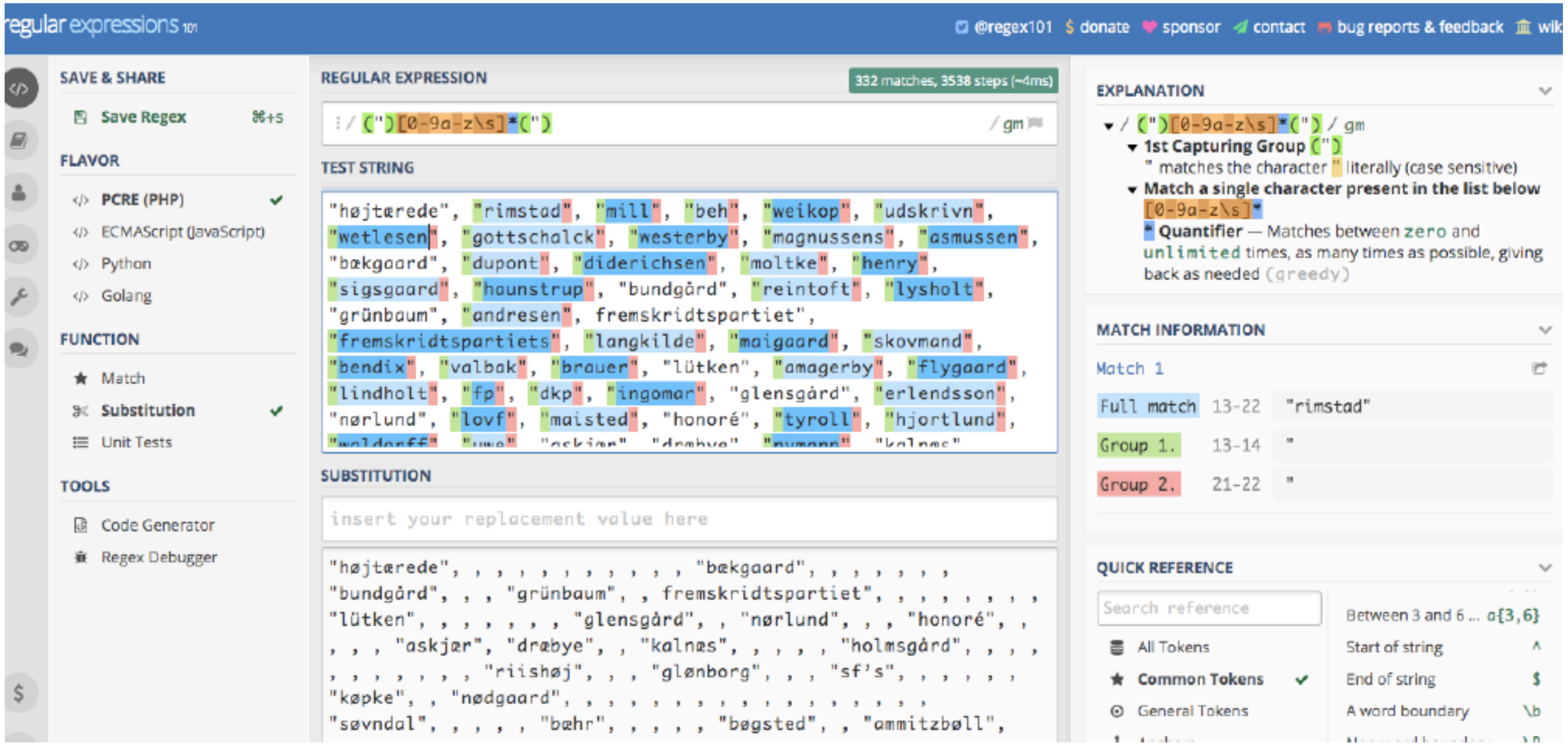
Helene Fonnesbæk Andersen 14.09.2020

W1: Regular expressions

Here are the links to my two solved exercises:
1: https://regex101.com/r/p5096s/1
2: https://regex101.com/r/COHXGB/1

EXERCISE 1

This was the first time for me working with regular expressions, so I tried a lot of things to make it work. I wanted to isolate the "", so I tried to do that, but they went into two different groups



REGULAR EXPRESSION — 332 matches, 3538 steps (~4ms)

```
/ (")[0-9a-z\s]*(") / gm
```

TEST STRING

```
"højtærede", "rimstad", "mill", "beh", "weikop", "udskrivn",
"wetlesen", "gottschalck", "westerby", "magnussens", "asmussen",
"bækgaard", "dupont", "diderichsen", "moltke", "henry",
"sigsgaard", "haunstrup", "bundgård", "reintoft", "lysholt",
"grünbaum", "andresen", fremskridtspartiet",
"fremskridtspartiets", "langkilde", "maigaard", "skovmand",
"bendix", "valbak", "brauer", "lütken", "amagerby", "flygaard",
"lindholt", "fp", "dkp", "ingomar", "glensgård", "erlendsson",
"nørlund", "lovf", "maisted", "honoré", "tyroll", "hjortlund",
"waldorff", "uwe", "askjær", "drøbye", "nymann", "kalnæs",
```

