קובץ תיעוד לפרוייקט 1 - ניהול נתונים באינטרנט

לירון כהן 207481268, יובל מור 209011543

תיאור הקוד

חלק א' - ייצור קובץ האונטולוגיה (create_graph)

- ייצור רשימת ה-urls של המדינות מהעמוד הנתון (get_countries_urls)
  - הורדת תוכן העמוד הנתון באמצעות requests, lxml.
  - שימוש בשאילתת xpath (COUTNRIES_XPATH_QUERY) על העמוד הנתון להוצאת המדינות.
  - שמירת שמות המדינות בקבוצה (countries_set).
  - הוספה ידנית של שלוש מדינות באמצעות שאילתות ייעודיות.

- הוספת שלישיות לגרף עבור כל מדינה (add_triplets_to_graph)
  - פרסור שם המדינה והוספתו לקבוצת המדינות.
  - הורדת תוכן עמוד המדינה באמצעות requests, lxml.
  - הוספת שלישייה לגרף באמצעות הרצת שאילתת ה-xpath המתאימה עבור ה-relation המתאים (add_country_triplet_to_graph).
  - אם מדובר בשאילתת president או prime minister מתבצע מעבר לעמוד האדם והוספת השלישיות המתאימות לו (add_person_triplets_to_graph), תוך התייחסות למדינת הלידה בהשוואה לרשימת המדינות שהתקבלו מהעמוד הנתון.
  - הוספה ידנית של שאילתות Population ו-Area מיוחדות עבור מקרי קצה.

- יצירת האונטולוגיה ושמירתה
  - מתבצע באמצעות פקודה serialize על הגרף שהתקבל.

חלק ב' - מציאת תשובה לשאלה שנשאלה (ask_question)

- המרת השאלה לשאילתת sparql (parse_question_to_query)
  - מציאת סוג השאלה לפי מילת השאלה ומילות מפתח נוספות.
  - פרסור הפרמטרים הנחוצים לשאילתא מתוך השאלה.
  - ייצור השאילתא המתאימה (generate_<x>_query) והחזרתה.

- ייבוא הגרף מהקובץ והרצת השאילתא
  - פרסור המשתנה המתאים שהגיע מהשאילתא.
  - סידור התשובה ברשימה ממוינת לפי הצורך.
  - התייחסות מיוחדת לשאילתת who is שנכתבה כאיחוד של שתי שאילתות (אחת ל-president ואחת ל-prime minister).
  - הדפסת התשובה ויציאה.

תיאור השאלה שהוספנו

השאלה שהוספנו היא:

List all countries that their names and their capitals' names end with the string <str>

דוגמאות לתשובות אפשריות

- עבור המחרוזת "ia" התקבלו התוצאות Bulgaria (שעיר בירתה Sofia) ו-Liberia (שעיר בירתה Monrovia).
- עבור המחרוזת "e" התקבלה התוצאה Zimbabwe (שעיר בירתה Harare).

תיאור מקרי קצה

1. חילוץ מדינות מהעמוד הנתון

כאשר ניגשנו לחלץ את רשימת ה-urls של המדינות מהעמוד הנתון, היו שלוש מדינות שלא חולצו (Afghanistan ,Western Sahara ,Channel Islands).

הוספנו שלוש שאילתות xpath ייעודיות שמחפשות את ה-title של המדינה הרלוונטית ומחלצות את הקישור המתאים. את תוצאות השאילתות הוספנו למקומות המתאימים ברשימת ה-urls.

2. שאילתת Population נוספת

כאשר ניגשנו לחלץ את שדה ה-population מעמודי המדינות, היו מספר מדינות שהשדה שלהן לא חולץ (Belarus, Dominican Republic, Malta, Russia).

המבנה ההיררכי של tr/td/text() לא התאים לעמודים שלהן ולכן הוספנו שאילתא המתאימה למבנה ההיררכי tr/td/span/text(). דרכים אחרות, כגון שימוש ב-//, הובילו לזיבול התוצאות והחלטנו ששאילתא ייעודית היא הדרך הנכונה לפתור את החוסר.ד

3. שאילתת Birth Place נוספת

כאשר ניגשנו לחלץ את שדה ה-Place of Birth של נשיאים וראשי ממשלות, היו אישים (דוגמת יצחק הרצוג ונפתלי בנט) אשר החזירו תוצאות לא נכונות עבור השאילתא שכתבנו, או שלא החזירו תוצאות כלל.

ההבדל נבע מהצגת התשובה בקישור או בטקסט (לעיתים שם העיר הוצג בקישור ושם המדינה בטקסט), ולכן המבנה ההיררכי tr/td/a/@href הומר ל-tr/td/text().