ניהול נתונים באינטרנט – תרגיל מסכם – Question Answering

הוראות:
יש לעלות את הפתרונות בקובץ ZIP לMOODLE, שכולל קובץ PDF בשם answers.pdf ובו הפתרון וקבצי קוד נוספים (HTML, XML או Python) לפי הדרישה של כל סעיף וסעיף. ההגשה היא בזוגות, ורק אחד מבני הזוג יגיש את התרגיל, אך יש להקפיד לכתוב את השמות והת.ז. של שני בני הזוג בתוך הקובץ.
שם של הקובץ ZIP צריך לכלול את הת.ז. של אחד מהמגישים (למשל: HWqa_123.zip).
תאריך הגשה: **17.06.2018**

רקע:
זהו התרגיל המסכם של הקורס, בו תיישמו את מה שלמדתם בתרגילים הקודמים (1-3) כדי לבנות מערכת למענה על שאלות בשפה טבעית (Question Answering) בעזרת Wikipedia. עליכם להעזר בידע על HTML, Xpath, IE, SPARQL, Ontology לכתיבת המערכת.
התרגיל להגשה עד היום הראשון של חופשת סמסטר ב' (17.06.2018) וכמו תרגילים קודמים, ניתן להגישו בזוגות או ביחידים. תרגיל זה מהווה 4 נק' מהציון הסופי בקורס.

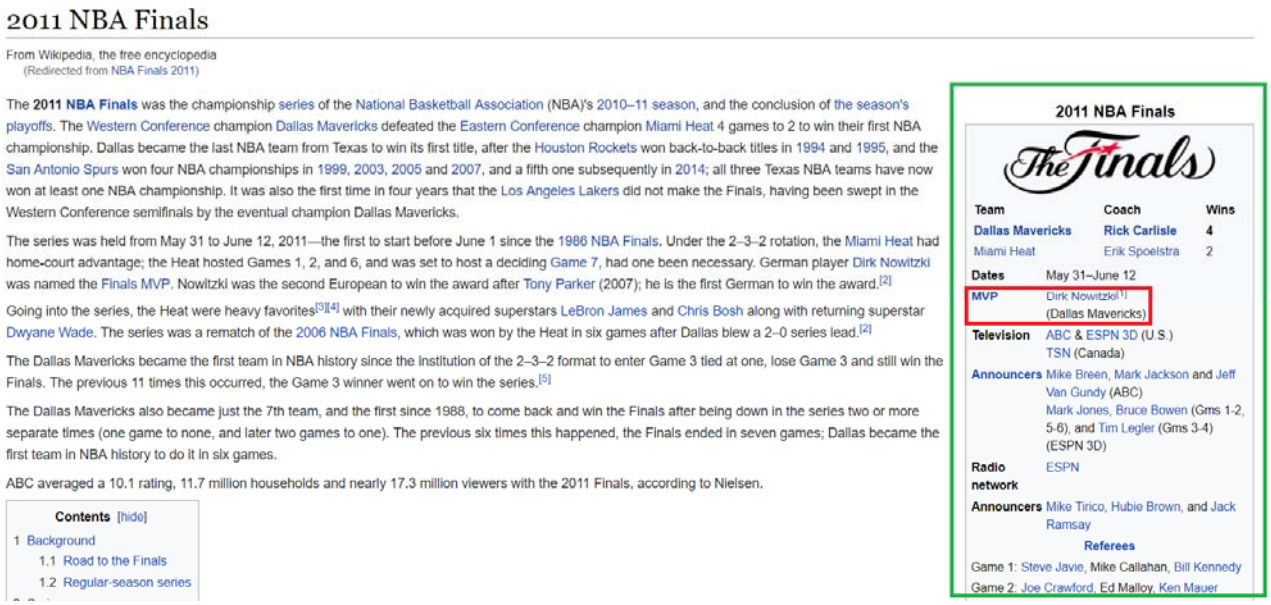
משימה:
בנו מערכת המקבלת כקלט שאלה בשפה טבעית ומחזירה:
1. את התשובה לשאלה.
2. את תרגום השאלה לשפת SPARQL.
3. אונטולוגיה מתאימה (יוסבר בהמשך) שהרצת שאילתת ה-SPARQL עליה תחזיר את התשובה.

<u>סוגי שאלות</u>:
כל השאלות יהיו בשפה האנגלית ויכללו <u>תמיד</u> את אחד מ-3 המבנים הבאים,
(i) Who is the <relation> of [the] <entity>?
(ii) What is the <relation> of [the] <entity>?
(iii) When was <entity> born?

