים "אמצעי זיהוי" בכדי שבהמשך הוא יזוהה כ-token בפני עצמו.
- **<u>ראשי תיבות</u>**- בערכים רבים זיהיתי מקרים של ראשי תיבות (Initials) ולכן היה צורך לטיפול מיוחד והשמתם בתור tokens בפני עצמם, דוגמאות בהן טיפלתי: H.C.A.Harrison, P.c.cinerus, a.k.a., v., U.S.
חשוב לציין שהטיפול בנושא יצר מצב שקשה לי לטפל במקרים כמו " XI. A player" שנחשב כעת ל-token בפני עצמו ולא זה הרצון, לכן הטיפול שלי היה חלקי ומה החלטתי לעשות זה לרדת שורה אחרי ".XI", מקרים נוספים כאלה גם קרו אבל החלטתי לזנוח אותם שכן הם לא רבים מדי.
- **<u>גרשיים</u>**- פסקאות שונות היו מורכבות אך ורק מציטוטים או ביטויים שונים בין גרשיים, במקרים כאלה חשוב היה להפריד את הגרשיים מתוכן המשפט עצמו כדי שיחשבו tokens בפני עצמם.
- **<u>סימני פיסוק (Punctuation)</u>** – סימני הפיסוק היו אתגר לא קטן, קרוב לוודאי שקיימים עוד צירופים שונים שלא מצאתי, אך אחרי מעבר על ערכים רבים טיפלתי במקרים הבאים:
  - <u>סוגריים</u>- יופיעו ב-token בפני עצמם, סוגר ב-token ופותח ב-token נפרד.
  - <u>גרשיים</u>- בנוסף למקרה הנ"ל של הגרשיים, אם קיימות גרשיים באמצע משפט הן יציינו לרוב ציטוט, שם נרדף, ביטוי ועוד פעולות- בכל מקרה גם הן יהיו ב-tokens משל עצמם שכן משמעות של מילה לא משתנה בגללן (בשונה מעברית שבה זה יכול להעיד על ראשי תיבות- במקרה הזה הגרשיים יהיו ביחד עם המילה כ-token)
  - <u>גרש</u>- בדומה מאוד לגרשיים, באנגלית גם כן יש לו משמעות כמו ב-"We're", אמנם זה בעצם צירוף של המילים "We" ו-"Are" אבל נכתב שאין לנו צורך לעשות ניתוח מורפולוגי ולכן אין צורך שהגרש יהיה בטוקן משלו ואכן התוכנית דואגת לכך. גרש שמופיע בצד מאוד דומה לגרשיים ומטופל בצורה דומה.
  - <u>קו</u>- הסימן "-" מופיע בצורות רבות ומשמעויות רבות: במשמעות דומה לנקודותיים, כלומר פירוט של כמה דברים אחריו (ולכן זה יכול להיות בשורה חדשה או בהמשך המשפט)- במקרה זה החלטתי שזה ישאר באותו משפט ולא ירד שורה כי זה המקרה השכיח יותר. במשמעות של צירוף מילים כמו "e-mail" שמבטא Electronic Mail, במקרה זה החלטתי שהצירוף ישאר כפי שהוא כלומר כ-token אחד.
במשמעות של קישור להמשך משפט ומשמעות של טווח בין זמנים ומספרים, כאשר במקרים אלו זה שמתי אותו ב-token משלו.
  - <u>נקודה</u>- ברוב המקרים נקודה מסמנת סוף משפט וצריכה להיות ב-token משלה, כאשר זהו אכן תפקידה, זיהיתי שלאחריה מופיע שורה חדשה (\n) או רווח, במקרה שהייתה שורה חדשה טיפלתי בהתחלה כאשר הוספתי רווח לנקודה בסוף פסקה שכן אני מיינתי את הערך לתאים על פי \n ולכן הוא לא באמת יופיע והרווח יחליף אותו בהבנת המשמעות. במקרה אחרים הנקודה הייתה כחלק מראשי תיבות (כפי שטופל למעלה) וסימן התחלה של שורה חדשה בקטע ממוספר כמו למשל ב-"1. 2. 3." שגם בו הנקודה צריכה לא צריכה להופיע ב-token משלה אלא ביחד עם המילה (או מספר), את המקרה הזה החלטתי להזניח שכן הוא פחות שכיח וטיפול בו יפגע בזיהוי מקרים אחרים שכיחים הרבה יותר.
  - <u>סימן שאלה וסימן קריאה</u>- הטיפול בהם מאוד דומה לנקודה, רק שהם פחות שכיחים. מציינים הרבה פעמים סוף משפט, כשזה קורה יופיע אחריהם רווח או \n וזה טופל כמו בנקודה.
  - <u>סוגריים מסולסלים ומרובעים</u>- לפעמים הופיעו כחלק מביטויים מתמטיים או תגיות HTML, במקרי התגיות הסרתי אותם לגמרי, במקרה של הביטויים מתמטיים החלטתי במקרים הנדירים שהם מופיעים להשאירם ולהתייחס אליהם כמו בשימוש הנפוץ יותר של סוגריים מסולסלים ולשים על סוגר ופותח ב-token משלו.
  - <u>נקודתיים</u>- הנקודתיים לרוב באות לפני פירוט של דברים, לפעמים בסוף משפט ואז הפירוט של הדברים בנקודות, ולפעמים כחלק ממשפט ללא ירידת שורה. החלטתי

