Fun Times with Regular Expressions

PyLing Meeting

10/10/2018

Na-Rae Han

Disclaimer!

- This session is NOT about:
  - teaching you all about regular expressions.
  - giving you a remotely full picture of regular expressions.

  🡐 Because these are impossible goals, in 1.5 hours, 10 hours, or a lifetime.

- This session is about:
  - demonstrating WHAT regular expressions are
  - demonstrating in what ways they are SUPER USEFUL
  - share what we all know and learn from each other
  - having FUN with said Regex!

WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

https://xkcd.com/208/

Setup

- Terminal (Mac users)
- Git bash (Windows users) from https://git-scm.com/downloads

- Unix commands:
  - `grep`

- Download these files:
  - `enable1.txt` from http://norvig.com/ngrams/
  - `austen-emma.txt` inside Project Gutenberg Selections corpus, from http://www.nltk.org/nltk_data/

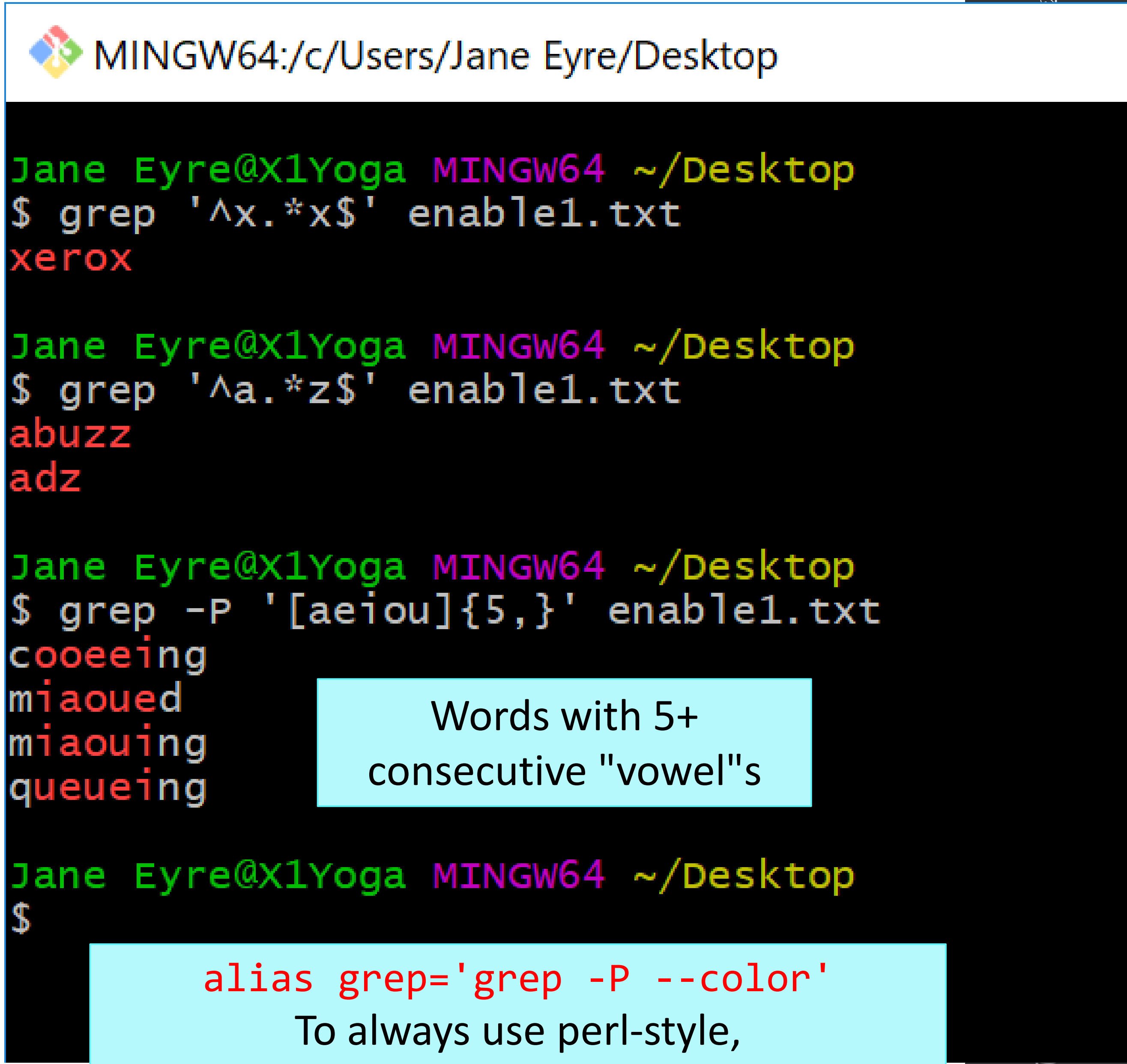
  ← Put them on your Desktop

grep

- grep
  - Searches each line in text for regular expression match
- grep -P
  - Accepts perl-style regular expressions
  - Perl-style == Python-style

Mac OS: grep -P is not available. Subsitute egrep or grep -e.

Alternatively: install GNU grep or Pcre grep.


MINGW64:/c/Users/Jane Eyre/Desktop

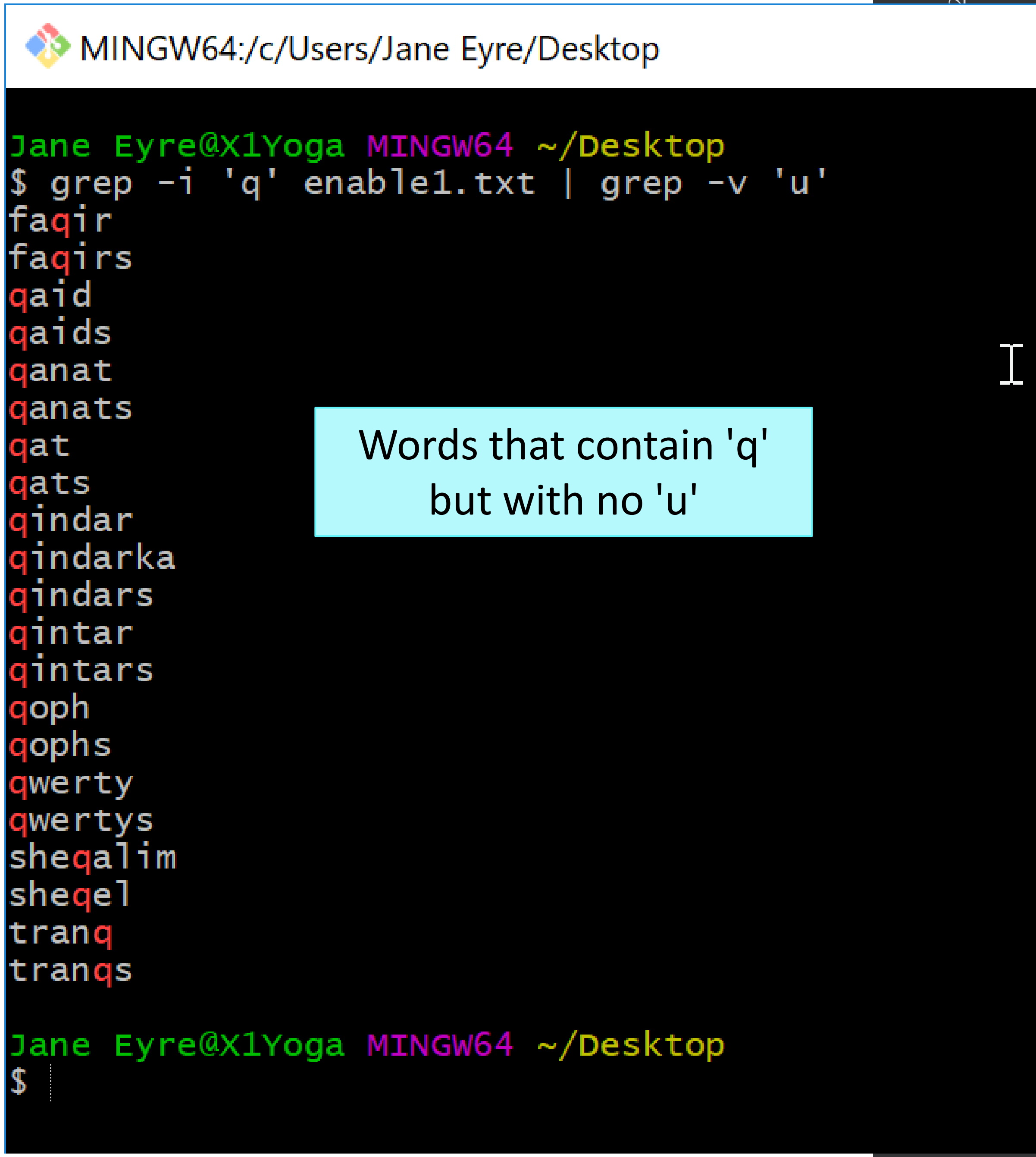
```
Jane Eyre@X1Yoga MINGW64 ~/Desktop
$ grep '^x.*x$' enable1.txt
xerox

Jane Eyre@X1Yoga MINGW64 ~/Desktop
$ grep '^a.*z$' enable1.txt
abuzz
adz

Jane Eyre@X1Yoga MINGW64 ~/Desktop
$ grep -P '[aeiou]{5,}' enable1.txt
cooeeing
miaoued
miaouing
queueing

Jane Eyre@X1Yoga MINGW64 ~/Desktop
$
```

