

Lecture 7

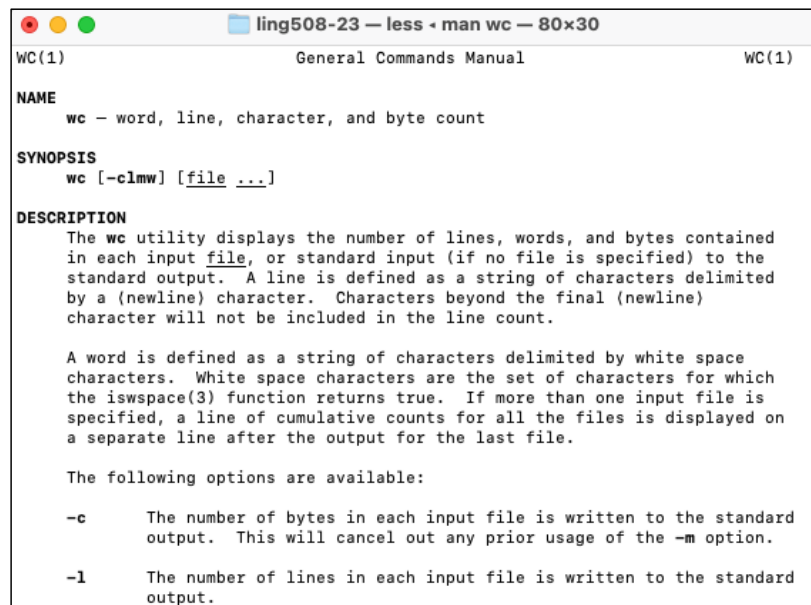
408/508 Computational Techniques for Linguists

Today's Topics

- Homework 3 review
 - *Step-by-step Bash shell exercises*
- Other things we can 'pipe' (|) into our workflow:
 - tail
 - awk
 - termgraph
- a note on file permissions

Homework 3: Exercise 1 Review

- Relevant bit:
 - The **wc** utility displays the number of **lines**, **words**, and **bytes**



```
ling508-23 — less - man wc — 80x30
WC(1)                                General Commands Manual                                WC(1)

NAME
    wc - word, line, character, and byte count

SYNOPSIS
    wc [-clmw] [file ...]

DESCRIPTION
    The wc utility displays the number of lines, words, and bytes contained
    in each input file, or standard input (if no file is specified) to the
    standard output. A line is defined as a string of characters delimited
    by a {newline} character. Characters beyond the final {newline}
    character will not be included in the line count.

    A word is defined as a string of characters delimited by white space
    characters. White space characters are the set of characters for which
    the iswspace(3) function returns true. If more than one input file is
    specified, a line of cumulative counts for all the files is displayed on
    a separate line after the output for the last file.

    The following options are available:

    -c      The number of bytes in each input file is written to the standard
            output. This will cancel out any prior usage of the -m option.

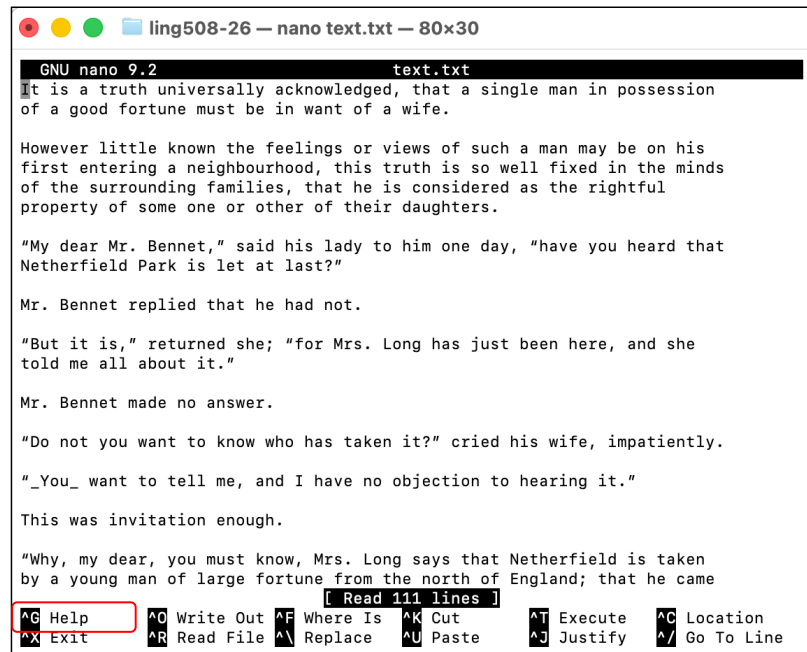
    -l      The number of lines in each input file is written to the standard
            output.
```