SUBSTITUTION

```
insert your replacement value here
```

```
"højtærede", , , , , , , , , , , "bækgaard", , , , , , ,
"bundgård", , , "grünbaum", , fremskridtspartiet", , , , , , , ,
"lütken", , , , , , , "glensgård", , "nørlund", , , "honoré", ,
, , , "askjær", "dræbye", , "kalnæs", , , , , "holmsgård", , , ,
, , , , , , , "riishøj", , , "glønborg", , , "sf's", , , , , ,
"køpke", , "nødgaard", , , , , , , , , , , , , , , , ,
"søvndal", , , , , "bæhr", , , , , "bøgsted", , "ammitzbøll",
```

EXPLANATION

- / (")[0-9a-z\s]*(") / gm
  - 1st Capturing Group (")
    - " matches the character " literally (case sensitive)
  - Match a single character present in the list below [0-9a-z\s]*
    - * Quantifier — Matches between zero and unlimited times, as many times as possible, giving back as needed (greedy)

MATCH INFORMATION

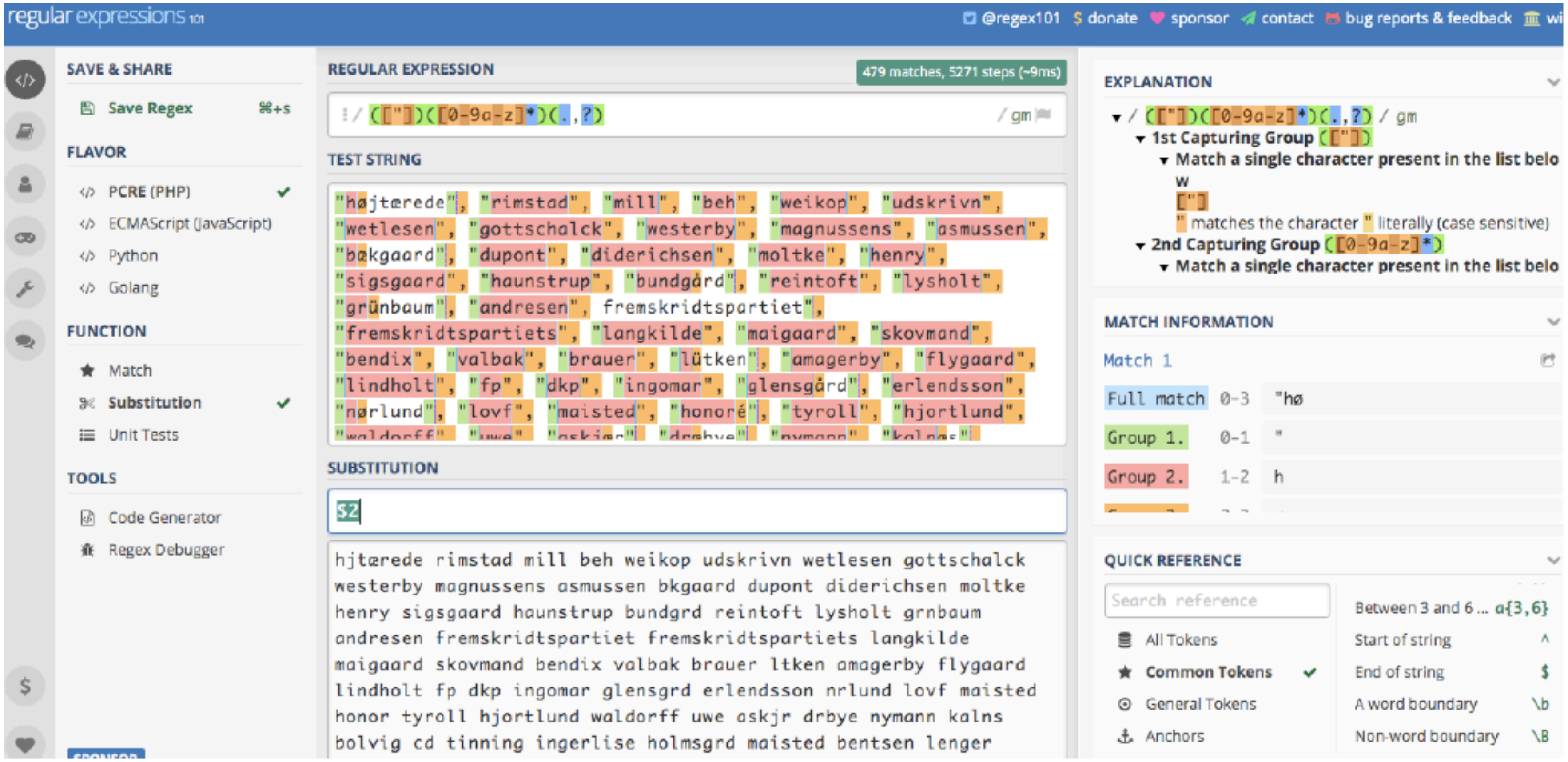
Match 1

Full match	13-22	"rimstad"
Group 1.	13-14	"
Group 2.	21-22	"

QUICK REFERENCE

Between 3 and 6 ...	a{3,6}
Start of string	^
End of string	$
A word boundary	\b

I added [], and then the “” went into the same group. Still, some words weren’t marked