Words with 5+ consecutive "vowel"s

alias grep='grep -P --color'
To always use perl-style,
& red color for matched portion!

grep -i, -v

- `grep -i`
  - ignores case
- `grep -v`
  - prints lines that DO NOT match



MINGW64:/c/Users/Jane Eyre/Desktop

```
Jane Eyre@X1Yoga MINGW64 ~/Desktop
$ grep -i 'q' enable1.txt | grep -v 'u'
faqir
faqirs
qaid
qaids
qanat
qanats
qat
qats
qindar
qindarka
qindars
qintar
qintars
qoph
qophs
qwerty
qwertys
sheqalim
sheqel
tranq
tranqs

Jane Eyre@X1Yoga MINGW64 ~/Desktop
$
```

Words that contain 'q' but with no 'u'

Grepping through words

- What are words that do not have any 'vowel's?
- Which words have 'wkw' in them?
- Words that are 25+ characters? Exactly 25 chars?
- [ADVANCED] Which words have …xxyyzz… pattern?
- [ADVANCED] 4-letter palindromes? 5- 6- 7-, letter?

Pipelines and I/O redirections

|

- Pass the output of one command to another for further processing

>

- Redirect the command-line output to a file. (Overwrites any existing file.)

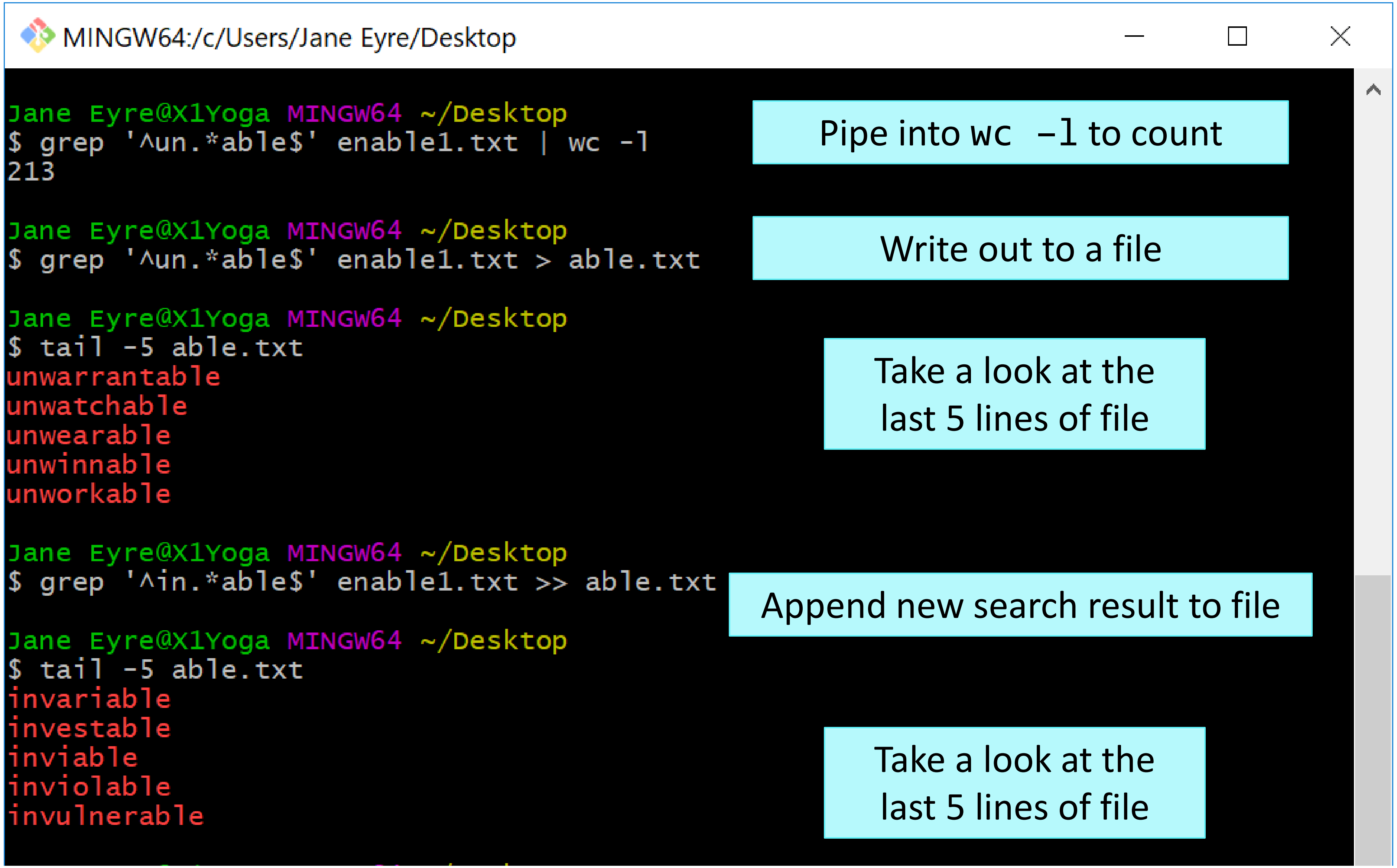
>>

- Append the output to the end of an existing file.

<

- Read from a file and feed the content as the command-line input.

grep and pipelines



```
Jane Eyre@X1Yoga MINGW64 ~/Desktop
$ grep '^un.*able$' enable1.txt | wc -l
213

Jane Eyre@X1Yoga MINGW64 ~/Desktop
$ grep '^un.*able$' enable1.txt > able.txt

Jane Eyre@X1Yoga MINGW64 ~/Desktop
$ tail -5 able.txt
unwarrantable
unwatchable
unwearable
unwinnable
unworkable

Jane Eyre@X1Yoga MINGW64 ~/Desktop
$ grep '^in.*able$' enable1.txt >> able.txt

Jane Eyre@X1Yoga MINGW64 ~/Desktop
$ tail -5 able.txt
invariable
investable
inviable
inviolable
invulnerable
```

