

פרוייקט #1 – Information Extraction

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

תיאור המערכת

על המערכת לדעת לענות על שאלות בשפה האנגלית, כאשר כל השאלות יהיו תמיד מאחת התבניות הבאות:

1. Who is the **president of** <country>?
2. Who is the **prime minister of** <country>?
3. What is the **population of** <country>?
4. What is the **area of** <country>?
5. What is the form of **government in** <country>?
6. What is the **capital of** <country>?
7. When was the **president of** <country> born?
8. Where was the **president of** <country> born?
9. When was the **prime minister of** <country> born?
10. Where was the **prime minister of** <country> born?
11. Who is <entity>?
12. How many <government_form1> are also <government_form2>?
13. List all **countries** whose **capital** name contains the string <str>
14. How many **presidents** were **born in** <country>?


בנוסף, עליכם להגדיר תבנית נוספת לבחירתכם שמסתמכת על המידע הקיים במערכת, ולאפשר למשתמש לשאול שאלות בתבנית זאת. על התבנית להכיל לפחות משתנה אחד.

השאלות יכולות להכיל התייחסויות לשלושה סוגי משתנים:

- **Substring**: תת מחרוזת כלשהי, מופיעה בשאלה 13 בלבד.
- **Entity**: ישות שיש לה ערך בויקיפדיה. לדוגמא לישות Emmanuel Macron יש את הדף https://en.wikipedia.org/wiki/Emmanuel_Macron. שם הישות יהיה זהה לשמה ב-URL של דף הויקיפדיה שלה עם רווח במקום קו תחתון.
- **Relation**: כל יחס הוא שדה ב-Wikipedia Infobox של הישות.

למשל התשובה לשאלה: What is the **capital of** Vietnam? תהיה **Hanoi**, כאשר המידע על היחס מגיע מהשדה המסומן ב-Infobox.

Not logged in | [Talk](#) | [Contributions](#) | [Create account](#) | [Log in](#)


WIKIPEDIA
 The Free Encyclopedia

Article Talk

Vietnam

From Wikipedia, the free encyclopedia

Coordinates: 16°N 108°E

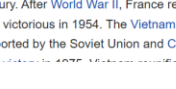
For other uses, see [Vietnam \(disambiguation\)](#).


Vietnam (Vietnamese: *Việt Nam*, [vjet̚nəm] (listen)), officially the **Socialist Republic of Vietnam**,^[n][]] is a country in Southeast Asia^[n][]]. Located at the eastern edge of mainland Southeast Asia, it covers 311,699 square kilometres. With a population of over 96 million, it is the world's fifteenth-most populous country. Vietnam borders China to the north, Laos and Cambodia to the west, and shares maritime borders with Thailand through the Gulf of Thailand, and the Philippines, Indonesia, and Malaysia through the South China Sea. Its capital is Hanoi and its largest city is Ho Chi Minh City.^[n][]]

Vietnam was inhabited as early as the Paleolithic age. The first known Vietnamese nation during the first millennium BC centred on the Red River Delta, located in modern-day northern Vietnam. The Han dynasty annexed and put the Vietnamese under Chinese rule from 111 BC, until the first independent dynasty emerged in 939. Successive monarchical dynasties absorbed Chinese influences through Confucianism and Buddhism, and expanded southward to the Mekong Delta. The Nguyễn—the last imperial dynasty—fell to French colonisation in 1887. Following the August Revolution, the nationalist Viet Minh under the leadership of communist revolutionary Ho Chi Minh proclaimed independence from France in 1945.

Vietnam went through prolonged warfare through the 20th century. After World War II, France returned to reclaim colonial power in the First Indochina War, from which Vietnam emerged victorious in 1954. The Vietnam War began shortly after, during which the nation was divided into communist North supported by the Soviet Union and China, and anti-communist South supported by the United States. Near North Vietnam's collapse in 1975, Vietnam unified as a socialist state.

Socialist Republic of Vietnam
Cộng hòa Xã hội chủ nghĩa Việt Nam (Vietnamese)


 Flag


 Emblem

Motto: "Độc lập – Tự do – Hạnh phúc"
"Independence – Liberty – Happiness"

Anthem: Tiến Quân Ca
"Army March"

0:00

MENU

Show globe

Show map of ASEAN

Show all

Location of Vietnam (green)
in ASEAN (dark grey) (Expand)

Capital	Hanoi 21°2′N 105°51′E
Largest city	Ho Chi Minh City 10°48′N 106°39′E

איסוף המידע ובניית האונטולוגיה

עליכם לאסוף מידע על המדינות המופיעות בעמוד הזה:

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

שימו לב שעליכם לחלץ מידע לא רק מה-infobox בעמודי המדינות, אלא גם מהעמודים של ראשיות והמשלה והנשיאות/ים. השתמשו בידע שלכם על Crawlrs ו-Xpath כדי לעבור בצורה אוטומטית על הדפים הרלוונטים ולחלץ משם את המידע הדרוש.

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

מענה על שאלות בשפה טבעית

בהנתן שאלה באנגלית, על התוכנית לתרגם את השאלה לשאילתת SPARQL שתרוץ מעל האונטולוגיה שבניתם ותחזיר את התשובה. התשובה לא צריכה להיות "תשובה מלאה", אלא רק להכיל את הערך הנדרש.

- למשל עבור השאלה: **Where** was **Justin Trudeau** **born**?
התשובה תהיה: **Canada**
אין צורך לציין מידע מעבר לשם המדינה. אם שם המדינה לא מצויין, אין צורך לאסוף מידע אחר.
- עבור השאלה: Who is **Pedro Castillo**?
נחזיר תשובה המורכבת מתפקידו (Prime Minister או President) ושם המדינה בה הוא מחזיק בתפקידו:
President of Peru
שימו לב ששאלות מתבנית זאת ישאלו תמיד על ראשי/ות ממשלה או נשיאות/ים.