REGULAR EXPRESSION

479 matches, 5271 steps (~9ms)

```
/ (["])([0-9a-z]*)(.,?) / gm
```

TEST STRING

```
"højterede", "rimstad", "mill", "beh", "weikop", "udskrivn",
"wetlesen", "gottschalck", "westerby", "magnussens", "asmussen",
"bækgaard", "dupont", "diderichsen", "moltke", "henry",
"sigsgaard", "haunstrup", "bundgård", "reintoft", "lysholt",
"grünbaum", "andresen", fremskridtspartiet",
"fremskridtspartiets", "langkilde", "maigaard", "skovmand",
"bendix", "valbak", "brauer", "lütken", "amagerby", "flygaard",
"lindholt", "fp", "dkp", "ingomar", "glensgård", "erlendsson",
"nørlund", "lovf", "maisted", "honoré", "tyroll", "hjortlund",
```

SUBSTITUTION

```
$2
```

```
hjterede rimstad mill beh weikop udskrivn wetlesen gottschalck
westerby magnussens asmussen bkgaard dupont diderichsen moltke
henry sigsgaard haunstrup bundgrd reintoft lysholt grnbaum
andresen fremskridtspartiet fremskridtspartiets langkilde
maigaard skovmand bendix valbak brauer ltken amagerby flygaard
lindholt fp dkp ingomar glensgrd erlendsson nrlund lovf maisted
honor tyroll hjortlund waldorff uwe askjr drbye nymann kalns
bolvig cd tinning ingerlise holmsgrd maisted bentsen lenger
```

EXPLANATION

- / (["])([0-9a-z]*)(.,?) / gm
  - 1st Capturing Group (["])
    - Match a single character present in the list below
      ["]
      " matches the character " literally (case sensitive)
  - 2nd Capturing Group ([0-9a-z]*)
    - Match a single character present in the list below