לדוגמה:
1. Who is the president of Italy? Sergio Mattarella
2. Who is the spouse of Gal Gadot? Yaron Varsano
3. What is the alma mater of Gal Gadot? IDC Herzliya
4. Who is the MVP of the 2011 NBA Finals? Dirk Nowitzki
5. What is the best picture of the 90th Academy Awards? The Shape of Water
6. What is the capital of Canada? Ottawa
7. When was Theodor Herzl born? 2 May, 1860 Pest, Kingdom of Hungary, Austrian Empire
8. Who is the parent of Barack Obama? Barack Obama Sr., Ann Dunham

כל אחד מ-3 המבנים של האפשריים לשאלה עשוי להכיל משתנים משני סוגים:

1. Entity – זו ישות שיש לה דף בWikipedia האנגלית.
   <u>דוגמה</u>: לישות 2011 NBA Finals ישנו ה-URL,
   https://en.wikipedia.org/wiki/2011_NBA_Finals
2. Relation – כל יחס הוא שדה ב-Wikipedia infobox של הישות שלו.
   <u>דוגמה</u>:
   Who is the MVP of the 2011 NBA Finals?
   היחס MVP הוא שדה ב-infobox של עמוד הויקיפדיה של 2011 NBA Finals.



2011 NBA Finals

From Wikipedia, the free encyclopedia
(Redirected from NBA Finals 2011)

The **2011 NBA Finals** was the championship series of the National Basketball Association (NBA)'s 2010–11 season, and the conclusion of the season's playoffs. The Western Conference champion Dallas Mavericks defeated the Eastern Conference champion Miami Heat 4 games to 2 to win their first NBA championship. Dallas became the last NBA team from Texas to win its first title, after the Houston Rockets won back-to-back titles in 1994 and 1995, and the San Antonio Spurs won four NBA championships in 1999, 2003, 2005 and 2007, and a fifth one subsequently in 2014; all three Texas NBA teams have now won at least one NBA championship. It was also the first time in four years that the Los Angeles Lakers did not make the Finals, having been swept in the Western Conference semifinals by the eventual champion Dallas Mavericks.

The series was held from May 31 to June 12, 2011—the first to start before June 1 since the 1986 NBA Finals. Under the 2–3–2 rotation, the Miami Heat had home-court advantage; the Heat hosted Games 1, 2, and 6, and was set to host a deciding Game 7, had one been necessary. German player Dirk Nowitzki was named the Finals MVP. Nowitzki was the second European to win the award after Tony Parker (2007); he is the first German to win the award.[2]

Going into the series, the Heat were heavy favorites[3][4] with their newly acquired superstars LeBron James and Chris Bosh along with returning superstar Dwyane Wade. The series was a rematch of the 2006 NBA Finals, which was won by the Heat in six games after Dallas blew a 2–0 series lead.[2]

The Dallas Mavericks became the first team in NBA history since the institution of the 2–3–2 format to enter Game 3 tied at one, lose Game 3 and still win the Finals. The previous 11 times this occurred, the Game 3 winner went on to win the series.[5]

The Dallas Mavericks also became just the 7th team, and the first since 1988, to come back and win the Finals after being down in the series two or more separate times (one game to none, and later two games to one). The previous six times this happened, the Finals ended in seven games; Dallas became the first team in NBA history to do it in six games.

ABC averaged a 10.1 rating, 11.7 million households and nearly 17.3 million viewers with the 2011 Finals, according to Nielsen.

Contents [hide]
1 Background
1.1 Road to the Finals
1.2 Regular-season series

2011 NBA Finals

The Finals

Team	Coach	Wins
Dallas Mavericks	Rick Carlisle	4
Miami Heat	Erik Spoelstra	2

Dates May 31–June 12
MVP Dirk Nowitzki[1] (Dallas Mavericks)
Television ABC & ESPN 3D (U.S.) TSN (Canada)
Announcers Mike Breen, Mark Jackson and Jeff Van Gundy (ABC) Mark Jones, Bruce Bowen (Gms 1-2, 5-6), and Tim Legler (Gms 3-4) (ESPN 3D)
Radio network ESPN
Announcers Mike Tirico, Hubie Brown, and Jack Ramsay

Referees
Game 1: Steve Javie, Mike Callahan, Bill Kennedy
Game 2: Joe Crawford, Ed Malloy, Ken Mauer