Homework 3: Exercise 1

```
ling508-26$ file text.txt
text.txt: Unicode text, UTF-8 text
ling508-26$ ls -l text.txt
-rw-r--r--@ 1 sandiway  staff  4616 Sep  8 13:00 text.txt
ling508-26$ wc text.txt
    111     847    4616 text.txt
ling508-26$ wc -w text.txt
    847 text.txt
```

-w The number of words in each input file is written to the standard output.

Homework 3: Exercise 1



```
GNU nano 9.2 text.txt
It is a truth universally acknowledged, that a single man in possession
of a good fortune must be in want of a wife.

However little known the feelings or views of such a man may be on his
first entering a neighbourhood, this truth is so well fixed in the minds
of the surrounding families, that he is considered as the rightful
property of some one or other of their daughters.

"My dear Mr. Bennet," said his lady to him one day, "have you heard that
Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she
told me all about it."

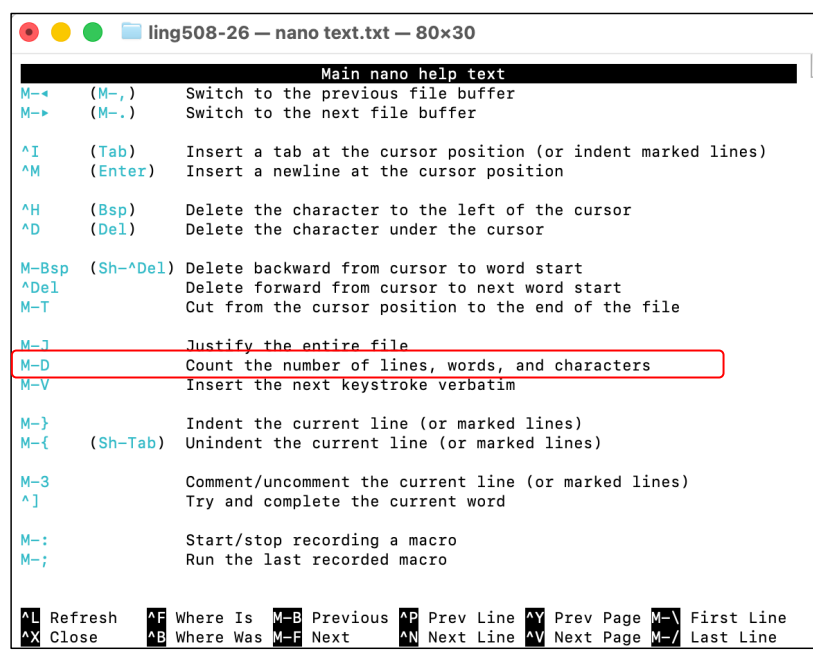
Mr. Bennet made no answer.

"Do not you want to know who has taken it?" cried his wife, impatiently.

"_You_ want to tell me, and I have no objection to hearing it."

This was invitation enough.

"Why, my dear, you must know, Mrs. Long says that Netherfield is taken
by a young man of large fortune from the north of England; that he came
[ Read 111 lines ]
^G Help      ^O Write Out  ^F Where Is   ^K Cut        ^T Execute    ^C Location
^X Exit      ^R Read File  ^\ Replace    ^U Paste      ^J Justify    ^_ Go To Line
```



```
ling508-26 — nano text.txt — 80x30
Main nano help text
M-^ (M-,)  Switch to the previous file buffer
M-> (M-.)  Switch to the next file buffer

^I (Tab)   Insert a tab at the cursor position (or indent marked lines)
^M (Enter) Insert a newline at the cursor position

^H (Bsp)   Delete the character to the left of the cursor
^D (Del)   Delete the character under the cursor

M-Bsp (Sh-^Del) Delete backward from cursor to word start
^Del     Delete forward from cursor to next word start
M-T      Cut from the cursor position to the end of the file

M-~      Justify the entire file
M-D      Count the number of lines, words, and characters
M-V      Insert the next keystroke verbatim

M-}      Indent the current line (or marked lines)
M-{ (Sh-Tab) Unindent the current line (or marked lines)

M-3      Comment/uncomment the current line (or marked lines)
^]        Try and complete the current word

M-:      Start/stop recording a macro
M-;      Run the last recorded macro

^L Refresh  ^F Where Is  M-B Previous ^P Prev Line ^Y Prev Page M-_ First Line
^X Close    ^B Where Was M-F Next   ^N Next Line ^V Next Page M-/ Last Line
```

