# Lab: Regular Expressions

In the current lab we are going to introduce some restrictions in the data for our database. Below you can see the constraints to validate the input entry before adding it to the data structure.

**Input format** – the format for the input should be the following:

**{Course Name}\_{Course Instance}{One or more white spaces}{Username}{One or more white spaces}{Score}**

Our task now is to write a regular expression that matches only valid entries, so we can add them to our data structure safely. Here is some example input data that may be given:

C#\_Feb\_2015 Kiko23\_4144 69

JSApps\_Dec\_2014 Ivo42\_230 17

C#\_Jul\_2016 Kiko23\_4144 94

JSApps\_Dec\_2014 Sten16\_96 41

C#\_Feb\_2015 Desi12\_2001 77

WebFundamentals\_Oct\_2015 Ivo42\_230 238

DataStructures\_Apr\_2016 Ivan23\_923 94

C#\_July\_2016 Rdsauja16\_23 71

JSApps\_Dec\_2014 NiK68\_0192 1

Unity\_Jan\_2016 Sten16\_96 56

unity\_Jan\_2016 Sten16\_96 53

JSApps\_Dec\_2014 Stan21\_23 46

C#\_Feb\_2015 NiK68\_0192 53

DataStructures\_Apr\_2016 Stan21\_23 93

WebFundamentals\_Oct\_2015 Desi12\_2001 81

Java\_May\_2015 Ivo12\_2341 77

C#\_Feb\_15 Sten16\_96 12

C#\_Feb\_2015 Desi12\_2001 93

WebFundamentals\_Oct\_2015 Kiko23\_4144 87

**Course name –** starts with a capital letter and may contain only small and capital English letters, plus ‘+’ or hashtag ‘#’.

**Course instance –** should be in the format ‘Feb\_2015’, e.g. containing the first three letters of the month, starting with a capital letter, followed by an underscore and the year. The year should be between 2014 and the current year.

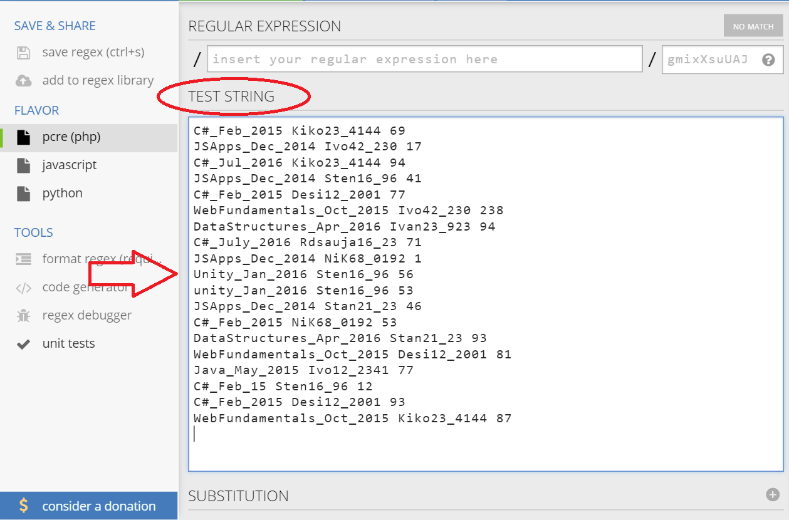
**Username** – starts with a capital letter and should be followed by at most 3 small letters after that. Then it should have 2 digits, followed by an underscore, followed by two to four digits. **Correct:** Ivan23\_234, Nas12\_4215, Re14\_203. **Incorrect:** Ivana33\_123, Stan\_12, Мари31\_421

**Score** – should be in the range from 0 to 100.

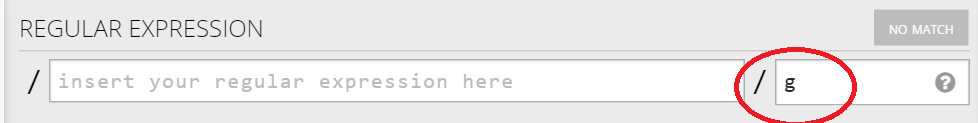
We are going to write a regular expressing for validating the input and implement it in the method for reading data from a file for the database of the university.

