פרויקט 1 – ניהול נתונים באינטרנט

שמות המגישים:

לירי נורקין

208788448

תמיר סדובסקי

315316612

תיאור של הקוד שבונה את האונטולוגיה

נתחיל בתיאור כללי של הflow, לאחר מכן פירוט פרטני של הפונקציות המרכזיות במימוש וכן פונקציות עזר.

תיאור כללי:
ראשית, אנחנו מכניסים את המדינות המופיעות בעמוד הנתון במטלה לתור שמכיל את הקישורים לדפי הויקיפדיה של כל אחת מהמדינות בטבלה הנתונה בעמוד.
לאחר מכן, באופן איטראטיבי מוציאים כל אחת מהמדינות מהתור ומוצאים עבורה את כל הנתונים הנדרשים לבניית האונטולוגיה, מכל עמוד ויקיפדיה של מדינה, אנו מחלצים את עיר הבירה, גודל האוכלוסייה, שטח המדינה, צורות ממשל, ראש ממשלה ונשיא מדינה. בשלב הבא, כל אחד מהנתונים האלו מוכנסים לאונטולוגיה עם הקשר המתאים למדינה.
בשלב הבא, עוברים לעמודי הויקיפדיה של ראש הממשלה ונשיא המדינה ומחלצים משם את תאריך הלידה וארץ הלידה עבורם, מכניסים עם קשרים מתאימים לאונטולוגיה. לאחר מכן, עוברים למדינה הבאה וחוזרים על השלבים הקודמים עד שהתור ריק וסיימנו לעבור על כל המדינות ולחלץ את הנתונים לבניית האונטולוגיה.

פונקציות מרכזיות (לפי סדר כרונולוגי):

initialize_crawl – פונקציה זאת עוברת בעזרת לולאה בסיסית על התור של הקישורים לדפי הויקיפדיה של המדינות ותיאור אינדיקטיבי של מדינה (שנוצר בפונקציה from_source_url_to_queue). בתהליך זה, הטיפול מתבצע על כל צמד כזה בנפרד עד שהתור ריק. הטיפול בכל צמד נעשה על ידי הפונקציה get_from_url שעוברת על דף הויקיפדיה הנתון ומחלצת ממנו את המידע הנדרש.

add_to_ontology – פונקציה זאת מקבלת שני פריטי מידע ואת הקשר ביניהם.
מוסיפה אותם לאונטולוגיה עם תחיליות מתאימות בהתאם לסטנדרט שראינו בתרגול.

from_source_url_to_queue – פונקציה זאת בונה את התור של ה tuples "Country", "href".
היא מקבלת את האתר הראשוני בויקיפדיה (של טבלת המדינות) שממנו נחלץ את המדינות המופיעות בטבלה. בעזרת lxml מתבצע החילוץ של המדינות באמצעות שאילתת XPATH, עם הלולאה רצים על כל הטבלה, במהלך הריצה מכניסים לתור url_queue. בנוסף במהלך ריצה זאת אנו יוצרים את הרשימה countries שתשמש אותנו עבור מציאת ארץ הלידה של נשיא / ראש ממשלה בהמשך. כמו כן בפונקציה זאת יש טיפול מיוחד במספר מדינות שעבורן ההכנסה לתור ולרשימה התבצעו באופן שונה כיוון שהlink שלהן היה מיוחד, לשם המדינה נוספו סימנים (מוסבר בפירוט במקרי קצה).

get_from_url – פונקציה זאת מחלצת את שם המדינה והקישור לעמוד הויקיפדיה שלה מתוך הtuple שקיבלה מהפונקציה Initialize_crawl.
היא נעזרת בפונקציות כמו data_spaces_to_bottom_line (מוסברת בפונקציות עזר). לאחר מכן נעזרת בפונקציות בת: add_capital, add_area, add_government, add_population שכל אחת מהן אחראית לחלץ מידע אחר מעמוד הויקיפדיה של המדינה (יפורט על פונקציות אלה בהמשך). בשלב הבא, יש קריאות לפונקציה add_president_or_prime_minister שמטפלות בחילוץ נשיא וראש ממשלה של המדינה (פירוט בהמשך). שאילתות הXPATH עבור כל אחת מהפונקציות נוצרו וישלחו לכל אחת מהן, בכל אחת מפונקציות הבת נתאר את השימוש בשאילתא שהתקבלה.

add_president_or_prime_minister – פונקציה זאת מתאימה לחילוץ מידע על נשיא וראש ממשלה מעמוד ויקיפדיה של מדינה (כאשר המידע מופיע בתוך הinfobox). על היישות שהתקבלה מפונקציית האב נבצע התאמה עבור הכנסה תקינה לאונטולוגיה בעזרת פונקציות העזר (מתוארות בהמשך) וכן לפי קידוד של 'utf-8'. לאחר מכן ישנן קריאות לפונקציות נוספות add_birthday, add_birth_location שבהן נוסיף את ארץ הלידה ותאריך הלידה של נשיא / ראש ממשלה לאונטולוגיה.