Pipe into wc –l to count

Write out to a file

Take a look at the last 5 lines of file

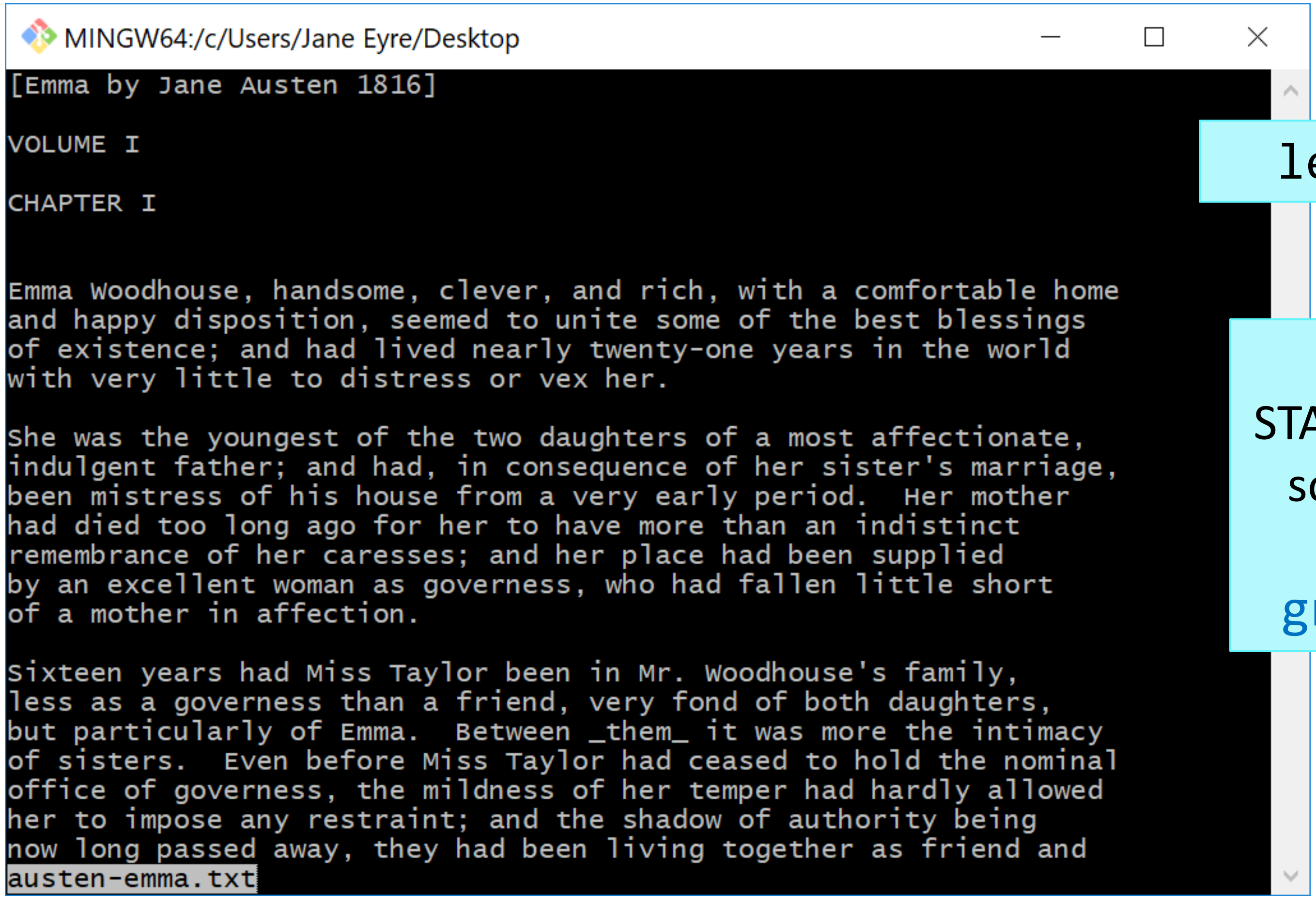
Append new search result to file

Take a look at the last 5 lines of file

more, less

Only less is available on git bash.

- more and less page through a text file content, one screen-full at a time. Press **SPACE** for next page, **q** to quit.



```
[Emma by Jane Austen 1816]

VOLUME I

CHAPTER I


Emma Woodhouse, handsome, clever, and rich, with a comfortable home
and happy disposition, seemed to unite some of the best blessings
of existence; and had lived nearly twenty-one years in the world
with very little to distress or vex her.

She was the youngest of the two daughters of a most affectionate,
indulgent father; and had, in consequence of her sister's marriage,
been mistress of his house from a very early period.  Her mother
had died too long ago for her to have more than an indistinct
remembrance of her caresses; and her place had been supplied
by an excellent woman as governess, who had fallen little short
of a mother in affection.

Sixteen years had Miss Taylor been in Mr. Woodhouse's family,
less as a governess than a friend, very fond of both daughters,
but particularly of Emma.  Between _them_ it was more the intimacy
of sisters.  Even before Miss Taylor had ceased to hold the nominal
office of governess, the mildness of her temper had hardly allowed
her to impose any restraint; and the shadow of authority being
now long passed away, they had been living together as friend and
austen-emma.txt
```

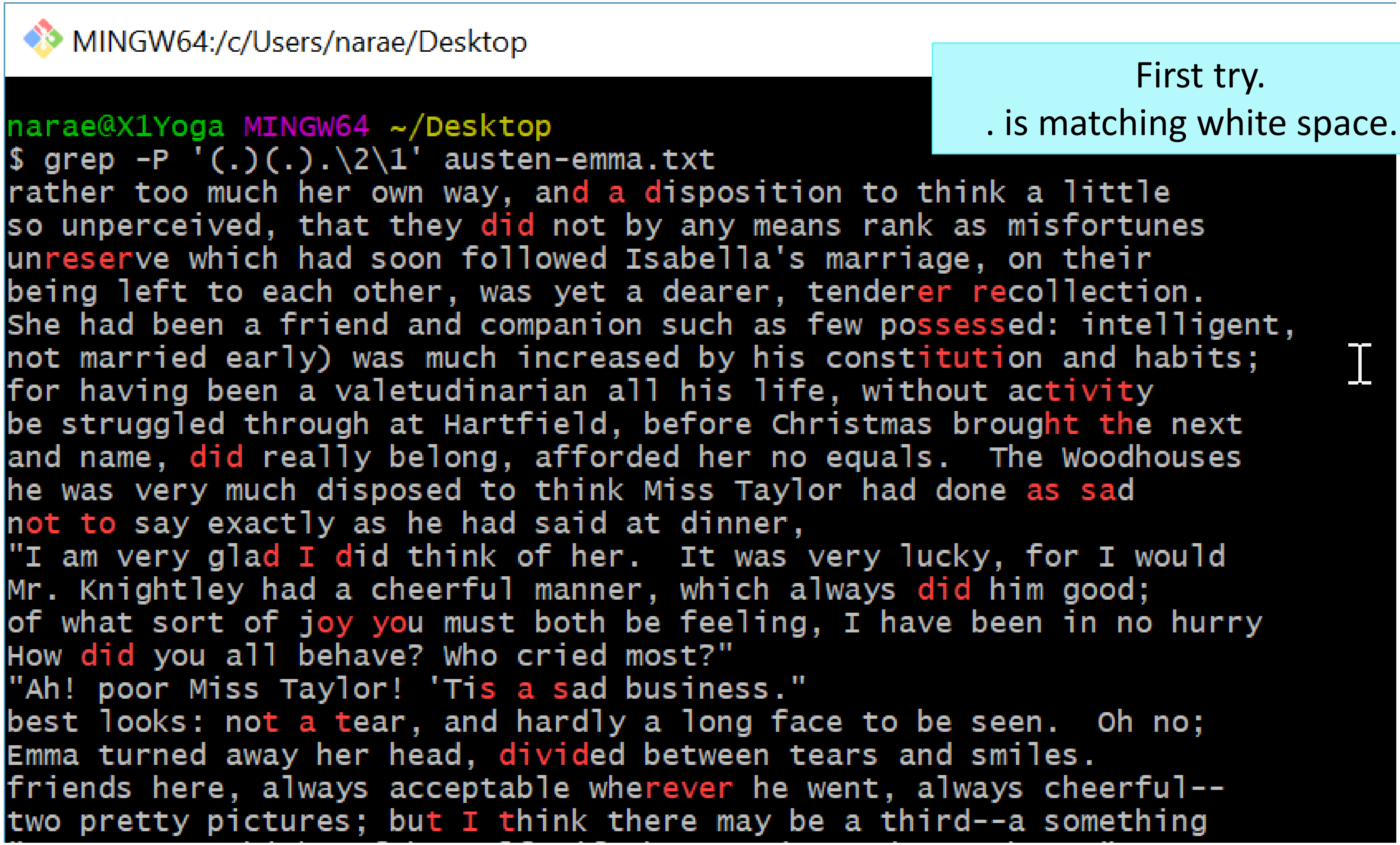
```
less austen-emma.txt
```

Often, you **pipe** your STANDARD OUTPUT into more, so you can look through the result, e.g.,

```
grep 'q' words | more
```