## Start Using a Regex Editor

First we want to open some regex editor that will help us to complete our task. You can use whatever editor you like but you should be already familiar with <https://regex101.com/> so we give you its link here. Next you may want to paste the sample data given above in the TEST STRING box:



Next you need to include the global modifier by simply adding a ‘g’ in the upper right corner:



Ok, that was pretty easy. Let’s proceed with the next task.

## Using Groups

We are going to have three groups, which are as follows:

1: Course name and instance

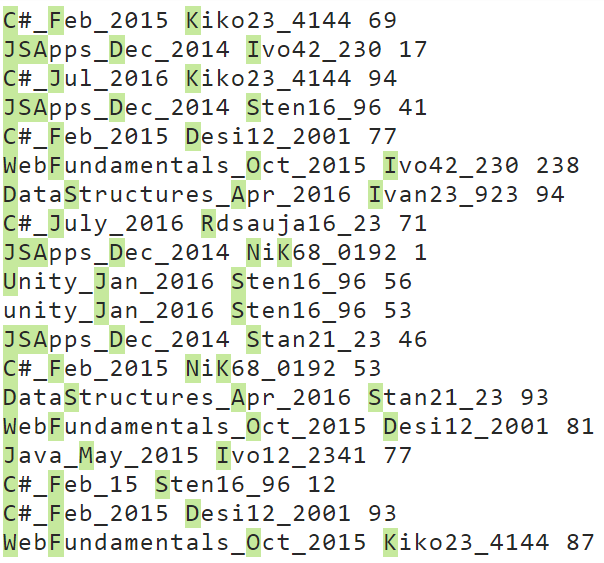
2: Student user name

3: Student score on task

First we will start with the first one and to be more specific with the course name. It should **start** with a **capital letter** so this is the first thing to add and you will be able to see as we go, how some data does not meet the conditions of the regex. So our regular expression so far should look like this:

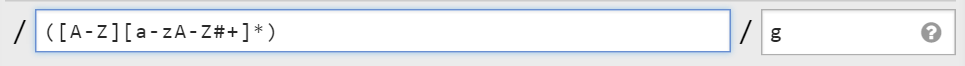


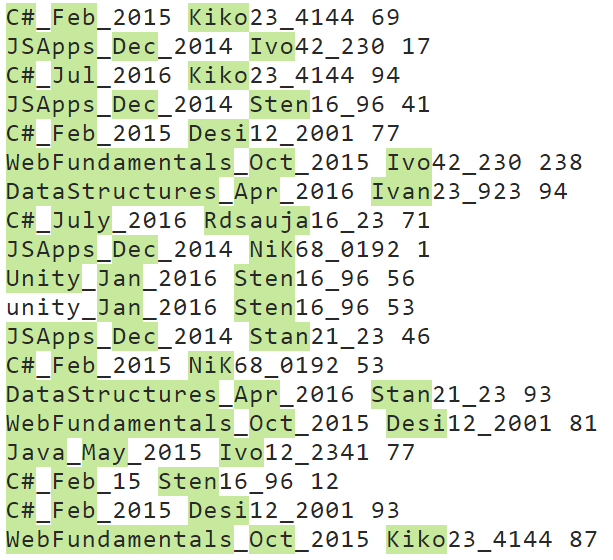
And the matches are still quite unclear:



As you can see even from the first condition we don’t catch the **unity** course written with small letters.

