

ניהול נתונים באינטרנט – תרגיל מסכם – Information Extraction

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 פרסום: 25.03.2019 תאריך הגשה: 16.06.2019

רקע:

זהו התרגיל המסכם של הקורס בנושא Information Extraction. בתרגיל זה תבנו מערכת למענה על שאלות בשפה טבעית (Question Answering) בנושאי גיאוגרפיה. עליכם להעזר בידע על HTML, Xpath, IE, SPARQL, Ontology לכתובת המערכת. התרגיל להגשה עד היום הראשון של חופשת סמסטר ב' (16.06.2019) וכמו תרגילים קודמים, ניתן להגישו בזוגות או ביחידים. תרגיל זה מהווה 4 נק' מהציון הסופי בקורס.

משימות התרגיל המסכם:

1. כתבו תוכנית היוצרת אונטולוגיה המכילה מידע על מדינות העולם.
2. כתבו שאלות SPARQL לארבע השאלות הבאות. הריצו את השאלות על האונטולוגיה מסעיף 1 והחזירו את תוצאותיהן:
 - a. כמה ראשי ממשלה יש בעולם?
 - b. כמה מדינות יש בעולם?
 - c. כמה מדינות בעולם הן רפובליקה (republic)?
 - d. כמה מדינות בעולם הן מונרכיות (monarchy)?
3. כתבו תוכנית שמאפשרת לתשאל את האונטולוגיה בשפה טבעית (אנגלית) ולקבל חזרה את התשובה (דוגמאות לשאלות בהמשך).

מדינות העולם:

עליכם להתייחס אך ורק למדינות שבעמוד:

[https://en.wikipedia.org/wiki/List_of_countries_by_population_\(United_Nations\)](https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations))

סוגי שאלות:

כל השאלות יהיו בשפה האנגלית ויכללו תמיד את אחד מ-9 המבנים הבאים,

- (i) Who is the **president** of <country>?
- (ii) Who is the **prime minister** of <country>?
- (iii) What is the **population** of <country>?
- (iv) What is the **area** of <country>?
- (v) What is the **government** of <country>?
- (vi) What is the **capital** of <country>?
- (vii) When was the **president** of <country> born?
- (viii) When was the **prime minister** of <country> born?
- (ix) Who is <entity>?

1. Who is the **president** of **Italy**? Sergio Mattarella
2. Who is the **prime minister** of **United Kingdom**? Theresa May
3. What is the **population** of **Democratic Republic of the Congo**?
78,736,153
4. What is the **area** of **Fiji**? 18,274 km²
5. What is the **government** of **Eswatini (Swaziland)**? Unitary
parliamentary absolute diarchy
6. What is the **capital** of **Canada**? Ottawa
7. When was the **president** of **South Korea** born? 1953-01-24
8. When was the **prime minister** of **New Zealand** born? 1980-07-26
9. Who is **Donald Trump**? President of United States
Who is **Alexis Tsipras**? Prime minister of Greece

כל אחד מ-9 המבנים האפשריים לשאלה יכול משתנים משני סוגים:

1. **Entity** – ישות שיש לה דף ב-Wikipedia.

דוגמה:

לישות **Jacinda Ardern** ישנו ה-URL,

https://en.wikipedia.org/wiki/Jacinda_Ardern

2. **Relation** – כל יחס הוא שדה ב-Wikipedia infobox של הישות שלו.

דוגמה:

Who is the **prime minister** of **New Zealand**?

היחס **prime minister** הוא שדה ב-Wikipedia infobox של עמוד הויקיפדיה **New Zealand**.

New Zealand

From Wikipedia, the free encyclopedia

Coordinates: 42°S 174°E﻿•﻿42°S 174°E

This article is about the country. For other uses, see New Zealand (disambiguation).
 "NZ" redirects here. For other uses, see NZ (disambiguation).

New Zealand (Māori: *Aotearoa* (aɔˈtɛaɾɔa)) is a sovereign island country in the southwestern Pacific Ocean. The country geographically comprises two main landmasses—the North Island (*Te Ika-a-Māui*), and the South Island (*Te Waipounamu*)—and around 600 smaller islands. New Zealand is situated some 2,000 kilometres (1,200 mi) east of Australia across the Tasman Sea and roughly 1,000 kilometres (600 mi) south of the Pacific island areas of New Caledonia, Fiji, and Tonga. Because of its remoteness, it was one of the last lands to be settled by humans. During its long period of isolation, New Zealand developed a distinct biodiversity of animal, fungal, and plant life. The country's varied topography and its sharp mountain peaks, such as the Southern Alps, owe much to the tectonic uplift of land and volcanic eruptions. New Zealand's capital city is Wellington, while its most populous city is Auckland.