add_capital – פונקציה זאת מוסיפה עיר בירה של מדינה לאונטולוגיה.
בעזרת lxml מתבצע החילוץ של העיר בירה באמצעות שאילתת XPATH מתאימה שהתקבלה מפונקציית האב. על היישות שהתקבלה מפונקציית האב נבצע התאמה עבור הכנסה תקינה לאונטולוגיה בעזרת פונקציות העזר וכן לפי קידוד של 'utf-8'.
טיפול במקרה קצה על המדינה Channel_Islands מתואר בחלק של מקרי הקצה.

add_area – פונקציה זאת מוסיפה את שטח המדינה לאונטולוגיה.
בעזרת lxml מתבצע החילוץ של שטח המדינה באמצעות שאילתת XPATH מתאימה שהתקבלה מפונקציית האב. לאחר מכן, ניקח את האיבר הראשון ברשימה המוחזרת.

add_government – פונקציה זאת מוסיפה את צורות הממשל של המדינה לאונטולוגיה.
בעזרת lxml מתבצע החילוץ של צורות הממשל באמצעות שאילתת XPATH מתאימה שהתקבלה מפונקציית האב. היישות שהתקבלה מנוקה מערכים שאינם צורות ממשל ותווים מיותרים לפי קידוד של 'utf-8' (כנאמר בפורום ע"י דנה). לאחר מכן מתבצע מיון אלפבתי של הרשימה ואז כל צורת ממשל ברשימה מוכנסת כאונטולוגיה נפרדת.

add_population – פונקציה זאת מוסיפה את גודל אוכלוסיית המדינה לאונטולוגיה.
בעזרת lxml מתבצע החילוץ של גודל האוכלוסייה באמצעות שאילתת XPATH מתאימה שהתקבלה מפונקציית האב. על היישות שהתקבלה מפונקציית האב נבצע התאמה עבור הכנסה תקינה לאונטולוגיה בעזרת פונקציות העזר (מתוארות בהמשך).
טיפול במקרה קצה על מדינות בודדות מתואר בחלק של מקרי הקצה.

add_birth_location – פונקציה זאת מוסיפה את ארץ הלידה של נשיא / ראש ממשלה לאונטולוגיה. בעזרת lxml מתבצע החילוץ של ארץ הלידה באמצעות שאילתת XPATH מתאימה שהתקבלה מפונקציית האב. מהרשימה שהתקבלה נחפש האם קיימת מחרוזת (לאחר התאמה של רווחים) שתואמת את רשימת המדינות שנוצרה קודם לכן. באופן דומה מתבצע חיפוש הפוך, *מוסבר בפירוט במקרי קצה.*
במידה ולא מתבצעת בדיקה נוספת עבור מקרים חריגים שמתוארת במקרי הקצה באופן מקיף. ולבסוף התאמה עבור הכנסה תקינה לאונטולוגיה.

add_birthday – פונקציה זאת מוסיפה את תאריך הלידה של נשיא / ראש ממשלה לאונטולוגיה.
בעזרת lxml מתבצע החילוץ של ארץ הלידה באמצעות שאילתת XPATH מתאימה שהתקבלה מפונקציית האב. על היישות שהתקבלה מפונקציית האב נבצע התאמה עבור הכנסה תקינה לאונטולוגיה בעזרת פונקציות העזר.

- נציין כי הסקנו מהפורום שיש לחלץ תאריך לידה אך ורק אם קיים עבור התגית bday, כפי שנכתב ע"י דנה.

פונקציות עזר:

data_spaces_to_underlines - פונקציה זאת מקבלת מחרוזת ומחליפה את הרווחים במחרוזת בקווים תחתונים. השימוש של פונקציה זאת הוא עבור הכנסת מידע באופן תקני לאונטולוגיה.

data_hyphens_to_underlines – פונקציה זאת מקבלת מחרוזת ומחליפה את המקפים במחרוזת הנתונה בקווים תחתונים. השימוש של פונקציה זאת הוא עבור הכנסת מידע באופן תקני לאונטולוגיה.

באופן דומה הפונקציות remove_hyphens ו – remove_underlines מורידות מקפים וקווים תחתוני שלא נחוצים ממחרוזת נתונה.

extract_country_from_url – פונקציה זאת מקבלת מחרוזת שמייצגת קישור לעמוד ויקיפדיה ומוציאה ממנה מחרוזת שמייצגת את המדינה אליה שייך הקישור, בעזרת חילוץ המחרוזת לאחר המקף האחרון בקישור.

שאלה נוספת:

Does <prime minister> born in <country>?

השאלה בודקת האם נשיא מסוים נולד במדינה מסוימת. התשובות האפשריות הן: True/False.
כמה דוגמאות:

שאלה 1:

Does Christian Ntsay born in Madagascar?

תשובה 1:

True

שאלה 2:

Does Naftali Bennett born in Israel?

תשובה 2:

True

שאלה 3:

Does Robert Abela born in France?