- בשאלות מסוימות יכולה להיות יותר מתשובה אחת, כמו למשל: What is the form of **government in Argentina**? במקרה כזה נציג את כל התשובות מופרדות פסיקים (רווח אחרי כל פסיק), וממיינות בסדר לקסיקוגרפי: **Federal republic, Presidential system, Republic**
- עבור שאלות מהסוג: List all **countries** whose **capital** name contains the string **free**, תוכלו להניח שהמחרוזת לחיפוש תהיה מורכבת מאותיות קטנות בלבד (במקרה הזה, **free**). יש לבצע חיפוש case insensitive. תוכלו לבצע זאת ב- SPARQL באמצעות הפונקציה lcase. אפשר לראות דוגמא [כאן](#).

הרצת הקוד

- על הקוד להיות כתוב בפייתון 3 ולרוץ באופן תקין בנובה.
- התכנית תיקרא geo_qa.py ותרוץ משורת הפקודה באופן הבא:
 - python3 geo_qa.py create
 במצב create התכנית תייצר את הקובץ ontology.nt שיכיל את האונטולוגיה שבניתם ותסיים לרוץ.
 - python3 geo_qa.py question "<question>"
 במצב question התכנית תקבל שאלה בשפה טבעית, תדפיס למסך את התשובה לשאלה ותסיים לרוץ. השאלה ניתנת במחרוזת 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 התז של המגשים
- תיאור של הקוד שבונה את האונטולוגיה – flow וחלקים חשובים.
- תיאור של השאלה שהוספתם למערכת ודוגמאות לתשובות אפשריות.
- תיאור של שלושה מקרי קצה שהתמודדתם איתם באיסוף המידע. כמו שראינו בתרגול, יתכנו מקרים ספציפיים שידרשו טיפול מיוחד – מקרים בהם הייתם צריכים לכתוב שאלות נוספות כדי להתמודד עם חלק קטן ב-data שמופיע בפורמט שונה מהשאר. תארו 3 מקרים כאלו שנתקלתם בהם במהלך העבודה - הסבירו אילו חיפוש 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. geo_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 - הספריות הנדרשות להרצת הפרוייקט, ראו פירוט ב"עבודה עם ספריות חיצוניות".

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

בדיקת הפרוייקט

ציון הפרוייקט מחושב באופן הבא:

- 74% - בדיקות גליות
- 16% - בדיקות נסתרות
- 10% - תיאור הקוד, מקרי הקצה, והשאלה שהוספתם למערכת

יבדקו 45 שאלות בשפה טבעית. 37 מהשאלות זמינות לכם במודל ותוכלו לבדוק את הקוד שלכם עליהן. 8 השאלות הנוספות נסתרות.

הפרוייקט יבדק באופן אוטומטי בנובה, לכן אנחנו ממליצים להקפיד שהקוד רץ ללא שגיאות ועומד במבנה התשובות הנדרש. 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 הגליות השתנתה, בבקשה תכתבו לנו בפורום כדי שנוכל לעדכן. בכל מקרה, הקוד שלכם יבדק מול גרסת האונטולוגיה שהגשתם, כך שאם סיימתם לפני הזמן תוכלו להגיש את הפרוייקט ללא חשש מעדכונים בויקיפדיה.

דוגמאות הרצה

מצורפות מספר דוגמאות הרצה:

```
nova:~> python3 geo_qa.py question "Who is the president of Portugal?"
Marcelo Rebelo de Sousa
```

```
nova:~> python3 geo_qa.py question "What is the form of government in Sweden?"
Constitutional monarchy, Parliamentary system, Unitary state
```

```
nova:~> python3 geo_qa.py question "List all countries whose capital name contains the string hi"
Bhutan, India, Moldova, Sint Maarten, United States
```

עבודה עם ספריות חיצוניות

כדי לאפשר עבודה עם ספריות חיצוניות בנובה, יש לעבוד עם סביבה וירטואלית. אפשר למצוא הוראות להרמת סביבה [כאן](#).

צור סביבה וירטואלית (כאן נשתמש בפיתון 3.7 ופיפ 3.7, ניתן לשנות זאת אם רוצים) (זה צריך להתבצע רק כשאתם בסביבה)

```
virtualenv --prompt=<env-prefix> --python=python3.7 <env-path>
```

דוגמה:

```
virtualenv --prompt=<my-env> --python=python3.7 .env
```

הפעל את הסביבה (זה מתבצע כל פעם שמשתמשים בשרת! בלי הפעלה של הסביבה צריך להתקין לפי המדריך למטה, לא מומלץ)

```
source <env-path>/bin/activate.csh
```

התקנת flask (רק כשמייצרים סביבה חדשה) (רלוונטי לכל ספרייה שתמצאו להתקין)

```
pip3.7 install flask
```

יש לצרף להגשה קובץ בשם requirements.txt, שיכיל את שמות כל הספריות הנדרשות להרצת הפרוייקט, כל ספריה בשורה נפרדת. אפשר לקרוא עוד ולראות דוגמא [כאן](#). אפשר לייצר את הקובץ בעזרת הפקודה freeze:

- היכנסו לסביבת הפיתון בה כתבתם את התרגיל

- הריצו את השורות הבאות:

```
from pip._internal.operations import freeze
print('\n'.join(freeze.freeze()))
```

- תוכן הקובץ יהיה הפלט של ההדפסה

בהצלחה!