Sometime between 1250 and 1300, Polynesians settled in the islands that later were named New Zealand and developed a distinctive Māori culture. In 1642, Dutch explorer Abel Tasman became the first European to sight New Zealand. In 1840, representatives of the United Kingdom and Māori chiefs signed the Treaty of Waitangi, which declared British sovereignty over the islands. In 1841, New Zealand became a colony within the British Empire and in 1907 it became a dominion; it gained full statutory independence in 1947 and the British monarch remained the head of state. Today, the majority of New Zealand's population of 4.9 million is of European descent; the indigenous Māori are the largest minority, followed by Asians and Pacific Islanders. Reflecting this, New Zealand's culture is mainly derived from Māori and early British settlers, with recent broadening arising from increased immigration. The official languages are English, Māori, and NZ Sign Language, with English being very dominant.

A developed country, New Zealand ranks highly in international comparisons of national performance, such as quality of life, health, education, protection of civil liberties, and economic freedom. New Zealand underwent major economic changes during the 1980s, which transformed it from a protectionist to a liberalised free-trade economy. The service sector dominates the national economy, followed by the industrial sector, and agriculture; international tourism is a significant source of revenue. Nationally, legislative authority is vested in an elected, unicameral Parliament, while executive political power is exercised by the Cabinet, led by the Prime Minister, who is currently Jacinda Ardern. Queen Elizabeth II is the country's head of state and is represented by a governor-general, currently Dame Patsy Reddy. In addition, New Zealand is organised into 11 regional councils and 67 territorial authorities for local government purposes. The Realm of New Zealand also includes Tokelau (a dependent territory), the Cook Islands and Niue (self-governing states in free association with New Zealand), and the Ross Dependency, which is New Zealand's territorial claim in Antarctica. New Zealand is a member of the United Nations, Commonwealth of Nations, ANZUS, Organisation for Economic Co-operation and Development, ASEAN Plus Six, Asia-Pacific Economic Cooperation, and the Pacific Islands Forum.

Contents

- 1 Etymology
- 2 History
- 3 Government and politics
 - 3.1 Foreign relations and military
 - 3.2 Local government and external territories
- 4 Environment
 - 4.1 Geography
 - 4.2 Climate
 - 4.3 Biodiversity
- 5 Economy
 - 5.1 Trade
 - 5.2 Infrastructure
- 6 Demography
 - 6.1 Ethnicity and immigration
 - 6.2 Language
 - 6.3 Religion
 - 6.4 Education
- 7 Culture

<div>New Zealand</div> <div>Aotearoa (Māori)</div>
<div> <div> <div></div> <div>Flag</div> </div> <div> <div></div> <div>Coat of arms</div> </div> </div>
Anthem: <div>"God Defend New Zealand"<div><div></div></div></div> <div>"God Save the Queen"^(H) </div>
 <div>A map of the world showing the location of New Zealand, including outlying islands, its territorial claim in the Antarctic, and Tokelau.</div>
Capital
Largest city
Official languages
Ethnic groups <div> (2013)</div>
Demonym(s)
Government
<div>• Monarch </div>
<div>• Governor-General </div>
<div>• Prime Minister </div>
Legislature

שילבי המערכת:

המערכת מורכבת משני שלבים עיקריים:

1. **בניית האונטולוגיה.** אתם נדרשים לבצע IE על דפי הויקיפדיה כל מדינות העולם (שנמצאות בטבלה שבעמוד [https://en.wikipedia.org/wiki/List_of_countries_by_population_\(United_Nations\)](https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations))). השתמשו בידע שלכם ב-XPath ו-SPARQL על מנת לעבור בצורה אוטומטית על דף הויקיפדיה של כל מדינה בטבלה ולחלץ מה-infobox של כל מדינה את המידע הרלוונטי כדי לענות על 9 סוגי השאלות שמופיעות למעלה. שימו לב שעליכם לחלץ מידע לא רק מה-infobox של כל מדינה אלא גם מה-infobox של מנהיגי מדינות מוֹג (King, Queen, etc. (אין צורך לחלץ מידע על .