להתייחס לנקודותיים ללא ירידת שורה אלא כאשר הן מופיעות הן יהיו חלק מהמשפט הנוכחי. במקרים מסוימים הופיעו הנקודותיים בביטויים מתמטיים או עם ירידת שורה מובנת בטקסט והשארתי אותם כפי שהם.

- נקודה ופסיק- השימוש בנקודה פסיק הוא מאוד מגוון בשפות שונות, בניהן שפות תיכנות רבות. לפעמים משמש כסוף משפט, לפעמים מאוד דומה לנקודותיים ובשפות אחרות בשימושים נוספים. החלטתי להתייחס לנקודה פסיק בדומה לנקודותיים שכן ברוב המקרים שראיתי זה אכן היה הפתרון המתאים.

היו צירופי "מילים-סימני פיסוק" נוספים שנפלו בין הכיסאות וטיפלתי בהם באופן ספציפי, עיקר הבעיות היו עם הנקודה לה משמעויות רבות ולכן היא יצרה הרבה מצבים נדירים שהעדפתי לטפל באופן ספציפי ופחות כוללני, מה שהיה יוצר לי בעיות חדשות (כמו שכתבתי בראשי תיבות)

בסוף התוכנית הוספתי כמה קיצורים באנגלית שהיה קשה לי לתפוס בצורה כוללת, כדוגמת .Dr, .Mr וכדומה.. ניסיון לתפוס אותם בצורה כוללת גרם לשגיאות רבות אחרות.

מקרים לא מטופלים:

ישנם מקרים שהעדפתי להזניח שכן הם שכיחים פחות וטיפול בהם יגרום לבעיות חדשות עם שכיחות גבוהה יותר, בדומה לבעיה המתוארת בראשי תיבות הלעיל.

- בשפה הספרדית, ערך שבחלקים מהדף שלו יש קישורים בתחתית הדף למקורם של חלקים אלו מופיעים בצורה של [number], לפעמים עם ירידת שורה ולפעמים לא, הצירוף הזה מכיל בתוכו עוד תווים נסתרים שהיה קשה לי לתפוס ולכן השארתי אותם כמו שהם.
- בקיצורים בשפה האנגלית בצורת 'runnin' או 'peelin' הגרש הופרד לטוקן משלו, טיפול בבעיה זו גרר בעיות שכיחות הרבה יותר מהמקרים הללו לכן החלטתי לספוג את זה.
- ביטויים מתמטיים הרבה פעמים הופיעו לי בצורות שונות ולכן אין לי דרך ספציפית שבה בחרתי להציג אותם, בכלליות כל אות בביטויים מתמטיים תופיע בדרך כלל עם רווח בינה לבין שכנתה וניתן יהיה להבחין שאין טקסט סדור אלא "יצור" אחר.
- במקרים בהם יש לנו נקודה באמצע משפט בתוך סוגריים או גרשיים תתבצע ירידת שורה- אין מקרים שכיחים מדי של המקרה והיה לי קשה לאתר את העניין בצורה כזאת שתטפל בכלל המקרים ולכן החלטתי לספוג את זה.