

## 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. לדוגמה לישות Emmanuel Macron יש את הדף [https://en.wikipedia.org/wiki/Emmanuel\\_Macron](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.



- 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**

## הרצת הקוד

- על הקוד להיות כתוב בפייתון 3 ולרוץ באופן תקין בנובה.
- התכנית תיקרא geo\_qa.py ותרץ 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 במחרוזת 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 שאילתות נוספות כדי להתמודד עם חלק קטן ב-data שמופיע בפורמט שונה מהשאר. תארו 3 מקרים כאלו שנתקלתם בהם במהלך העבודה - 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. 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 כשאינן סביבה)

```
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!