תרגיל בית 4 – bash scripting

מועד ההגשה:	יום ה', 10/12/20, בשעה 23:55	
האחראי על התרגיל:	פבל ליפשיץ	pavel@ee.technion.ac.il

בתרגיל זה ננסה לבצע אוטומציה לעבודתו של יועץ תקשורת לאיש ציבור, אשר מחפש כתבות בעיתון בהן איש הציבור מוזכר.

לצורך כך, נכתוב סקריפט bash (ראו הסבר בסוף המסמך וסדנא מס' 6) שיעשה שימוש בכלי command line שנלמדו בתרגול מס' 4 ומס' 5 ובסדנא מס' 5 במטרה למצוא את כל הכתבות היומיות באתר החדשות ynetnews.com בהן מוזכרים נתניהו ו/או גנץ, ולייצר סיכום בקובץ טקסט הכולל:

1. מספר הכתבות שנבדקו
2. קישור לכל כתבה + מספר הפעמים ש- Netanyahu מוזכר בכתבה ומספר הפעמים ש- Gantz מוזכר בכתבה באופן הבא:

60

https://www.ynetnews.com/article/SklUnwXow, Netanyahu, 18, Gantz, 21

https://www.ynetnews.com/article/rJBFpUicD, -

https://www.ynetnews.com/article/Hy3I0PGow, Netanyahu, 6, Gantz, 2

…

את התוצאה יש לשמור בקובץ results.csv.

שימו לב ש- Netanyahu ו- Gantz מתחילים באות גדולה. במידה ואחד מהם לא מופיע והאחר כן, נכתוב 0 ליד זה שאינו מופיע. במידה ואף אחד מהם לא מופיע, נכתוב את הסימן -.

יש להגיש את הסקריפט בשם scrape_news.sh יחד עם קובץ התוצאה ב-git ולהגיש את הלינק במערכת ה-moodle.

הקוד הקצר והיעיל יזכו את המגישים ב-0.5 נקודה בונוס לציון הסופי.

הכוונה 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.ynetnews.com/category/3082

```
wget https://www.ynetnews.com/category/3082
```

תוצר הפקודה יהיה קובץ 3082, זהו למעשה קובץ html שהדפדפן שלכם מציג את תוכנו בצורה גרפית.

(נסו לגלוש לאתר באמצעות הדפדפן, לחצו על כפתור עכבר-ימני, ובחרו view-source).

<u>שלב שני</u>

כעת נעבור על הקובץ שהתקבל מהשלב הראשון (זהו קובץ העמוד החדשות הראשי ובו קישורים לכל כתבה/ידיעה חדשותית). אנו נחפש בו קישורים מהצורה: https://www.ynetnews.com/article/XXXXXXXXX

כאשר X יכול להיות ספרה או אות גדולה/קטנה.

ייתכן ותרצו לשמור רשימה זאת.

<u>שלב שלישי</u>

כעת ניתן לספור כמה קישורים *ייחודיים* יש לכם.

<u>שלב רביעי</u>

כעת נרצה "לגלוש" (בפקודת wget) לכל קישור כזה לידיעה חדשותית. ולבצע מעבר על התוכן מתוך מטרה לספור כמה פעמים (אם בכלל) מופיעים השמות Netanyahu ו/או Gantz.

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

הגישו את קובץ התוצאה יחד עם קובץ/י הסקריפט שתכתבו.

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 כתבות שעדיין מופיעות 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 המכיל את הקבצים scrape_news.sh results.csv answer.txt (שימו לב לשמות הקבצים עם lower case). על ה- repository להיות בעל הרשאות public. בעת בדיקת התרגיל, אנו נבצע clone ל- repository שלכם, נריץ את הסקריפט scrape_news.sh ונבדוק את קובץ ה-results.csv שיתקבל בהרצה.
   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 ה', 10/12/2020 בשעה 23:55
המתרגל האחראי על התרגיל	פבל ליפשיץ pavel@ee.technion.ac.il
קבצי הקוד הנתונים	
קבצי הקלט והפלט הנתונים	
הקבצים שיש להגיש	answers.txt results.csv scrape_news.sh

בהצלחה!