תרגיל בית 1 – bash scripting

מועד ההגשה:	יום ב', 12/04/20, בשעה 23:55	
האחראי על התרגיל:	איגור דה פאולה	De-paula@campus.technion.ac.il

בתרגיל זה ננסה לעזור למשה חוגג הבעלים של קבוצת הכדורגל ביתר ירושלים במציאת שחקני רכש מתאימים לקבוצה, אשר נמצאת בצרות בליגת העל.

לצורך כך נכתוב סקריפט bash שיעשה שימוש בכלים שלמדנו בתרגולים וסדנאות בשביל לסרוק רשימה של דפי שחקנים ברשת. עבור כל דף שחקן נחפש את המילה "midfielder" (יכול להיות עם אות גדולה או קטנה) אשר מעידה על התפקיד של השחקן אותו אנו מחפשים. נסרוק את הרשימת שחקנים מאתר https://www.premierleague.com ונייצר קובץ טקסט הכולל:

1. מספר השחקנים שנבדקו בסך הכל
2. קישור לכל דף של שחקן אשר נמצא משחק בתפקיד אותו אנחנו מחפשים + מספר הפעמים שהתפקיד של השחקן הוזכר בדף שלו:

Total Players: 25

```
https://www.premierleague.com/players/9167/Dele-Alli/overview, Midfielder, 5
https://www.premierleague.com/players/15994/Allan/overview, Midfielder, 6
```

. . .

. . .

את התוצאה יש לשמור בקובץ results.csv. מצורף קובץ לדוגמה הבדק סך הכל 25 שחקנים ומצא 2 שחקנים מתאימים.

שימו לב שיש לרשום בקובץ הסופי Midfielder עם M גדולה, אך בדפי השחקן עצמו Midfielder יכולה להופיע עם אות גדולה בהתחלה או קטנה.

שימו לב, בקובץ results.csv צריך להיות רק שחקנים שהם Midfielder (כלומר שמצאנו לפחות פעם אחת את המילה Midfielder בדף שחקן שלהם).

יש להגיש את הסקריפט בשם scrape_players.sh יחד עם קובץ התשובות (ראו שאלות למטה) ב-git ולהגיש את הלינק במערכת ה-moodle.

הכוונה, דגשים ורמזים:

עבדו בשלבים. פצלו כל שלב ובדקו את תקינותו ורק אח"כ חברו בין כל הדברים.

ייתכן ותרצו לעשות את חלקי המשימה באופן "ידני" על מנת להבין טוב יותר מה מנסים לעשות.

<u>שלב ראשון</u>

הדבר הראשון שתבצעו יהיה "לגלוש" באמצעות פקודה wget לאתר

```
https://www.premierleague.com/players
```

```
wget https://www.premierleague.com/players
```

תוצר הפקודה יהיה קובץ ששמו players, זהו למעשה קובץ html שהדפדפן שלכם מציג את תוכנו בצורה גרפית.

(נסו לגלוש לאתר באמצעות הדפדפן, לחצו על כפתור עכבר-ימני, ובחרו view-source). תיכנסו לאחד השחקנים באתר הנתון ותראו איך נראה קישור לדף של השחקן עצמו.

<u>שלב שני</u>

כעת נעבור על הקובץ שהתקבל מהשלב הראשון (זהו קובץ עם הקישורים לשחקנים שיש באתר עם קישורים לכל דפי השחקנים).

אנו נחפש בו קישורים מהצורה: /players/XXXX/YYYY-YYYY/overview כאשר X יכול להיות כל מספר [0-9], כמו כן לא בהכרח צריך להופיע 4 פעמים כמו בדוגמא הנ"ל, יכול להיות יותר או פחות. Y זהו השם של השחקן כאשר גם כאן אין התחייבות על האם השם המלא שלו מורכב ממילה אחת או יותר. אם השם מורכב מיותר ממילה אחת יש "-" שמפריד בין השמות.

כלומר שלושת הדוגמאות הללו מתאימות:

/players/15994/Allan/overview

/players/4093/Marcos-Alonso/overview

/players/14732/Trent-Alexander-Arnold/overview

הערה: בשמות יהיו רק אותיות גדולות או קטנות. יש להתעלם מכל שם שמכיל תו שהוא לא [a-z A-Z]. לדוגמה:

players/54312/Miguel-Almirón/overview/

מכיוון שהתו ó אינו חלק מ[a-z A-Z] אין לקחת את השחקן הזה בחשבון.

<u>שלב שלישי</u>

כעת ניתן לספור כמה קישורים *יחודיים* יש לכם.

<u>שלב רביעי</u>

כעת נרצה "לגלוש" (בפקודת wget) לכל קישור כזה לדף שחקן לבצע מעבר על התוכן ולראות כמה פעמים (אם בכלל) מופיעה המילה "midfielder". אם המילה כן מופיעה אנו נסיק שהוא אכן בתפקיד אותו אנחנו מחפשים.

נשים לב שכדי לגלוש צריך להוסיף את הקידומת https://www.premierleague.com/ לכל התאמה שמצאנו, אחרת wget לא יעבוד שכן הקישורים שמצאנו מהקובץ HTML כמו שהם לא תקינים לגלישה באינטרנט. ניתן לעשות זאת על ידי הפקודה:

```
sed -i 's/\/players/https:\/\/www\.premierleague\.com\/players/' urls.txt
```

אם קובץ urls.txt זה הקובץ בו יש את מה שמצאנו משלב 2 אז הפקודה מחפשת מופע של /players ומוסיפה לו את השם של האתר.

שימו לב שלאחר הגלישה ומציאת מספר הפעמים שמופיע המילה "midfielder" יהיו שחקנים שהם לא בתפקיד אותו אנו מחפשים (כלומר שמספר הפעמים הינו 0). ולא נרצה את הקישור שלהם בקובץ הסופי.