<u>שלבי המערכת:</u>

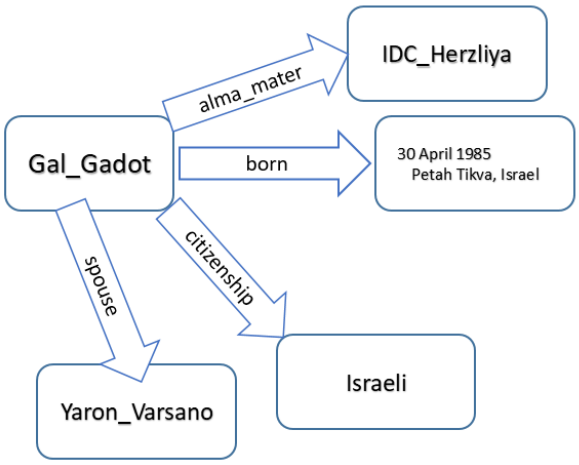
המערכת מורכבת מכמה שלבים, שבסיומם תוחזר התשובה הנכונה, בנוסף לאונטולוגיה ושאילת SPARQL שהיא תרגום של השאלה המקורית. נתאר את פעולת המערכת.

1. קבלת השאלה בשפה טבעית וחילוץ הישות והיחס הרלוונטים מתוך השאלה.
2. עליכם למצוא דרך לעבור משם הישות בשאלה לעמוד הויקיפדיה המתאים לה, למשל לעבור מ- Gal Gadot אל https://en.wikipedia.org/wiki/Gal_Gadot.
3. השתמשו ב-XPATH ובידע על HTML כדי לחלץ את היחס הרלוונטי מתוך ה-infobox על הישות בויקיפדיה.
4. החזירו את תוצאת היחס הרלוונטי כתשובה לשאלה.
5. חלצו את **כל** היחסים מתוך ה-infobox של הישות ובנו אונטולוגיה בעזרת rdflib.
   a. על entities, relations באונטולוגיה להיות URIRefs.
6. תרגמו את השאלה בשפה טבעית לשאילתת sparql שאם תשוערך על האונטולוגיה תחזיר את התשובה לשאלה (rdflib).
7. שמרו את האונוטולוגיה בקובץ ontology.nt בעזרת rdflib.
8. שמרו את שאילתת ה-sparql בקובץ query.sparql.

דוגמה:

Who is the spouse of Gal Gadot?

- היחס spouse הוא שדה ב-infobox של עמוד הויקיפדיה של Gal Gadot.
- נחלץ את היחס ונחזיר את "Yaron Varsano" כתשובה.
- נבנה אונטולוגיה מכל היחסים ב-infobox כולל שם הישות עצמה (דוגמה חלקית):



- נתרגם את השאלה בשפה טבעית לשאילתת ה-sparql:
  <Gal_Gadot> <spouse> ?p .
- נשמור את התוצאות ונחזיר את התשובה לשאלה כפלט.

הרצת הקוד:

- על הקוד שלכם להיות כתוב בפייתון (גרסה 2 או 3, אנא ציינו איזו).
- התוכנית תקרא wiki_qa.py, ותרוץ משורת הפקודה באופן הבא:
  python wiki_qa.py <natural language question string>
- על התוכנית להחזיר כפלט מחרוזת שתהא התשובה לשאלה.
- בנוסף על התוכנית לייצר שני קבצים:
  - קובץ אונטולוגיה ontology.nt (באמצעות rdflib), שמכיל את האונטולוגיה של דף הויקיפדיה של הישות.
  - קובץ query.sparql שמכיל את תרגום השאלה בשפה טבעית לשאילתת SPARQL, שהרצתה על ontology.nt מחזירה את התשובה לשאלה. (ודאו זאת באמצעות rdflib).

הערות:

- הקוד ייבדק בבדיקה אוטומטית על מספר שאלות בשפה טבעית (כמו בדוגמאות).
- על הקוד לרוץ ללא כל שגיאות על שרת nova ולסיים לרוץ תוך פחות מ-4 דקות.
- על המערכת לדעת להתמודד עם שאלות על ישויות ויקיפדיה שונות (שחקנים, נשיים, מדינות, סרטים, טקסים, גמר ליגת האלופות ועוד).
- ניתן להניח שכל השאלות בשפה טבעית יהיו תמיד מאחד מ-3 המבנים שצוינו.
- ניתן להניח שהישויות בשאלה בשפה טבעית תמיד יהיו ישויות שקיים עבורן דף ויקיפדיה. וכי ה-relation בשאלה תמיד יופיע ב-infobox של אותו דף ויקיפדיה.
  - עליכם לדאוג להמרה של הישויות והיחסים בשפה טבעית לפורמט ויקיפדיה.
- תרגיל הבית המסכם מהווה 4 נק' מהציון הסופי בקורס.

- אנא השתדלו להגיש את התרגיל המסכם בזוגות ולא ביחידים.