תשובה 3:

False

תיאור מקרי קצה:

.1

בפונקציה add_capital נתקלנו במקרה מיוחד בחיפוש עיר הבירה עבור המדינה Channel Islands, בinfoxbox שלה הכותרת עבור עיר הבירה היא Capital and largest settlement, לכן השאילתא היא בהתאם:

```
capital = doc.xpath('//table[contains(@class, "infobox")]/tbody/tr[th//text()="Capital and largest settlement"]//@title')
```

לאחר מכן, לקחנו את התא הראשון מהרשימה שהוחזרה מהשאילתא וקיבלנו את העיר בירה המבוקשת.
חוץ ממדינה זאת, בכל המדינות הinfobox ובפרט הכותרת של העיר בירה היא Capital.

Capital and largest city	Jerusalem (limited recognition)[fn 1][fn 2] 31°47′N 35°13′E
Official languages	Hebrew
Recognized languages	Arabic[fn 3]
Ethnic groups (2019)[13]	74.2% Jews 20.9% Arabs 4.8% Others

Infobox from Israel Wikipedia site

(For example)

Administration	
Bailiwick of Guernsey	
Capital and largest settlement	Saint Peter Port, Guernsey
Area covered	78 km^2 (30 sq mi; 39.4%)
Bailiwick of Jersey	
Capital and largest settlement	Saint Helier, Jersey
Area covered	118 km^2 (46 sq mi; 59.6%)

Infobox from Channel Islands Wikipedia site

.2

בפונקציה add_population נתקלנו במקרים מיוחדים בחיפוש גודל האוכלוסייה עבור המדינות Russia, Dominican Republic, Channel Islands. בinfobox שלהן, הpopulation מסודר באופן היררכי שונה מהמדינות האחרות לכן שאילתות הXPATH הותאמו להן ידנית עבור כל אחת מהן.

```
population = doc.xpath('//table[contains(@class,"infobox")]/tbody//tr[contains(.//text(),"Population")]/following-sibling::tr/td//text()')
if country == "Russia":
    population = doc.xpath('//table[contains(@class,"infobox")]/tbody//tr[contains(.//text(),"Population")]/following-sibling::tr/td/div/ul/li/text()')
elif country == "Dominican_Republic":
    population = doc.xpath('//*[@id="mw-content-text"]/div[1]/table[1]/tbody/tr[37]/td/span/text()')
elif country == "Channel_Islands":
    population = doc.xpath('//*[@id="mw-content-text"]/div[1]/table[1]/tbody/tr[21]/td/text()[1]')
```

.3

בפונקציה add_birth_location נתקלנו במקרה מיוחד בחיפוש ארץ הלידה עבור ראש ממשלה / נשיא של מדינה. כיוון שהחילוץ מהשאילתא עם התגית Born מהinfobox נותן תוצאות בצורות שונות, נטפל בהם באופן שונה. ראשית, כיוון שפעמים רבות החילוץ בעזרת שאילתת הXPATH מניב רשימות שהערכים בהן עם סימנים שמקשים על הבדיקה (פסיקים, סוגריים, נקודות ורווחים – אלה הסימנים הנפוצים ביותר) נוריד אותם. לאחר מכן, ישנן שתי בדיקות כדי לתפוס יותר מקרים, קודם כל נבדוק האם לאחר ההורדה קיימת ברשימה מדינה מרשימת המדינות שנוצרה בהתחלה, אם כן נוסיפה וסיימנו, אחרת, נבצע בדיקה נוספת כי מדינות רבות עדיין מתפספסות באופן זה, נבדוק באופן ההפוך, האם קיימת מדינה מרשימת המדינות ההתחלתית שנמצאת ברשימה שהוחזרה מהשאילתא, אם כן נוסיפה וסיימנו.
אם עדיין לא מצאנו, בדיקה נוספת בעזרת השאילתא, נחלץ את שם המדינה הנמצא בכותרת שקשורה לתגית Born.

```
list_from_title = doc.xpath('//table[contains(@class, "infobox")]/tbody/tr[th//text()="Born"]/td//@title')
```

כמו כן, נוצרה תוספת ידנית עבור המקרה של Jorge Bom Jesus כיוון שבinfobox בעמוד שלו, מקום הלידה עבורו נכתב ללא התויים המיוחדים שיש ברשימת המדינות.

.4

טיפול בפונקציה from_source_url_to_queue עבור מדינות שה url שלהן מכיל סימנים נוספים בשם המדינה שמקשים על חילוץ נקי של שם המדינה. מקרה זה זוהה ע"י בדיקה האם התו "%" נמצא בשם המדינה, זהו אחד מהתווים שהיו בכל 3 המדינות האלו. לאחר מכן, הטיפול נעשה בעזרת urllib unquote.

.5

ראש ממשלת The Bahamas, Philip "Brave" Davis, שמור בויקיפדיה עם מרכאות. לא ניתן להכניס מרכאות לאונטולוגיה ומכיוון שזה המקרה היחיד בו נתקלנו בסוגיה, החלטנו לטפל בה ידנית. לכן החלפנו את המרכאות בסימן '@', והכנסנו את היישות הבאה -
<http://example.org/Philip_@Brave@_Davis> בכל אחת מן האנטולוגיות בהן הופיע Philip "Brave" Davis. בזמן מענה של שאלות, שינינו את הסימן '@' בחזרה למרכאות ע"י replace("@","")'.