HTML וחבילת BeautifulSoup4

HTML היא השפה העיקרית בה משתמשים כיום לתצוגה ולעיצוב של דפי אינטרנט. השפה היא מבוססת תגיות, כאשר לכל תגית תפקיד עיצובי אחר בעמוד. בקוד HTML, החלקים העיקריים של כל תגית לרוב הם שם, תכונות, ותוכן, והמבנה התחבירי הכללי נראה כך:

```
<tag prop1=val1, prop2=val2, … prop=valN> content </tag>
```

כאשר:

- tag – שם התגית
- prop1,…,propN – התכונות של התגית
- val1,…,valN – ערכי התכונות בהתאמה
- content – התוכן של התגית

שמות התגיות, וכן התכונות של כל תגית, הם לרוב מוגדרים מראש. תגיות יכולות אף להיות מקוננות, כלומר חלק התוכן של כל תגית יכול להכיל תגיות פנימיות נוספות. לצורך התרגיל, אנו נתמקד ב-2 תגיות שמעוניינות אותנו בלבד:

- התגית text המייצגת פסקה של טקסט
- התגית link המשמשת לצורך הגדרת קישורים (לינקים), כאשר href הוא התכונה שהערך שלה "url" מכיל את כתובת האתר אליו הלינק מקשר, ו-link הוא הטקסט המוצג למשתמש שישמש כלינק עצמו.

קוד HTML שמור לרוב בקבצי HTML, שלהם הסיומת html.

בפייתון קיימת ספריה בשם BeutifulSoup4 שמקלה מאוד על העבודה עם קבצי HTML. היא מאפשרת לגשת לתגיות השונות של קוד HTML ראשית יש לייבא אותה בראש הקובץ כך:

```
Import bs4
```

לאחר מכן, כדי להשתמש בספריה כדי לגשת לתגיות השונות של קוד HTML כלשהו, יש ליצור אוביקט BeautifulSoup באופן הבא:

```
soup = bs4.BeautifulSoup(html)
```

כאשר html הוא משתנה מטיפוס מחרוזת, המכיל את כל קוד ה-HTML של קובץ כלשהו. לאחר מכן תוכלו למצוא את כל התגיות מסוג מסוים בעזרת

הפונקציה find_all. למשל, כדי למצוא את כל תגיות הפסקאות , יש לכתוב, לאחר שורת הקוד הקודמת, את השורה הבאה:

```
paragraphs = soup.find_all("p")
```

על המשתנה paragraphs, שמכיל את כל הפסקאות שבקוד השמור במשתנה html, אפשר לעבור בעזרת לולאת for, על כל פסקה, ולמצוא בה, למשל, את כל תגיות <a> המוכלות בה, כך:

```
for p in paragraphs:
   links = p.find_all("a")
```

בדוגמה זו, עבור כל פסקה (השמורה במשתנה הלולאה p), נקבל את רשימת כל התגיות a המוכלות בפסקה לתוך המשתנה links, שגם עליו אפשר לעבור בלולאה פנימית עבור כל תגית <a> בנפרד.

פונקציה נוספת, שיכולה להיות שימושית עבורכם, היא הפונקציה get המאפשרת לקבל את הערך של תכונה מסוימת של תגית. אם למשל l הוא משתנה המייצגת תגית <a> כלשהי (למשל אחד האיברים של המשתנה links מהדוגמה הקודמת), אפשר לגשת לערך התכונה href שלו כך:

```
target = l.get("href")
```

תוכלו לקרוא עוד על הספריה BeautifulSoup4 בלינק הבא:

https://www.crummy.com/software/BeautifulSoup/bs4/doc/