MATCH INFORMATION

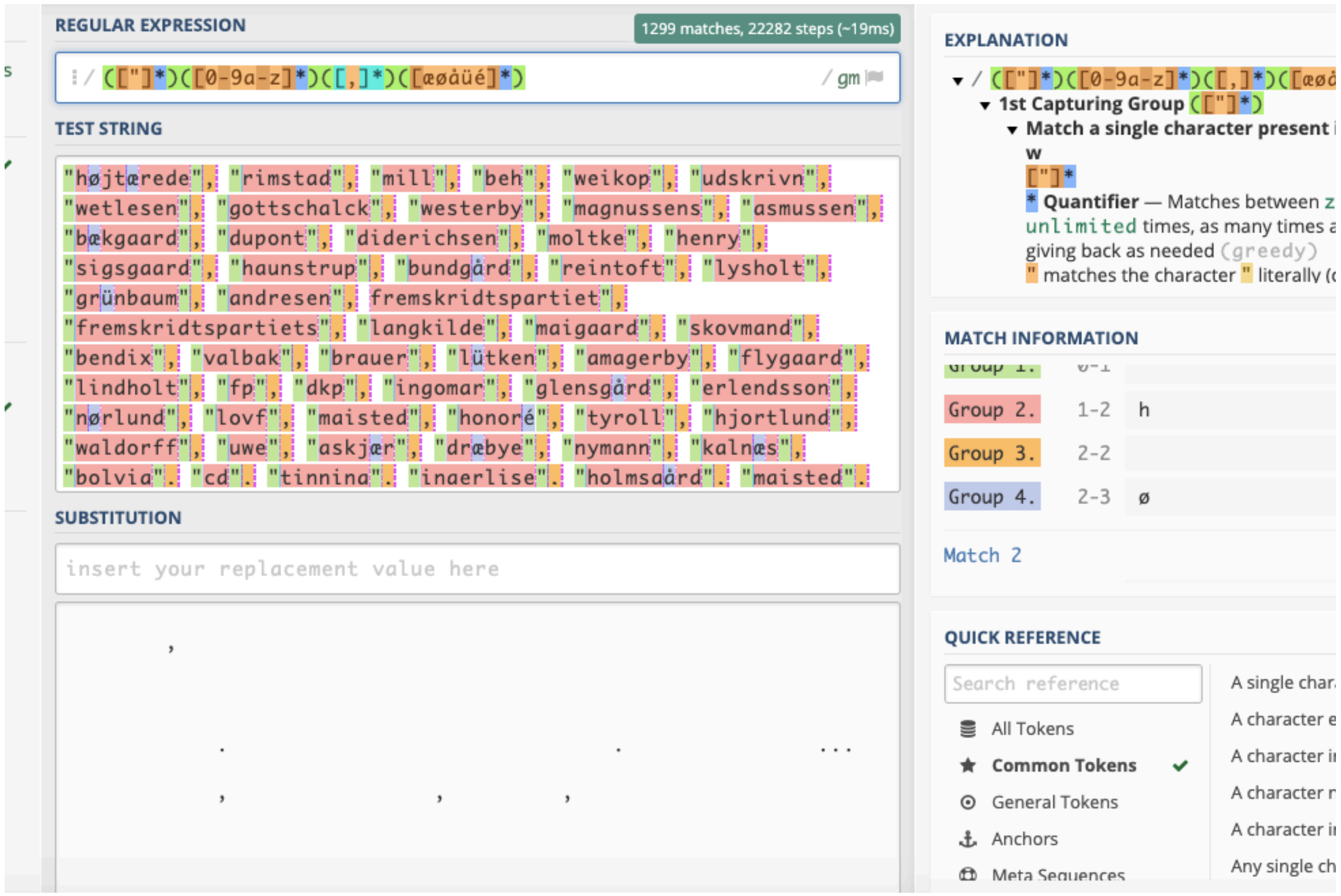
Match 1

Full match	0-3	"hø
Group 1.	0-1	"
Group 2.	1-2	h

QUICK REFERENCE

Between 3 and 6 ...	a{3,6}
Start of string	^
End of string	$
A word boundary	\b
Non-word boundary	\B

I could see that something was wrong in the words who had characters like "æøåüé". Therefore, I included these in my regular expression. They were marked and I was happy, but still, they were marked as their own group instead of part of the word



REGULAR EXPRESSION

1299 matches, 22282 steps (~19ms)

```
/ (["]*)([0-9a-z]*)([,]*)([æøåüé]*) / gm
```

TEST STRING

```
"højtærede", "rimstad", "mill", "beh", "weikop", "udskrivn",
"wetlesen", "gottschalck", "westerby", "magnussens", "asmussen",
"bækgaard", "dupont", "diderichsen", "moltke", "henry",
"sigsgaard", "haunstrup", "bundgård", "reintoft", "lysholt",
"grünbaum", "andresen", fremskridtspartiet",
"fremskridtspartiets", "langkilde", "maigaard", "skovmand",
"bendix", "valbak", "brauer", "lütken", "amagerby", "flygaard",
"lindholt", "fp", "dkp", "ingomar", "glensgård", "erlendsson",
"nørlund", "lovf", "maisted", "honoré", "tyroll", "hjortlund",
"waldorff", "uwe", "askjær", "dræbye", "nymann", "kalnæs",
"bolvia". "cd". "tinnina". "inaerlise". "holmsaård". "maisted".
```

SUBSTITUTION

insert your replacement value here

EXPLANATION

▾ / (["]*)([0-9a-z]*)([,]*)([æøå

▾ 1st Capturing Group (["]*)

▾ Match a single character present i

w

["]*

* Quantifier — Matches between z
unlimited times, as many times a
giving back as needed (greedy)
" matches the character " literally (c