Homework 3: Exercise 1

```
$ wc text.txt
111 847 4616 text.txt
```

UW PICO 5.09 File: text.txt

GNU nano 9.2 text.txt

It is a truth universally acknowledged, of a good fortune must be in want of a w

However little known the feelings or view first entering a neighbourhood, this truth of the surrounding families, that he is property of some one or other of their daughters.

"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she told me all about it."

Mr. Bennet made no answer.

"Do not you want to know who has taken it?" cried his wife, impatiently.

"_You_ want to tell me, and I have no objection to hearing it."

This was invitation enough.

"Why, my dear, you must know, Mrs. Long says that Netherfield is taken

"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she told me all about it."

Mr. Bennet made no answer.

"Do not you want to know who has taken it?" cried his wife, impatiently.

"_You_ want to tell me, and I have no objection to hearing it."

This was invitation enough.

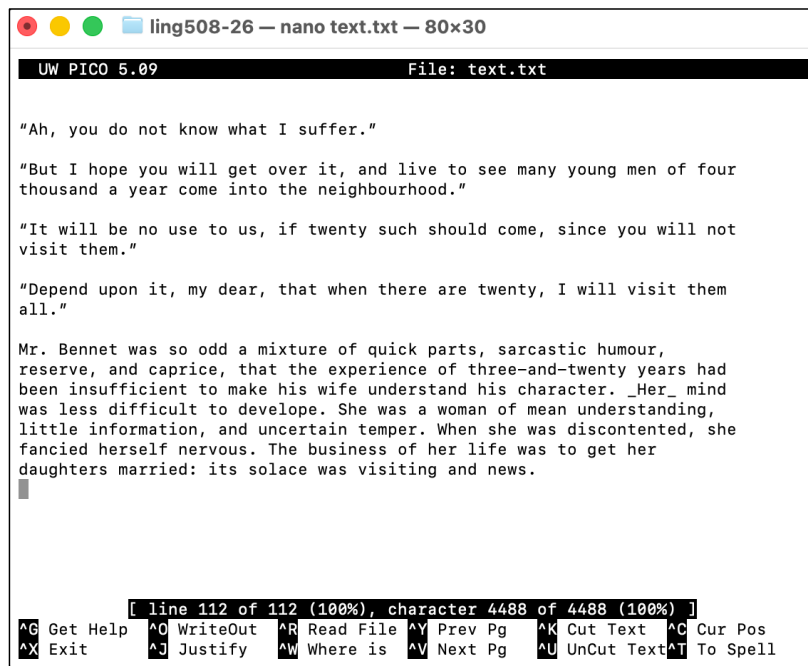
"Why, my dear, you must know, Mrs. Long says that Netherfield is taken by a young man of large fortune from the north of England; that he came

[111 lines, 853 words, 4488 characters]

^G Get Help ^O WriteOut ^R Read File ^Y Prev Pg ^K Cut Text ^C Cur Pos
^X Exit ^J Justify ^W Where is ^V Next Pg ^U UnCut Text ^T To Spell

^G Help ^O Write Out ^F Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^_ Go To Line

Homework 3: Exercise 1



```
ling508-26 — nano text.txt — 80x30
UW PICO 5.09 File: text.txt

"Ah, you do not know what I suffer."

"But I hope you will get over it, and live to see many young men of four
thousand a year come into the neighbourhood."

"It will be no use to us, if twenty such should come, since you will not
visit them."

"Depend upon it, my dear, that when there are twenty, I will visit them
all."

Mr. Bennet was so odd a mixture of quick parts, sarcastic humour,
reserve, and caprice, that the experience of three-and-twenty years had
been insufficient to make his wife understand his character. _Her_ mind
was less difficult to develope. She was a woman of mean understanding,
little information, and uncertain temper. When she was discontented, she
fancied herself nervous. The business of her life was to get her
daughters married: its solace was visiting and news.

[ line 112 of 112 (100%), character 4488 of 4488 (100%) ]
^G Get Help ^O WriteOut ^R Read File ^Y Prev Pg ^K Cut Text ^C Cur Pos
^X Exit ^J Justify ^W Where is ^V Next Pg ^U UnCut Text ^T To Spell
```

```
$ wc text.txt
    111      847
 4616 text.txt
```