SPACE for next page
q to quit

5-letter palindromes in Emma



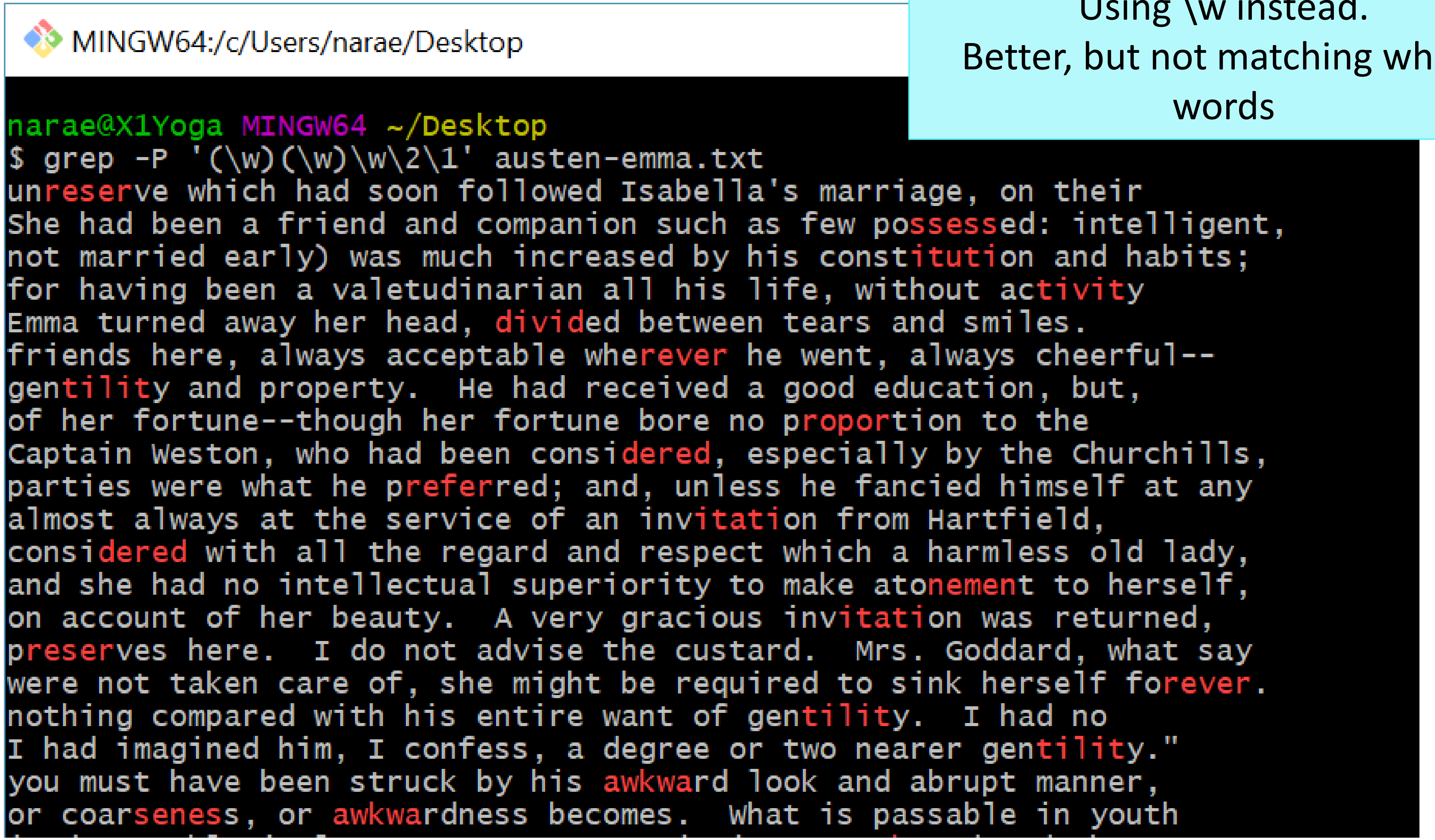
MINGW64:/c/Users/narae/Desktop

```
narae@X1Yoga MINGW64 ~/Desktop
$ grep -P '(.)(.).\2\1' austen-emma.txt
rather too much her own way, and a disposition to think a little
so unperceived, that they did not by any means rank as misfortunes
unreserve which had soon followed Isabella's marriage, on their
being left to each other, was yet a dearer, tenderer recollection.
She had been a friend and companion such as few possessed: intelligent,
not married early) was much increased by his constitution and habits;
for having been a valetudinarian all his life, without activity
be struggled through at Hartfield, before Christmas brought the next
and name, did really belong, afforded her no equals.  The Woodhouses
he was very much disposed to think Miss Taylor had done as sad
not to say exactly as he had said at dinner,
"I am very glad I did think of her.  It was very lucky, for I would
Mr. Knightley had a cheerful manner, which always did him good;
of what sort of joy you must both be feeling, I have been in no hurry
How did you all behave? Who cried most?"
"Ah! poor Miss Taylor! 'Tis a sad business."
best looks: not a tear, and hardly a long face to be seen.  Oh no;
Emma turned away her head, divided between tears and smiles.
friends here, always acceptable wherever he went, always cheerful--
two pretty pictures; but I think there may be a third--a something
```

First try.
. is matching white space…

5-letter palindromes, again

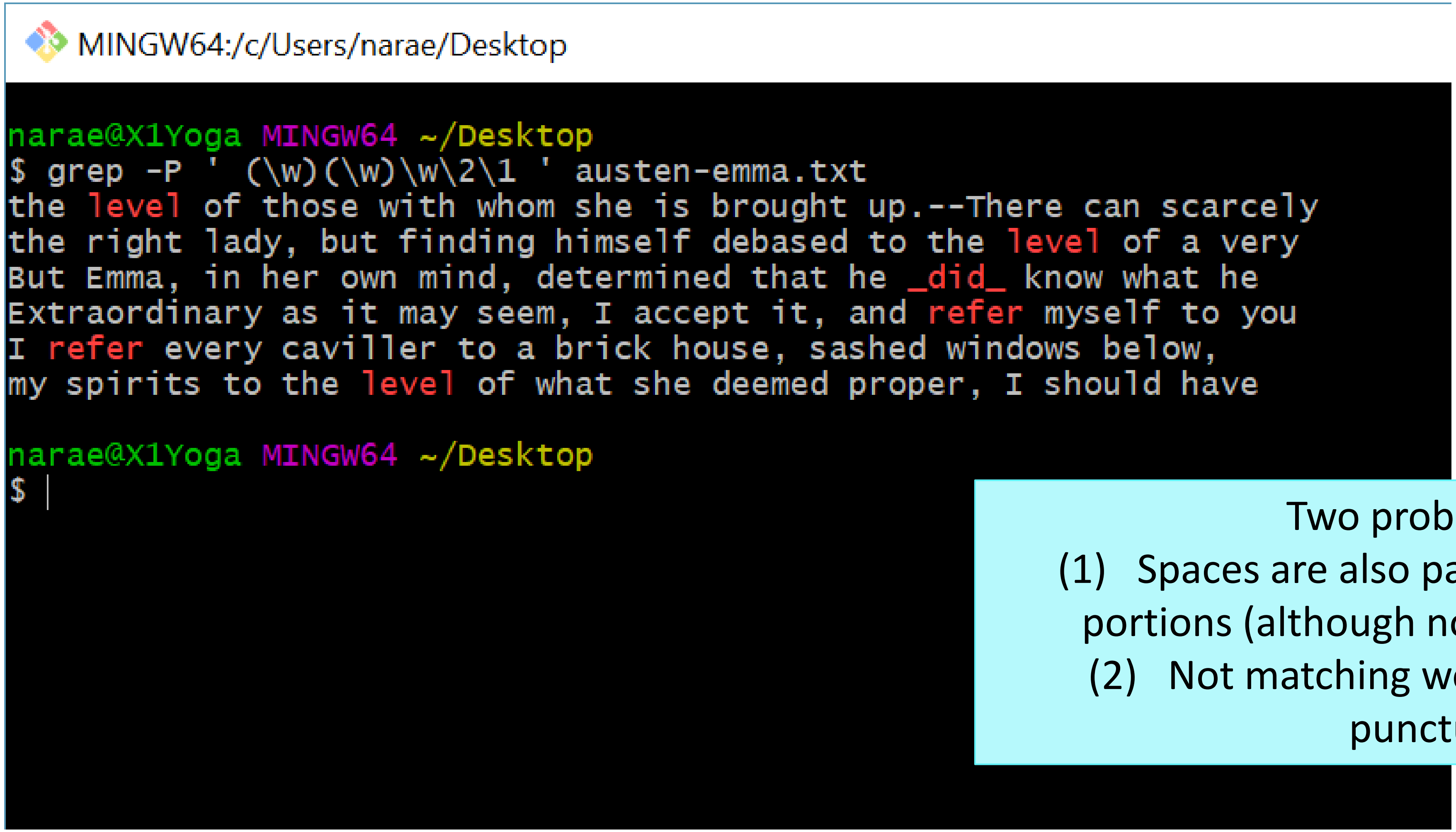
Using \w instead.
Better, but not matching whole words



MINGW64:/c/Users/narae/Desktop

```
narae@X1Yoga MINGW64 ~/Desktop
$ grep -P '(\w)(\w)\w\2\1' austen-emma.txt
unreserve which had soon followed Isabella's marriage, on their
She had been a friend and companion such as few possessed: intelligent,
not married early) was much increased by his constitution and habits;
for having been a valetudinarian all his life, without activity
Emma turned away her head, divided between tears and smiles.
friends here, always acceptable wherever he went, always cheerful--
gentility and property.  He had received a good education, but,
of her fortune--though her fortune bore no proportion to the
Captain Weston, who had been considered, especially by the Churchills,
parties were what he preferred; and, unless he fancied himself at any
almost always at the service of an invitation from Hartfield,
considered with all the regard and respect which a harmless old lady,
and she had no intellectual superiority to make atonement to herself,
on account of her beauty.  A very gracious invitation was returned,
preserves here.  I do not advise the custard.  Mrs. Goddard, what say
were not taken care of, she might be required to sink herself forever.
nothing compared with his entire want of gentility.  I had no
I had imagined him, I confess, a degree or two nearer gentility."
you must have been struck by his awkward look and abrupt manner,
or coarseness, or awkwardness becomes.  What is passable in youth
```