כמובן שזוהי רק אפשרות אחת מני רבות לבצע את המשימה. כל גישה/שיטה/רעיון שלכם יתקבל בברכה.

<u>הגשה:</u>

יש לצרף את הscript שכתבתם ביחד עם קובץ תשובות (ראו שאלות למטה) בגיט שלכן ולהגיש את הקישור במודל.

הערות ודגשים כללים:

- אם wget מהדוגמה לא עובד לכם כמצופה תבדקו אם יש לכם חיבור לאינטרנט תקין. ניתן לעשות

wget google.com

אם יש לכם אינטרנט זה אמור לעבוד.

- שימו לב לדגשים לגבי אותיות גדולות וקטנות. מבחינת הגלישה אין הבדל בין אות גדולה או קטנה, google.com==GOOGLE.COM אך בbash כן יש הבדל בין השניים.

- יש להגיש רק את הסקריפט (אין צורך להוסיף את הקובץ results.csv) להגשה בגיט. שימו לב שהסקריפט כן צריך לייצר קובץ כזה כאשר מריצים אותו.

- אפשר להניח שנריץ את הסקריפט שלכם בתיקייה ריקה כך שאם אתם יוצרים קבצים תוך כדי הריצה אין חשש מכפילות בשם הקובץ.

- אם אתם יוצרים קבצים זמניים תוך כדי התרגיל מומלץ גם לנקות אחר כך אך זה לא חובה.

- פורמט הדפסה לresults.csv:

```
https://www.premierleague.com/players/11357/Adrien-Silva/overview, Midfielder, 5
```

קישור, פסיק, רווח, Midfielder ומספר הפעמים שהופיע "midfielder" בדף השחקן.

- בקובץ הסופי לא נוסיף את הקישור עבור שחקנים שהם לא "midfielder" אך כן נחשיב אותם בסך כל השחקנים שבדקנו.

- אין חשיבות לסדר הצגת הקישורים בקובץ הסופי.

- מה שחשוב הוא החיפוש בקבצים המתקבלים מwget, אמנם נכון יהיה להיכנס לאתרים ידנית ולוודא את עצמו אך אין בהכרח התאמה. כלומר אם ניכנס באופן ידני לקישור:

```
https://www.premierleague.com/players/11357/Adrien-Silva/overview
```

ובאמצעות ctrl-f נחפש midfielderו לא נמצא 5. אך אם נעשה זאת בקובץ אשר מתקבל מwget עבור הקישור הזה כן יהיו 5 וזה מה שאנו מחפשים.

ענו בכתב בקצרה באנגלית בקובץ answers.txt:
א. כמה זמן להערכתכם היה לוקח לעשות זאת באופן ידני?

ב. לאיזה מסקנות אתם מגיעים בעקבות התרגיל? האם יש לכם רעיונות באלו עוד משימות/יישומים ניתן ליישם רעיון מסוג זה?

ג. במידה והייתי רוצה לחזור על הפעולה כל שעה? מה היה נדרש ממני? האם וכיצד הייתי יכול לבצע גם את זה באופן אוטומטי? כיצד נתמודד עם שחקני כדורגל שעדיין מופיעים וכבר נסרקו על ידי הקוד שלנו?

מהו סקריפט bash?

עד כה, ראיתם בתרגול ובסדנאות סדרה של פקודות command line אשר ניתן לשלבן ביחד (למשל על ידי pipe - | או הפניית קלט/פלט). אך מה קורה במידה ואנחנו רוצים להריץ סדרה של פקודות? לצורך כך ניתן לערוך קובץ טקסט. הקבוץ חייב להתחיל בשורה #!/bin/bash. לאחר מכן ניתן לתת לקובץ הרשאות הרצה:

```
$ chmod +x ./my_script.sh
```

ואם תוכן הקובץ נראה כך:

```
#!/bin/bash

echo hello everyone
# this is a comment
echo there are `ls -1 | wc –l` entries in this directory
```

הרצתו על ידי:

```
$ ./my_script.sh
```

תגרום להדפסה של:

```
hello everyone

the are 5 entries in this directory
```

הוראות הגשה:

1. עברו היטב על הוראות ההגשה של תרגילי הבית המופיעים באתר טרם ההגשה! ודאו כי התכנית שלכם עומדת בדרישות הבאות:
   א. התכנית קריאה וברורה
   ב. התכנית מתועדת היטב לפי דרישות התיעוד המופיעות באתר
2. יש להגיש לינק ל-repository ב-Github המכיל את הקבצים scrape_players.sh answer.txt (שימו לב לשמות הקבצים עם lower case). שימו לב להגיש לפי הפורמט הבא:

```
https://github.com/your-username/repository-name

0123456789 student_1_mail@campus.technion.ac.il first_name_1 last_name_1

0123456789 student_2_mail@campus.technion.ac.il first_name_2 last_name_2
```

3. שאלות בנוגע לתרגיל יש להפנות לפורום התרגיל ב-moodle בלבד – ניתן לשלוח שאלות במייל **למתרגל האחראי על התרגיל בלבד**, ורק במידה והשאלה מכילה פתרון חלקי.
4. סיכום מפרט התרגיל:

סעיף	תיאור
נושא התרגיל	Bash scripting
תאריך ההגשה	יום ב', 12/04/2021 בשעה 23:55
המתרגל האחראי על התרגיל	איגור דה פאולה De-paula@campus.technion.ac.il
קבצי הקוד הנתונים	
קבצי הקלט והפלט הנתונים	Results.csv
הקבצים שיש להגיש	answers.txt scrape_players.sh

בהצלחה!