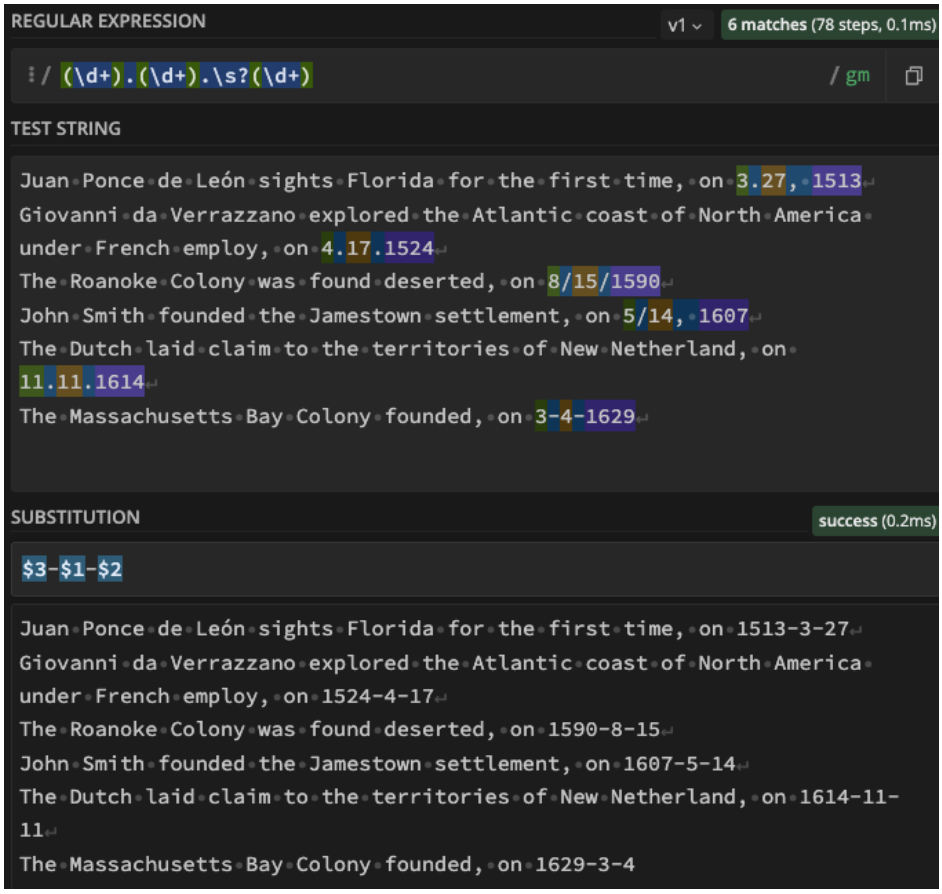
Emma-Marie Vitskov Jørgensen, Cultural Data Science 2022

Homework 1 week 35: regular expressions and spreadsheets

For this homework assignment I use the digital tool regex101 https://regex101.com/