Space doesn't cut it



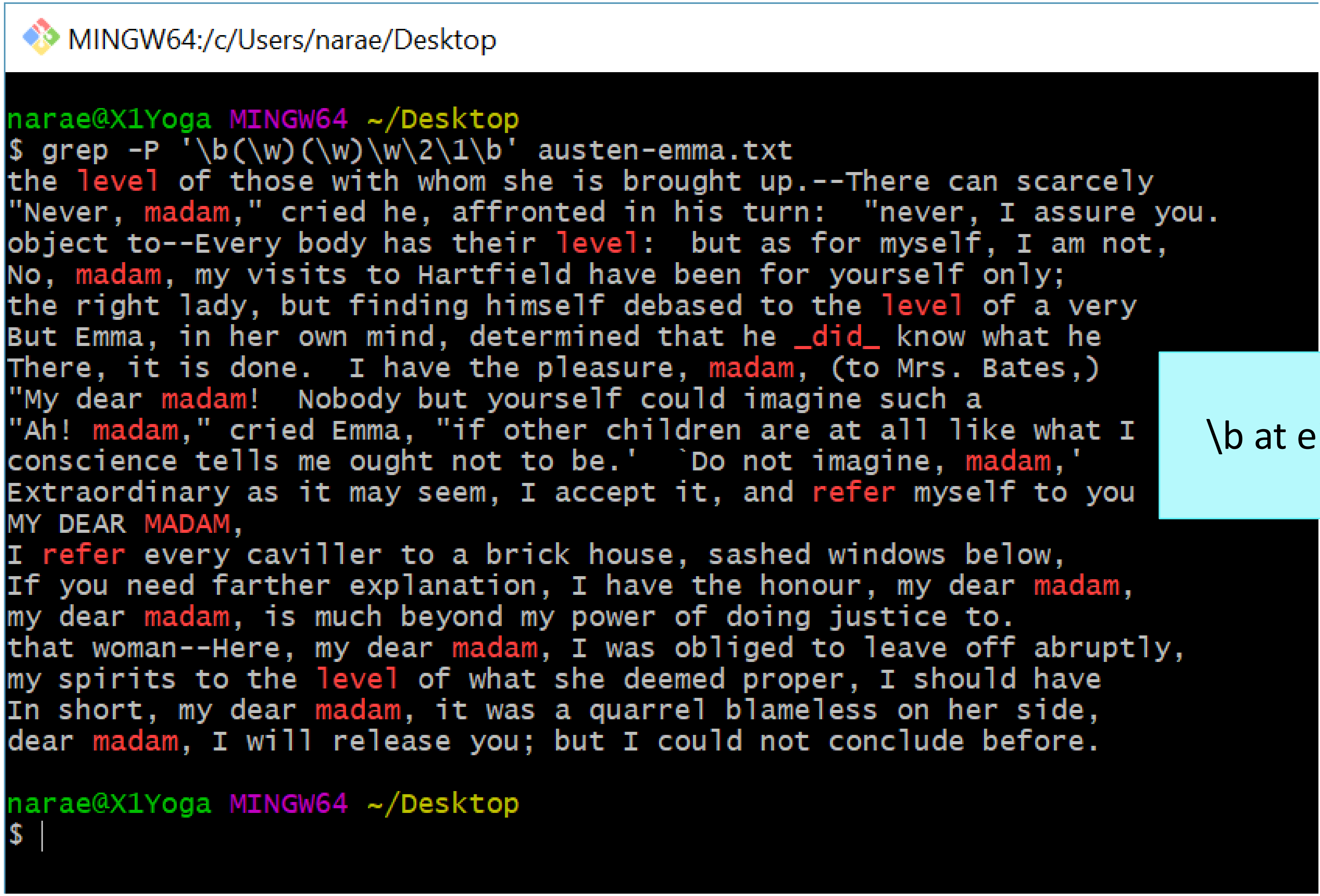
```
narae@X1Yoga MINGW64 ~/Desktop
$ grep -P ' (\w)(\w)\w\2\1 ' austen-emma.txt
the level of those with whom she is brought up.--There can scarcely
the right lady, but finding himself debased to the level of a very
But Emma, in her own mind, determined that he _did_ know what he
Extraordinary as it may seem, I accept it, and refer myself to you
I refer every caviller to a brick house, sashed windows below,
my spirits to the level of what she deemed proper, I should have

narae@X1Yoga MINGW64 ~/Desktop
$ 
```

Two problems:

(1) Spaces are also part of the matched portions (although not showing up as red)

(2) Not matching words followed by punctuation

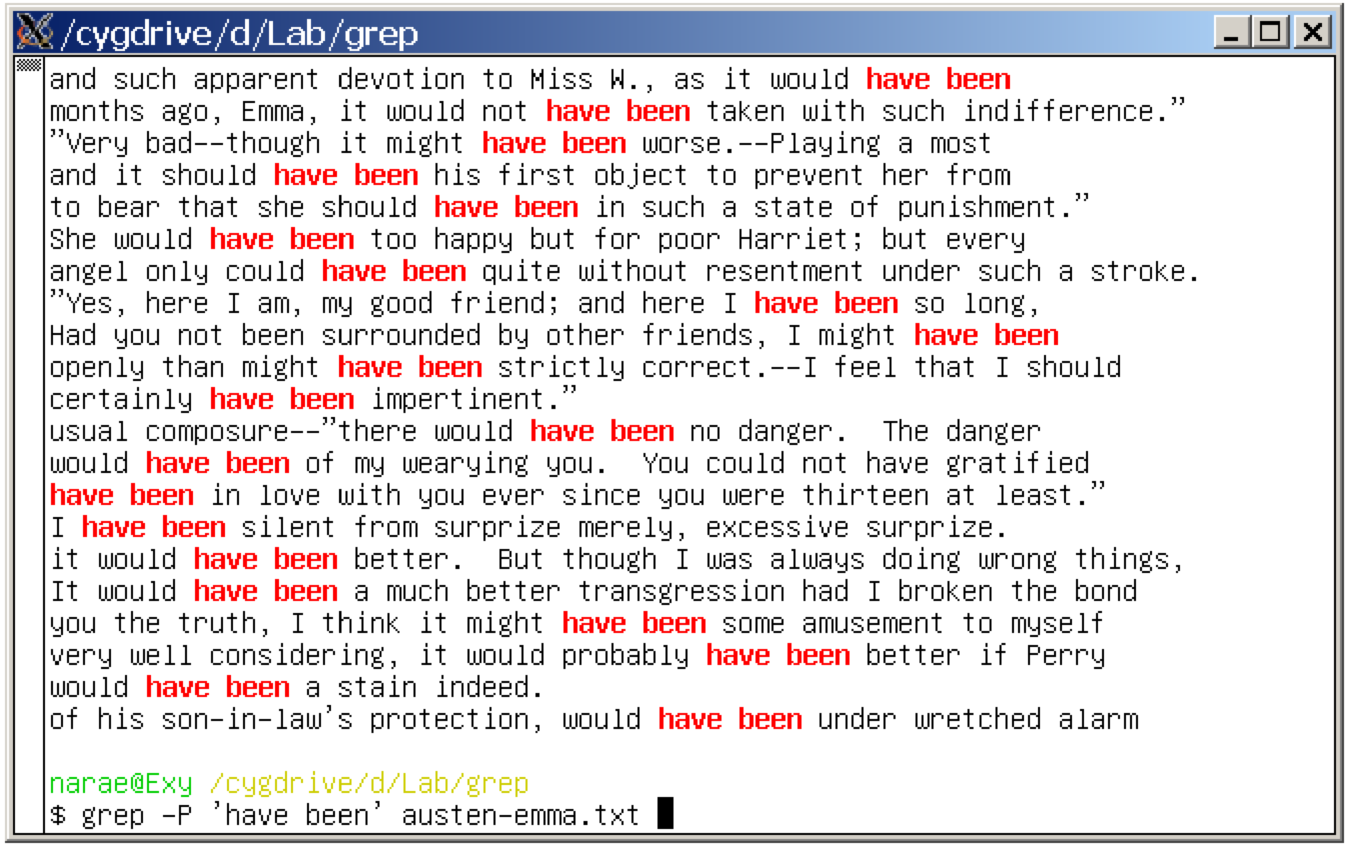
Proper word boundary: \b



```
narae@X1Yoga MINGW64 ~/Desktop
$ grep -P '\b(\w)(\w)\w\2\1\b' austen-emma.txt
the level of those with whom she is brought up.--There can scarcely
"Never, madam," cried he, affronted in his turn:  "never, I assure you.
object to--Every body has their level:  but as for myself, I am not,
No, madam, my visits to Hartfield have been for yourself only;
the right lady, but finding himself debased to the level of a very
But Emma, in her own mind, determined that he _did_ know what he
There, it is done.  I have the pleasure, madam, (to Mrs. Bates,)
"My dear madam!  Nobody but yourself could imagine such a
"Ah! madam," cried Emma, "if other children are at all like what I
conscience tells me ought not to be.'  `Do not imagine, madam,'
Extraordinary as it may seem, I accept it, and refer myself to you
MY DEAR MADAM,
I refer every caviller to a brick house, sashed windows below,
If you need farther explanation, I have the honour, my dear madam,
my dear madam, is much beyond my power of doing justice to.
that woman--Here, my dear madam, I was obliged to leave off abruptly,
my spirits to the level of what she deemed proper, I should have
In short, my dear madam, it was a quarrel blameless on her side,
dear madam, I will release you; but I could not conclude before.

narae@X1Yoga MINGW64 ~/Desktop
$ 
```

