תרגיל בית מספר 1

חומר הלימוד לתרגיל: ביטויים רגולריים, פיית'ון

בתרגיל זה נעבוד עם הקובץ הנתון לכם באתר: TheLittlePrinceRegexExercise.txt.

שימו לב שעדכנתי את הקובץ ב24.5.24 בשעה 15:20. אנא הורידו אותו מחדש.

חלק א'

קראו את הקובץ וממשו את המשימות הבאות בעזרת ביטויים רגולריים ובשפת פיית'ון:

1. ספרו כמה פעמים בקובץ מופיעה המילה I (באות גדולה).
2. מצאו את המילים שמתחילות באות גדולה – ספרו כמה מילים כאלה יש בקובץ.
3. מצאו צמדי מילים שמופרדות במקף (למשל air-planes, grown-ups).
   א. כמה צמדי מילים כאלה מופיעים בקובץ?
   ב. שמרו אותם במילון שסופר את הצירופים ודווחו בקובץ התשובות (המפתח הוא צירוף המילים והערך הוא מספר ההופעות).
   למשל {'air-planes': 1, 'grown-ups': 6}.
4. מצאו מספרים בקובץ.
   א. כמה מופיעים בקובץ?
   ב. שמרו רשימה (list) של כל ההופעות לפי סדר ההופעה בקובץ
5. מצאו מילים כפולות סמוכות (למשל and and).
   א. כמה צירופים כאלה מופיעים?
   ב. שמרו אותם במילון הסופר מספר הופעות לכל צירוף (כמו בסעיף 3ב')
   ג. הדפיסו אותם לקובץ בשם duplications.txt
6. מצאו ציטוטים בקובץ. כלומר משפטים המוקפים ב"" (גרשיים כפולים) או ב " (גרש יחיד). למשל "hello world" או 'hello world'.
   א. כמה ציטוטים מצאתם?
   ב. הדפיסו אותם לקובץ בשם quotations.txt

חלק ב'

ממשו פונקציה בשם clean_text() המייצרת את הקובץ TheLittlePrinceCleaned.txt.

הקובץ החדש מבוסס על הקובץ המקורי לאחר השינויים הבאים:

א. כל האותיות הן lower case
ב. אין סימני פיסוק
ג. אין שורות חדשות (new line)
ד. מילים סמוכות יש להמיר בהופעה אחת. למשל and and יומר ב and.

הנחיות הגשה:

1. את הקוד יש לממש בקובץ בשם ex1.py או ex1.ipynb.
2. את התשובות המילוליות יש לכתוב בקובץ ex1.docx (תשובות לגבי מספר המופעים, מילונים שהתבקשתם לשמור וכו'). ציינו בראש הקובץ שם ות"ז.
3. ארזו את הקבצים מסעיפים 1+2 וכן את קבצי הפלט שיצרתם ב- **ex1.zip** והגישו במערכת המטלות של המכללה.
4. ניתן לעבוד בצוותים של עד שלושה אנשים. בהצלחה.