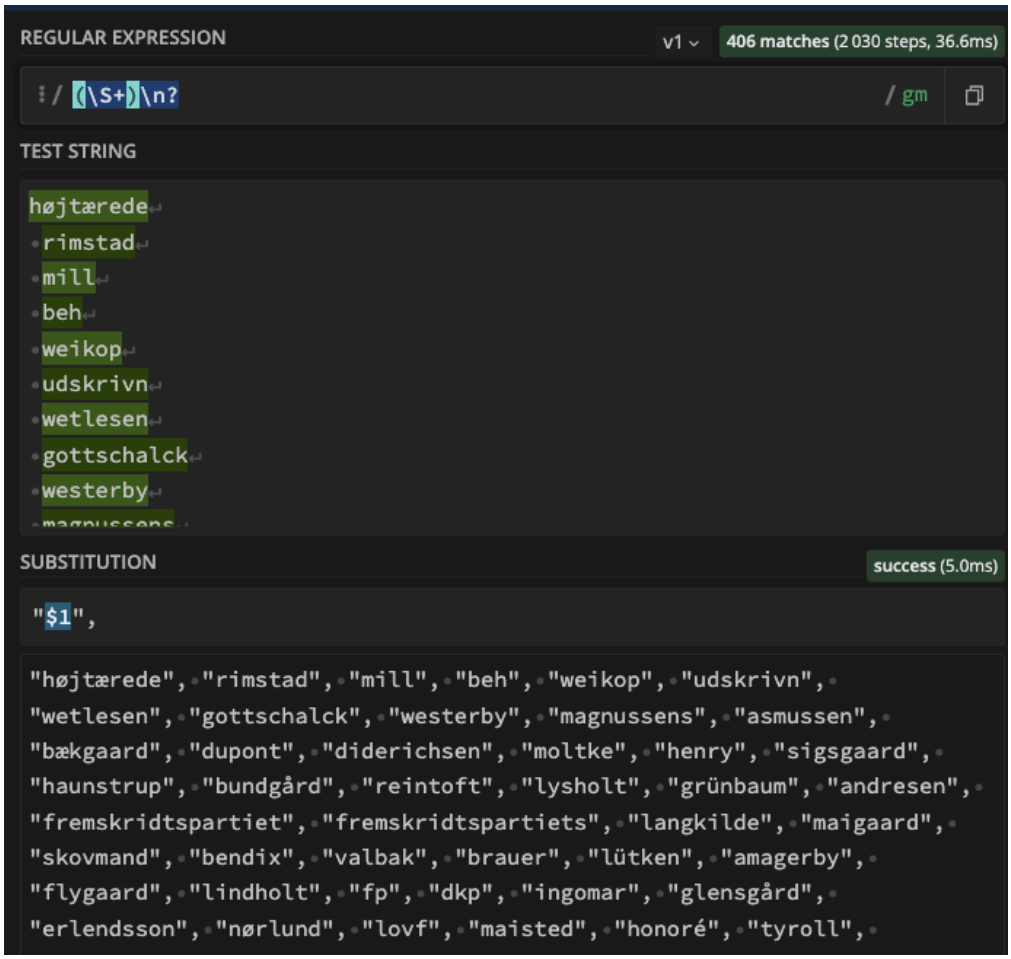
Task 1: extracting the dates and turn them into a YYYY-MM-DD format

To extract all the dates in the blurb http://bit.ly/regexexercise2, I used the tool regex101 and typed the regular expression: `(\d+).(\d+).\s?(\d+)`. This regular expression matches any single digit which appear one or more times `\d+`. I write this to times to catch both the day and month of the dates. Then I match any space, tab or newline which occurs zero or one time `\s?`. Finally, I match any single digit which appear one or more times to catch the year of the date `\d+`. The full stops match any character, and I use them in my regular expression, because a date often contains either full stops, slashes, or hyphens. I surround the dates with parenthesises. The first parenthesises can be represented by `$1`, the second by `$2` and the third by `$3`. I click on "Substitution" in my and type `$3-$1-$2`. This changes the order of the dates into YYYY-MM-DD.