SUCCESS!
\b at either end, which marks word boundary

Searching for a phrase, literally

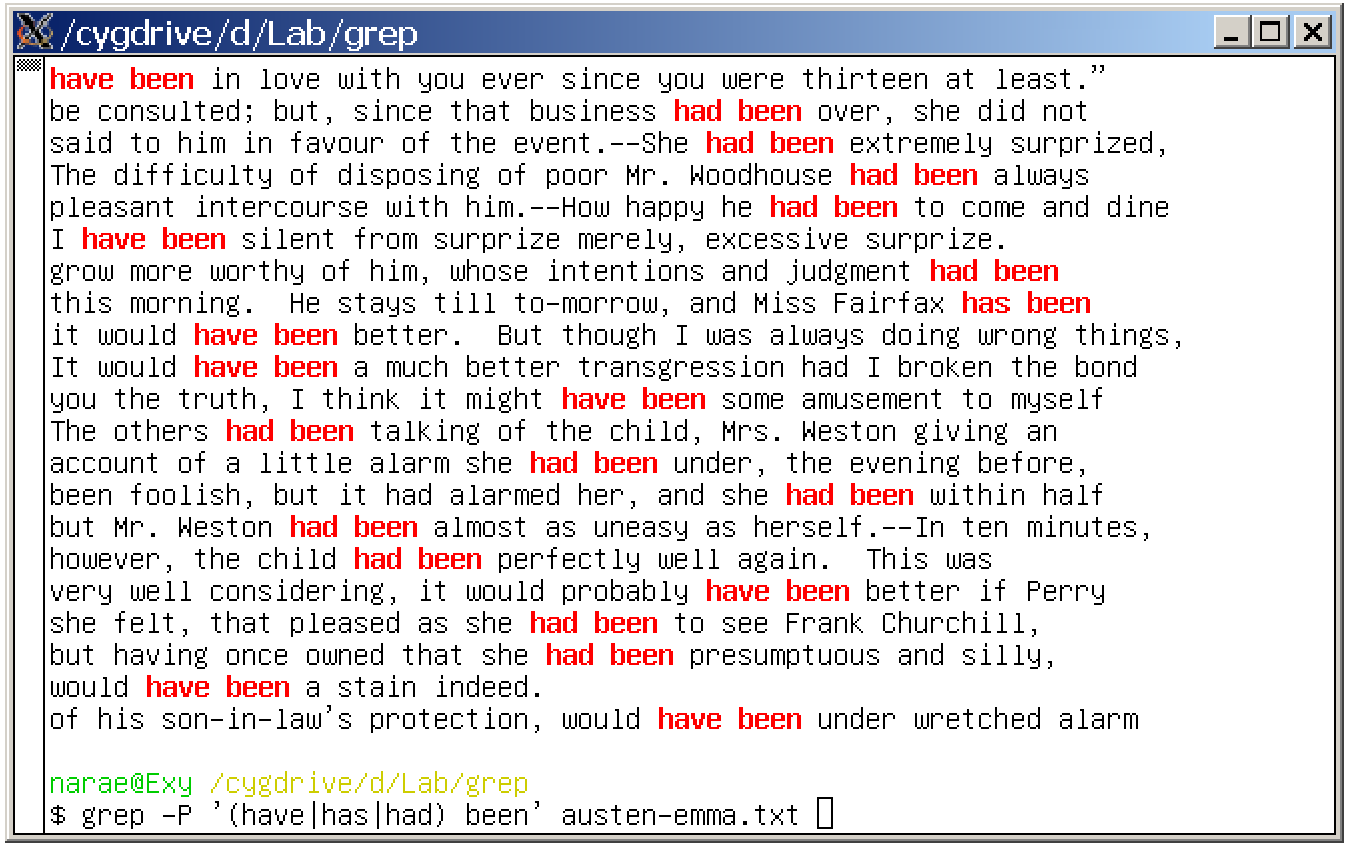


```
and such apparent devotion to Miss W., as it would have been
months ago, Emma, it would not have been taken with such indifference."
"Very bad--though it might have been worse.--Playing a most
and it should have been his first object to prevent her from
to bear that she should have been in such a state of punishment."
She would have been too happy but for poor Harriet; but every
angel only could have been quite without resentment under such a stroke.
"Yes, here I am, my good friend; and here I have been so long,
Had you not been surrounded by other friends, I might have been
openly than might have been strictly correct.--I feel that I should
certainly have been impertinent."
usual composure--"there would have been no danger.  The danger
would have been of my wearying you.  You could not have gratified
have been in love with you ever since you were thirteen at least."
I have been silent from surprize merely, excessive surprize.
it would have been better.  But though I was always doing wrong things,
It would have been a much better transgression had I broken the bond
you the truth, I think it might have been some amusement to myself
very well considering, it would probably have been better if Perry
would have been a stain indeed.
of his son-in-law's protection, would have been under wretched alarm

narae@Exy /cygdrive/d/Lab/grep
$ grep -P 'have been' austen-emma.txt
```

/have been/

- *have been* as a literal string

'have been', 'has been', 'had been'

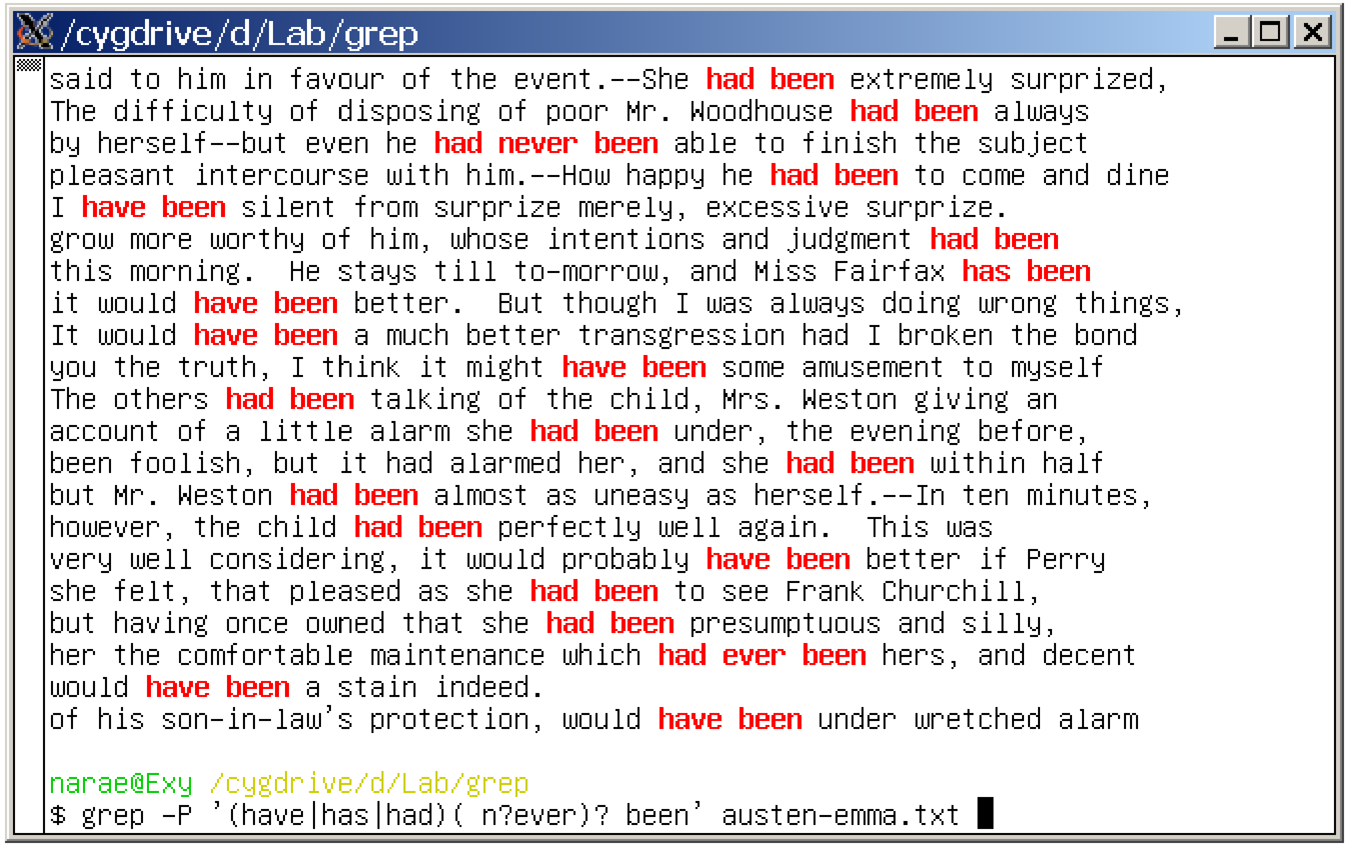


```
have been in love with you ever since you were thirteen at least.”
be consulted; but, since that business had been over, she did not
said to him in favour of the event.--She had been extremely surprized,
The difficulty of disposing of poor Mr. Woodhouse had been always
pleasant intercourse with him.--How happy he had been to come and dine
I have been silent from surprize merely, excessive surprize.
grow more worthy of him, whose intentions and judgment had been
this morning.  He stays till to-morrow, and Miss Fairfax has been
it would have been better.  But though I was always doing wrong things,
It would have been a much better transgression had I broken the bond
you the truth, I think it might have been some amusement to myself
The others had been talking of the child, Mrs. Weston giving an
account of a little alarm she had been under, the evening before,
been foolish, but it had alarmed her, and she had been within half
but Mr. Weston had been almost as uneasy as herself.--In ten minutes,
however, the child had been perfectly well again.  This was
very well considering, it would probably have been better if Perry
she felt, that pleased as she had been to see Frank Churchill,
but having once owned that she had been presumptuous and silly,
would have been a stain indeed.
of his son-in-law’s protection, would have been under wretched alarm

narae@Exy /cygdrive/d/Lab/grep
$ grep -P ’(have|has|had) been’ austen-emma.txt
```