Homework 3: Exercise 2

```
ling508-26 — -bash — 80x30

You are over scrupulous surely I dare say Mr Bingley will be very
glad to see you and I will send a few lines by you to assure him of my
hearty consent to his marrying whichever he chooses of the girlstthough
I must throw in a good word for my little Lizzy

I desire you will do no such thing Lizzy is not a bit better than the
others and I am sure she is not half so handsome as Jane nor half so
goodhumoured as Lydia But you are always giving her the preference

They have none of them much to recommend them replied he they are
all silly and ignorant like other girls but Lizzy has something more of
quickness than her sisters

Mr Bennet how can you abuse your own children in such a way You take
delight in vexing me You have no compassion on my poor nerves

You mistake me my dear I have a high respect for your nerves They
are my old friends I have heard you mention them with consideration
these twenty years at least

Ah you do not know what I suffer

But I hope you will get over it and live to see many young men of four
thousand a year come into the neighbourhood

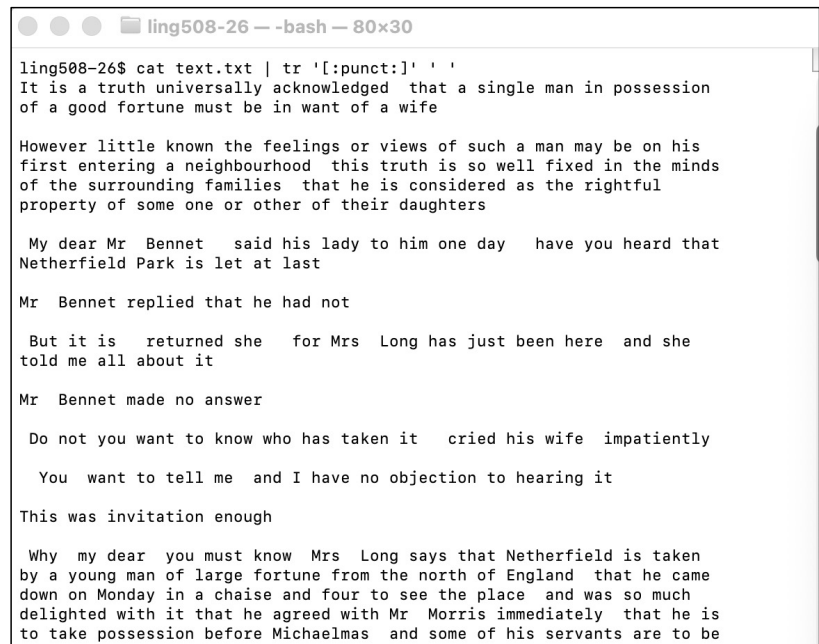
It will be no use to us if twenty such should come since you will not
visit them

Depend upon it my dear that when there are twenty I will visit them
all
```

- tr command:
 - `tr -d '[:punct:]'`
 - `-d` deletes punctuations, may combine words inadvertently

Homework 3: Exercise 2

- tr command:
 - tr '[:punct:]' ' ' replaces punctuation with a space character



```
ling508-26 — -bash — 80x30
ling508-26$ cat text.txt | tr '[:punct:]' ' '
It is a truth universally acknowledged that a single man in possession
of a good fortune must be in want of a wife

However little known the feelings or views of such a man may be on his
first entering a neighbourhood this truth is so well fixed in the minds
of the surrounding families that he is considered as the rightful
property of some one or other of their daughters

My dear Mr Bennet said his lady to him one day have you heard that
Netherfield Park is let at last

Mr Bennet replied that he had not

But it is returned she for Mrs Long has just been here and she
told me all about it

Mr Bennet made no answer

Do not you want to know who has taken it cried his wife impatiently

You want to tell me and I have no objection to hearing it

This was invitation enough

Why my dear you must know Mrs Long says that Netherfield is taken
by a young man of large fortune from the north of England that he came
down on Monday in a chaise and four to see the place and was so much
delighted with it that he agreed with Mr Morris immediately that he is
to take possession before Michaelmas and some of his servants are to be
```

Homework 3: Exercise 2

```
$ cat text.txt | tr '[:punct:]' ' ' | tr ' ' '\n' | sort | uniq -c | sort -rn
```

247	11 in	8 as	5 or	4 man
29 of	11 his	8 are	5 one	4 how
24 you	11 he	7 You	5 me	4 him
22 to	11 for	7 no	5 know	4 go
20 a	11 be	7 must	5 her	4 girls
17 the	10 my	7 have	5 has	4 do
17 I	10 Mr	6 she	4 with	4 daughters
17 and	9 will	6 But	4 wife	4 can
15 that	9 not	6 Bennet	4 twenty	4 by
12 is	8 was	5 visit	4 single	4 Bingley
11 them	8 so	5 such	4 on	3 your
11 it	8 dear	5 see	4 may	...