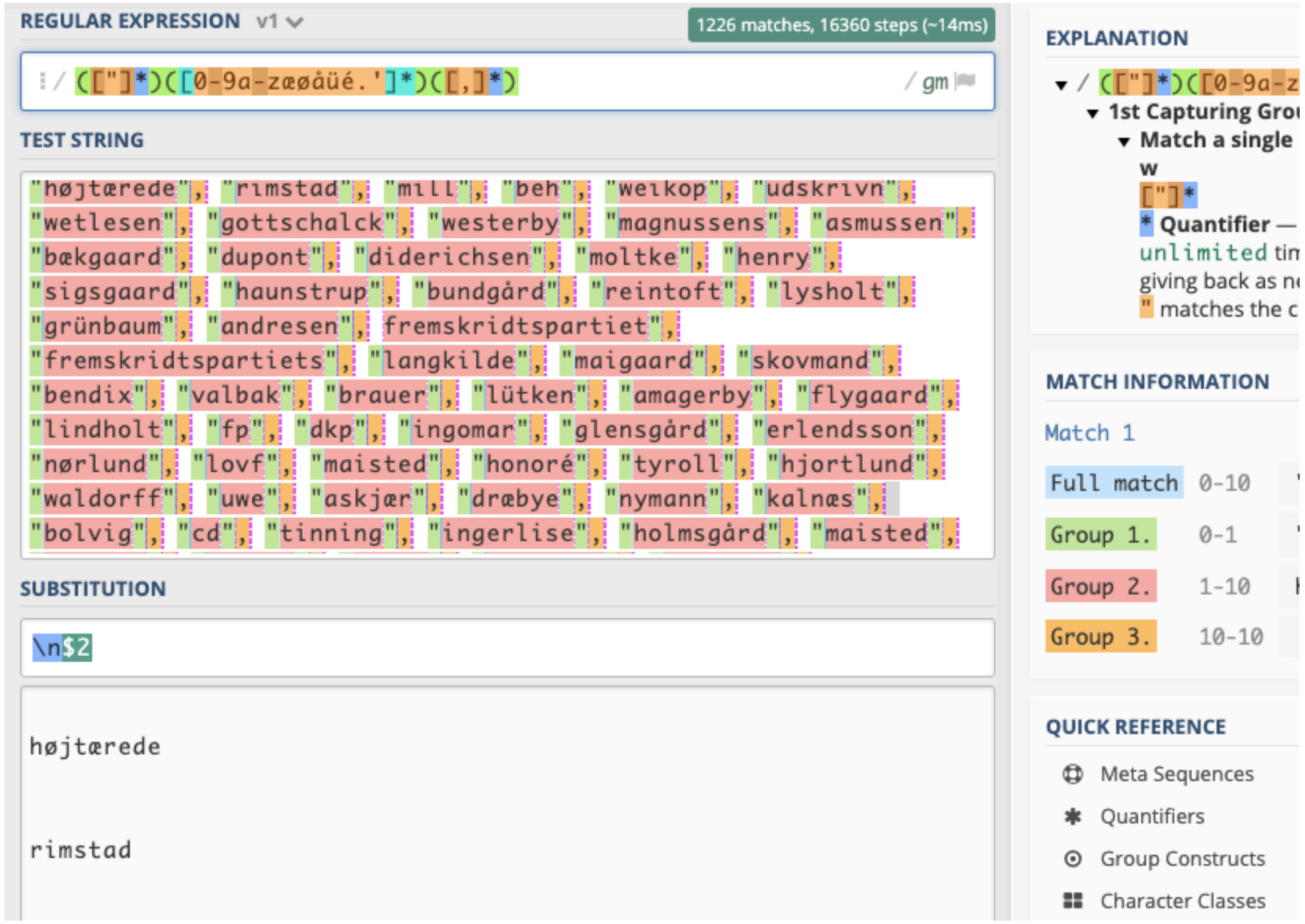
MATCH INFORMATION

Group 1.	0-1	
Group 2.	1-2	h
Group 3.	2-2	
Group 4.	2-3	ø

Match 2

QUICK REFERENCE

I tried to put these "æøåüé" a different place in the regular expression, in extension of the words, and suddenly it worked! Then I figured out that to finish my work in the substitution box, I needed to use "/n". Then my work was finished. But I did not manage to sort the words alphabetically, is that possible? And sure, there might be smarter ways to do it… Probably a way where you do not have to write all the things that you want it to mark, but rather: mark what's in between ""



REGULAR EXPRESSION v1 — 1226 matches, 16360 steps (~14ms)

```
/ (["]*)([0-9a-zæøåüé.']*)([,]*) / gm
```

TEST STRING

```
"højtærede", "rimstad", "mill", "beh", "weikop", "udskrivn",
"wetlesen", "gottschalck", "westerby", "magnussens", "asmussen",
"bækgaard", "dupont", "diderichsen", "moltke", "henry",
"sigsgaard", "haunstrup", "bundgård", "reintoft", "lysholt",
"grünbaum", "andresen", fremskridtspartiet",
"fremskridtspartiets", "langkilde", "maigaard", "skovmand",
"bendix", "valbak", "brauer", "lütken", "amagerby", "flygaard",
"lindholt", "fp", "dkp", "ingomar", "glensgård", "erlendsson",
"nørlund", "lovf", "maisted", "honoré", "tyroll", "hjortlund",
"waldorff", "uwe", "askjær", "dræbye", "nymann", "kalnæs",
"bolvig", "cd", "tinning", "ingerlise", "holmsgård", "maisted",
```

SUBSTITUTION

```
\n$2
```

```
højtærede

rimstad
```

EXPLANATION

```
/ (["]*)([0-9a-z
  1st Capturing Grou
    Match a single
    w
    ["]*
    * Quantifier —
    unlimited tim
    giving back as ne
    " matches the c
```

MATCH INFORMATION

Match 1

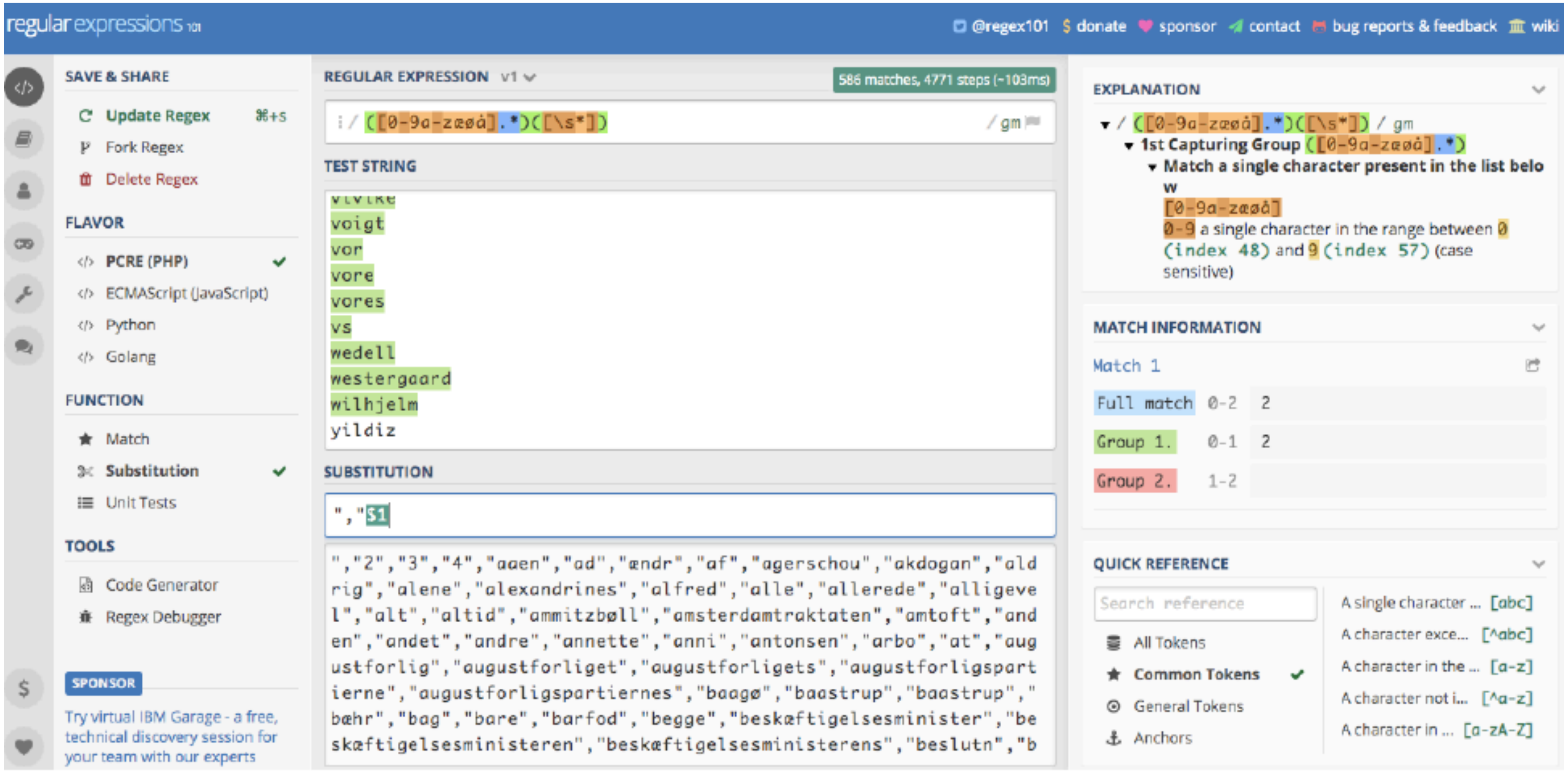
Full match	0-10
Group 1.	0-1
Group 2.	1-10
Group 3.	10-10