`/(have|has|had) been/`

- Allows inflected forms of *have*

Include *never* or *ever*

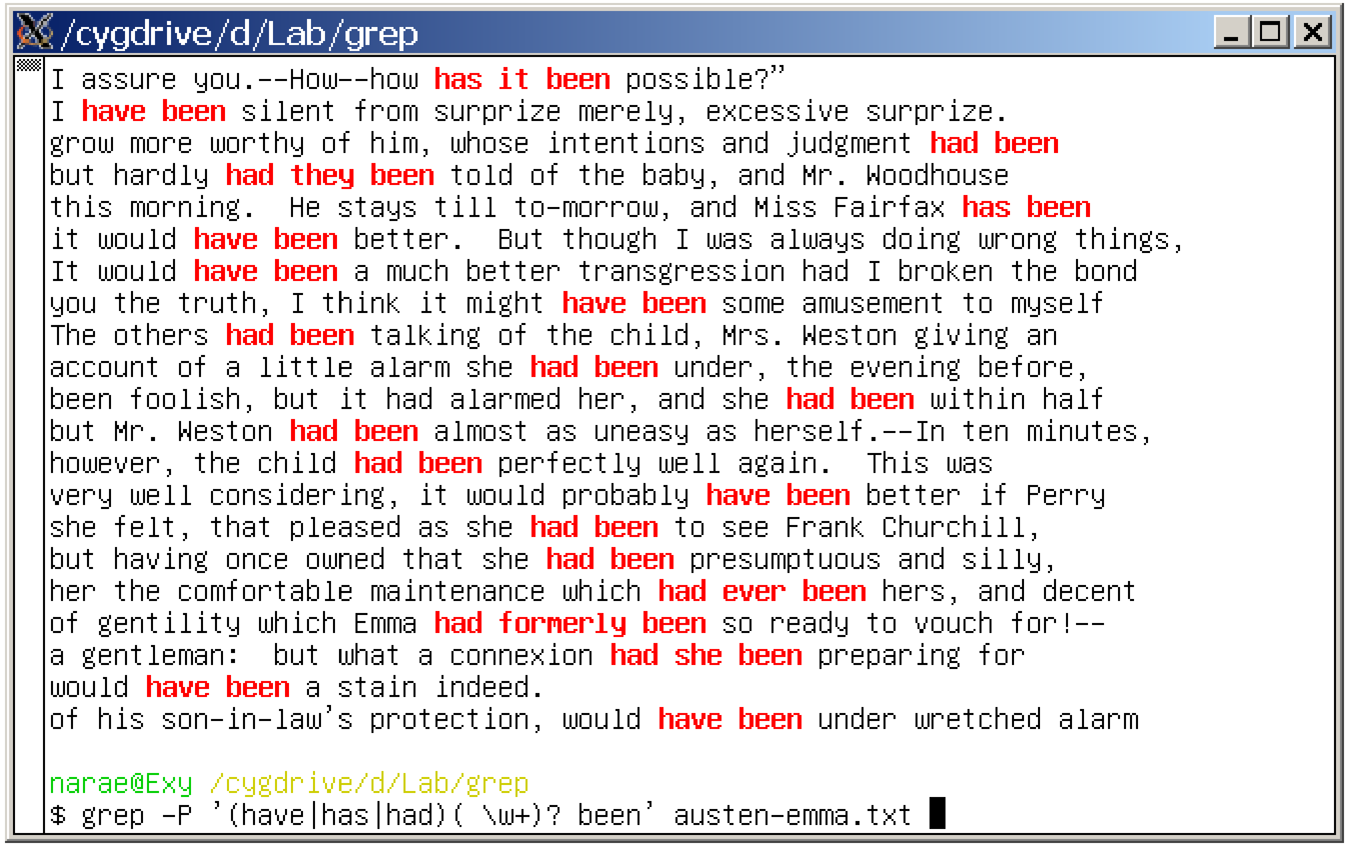


```
said to him in favour of the event.--She had been extremely surprized,
The difficulty of disposing of poor Mr. Woodhouse had been always
by herself--but even he had never been able to finish the subject
pleasant intercourse with him.--How happy he had been to come and dine
I have been silent from surprize merely, excessive surprize.
grow more worthy of him, whose intentions and judgment had been
this morning.  He stays till to-morrow, and Miss Fairfax has been
it would have been better.  But though I was always doing wrong things,
It would have been a much better transgression had I broken the bond
you the truth, I think it might have been some amusement to myself
The others had been talking of the child, Mrs. Weston giving an
account of a little alarm she had been under, the evening before,
been foolish, but it had alarmed her, and she had been within half
but Mr. Weston had been almost as uneasy as herself.--In ten minutes,
however, the child had been perfectly well again.  This was
very well considering, it would probably have been better if Perry
she felt, that pleased as she had been to see Frank Churchill,
but having once owned that she had been presumptuous and silly,
her the comfortable maintenance which had ever been hers, and decent
would have been a stain indeed.
of his son-in-law's protection, would have been under wretched alarm

narae@Exy /cygdrive/d/Lab/grep
$ grep -P '(have|has|had)( n?ever)? been' austen-emma.txt
```

/(have|has|had)(n?ever)? been/

- Allows *never* or *ever* to intervene

Any word in between

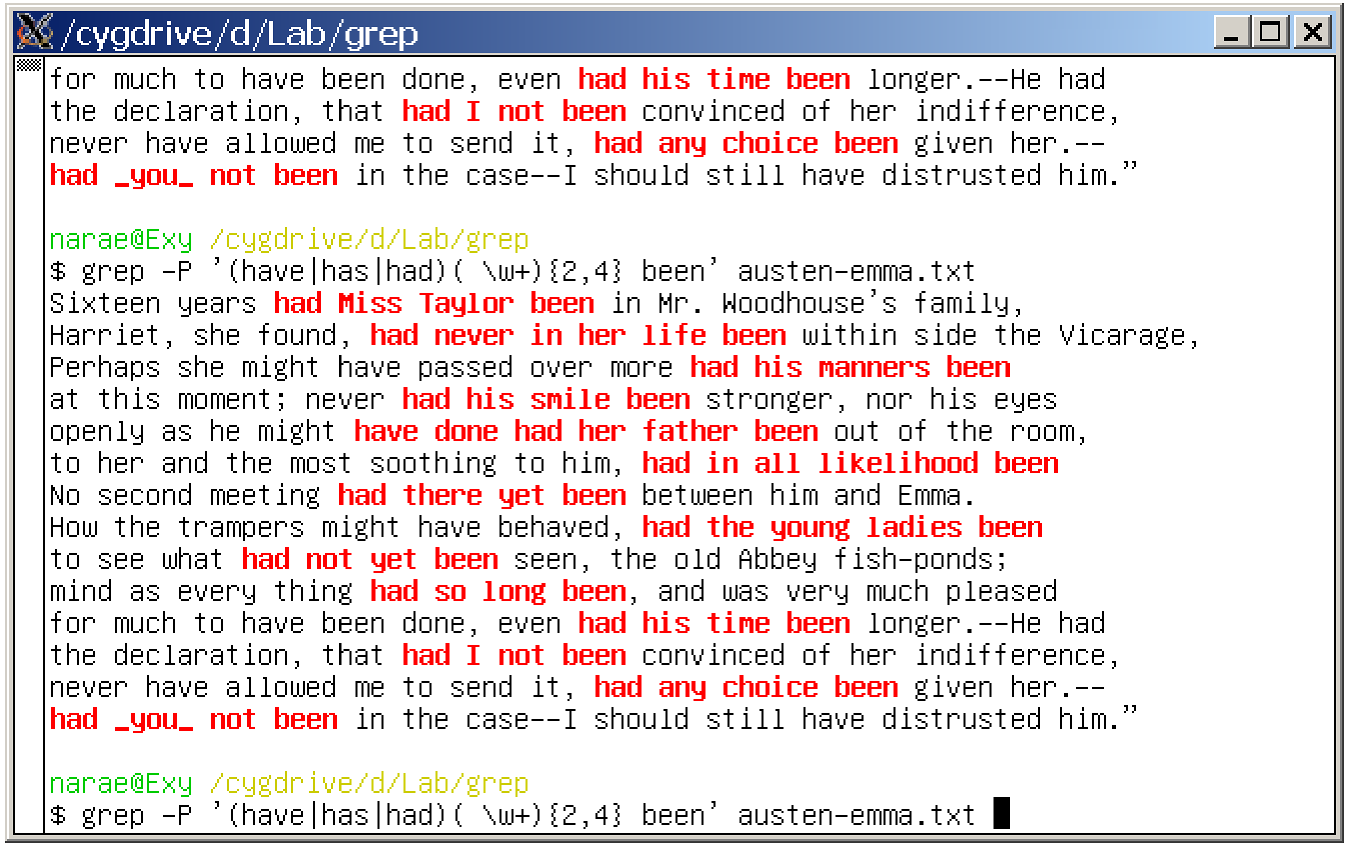


```
I assure you.--How--how has it been possible?"
I have been silent from surprize merely, excessive surprize.
grow more worthy of him, whose intentions and judgment had been
but hardly had they been told of the baby, and Mr. Woodhouse
this morning.  He stays till to-morrow, and Miss Fairfax has been
it would have been better.  But though I was always doing wrong things,
It would have been a much better transgression had I broken the bond
you the truth, I think it might have been some amusement to myself
The others had been talking of the child, Mrs. Weston giving an
account of a little alarm she had been under, the evening before,
been foolish, but it had alarmed her, and she had been within half
but Mr. Weston had been almost as uneasy as herself.--In ten minutes,
however, the child had been perfectly well again.  This was
very well considering, it would probably have been better if Perry
she felt, that pleased as she had been to see Frank Churchill,
but having once owned that she had been presumptuous and silly,
her the comfortable maintenance which had ever been hers, and decent
of gentility which Emma had formerly been so ready to vouch for!--
a gentleman:  but what a connexion had she been preparing for
would have been a stain indeed.
of his son-in-law's protection, would have been under wretched alarm

narae@Exy /cygdrive/d/Lab/grep
$ grep -P '(have|has|had)( \w+)? been' austen-emma.txt
```

