בס"ד

Regular expressions

ביטויים רגולריים – פעמים רבות מתייחסים אליהם בקיצור כאל RegEx, הם מחרוזות המייצגים תבניות חיפוש שישמשו אותנו למציאת התאמות ו/או אי התאמות בטקסט נתון.

לדוגמה, אם נרצה לבדוק שקלט שקלטנו הוא אימייל חוקי, נבדק אותו מול תבנית שמייצגת את התבנית של אימייל חוקי.

לכתובת אימייל חוקית יש כמה אותיות ו/או מספרים, אח"כ יבוא סימן השטרודל @ ואח"כ יבואו אותיות נוספות, שלאחריהם תופיע נקודה וסיומת com או סיומת אחרת. כלומר, לכתובת אימייל יש חוקיות – תבנית, ואנחנו רוצים לוודא שהקלט שקיבלנו מתאים לכתובת אימייל.

יש דרך להגדיר איך נראית תבנית. אנחנו ניצור תבניות באמצעות תווים שהוסכם שהם תווי הגדרה, וידוע לנו מה כל אחד מגדיר לטובת בדיקת ההתאמה.

התבנית היא כללית, סוג של תקן אם נרצה, ובשפות רבות מימשו אובייקט שיודע לבדוק התאמות על סמך ביטוי רגולרי נתון.
כל שפה לפעמים שינתה מעט את המימוש וייתכנו תבניות שיעבדו לכם בשפה אחת ולא יעבדו כפי שציפיתם בשפה אחרת, אך רובם מתייחסים לרוב התווים שמייצגים את התבנית שמורכבת מתווי הגדרה (Meta characters) בצורה זהה. כמו"כ כל שפה מימשה יכולות אחרות לאובייקט RegEx שלהם, אך בדרך כל היכולות די דומות.

נראה דוגמת קוד: במקום לכתוב קוד ייעודי שמספר זהות מכיל 9 ספרות, נוכל להשתמש ב Regex לוודא כי המחרוזת המייצגת את מס' הזהות, מכילה בדיוק 9 ספרות.

```
Regex regex = new Regex(@"^\d{9}$");
var match = regex.IsMatch (idString);
```

נהוג להשתמש ב RegEx לבדיקת אימייל, בדיקת Url's וכמעט כל מחרוזת בעלת תבנית (קרי, מתובנתת 😊).
ניתן להשתמש ב RegEx לטובת ולידיצות (בדיקות תקינות) של קלט או של כל מחרוזת נתונה, מציאת אוסף התאמות בטקסט, ולחלוקת טקסט ע"פ התבנית. לדוגמה עם תבנית שמייצגת אימייל נוכל לשלוף מקובץ את כל כתובות האימיל שבקובץ.

עכשיו נותר לנו ללמוד מה אומרת כל אות בתבנית: כל אות בתבנית. תמצאו לינקים להמון טבלאות של תווי ההגדרה, במסמך הזה הצגנו רק תווים מרכזיים יותר, כאלו שנוח לנו ללמוד דרכם את הנושא.

נתונה התבנית "\d\d" זה מייצג רצף של שני ספרות.
המחרוזת "01" תואמת לתבנית. לעומת זאת, אם נשאל אם כל המחרוזת "0136" תואמת לתבנית נקבל תשובה שלילית.
שימו לב! ב #C השאלה היא IsMatch היא האם ישנה התאמה לפחות אחת בתוך ה string ולא אם כל ה String תואם!
אבל, אם נשאל האם במחרוזת ישנם התאמות לתבנית (לא אם כל המחרוזת תואמת את התבנית, אלא האם בתוך המחרוזת ישנם תתי מחרוזת שעונים לתבנית), נקבל תשובה חיובית. ישנם 2 התאמות: "36","01".
שימו לב! היה ניתן לומר שישנם 3 התאמות ("36","13","01") אבל נהוג לחשב את ההתאמה הבאה מסיום כל התוים של ההתאמה הקודמת.

```
var r = new Regex(@"\d\d");
var b = r.IsMatch("0136"); // true  כי השאלה היא אם ישנה התאמה בתוך המחרוזת
var arr = r.Matches("0136"); // 2  התאמות
var match = r.Match("0136"); // "01" מחזיר את ההתאמה הראשונה
```