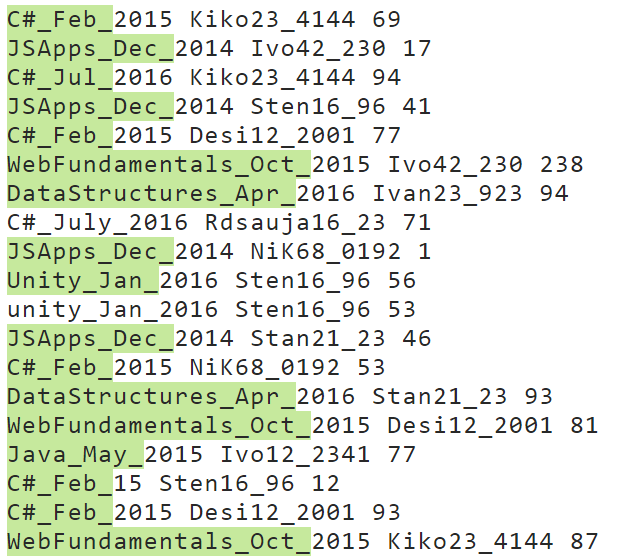
The next thing our regular expression should include is that the course name may contain small and capital letters as well as the symbols ‘+’ and ‘#’:

  
We have put an asterisk after the range, because the name of the course may be only of one capital letter. The result of the current modification looks like this:



Now it’s time to add an underscore and the condition for the month name that follows, followed by an underscore:   

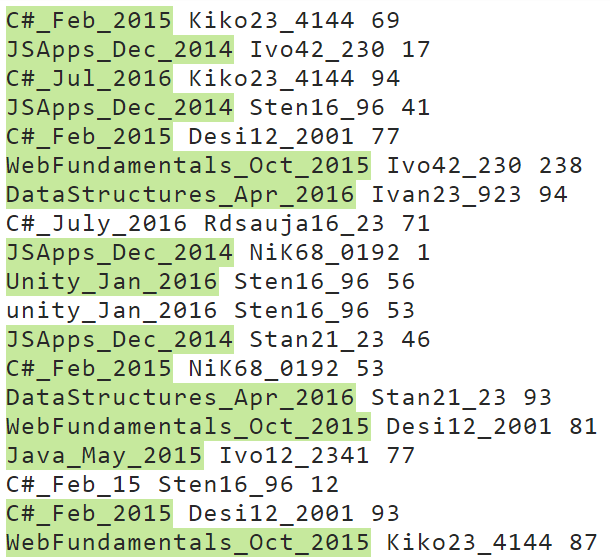

Here the condition after the range should be exactly two letters, because the total number of letters for the name of the month is 3 and the first one should be capital. The result after this addition clearly shows where we are headed:



As you can see, now the C# course in July is no longer valid because the month is written with four letters.

Finally, it is time to add the year to the matching regular expression:

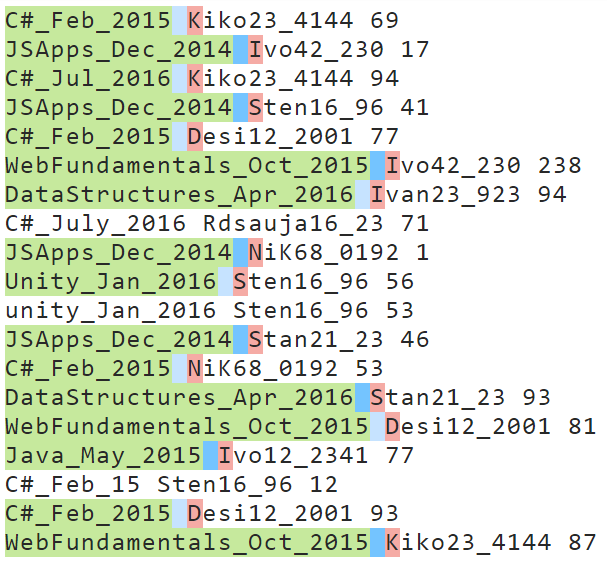
  
Now we should be ready with the group for the course and the result that we match by now should be as the one below:



As you can see, now we don’t catch the C#\_Feb\_15, because the year is not in the valid format.