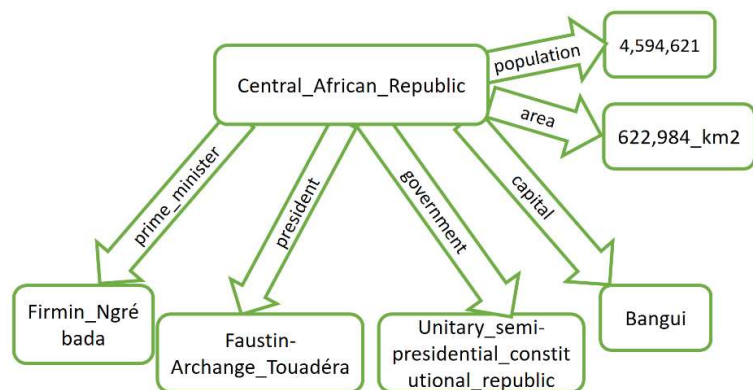
2. **מענה על שאלות בשפה טבעית.** לאחר בניית האונטולוגיה ושמירתה בקובץ ontology.nt על התוכנית לדעת להתמודד עם שאלות באנגלית על גבי האונטולוגיה. בהנתן שאלה באנגלית (מאחד מ-9 המבנים למעלה) על התוכנית לתרגם את השאלה לשאלתת SPARQL שתורץ מעל האונטולוגיה שבניתם ותחזיר את התשובה.

דוגמה:

- What is the **capital** of **Central African Republic**?
- היחס capital הוא שדה ב-infobox של עמוד הויקיפדיה של Central African Republic.
 - בזמן בניית האונטולוגיה חילצנו את השלשה, `<central_african_republic, catpital, bangui>`
 - כשקיבלנו את השאלה בשפה טבעית זיהינו שמדובר בשאלה על capital של המדינה Central African Republic.
 - הרצנו על האונטולוגיה את שאלתת ה- SPARQL המתאימה שתחזיר את התשובה, Bangui.



Government	Unitary semi-presidential constitutional republic
• President	Faustin-Archange Touadéra
• Prime Minister	Firmin Ngrébada
Legislature	National Assembly
Independence	
• from France	13 August 1960
• Central African Empire established	4 December 1976
• Republic restored	21 September 1979
Area	
• Total	622,984 km ² (240,535 sq mi) (44th)
• Water (%)	12
Population	
• 2016 estimate	4,594,621 ^[1] (119th)
• 2003 census	4,987,640 ^[2]
• Density	7.1/km ² (18.4/sq mi) (221st)



הרצת הקוד:

- על הקוד שלכם להיות כתוב בפייתון (גרסה 2 או 3, ציינו איזו).
- התוכנית תקרא `geo_qa.py`, ותרוץ משורת הפקודה באופן הבא:

```
python geo_qa.py create onotology.nt
```

```
python geo_qa.py question <natural language question string>
```
- על התוכנית להחזיר כפלט מחרוזת שתהא התשובה לשאלה.
 - בנוסף על התוכנית לייצר את קובץ האונטולוגיה `ontology.nt` (באמצעות `rdflib`), שמכיל את האונטולוגיה שבניתם על מדינות העולם ומנהיגיהם.

הוראות הגשה:

- על אחד המגשים להעלות ל- Moodle קובץ זיפ בשם `HWqa_123.zip`, כאשר במקום 123 יופיעו ת.ז של המגשים. על הקובץ לכלול:
 - קובץ האונטולוגיה שיצרתם בשם `ontology.nt`.
 - תוכנית הפייתון שיוצרת את האונטולוגיה ועונה על שאלות, `geo_qa.py`.
 - קובץ PDF בשם `answers` ובו התרגום ל- SPARQL של 4 השאלות בעברית מסעיף 2 של המשימות וכן תוצאות ההרצה שלהן על האונטולוגיה.

הערות:

- הקוד ייבדק בבדיקה אוטומטית על מספר שאלות בשפה טבעית (כמו בדוגמאות).
- על הקוד לרוץ ללא כל שגיאות ולסיים לרוץ תוך פחות מ-4 דקות.
- ניתן להניח שכל השאלות בשפה טבעית יהיו תמיד מאחד מ-9 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 מהשאלה בשפה טבעית לשמות בהם הם מופיעים באונטולוגיה שבניתם.
- תרגיל הבית המסכם מהווה 4 נק' מהציון הסופי בקורס.
- אנא השתדלו להגיש את התרגיל המסכם בזוגות ולא ביחידים.

בהצלחה! ☺