QUICK REFERENCE

- Meta Sequences
- Quantifiers
- Group Constructs
- Character Classes

EXERCISE 2

I tried to do the same as in exercise 1, and quickly all the words were marked. I also added "([\s]) to remove the spaces. In the substitution box, I added a comma so it would look like a stopword list. One thing I did not understand was why the world "yildiz" was not part of the list. Do you know why?



FLAVOR: PCRE (PHP) ✓

FUNCTION: Substitution ✓

REGULAR EXPRESSION v1 — 586 matches, 4771 steps (~103ms)

```
/ ([0-9a-zæøå].*)([\s*]) / gm
```

TEST STRING

```
vivike
voigt
vor
vore
vores
vs
wedell
westergaard
wilhjelm
yildiz
```

SUBSTITUTION

```
","$1
```

```
","2","3","4","aaen","ad","ændr","af","agerschou","akdogan","ald
rig","alene","alexandrines","alfred","alle","allerede","alligeve
l","alt","altid","ammitzbøll","amsterdamtraktaten","amtoft","and
en","andet","andre","annette","anni","antonsen","arbo","at","aug
ustforlig","augustforliget","augustforligets","augustforligspart
ierne","augustforligspartiernes","baagø","baastrup","baastrup","
bæhr","bag","bare","barfod","begge","beskæftigelsesminister","be
skæftigelsesministeren","beskæftigelsesministerens","beslutn","b
```