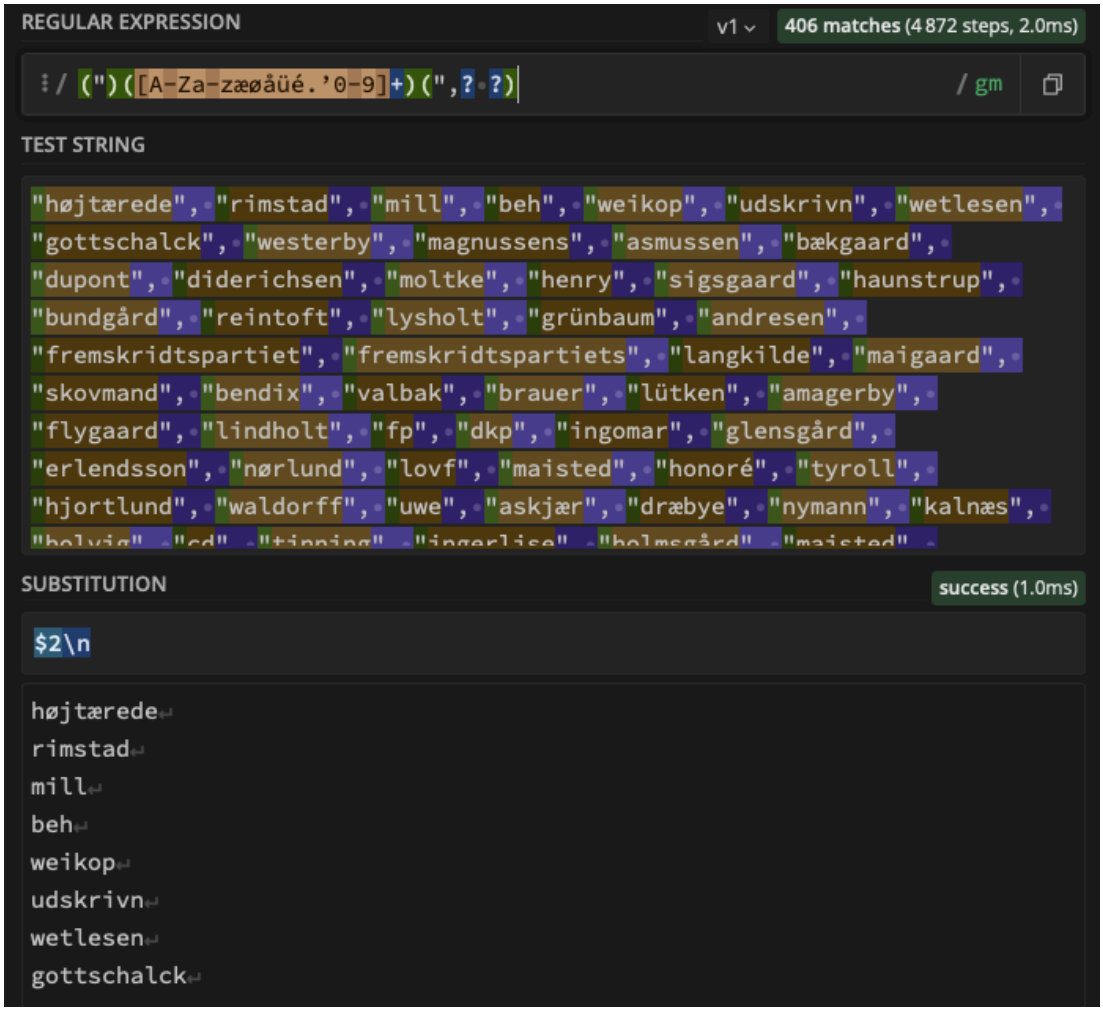
Task 2: converting stopwordlist from Voyant into stopwordlist for R and vice versa

To convert the stopwordlist from Voyant into a stopwordlist for R, I used the regular expression `(\S+)\n?`. The expression `\S+` matches any non-whitespace character which occur one or more times, and any new line characters which occur zero to one time are captures by the `\n?`. Under "substitution" I typed `"$1",` to separate the words in the hyphen with "" and dividing them by commas and a white space.



To convert the stopwordlist from R into a list for Voyant without punctuation, I used the regular expression: `(")([A-Za-zæøåüé.'0-9]+)(",? ? )`. The regex `([A-Za-zæøåüé.'0-9]+)` captures the actual words on the list, which must contain any capital or small letter and/or any digit one or more times. The words are separated from the punctuation by the regular expressions `(“)` and `(",? ?)` which matches any quotation marks followed by a comma appearing zero to one times and any following white space are appearing zero to one times . I wrote `$2\n` in the "substitution" bar to list the words inside

the second parenthesis, and this removes the punctuation.



Task 3: Answer the question "What are the basic principles for using spreadsheets for good data organisation?"

There are a few things to be aware of when using a spreadsheet for data organisation. You shouldn't colour your cells as part of your data because the computer will not interpret a colour as data. If a piece of data has the value 0, you should write a 0 in the cell and not leave the cell empty, because this can be interpreted by your software as missing data. If you have missing data, choose a name for it and be consistent. Remember to choose a name that your particular software understands e.g. NA or na if you use R or leaving the cell empty if you use R or Python. You must be consistent with variable names. If an object name is misspelled, it will not be recognized as the correct object, and it will get its own category when statistics are made. Don't put several observations in one cell but give each observation a column. Don't write notes outside the table but make the note into an

observation with its own column. It is a bad idea to create multiple tables in one spreadsheet. Instead, give each case its own row in the same table to keep the data together. Make consistent column names, and don't name some with underscore and others without. Be aware if you create multiple tabs in one spreadsheet, because if you save the spreadsheet as a csv, the only tab which will be saved is the one you are currently working in.