ניהול נתונים באינטרנט – תרגיל מס'3 – Information Extraction, Ontologies and POS tagging

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: HW1_123.zip).
תאריך הגשה: **14.05.2018**

שאלה 1: (Simple HMMs)
נתון ה- HMM הבא:
מצבים: adj,noun,verb

Pstart(noun)= Pstart(adj)=0.4
Pstart(verb)=0.2
Ptrans(noun,noun)= Ptrans(noun,verb)=0.3, Ptrans(noun,adj)= 0.4
Ptrans(adj,noun)= Ptrans(adj,adj)=0.5
Ptrans(verb,adj)=0.8, Ptrans(verb,noun)=0.2
Pout(noun,"sound")=0.3, Pout(noun,"sounds")=0.5, Pout(noun,"table")=0.2
Pout(verb,"sound")=0.4, Pout(verb,"sounds")=0.6
Pout(adj,"sound")=0.3, Pout(adj,"sounds")=0.5, Pout(adj,"nice")=0.2

עבור כל אחד מהמשפטים הבאים, מצאו את התיוג הסביר ביותר שלו תוך שימוש באלגוריתם Viterbi . פרטו את כל תוצאות הביניים של האלגוריתם (כל הערכים המחושבים בטבלת ה- Dynamic Programming):
א. sound sounds sound
ב. sounds sound nice

שאלה 2: (Data Mining an Ontology)
בשאלה זו עליכם לכתוב שאילתות SPARQL ולהריץ אותן ב:
https://dbpedia.org/sparql
יש להגיש את השאילתה ואת **10** התוצאות הראשונות של הפלט בתוך קובץ הPDF.

א. שאילתה שמחזירה את שמות כל העיתונים שמתפרסמים בשפה הספרדית.
ב. שאילתה שמחזירה שחקני כדורגל שנולדו במדריד ושיחקו בקבוצה שהאצטדיון שלה ממוקם באנגליה. החזירו את שם השחקן ואת שם הקבוצה.
ג. שאילתה שמחזירה נהרות (river) שעוברים בצרפת ובעוד מדינות נוספות.
ד. שחקני קולנוע שנולדו במדינה שמדברים בה גרמנית ושיחקו בסרט שצולם במדינה שמדברים בה אנגלית. החזירו את שם השחקן, את המדינה בה נולד ואת שם הסרט.

שאלה 3: (IE generation of an Ontology)

בשאלה זאת נתמקד ב INFORMATION EXTRACTION מויקיפדיה. כתבו תוכנית PYTHON המתחבר לעמוד של שחקן כדורגל מסויים בויקיפדיה. המטרה היא לבנות אונטולוגיה של קבוצות כדורגל ושחקני כדורגל שמחקים בהם, כולל מקום ותאריך לידת, ותפקידם במגרש. יש לבנות את הקשרים הבאים:

?league	<country>	?country
?team	<league>	?league
?team	<homeCity>	?city
?player	<playsFor>	?team
?player	<birthPlace>	?city
?city	<located_in>	?country
?player	<birthDate>	?date
?player	<position>	?position

יש להתחיל מהעמוד של הליגה האנגלית בעונת 2016-2017
https://en.wikipedia.org/wiki/2016%E2%80%9317_Premier_League
ולעבור על כל 20 הקבוצות, ואז בכל אחד מהקבוצות לעבור על השחקנים שלהם.

(א) על האונטולוגיה שבניתם בסעיף (א) ממשו קוד שמריץ בעזרת RDFLIB את השאילתות הבאות:

a. כל השחקנים שנולדו בספרד ומשחקים בליגה האנגלית ואת הקבוצה שלהם
b. כל השחקנים שנולדו אחרי 1990 ואת הקבוצה שלהם
c. שחקנים שנולדו בעיר בה הם משחקים היום
d. כל משחקי הדרבי האפשריים בליגה (משחק של 2 קבוצות מאותה העיר)

(ב) הריצו את הקוד שכתבתם בסעיף (ב) על האונטולוגיה שבניתם בסעיף (א) והגישו את התוצאה.

שאלה 4: (Fagin's algorithm & Rank Aggregation)

(א) ממשו בשפת PYTHON את האלגוריתם FAGIN שראינו בשיעור, כאשר הפונקציה אמור לקבל **<u>מערך של שמות קבצים</u>** (יכול להיות יותר מ2) ומחשב את הRANK המשותף של הITEMS, כאשר פונקציית האגרגציה היא <u>ממוצע</u>. ניתן להשתמש בפונקציות שכתבנו בכיתה. הגישו את התוצאה בקובץ question4a.py

(ב) הריצו את התוכנית מסעיף (א) על הקבצים הבאים:
http://slavanov.com/teaching/wdm1516b/files_fagin.zip
הקבצים הינם קבצי CSV - שם הITEM בעמודה מס'1, הRANK בעמודה מס'2 ומספר הקובץ בעמודה מס'3. הגישו את התוצאה בקובץ question4b.txt