`/(have|has|had)( \w+)? been/`

- Allows any single word (along with a space) to intervene

More intervening words

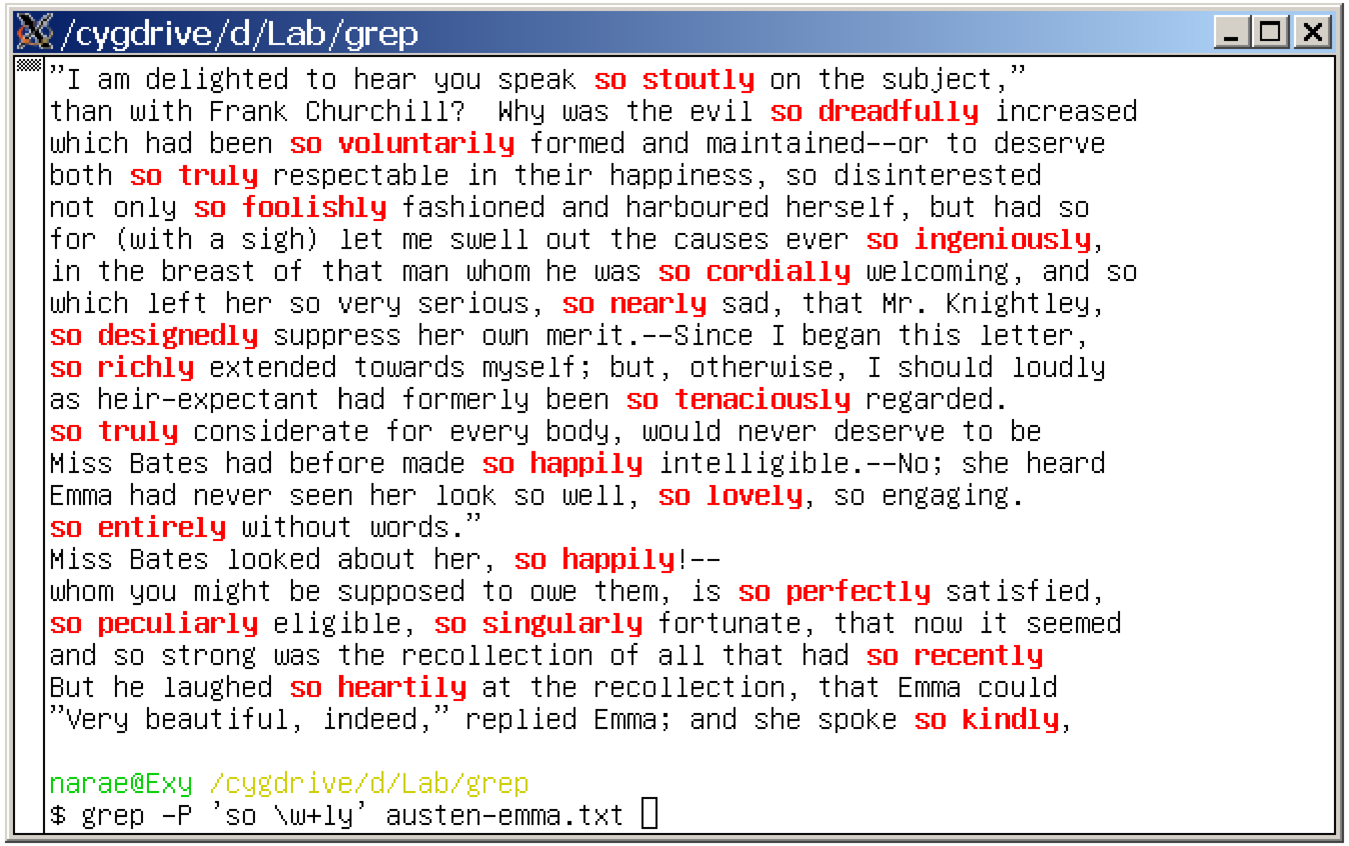


```
for much to have been done, even had his time been longer.--He had
the declaration, that had I not been convinced of her indifference,
never have allowed me to send it, had any choice been given her.--
had _you_ not been in the case--I should still have distrusted him.”

narae@Exy /cygdrive/d/Lab/grep
$ grep -P '(have|has|had)( \w+){2,4} been' austen-emma.txt
Sixteen years had Miss Taylor been in Mr. Woodhouse's family,
Harriet, she found, had never in her life been within side the Vicarage,
Perhaps she might have passed over more had his manners been
at this moment; never had his smile been stronger, nor his eyes
openly as he might have done had her father been out of the room,
to her and the most soothing to him, had in all likelihood been
No second meeting had there yet been between him and Emma.
How the trampers might have behaved, had the young ladies been
to see what had not yet been seen, the old Abbey fish-ponds;
mind as every thing had so long been, and was very much pleased
for much to have been done, even had his time been longer.--He had
the declaration, that had I not been convinced of her indifference,
never have allowed me to send it, had any choice been given her.--
had _you_ not been in the case--I should still have distrusted him.”

narae@Exy /cygdrive/d/Lab/grep
$ grep -P '(have|has|had)( \w+){2,4} been' austen-emma.txt
```

/(have|has|had)(\w+){2,4} been/

- With 2-4 intervening words (along with a space)

That is so …ly



```
"I am delighted to hear you speak so stoutly on the subject,"
than with Frank Churchill?  Why was the evil so dreadfully increased
which had been so voluntarily formed and maintained--or to deserve
both so truly respectable in their happiness, so disinterested
not only so foolishly fashioned and harboured herself, but had so
for (with a sigh) let me swell out the causes ever so ingeniously,
in the breast of that man whom he was so cordially welcoming, and so
which left her so very serious, so nearly sad, that Mr. Knightley,
so designedly suppress her own merit.--Since I began this letter,
so richly extended towards myself; but, otherwise, I should loudly
as heir-expectant had formerly been so tenaciously regarded.
so truly considerate for every body, would never deserve to be
Miss Bates had before made so happily intelligible.--No; she heard
Emma had never seen her look so well, so lovely, so engaging.
so entirely without words."
Miss Bates looked about her, so happily!--
whom you might be supposed to owe them, is so perfectly satisfied,
so peculiarly eligible, so singularly fortunate, that now it seemed
and so strong was the recollection of all that had so recently
But he laughed so heartily at the recollection, that Emma could
"Very beautiful, indeed," replied Emma; and she spoke so kindly,

narae@Exy /cygdrive/d/Lab/grep
$ grep -P 'so \w+ly' austen-emma.txt
```

`/so \w+ly/`

- *so* followed by a word ending in *-ly*

Learn more regex!

- Regular expressions tutorials:
  - https://www.regular-expressions.info/tutorial.html

- Regular expressions puzzles:
  - https://regexcrossword.com/

- Handy online tester:
  - https://regex101.com/