EXPLANATION

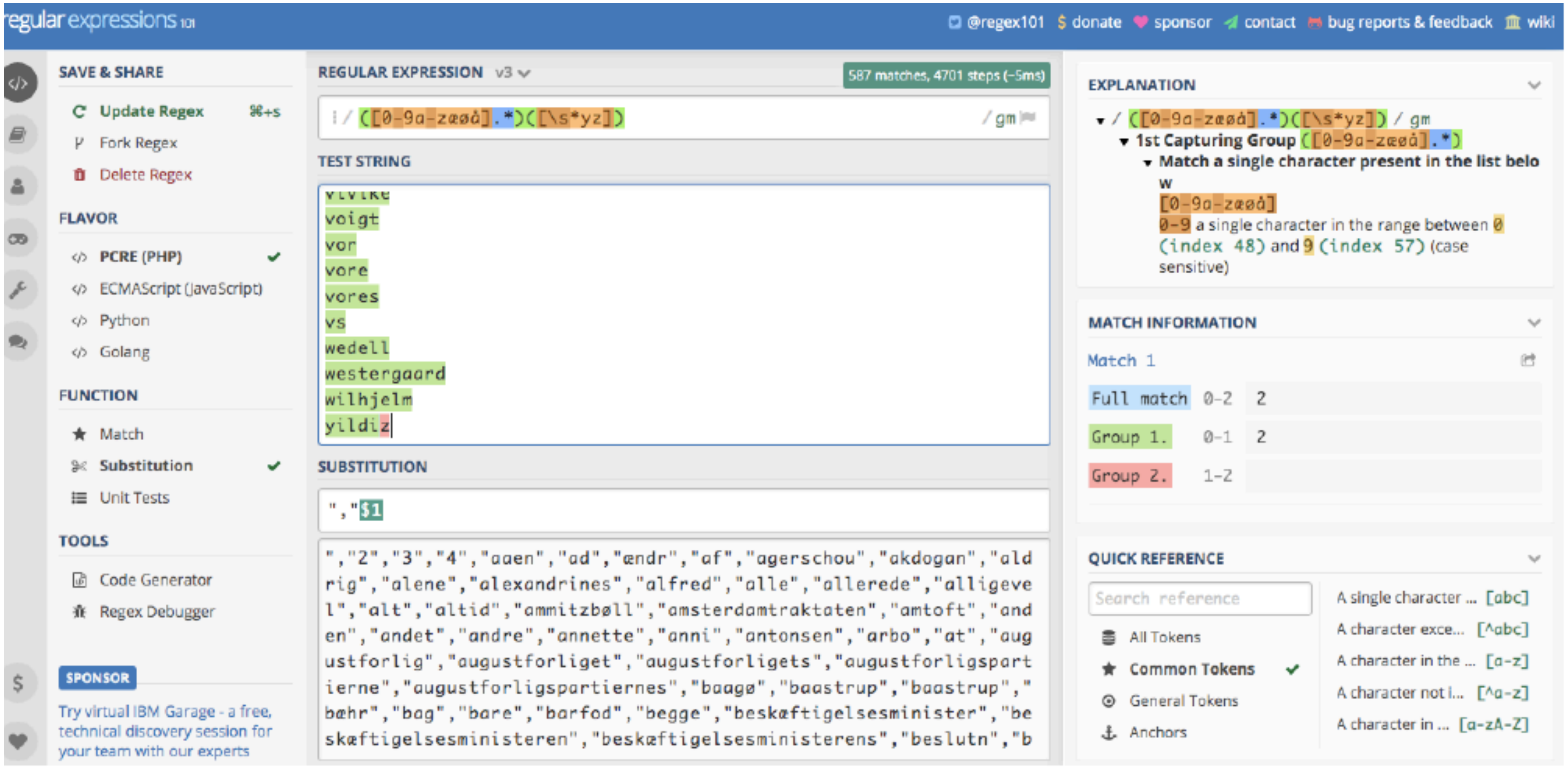
- / ([0-9a-zæøå].*)([\s*]) / gm
  - 1st Capturing Group ([0-9a-zæøå].*)
    - Match a single character present in the list below
      [0-9a-zæøå]
      0-9 a single character in the range between 0 (index 48) and 9 (index 57) (case sensitive)

MATCH INFORMATION

Match 1

Full match	0-2	2
Group 1.	0-1	2
Group 2.	1-2	

Anyway, I tried different things to include it, but it could probably have been done smarter. Here it still did not mark yildiz as one group



587 matches, 4701 steps (~5ms)

```
/ ([0-9a-zæøå].*)([\s*yz]) / gm
```

```
vivike
voigt
vor
vore
vores
vs
wedell
westergaard
wilhjelm
yildiz
```

```
","$1
```

```
","2","3","4","aaen","ad","ændr","af","agerschou","akdogan","ald
rig","alene","alexandrines","alfred","alle","allerede","alligeve
l","alt","altid","ammitzbøll","amsterdamtraktaten","amtoft","and
en","andet","andre","annette","anni","antonsen","arbo","at","aug
ustforlig","augustforliget","augustforligets","augustforligspart
ierne","augustforligspartiernes","baagø","baastrup","baastrup","
bæhr","bag","bare","barfod","begge","beskæftigelsesminister","be
skæftigelsesministeren","beskæftigelsesministerens","beslutn","b
```

