תרגיל תכנותי #1 – Information Extraction

בתרגיל זה תבנו מערכת למענה על שאלות בשפה טבעית בנושא סרטים זוכי אוסקר, תוך שימוש בידע שלכם על אונטולוגיות, HTML, SPARQL ו-Xpath. התרגיל להגשה עד ה-01.06, וכמו כל תרגילי הבית, יש להגישו בזוגות בלבד. תרגיל זה מהווה 11% מהציון הסופי בקורס.

תיאור המערכת

על המערכת לדעת לענות על שאלות מהסוגים המפורטים למטה. כל השאלות יהיו באנגלית ויהיו תמיד מאחת התבניות הבאות.

<u>שאלות על ישויות ספציפיות:</u>

1. Who directed <film>?
2. Who produced <film>?
3. Is <film> based on a book?
4. When was <film> released?
5. How long is <film>?
6. Who starred in <film>?
7. Did <person> star in <film>?
8. When was <person> born?
9. What is the occupation of <person>?

<u>שאלות כלליות:</u>

1. How many films are based on books?
2. How many films starring <person> won an academy award?
3. How many <occupation1> are also < occupation2>?

<u>שאלה לבחירתכם:</u>

עליכם להוסיף שאלה נוספת לבחירתכם, שמסתמכת על המידע הקיים במערכת.

השאלות יכולות להכיל התייחסויות לשני סוגי משתנים:

- Entity: ישות שיש לה ערך בויקיפדיה. לדוגמא לישות Emma Watson יש את הדף https://en.wikipedia.org/wiki/Emma_Watson. שם הישות יהיה זהה לשמה בURL של דף הויקיפדיה שלה עם רווח במקום קו תחתון.
- Relation: כל יחס הוא שדה ב-Wikipedia Infobox של הישות.
  למשל התשובה לשאלה: ?Who directed The Great Gatsby (2013 film)
  תהיה Baz Luhrmann, כאשר המידע על היחס מגיע מהשדה המסומן ב-Infobox.

איסוף המידע ובניית האונטולוגיה

עליכם לאסוף מידע על הסרטים המופיעים בעמוד הזה:

https://en.wikipedia.org/wiki/List_of_Academy_Award-winning_films

שימו לב שעליכם לחלץ מידע לא רק מה-infobox בעמודי הסרטים, אלא גם בעמודים של המפיקים/ות, במאים/ות והשחקנים/שחקניות. השתמשו בידע שלכם על Xpath ו-SPARQL כדי לעבור בצורה אוטומטית על הדפים הרלוונטים ולחלץ משם את המידע הדרוש.

העמוד מכיל מספר רב של סרטים, ואנחנו נבדוק את איסוף המידע רק על סרטים ששנת הזכיה שלהם (לפי הטבלה הראשית) היא 2010 (כולל).

את האונטולוגיה יש לשמור בקובץ בשם ontology.nt ולהגיש אותה.

מענה על שאלות בשפה טבעית

על התוכנית לדעת להתמודד עם שאלות באנגלית על גבי האונטולוגיה. בהנתן שאלה באנגלית (מאחד מ-12 המבנים למעלה) על התוכנית לתרגם את השאלה לשאילתת SPARQL שתורץ מעל האונטולוגיה שבניתם ותחזיר את התשובה. התשובה לא צריכה להיות "תשובה מלאה", אלא רק להכיל את הערך הנדרש. למשל עבור השאלה:

Is Little Women (2019 film) based on a book?

התשובה תהיה: Yes
אין צורך לציין את שם הספר.

בשאלות מסויימות יכולה להיות יותר מתשובה אחת, כמו למשל:

Who produced The Great Gatsby (2013 film)?

במקרה כזה נציג את כל התשובות מופרדות פסיקים (רווח אחרי כל פסיק), וממויינות בסדר לקסיקוגרפי:

Baz Luhrmann, Catherine Knapman, Catherine Martin (designer), Douglas Wick, Lucy Fisher

הרצת הקוד

- על הקוד להיות כתוב בפייתון 3 ולרוץ באופן תקין בנובה. יש לפרט בקובץ requirements.txt כל ספריה חיצונית שנעשה בה שימוש (שמות כל החבילות מופרדות ע"י ירידת שורה)
- התכנית תיקרא film_qa.py ותרוץ משורת הפקודה באופן הבא:
  - python film_qa.py create
    במצב create התכנית תייצר את הקובץ ontology.nt שיכיל את האונטולוגיה שבניתם ותסיים לרוץ.
  - python film_qa.py question "<question>"
    במצב question התכנית תקבל שאלה בשפה טבעית, תדפיס למסך את התשובה לשאלה ותסיים לרוץ.
    השאלה ניתנת כמחרוזת אחת, כלומר מועברת בשורת הפקודה כמחרוזת שמתחילה במרכאות ומסתיימת במרכאות.
- על התכנית להסתיים לאחר הרצת הפקודה (create או question). אין להשאיר את התכנית רצה.

תיאור הפרוייקט

עליכם להגיש קובץ נוסף בשם project.pdf שיכיל את הפרטים הבאים

- שמות ומספרי הת.ז של המגישים
- תיאור של הקוד שבונה את האונטולוגיה, איך הוא בנוי ומה עשיתם
- תיאור של השאלה שהוספתם למערכת ודוגמאות לתשובות אפשריות
- תיאור של שלושה מקרי קצה שהתמודדתם איתם באיסוף המידע. הסבירו אילו חיפושי xpath מיוחדים הייתם צריכים להוסיף ואיך המבנה של המקרה הזה היה שונה ממקרים אחרים.

הוראות הגשה

עליכם להגיש קובץ zip בשם hw1_<id1>_<id2>.zip שיכיל את הקבצים הבאים:

1. film_qa.py - הקובץ שמכיל את התכנית שבונה את האונטולוגיה ועונה על השאלות
2. ontology.nt – קובץ אונטולוגיה בנוי
3. project.pdf – תיאור הפרוייקט
4. requirements.txt – קובץ txt שמכיל את שמות כל החבילות החיצוניות שנחוצות להרצת הפרוייקט, מופרדות ע"י ירידת שורה.

אין בעיה לפצל את הקוד למספר קבצים ולהוסיף קבצי עזר כל עוד הקוד עובד כמצופה. במקרה כזה יש להגיש את כל הקבצים הרלוונטים. קבצים שאינם zip לא יבדקו.

בדיקת הפרוייקט

ייבדקו 20 שאלות בשפה טבעית. 12 מהשאלות זמינות לכם במודל כך שמובטח שאם התכנית שלכם עונה עליהן בצורה תקינה תקבלו ציון עובר. 8 השאלות הנוספות נסתרות.

הפרוייקט יבדק באופן אוטומטי לחלוטין בנובה, לכן אנחנו ממליצים להקפיד שהקוד רץ ללא שגיאות ועומד במבנה התשובות שנדרש. תשובות בפורמט אחר, או שונות מהתשובות המצופות אפילו בתווים בודדים ייחשבו כתשובה שגויה.

המידע בויקיפדיה משתנה עם הזמן ויכולים להיות שינויים בערכים שרלוונטים לפרוייקט. אם אתם חושבים שהתשובה באחת השאלות הגלויות השתנתה, בבקשה תכתבו לנו בפורום כדי שנוכל לעדכן.

בהצלחה!