Homework 3: Exercise 2

```
508 — less ◀ man uniq — 80×30

NAME
    uniq — report or filter out repeated lines in a file

SYNOPSIS
    uniq [-cdiu] [-D[septype]] [-f num] [-s chars] [input_file [output_file]]

DESCRIPTION
    The uniq utility reads the specified input_file comparing adjacent lines,
    and writes a copy of each unique input line to the output_file. If
    input_file is a single dash ('-') or absent, the standard input is read.
    If output_file is absent, standard output is used for output. The second
    and succeeding copies of identical adjacent input lines are not written.
    Repeated lines in the input will not be detected if they are not
    adjacent, so it may be necessary to sort the files first.

    The following options are available:

    -c, --count
        Precede each output line with the count of the number of times
        the line occurred in the input, followed by a single space.
```

- See adjacency requirement for **repeating lines!**

Homework 3: Exercise 2

```
ling508-26 — -bash — 80x30
[ling508-26$ cat text.txt | tr '[:punct:]' ' ' | tr ' '\n' | sort | uniq -c | pr -t5
r -t5
247
20 a          1 disconte  1 In          1 new         1 somethin
1 A          4 do          1 indeed     1 news        1 soon
1 about      1 Do          1 Indeed     1 next        1 still
1 abuse      1 down        1 informat   7 no          5 such
1 account    1 end         1 insuffic   1 none        1 suffer
1 acknowle   1 engage     2 into       1 Nonsense    2 sure
1 affect     1 England    1 invitati   1 nor         1 surely
1 agreed     1 enough     12 is        1 north       1 surround
1 Ah         1 entering   2 Is         9 not         2 take
3 all        1 establis   11 it        1 now         2 taken
1 always     1 experien   3 It         1 objectio    1 talk
2 am         1 extraord   1 its        1 occasio    1 tell
1 an         1 fall       1 Jane       1 odd         1 temper
17 and       1 families   1 just       29 of         3 than
1 answer     1 fancied    5 know       1 often       15 that
1 any        1 feelings   1 known      1 Oh          17 the
1 anything   1 few        1 lady       1 old         1 The
8 are        1 fine       1 Lady       4 on          1 their
8 as         1 first      2 large      5 one         11 them
2 assure     2 five       1 last       1 Only        1 themself
2 at         1 fixed      1 least      5 or          1 there
11 be        1 flatter    1 less       2 other       1 therefor
3 beauty     11 for       1 let        1 others      1 these
2 been       3 fortune    1 life       1 ought       2 they
1 before     3 four       2 like       1 our         2 They
6 Bennet     1 friends    1 likely     3 over        2 thing
1 best       1 from       1 lines      2 own         2 think
            1 general    3 little     1 Park        2 thinking
```

A step beyond Homework 3

Let's graph our homework result!

- There's something called **termgraph** (written in Python) but you can use it on the command line
- Assume you have python3 already installed
- Check whether it's already installed
 - `which termgraph`
 - `/Users/sandiway/opt/anaconda3/bin/termgraph`
- if not:
 - `pip3 install termgraph`

termgraph install

```
$ which termgraph      (no response means can't find the command)
```

```
$ pip3 install termgraph
```

```
Collecting termgraph
```

```
  Downloading termgraph-0.7.6-py3-none-any.whl.metadata (8.1 kB)
```

```
Collecting colorama>=0.4.6 (from termgraph)
```

```
  Using cached colorama-0.4.6-py2.py3-none-any.whl.metadata (17 kB)
```

```
Downloading termgraph-0.7.6-py3-none-any.whl (18 kB)
```

```
Using cached colorama-0.4.6-py2.py3-none-any.whl (25 kB)
```

```
Installing collected packages: colorama, termgraph
```

```
Successfully installed colorama-0.4.6 termgraph-0.7.6
```

```
$ which termgraph
```

```
1. /Library/Frameworks/Python.framework/Versions/3.14/bin/termgraph
```

```
2. /opt/miniconda3/bin/termgraph
```