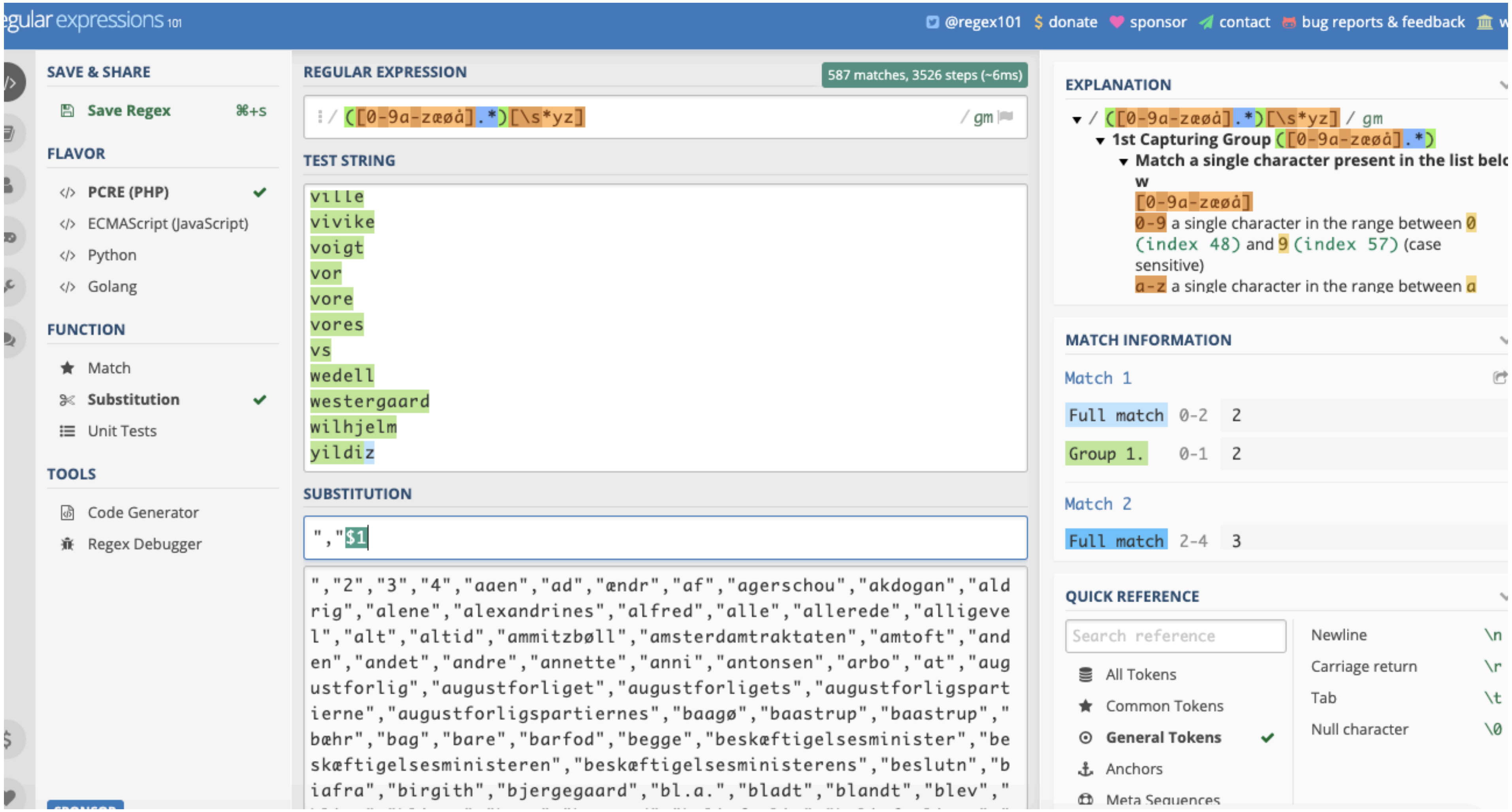
- / ([0-9a-zæøå].*)([\s*yz]) / gm
  - 1st Capturing Group ([0-9a-zæøå].*)
    - Match a single character present in the list below
      [0-9a-zæøå]
      0-9 a single character in the range between 0 (index 48) and 9 (index 57) (case sensitive)

Match 1

Full match	0-2	2
Group 1.	0-1	2
Group 2.	1-2	

Helene Fonnesbæk Andersen 14.09.2020

After I deleted the "()" in mt last regular expression, it worked. But there must be an easier way. Do you know that?



gular expressions 101

@regex101 $ donate sponsor contact bug reports & feedback

SAVE & SHARE

Save Regex ⌘+s

FLAVOR

- PCRE (PHP) ✓
- ECMAScript (JavaScript)
- Python
- Golang

FUNCTION

- Match
- Substitution ✓
- Unit Tests

TOOLS

- Code Generator
- Regex Debugger

REGULAR EXPRESSION 587 matches, 3526 steps (~6ms)

```
/ ([0-9a-zæøå].*)[\s*yz] / gm
```

TEST STRING

```
ville
vivike
voigt
vor
vore
vores
vs
wedell
westergaard
wilhjelm
yildiz
```

SUBSTITUTION

```
","$1
```

```
","2","3","4","aaen","ad","ændr","af","agerschou","akdogan","ald
rig","alene","alexandrines","alfred","alle","allerede","alligeve
l","alt","altid","ammitzbøll","amsterdamtraktaten","amtoft","and
en","andet","andre","annette","anni","antonsen","arbo","at","aug
ustforlig","augustforliget","augustforligets","augustforligspart
ierne","augustforligspartiernes","baagø","baastrup","baastrup","
bæhr","bag","bare","barfod","begge","beskæftigelsesminister","be
skæftigelsesministeren","beskæftigelsesministerens","beslutn","b
iafra","birgith","bjergegaard","bl.a.","bladt","blandt","blev","
```

EXPLANATION

- / ([0-9a-zæøå].*)[\s*yz] / gm
  - 1st Capturing Group ([0-9a-zæøå].*)
    - Match a single character present in the list belo
      w
      [0-9a-zæøå]
      0-9 a single character in the range between 0 (index 48) and 9 (index 57) (case sensitive)
      a-z a single character in the range between a

MATCH INFORMATION

Match 1

Full match	0-2	2
Group 1.	0-1	2

Match 2

Full match	2-4	3

QUICK REFERENCE

Search reference

- All Tokens
- Common Tokens
- General Tokens ✓
- Anchors
- Meta Sequences

Newline	\n
Carriage return	\r
Tab	\t
Null character	\0

As I am completely new to this and I mostly did this exercise by just trying different things, I have a few questions…

QUESTIONS

Related to exercise one:

- Is there a smarter way to just nominate all the characters in between “”, regardless of what they are?
- Is there a way to list the words alphabetically?

Related to exercise two:

- Why was the word yildiz not marked?