דוגמה נוספת:

```
var r = new Regex(@"^\d\d$");
var b = r.IsMatch("0136"); // false
```

```
var arr = r.Matches("0136"); // count of matches is 0
var match = r.Match("0136");  //match.success = false
```

בדוגמה זו, הוספנו לתבנית תו שמגדיר את תחילת המחרוזת ("^"), מיד אחרי תחילת המחרוזת בדיוק שתי ספרות ("\d\d") ואז מסתיימת המחרוזת ("$"), במקרה הז קיבלנו false כי אין התאמה, שהרי המחרוזת לא מסתיימת אחרי 2 הספרות.

כשמבינים את זה, נותר לנו רק לנסות שלב אחרי שלב תבניות שונות ולחפש אלו מחרוזות תואמות בדיוק לתבנית, ובאלו מחרוזות ישנם מופעים של התבנית.
מומלץ מאוד פעם אחת לעבור על האתר https://regexone.com/ (הם זכו בתרומתי בסך 4$ שמבקשים שם באתר - ביושר).

בינתיים לינק מעולה לרשימה שחלקה הועתקה כאן:
https://cheatography.com/davechild/cheat-sheets/regular-expressions/

עוגנים - Anchors

^	מציין את תחילת המחרוזת (או תחילת שורות בתבנית בעלת מס' שורות	
\A	תחילת מחרוזת	
$	מציין את סוף המחרוזת	
\Z	מציין את סוף המחרוזת	
\b	חיפוש במסגרת של מילה	\bdod מתאים למילה dod ולא נחשב תואם לאותיות dod במילה Ashdod
\B	חיפוש במסגרת שאינה מילה	
\<	התאמה בתחילת מילה	\<ba תואם לאותיות **ba**boon ולא ל Aba
\>	התאמה בסוף מילה	

קבוצות ותחומים - Groups and Ranges

.	מייצג כל תו (מלבד את התו \n – שורה חדשה).	
(a|b)	a או b	
(...)	קבוצה	
[abc]	תחום a or b or c	[abc] **a** matches **b** matches **f Does not match**
[^abc]	תואם לכל תו חוץ מ a, b ו c	f matches **a Does not match**
[a-q]	תואם לכל האותיות שבין a קטנה ל q קטנה	
[A-Q]	תואם לכל האותיות הגדולות שבין A ל Q	
[0-7]	תואם לכל הספרות שבין 0 ל 7	

תווים כמותיים - Quantifiers

את התו שמייצג כמות כותבים לאחר התו שמגדיר את ההתאמה שאנחנו מחפשים, למשל \d{6,} הכוונה היא שמחרושת שיש בה רצף של 6 ספרות או יותר תואמת לתבנית.

*	0 או יותר	**A*** **A appears ones or none times**

+	1 או יותר	
?	0 או 1 פעמים	
{3}	בדיוק 3 מופעים	A{3} AAA matches AAAA Does not match
{3,}	3 או יותר	
{3,5}	בין 3 ל 5 מופעים של התו	

ולטובת Reference כשתצטרכו, מצורפת צורה נוספת לטבלת ה Meta Characters (לא מתורגמת)

פשוט תעבדו עם מה שנוח לכם, מאחר וב RegEx אין צורך לזכור כל אות ותפקידה בתבנית (אלא אם אתם עובדים בזה הרבה ביום יום ואז באופן טבעי תזכרו את מה שנצרך לכם ואתם עושים בזה שימוש חוזר רב), צריך להכיר את הנושא, לדעת מה האפשרויות, ולחפש ולהתנסות בביטוי שנדרש לכם לכשיידרש.