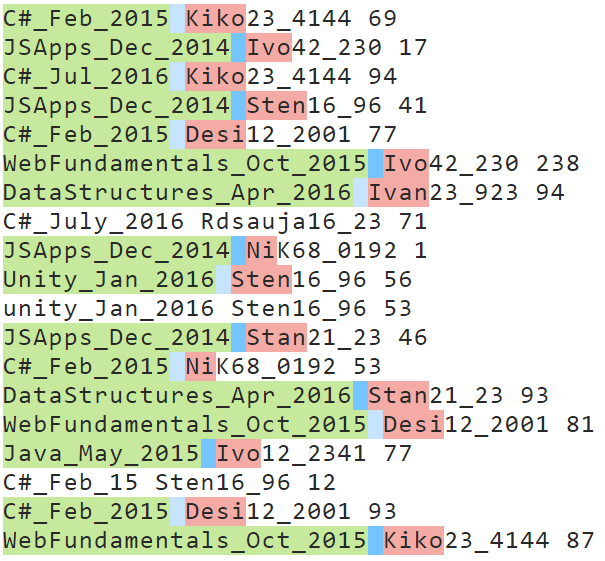
Now it is time to write the next group for the user name which is really similar to the one we just wrote.  
We have to put a separator for one or more spaces followed by the group starting with a first capital letter:



The result after this filter is pretty much the same for the input we’ve chosen, but there could have been a person whose name starts with lower letter or an entry where there is no space between the course and the user name:  


Now we should finish the regex for the rest of the name before the two numbers that follow. We should keep in mind that the we can have **at most** 3 letters after the first capital letter. This means that we may have 3 letters but there may be no letters after the first one. So this may be expressed as follows in the regular expression:



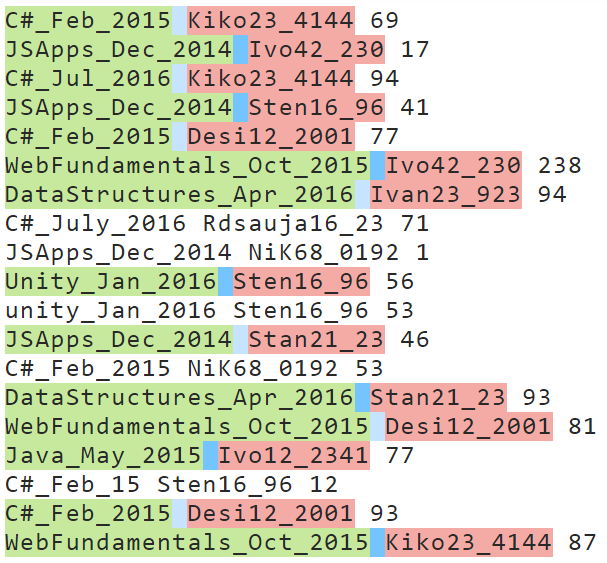


After this addition you can note that even if C#\_July\_2016 was written following the conditions, the user name would still be incorrect and of course the whole entry would be invalid.

So the only thing left for the username is the **two digits** that follow after the letters, followed by an **underscore**, which is also followed by **two to four digits:**

****

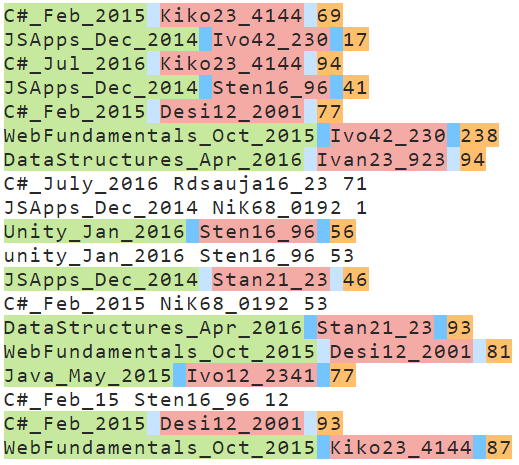
You may see below that 2 more matches are now invalid, because they don’t match the required format for the user name and more specifically for the numbers that are in the user name:



The final group we need to catch is the group for the task ‘s score for the given person and the given course:

