ניהול נתונים באינטרנט - עבודה מעשית 1

שמות המגישים:

שם: נתן בלוך.

שם: שי פוקס.

<u>תיאור הקוד שבונה את האונטולוגיה:</u>

- אתחול מבנה האונטולוגיה ע"י שימוש ב-rdflib.
- באמצעות xpath השגנו את הלינקים לכל הסרטים הזוכים בין השנים 2010-2020.
- בהינתן לינק של סרט מהשלב הקודם ובאמצעות xpath וע"י גישה ל-infobox בלבד השגנו את המפיקים של הסרט, הבמאים של הסרט, השחקנים/הכוכבים בסרט, אורך הסרט(בדקות), תאריכי בכורה של הסרט, והאם הסרט מבוסס על ספר.
- בהינתן לינק של אדם מהשלב הקודם ובאמצעות xpath וע"י גישה ל-infobox בלבד השגנו את המקצוע שלו, ותאריך הלידה שלו.
- בהינתן המידע שחילצנו באמצעות xpath בשלבים הקודמים הוספנו את הישויות ואת היחסים לאונטולוגיה שאיתחלנו קודם לפי הנדרש במטלה.
- בהתאם להנחיות ביצענו parsing לחלק מהמידע שנאסף , למשל עבור bday בחלק מהלינקים ה-bday הגיע בצורה שאינה תואמת ל-span bday כלומר שאינו מהצורה yyyy-mm-dd ולכן אינה תואמת להנחיות ולכן בהתאם להנחיות הוספנו לאונטולוגיה רק את השנה מכל התאריך שחילצנו.
- הישויות והיחסים באונטולוגיה הם מהצורה הבאה :
  - עבור יחס כלשהו בשם relation_name: <http://example.org/relation_name>.
  - עבור ישות כלשהי בשם entity_name: < http://example.org/entity_name>.
- שמות של בני אדם וסרטים הן capitalized כפי שמקובל ומקצועות ב-lowercase כפי שהתבקש.
- ישויות שיש להן לינק הוכנסו לאונטולוגיה על פי הלינק וישויות ללא לינק הוכנסו על פי טקסט לפי ההנחיות.

<u>תיאור של השאלה שהוספנו למערכת ותשובות אפשריות לגביה:</u>

- השאלה שהוספנו היא : **"Who was born in <bday>?"**, כאשר <bday> הינו מהאופנים שהוגדרו בעבודה: yyyy-mm-dd או yyyy (כלומר לפי האופן שבו הוספנו את ה-bday לאונטולוגיה שבכפוף להנחיות הפרויקט).

- נציג את שלושת הדוגמאות הבאות:
  - עבור "Who was born in 1969?" אנו מצפים לקבל את התשובה הבאה:

    Marshall Curry, Richard Suckle
  - עבור "Who was born in 1956-08-20?" אנו מצפים לקבל את התשובה הבאה:

    Joan Allen

<u>בעיות שהתמודדנו איתן בעת חילוץ המידע ע"י xpath:</u>

1. הבדלי קידוד בין אותיות לטיניות לבין הקידוד המתאים של לינק (כלומר קידוד שדפדפן יכול להריץ) , דוגמה:

```
<a href="/wiki/Chlo%C3%A9_Zhao" title="Chloé Zhao">Chloé Zhao</a>
```

כאן ניתן לראות כי השם של הבמאית בטקסט הוא באותיות לטיניות אך הלינק עצמו מקודד אחרת.

2. בחילוץ ה-occupation של persons חלק מהמקצועות הם מבוססי לינק וחלק מבוססי טקסט, אך חלק מהלינקים שכתובים מקיימים שה-href attribute שלהם אינו תואם לטקסט תחת הלינק.למשל בדוגמה הנ"ל:

```
class="infobox-label">Occupation</th><td class="infobox-data role"><a href="/wiki/Actor" title="Actor">Actress</a></td>
```

מופיע לינק עם טקסט actress אך הוא מצביע לדף הויקיפדיה של actor.

3. היינו צריכים לתחזק blacklist של שגיאות נפוצות. למשל בקידוד לינקים, קיימים לינקים פנימיים של ויקיפדיה שאינם לינקים אמיתיים למשל ציטוט או reference, וכדי לתחזק רשימה כזו היינו צריכים לעבור על כל המידע שחולץ מה-xpath באופן ידני ולמצוא את טעויות אלו.

חיפוש xpath שהיה שונה משאר החיפושים היה עבור חילוץ occupation , שם נדרשנו לעבור גם על גבי li element שהוא בעצם list של html וזהו המקום היחיד שבו נדרשנו לעבור על element מסוג כזה(בכל שאר החיפושים היינו עוברים לינקים או על טקסט).