Metacharacter	Description	Example
.	Matches individual characters.	x.y.z matches a string such as x1y0z or xaybz.
[ ]	Contains individual characters and value ranges to be matched.	[xyz] matches strings that contain x, y, or z.
^	Matches starting input when it is at the beginning of the expression. When inside brackets and followed by characters, it negates the characters that follow. Remarque : If followed by a bracketed group, the characters within the group are matched.	[^abc] matches strings that do not contain any combination of a, b, and c. Strings that would match include bat and bar, but not cab. ^[xyz] matches strings that start with x, y, or z.
-	Indicates a range of values to be matched. Remarque :	[1–5] matches strings such as 12345 or 26589, but not 6789.

	The range must be enclosed in brackets.	
?	Preceding characters or value ranges are an optional part of the expression to be matched.	Sept? matches Sept and September, but not December.
+	Preceding characters or value ranges can be matched one or more times.	[0–9]+ matches 1, 11, 456, and so forth.
*	Preceding characters or value ranges can be matched zero or more times.	12*3 matches 1223 and 123, but not 223 or 23.
??	Matches a minimal part of the optional characters or value ranges.	6(th)?? matches 6th.
+?	Matches a minimal part of the characters or range values that can be repeated.	Ju+? matches June and July, but not January.
*?	Matches a minimal part of the characters or range values that can be repeated.	ea*? matches strings such as each, era, and fare.
( )	Contains a group of expressions and values.	(cat) matches strings such as category and concatenate, but not cart.
\	Allows a metacharacter to be used as a literal character.	\+ allows the plus sign to be recognized as such.
$	Matches the input based on the last character.	[123]$ matches strings that end with 1, 2, or 3.
|	Matches an alternative phrase or spelling.	I|international matches International and international.

!	Indicates what characters are not included in the match.	c(a!b) matches cat or can, but not cab.

תרגילי כיתה ובית

1. כתוב ביטוי שיוודא שמחרוזת מתחילה באות D.
2. כתוב ביטוי שיוודא שמחרוזת מתחילה באותיות D או F או J.
3. הוסף בביטוי הקודם את התנאי שיש 6 ספרות אחרי האות הפותחת, ואח"כ קו תחתון ושני אותיות גדולות.
4. התבקשתם לשנות טקסט, כך שכל פעם שכתובה המילה זזזז או המילה yyy הן מוחלפות במחרוזת ריקה (במחלקת RegEx יש מתודה Replace שמחליפה במחרוזת כל טקסט שעונה על התבנית הנתונה).
5. כתוב ביטוי שיוודא שמחרוזת מתחילה באות גדולה.
6. כתוב תוכנית שבודקת באמצעות RegEx אם מחרוזת מכילה את הרצף abc או ABC.
7. כתוב תבנית הבודקת שמדובר בשם פרטי ושם משפחה (כל מילה לפחות 2 אותיות ומקסימום 18, ולפחות 2 מילים).
8. כתוב תוכנית שתוודא אם מחרוזת הינה מספר עשרוני.
9. כתוב ביטוי שבודק במס' עשרוני שיש בדיוק 3 ספרות אחרי הנקודה.
10. כתוב ביטוי המוודא שכתובת אימייל היא חוקית והיא בדומיין של Google או yahoo (יש תבניות אימייל באינטרנט, הן מסובכות, אולי עדיף להשתמש במשהו שלא סוגר את כל הפינות, אבל הוא שלכם. בחיים האמתיים תקחו ביטוי מהאינטרנט)!.
11. התבקשת לכתוב תוכנית פרוקסי בארגון, כך שכל תעבורת הרשת עוברת אצלך בתוכנית, וודא שה url הינה לדומיין במרחב .co.il (שאלת רשות: השאלה הזו, לא למדנו בשיעור. למתקדמים: קחו תבנית בדיקת URL ושנו אותה בהתאם למה שהתבקשנו. רמת קושי: גבוהה)

מתקדמים (שאלת רשות!!! ורק במידה ונשאר זמן): כתבו תוכנית עם ממשק משתמש שעוזרת ליצור RegEx (לא צריך לממש כל פיפס של השפה).