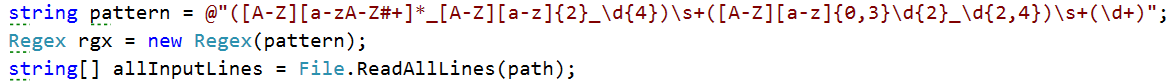


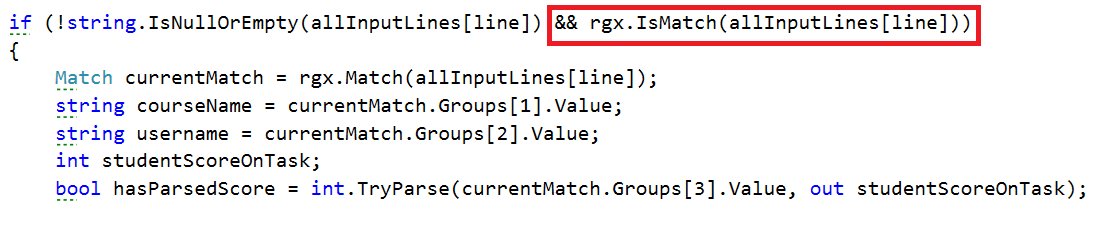
The result with the given matches from the example should now look like this:



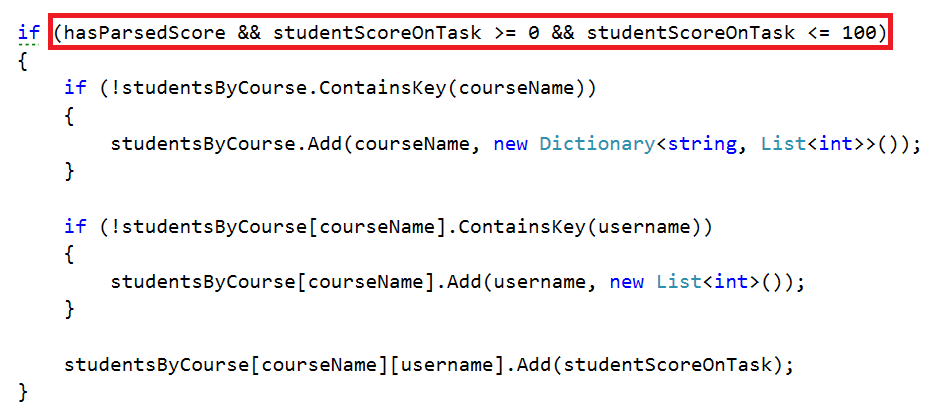
We’ve finally written the whole regular expression and it is time to implement it in C# in the method that reads all the students from the “database” file, but now you will be given a new file that will contain entries that do not match the given format.

We are going to refactor some code in the **ReadData** method in the **StudentRepository** class because we will now have to get the data from the groups of the current match.

The first thing is to copy the pattern of the match for the entries and also create a new Regex with the given pattern:   
  


Now that we have this instance of the Regex class, we can use it to check if there is a match with the current input line and if there is such, to get the data that we need from it in order to insert it in the data structure:  
  


So the only thing left is to check if the score has been parsed and whether it is in the range between 0 and 100 and if all the three conditions are true, we can insert the data that we have extracted.



By now we should be ready with the **implementation** of the **regular** **expressions**.

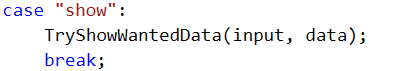
## Adding Features to the Command Interpreter

Just **before** **testing** our new functionality there is one little thing we can **add** **to the** **Command** **Interpreter**, because obviously we forgot adding it in the previous piece. We are going to add a new command which has the following format:

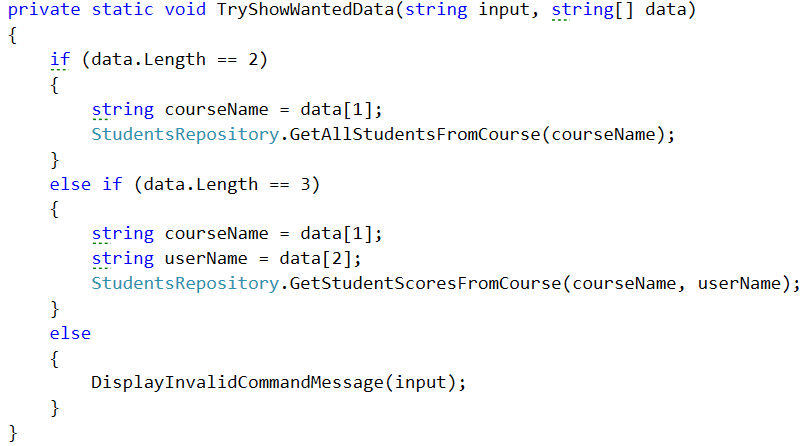
**show courseName (username) – user name may be omitted**

It’s purpose is to show information for the given course or the given username for a course from the database.

In the switch of the **Command Interpreter** the case looks like as follows:

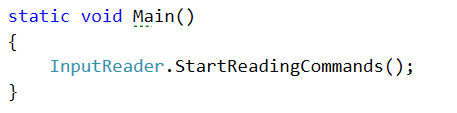


Now it’s time to add the **TryShowWantedData** method and implement its functionality. We should only check for the number of parameters and depending on this, call the corresponding method.

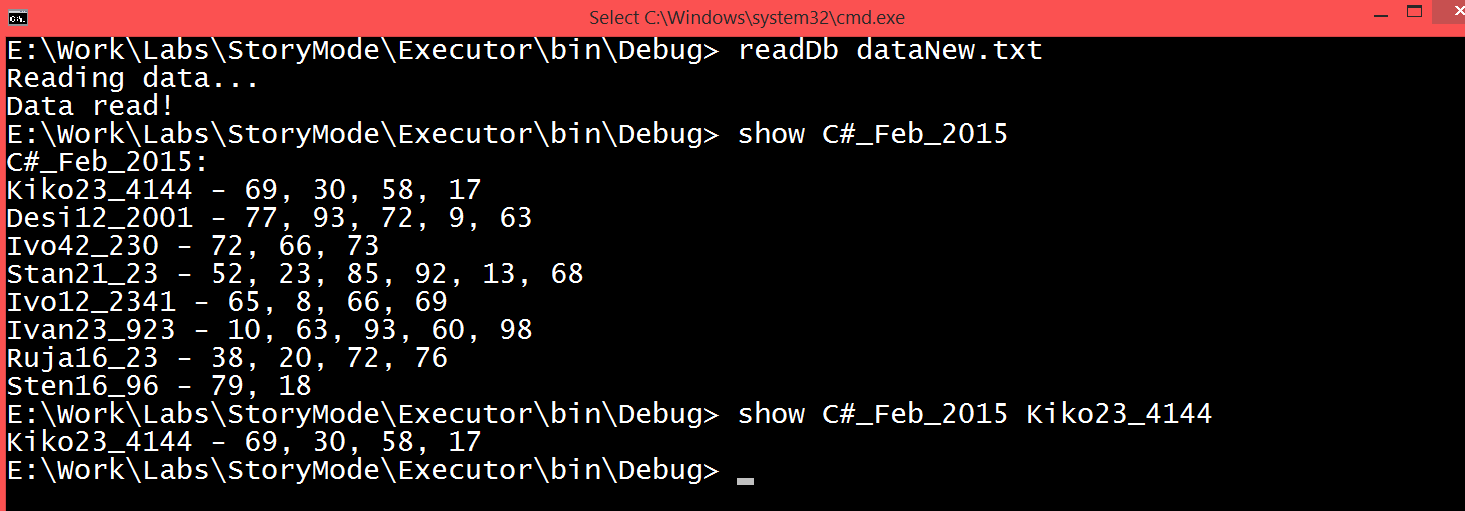


## Testing Your Code

Finally, we should be done with the corrections and now it’s time for testing. Тhe only thing we should call in the **main** **method** is **InputReader.StartReadingCommands();**



Since I’ve put the dataNew.txt in the Debug folder, I only read it, but you ’ve put it elsewhere, you’ll have to navigate to the folder first.



You should be ready with the testing! Next time expect to filter and order the data we have just read by some criteria using functional programming.

Congratulations! You’ve completed the lab exercise for Regular Expressions.