termgraph install

- It may place the executable in a directory that's not in your PATH.
- If so:
 - export PATH=**/home/yourname/.login/bin**:\$PATH
 - will prepend **/home/yourname/.login/bin** to your PATH
- To make the change permanent, you can add this line to your startup file, either `.bashrc` or `.bash_profile` or `.zprofile` (for zsh users) in your home directory (*depending on which one exists*)
 - `cd` *(goto home)*
 - `nano your_starup_file` *(save change and exit)*

termgraph

- Google termgraph
 - <https://github.com/mkaz/termgraph>

Termgraph

A command-line tool that draws basic graphs in the terminal, written in Python.

Graph types supported:

- Bar Graphs
- Color charts
- Multi-variable
- Stacked charts
- Histograms
- Horizontal or Vertical
- Emoji!

```
termgraph data/ex1.dat
```

Reading data from data/ex1.dat

A horizontal bar chart with a dark background. The y-axis lists years from 2007 to 2014. Each year has a corresponding light gray horizontal bar. To the right of each bar is its numerical value. The bars vary significantly in length, with 2011 having the longest bar and 2009 having the shortest.

Year	Value
2007:	183.32
2008:	231.23
2009:	16.43
2010:	50.21
2011:	508.97
2012:	212.05
2014:	1.00

An example using emoji as custom tick:

```
termgraph data/ex1.dat --custom-tick "人" --width 20 --title "Running Data"
```

Running Data

Year	Count (人)	Value
2007:	8	183.32
2008:	10	231.23
2009:	1	16.43
2010:	2	50.21
2011:	18	508.97
2012:	8	212.05
2014:	1	1.00

termgraph

```
cat text.txt | tr '[:punct:]' ' ' | tr ' ' '\n' | sort | uniq
-c | sort -rn | tail -n +2 | awk '{print $2, $1}' | termgraph
```

Word	Count
of	29.00
you	24.00
to	22.00
a	20.00
the	17.00
I	17.00
and	17.00
that	15.00
is	12.00
them	11.00
it	11.00
in	11.00
his	11.00
he	11.00
for	11.00
be	11.00
my	10.00
Mr	10.00
will	9.00

ASCII graphics!

tail -n +2

NAME

tail – display the last part of a file

SYNOPSIS

tail [-F | -f | -r] [-q] [-b number | -c number | -n number] [file ...]

DESCRIPTION

The **tail** utility displays the contents of file or, by default, its standard input, to the standard output.

Numbers having a leading plus ('+') sign are relative to the beginning of the input, for example, “-c +2” starts the display at the second byte of the input.

Numbers having a leading minus ('-') sign or no explicit sign are relative to the end of the input, for example, “-n 2” displays the last two lines of the input.

-n number, --lines=number

The location is number lines.

tail -n +2

```
cat text.txt | tr '[:punct:]' ' ' | tr ' ' '\n' | sort | uniq -c | sort -rn
```

221

56 the

26 and

22 of

20 a

19 it

18 was

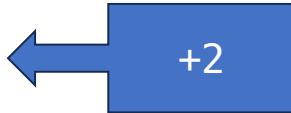
15 she

15 in

14 her

12 to

11 one



awk '{print \$2, \$1}'

NAME

awk – pattern-directed scanning and processing language

SYNOPSIS

```
awk [ -F fs ] [ -v var=value ] [ 'prog' | -f progfile ] [ file ... ]
```

DESCRIPTION

Awk scans each input file for lines that match any of a set of patterns specified literally in prog or in one or more files specified as **-f** progfile. With each pattern there can be an associated action that will be performed when a line of a file matches the pattern.

A pattern-action statement has the form:

```
pattern { action }
```

A missing { action } means print the line; **a missing pattern always matches**.

The **print** statement prints its arguments on the standard output

```
{ print $2, $1 }
```

Print first two fields in opposite order.

tail -n +2

```
cat text.txt | tr '[:punct:]' ' ' | tr ' ' '\n' | sort | uniq -c | sort -rn | awk '{print $2, $1}' | termgraph
```

221

56 the
26 and
22 of
20 a
19 it
18 was
15 she
15 in
14 her
12 to
11 one

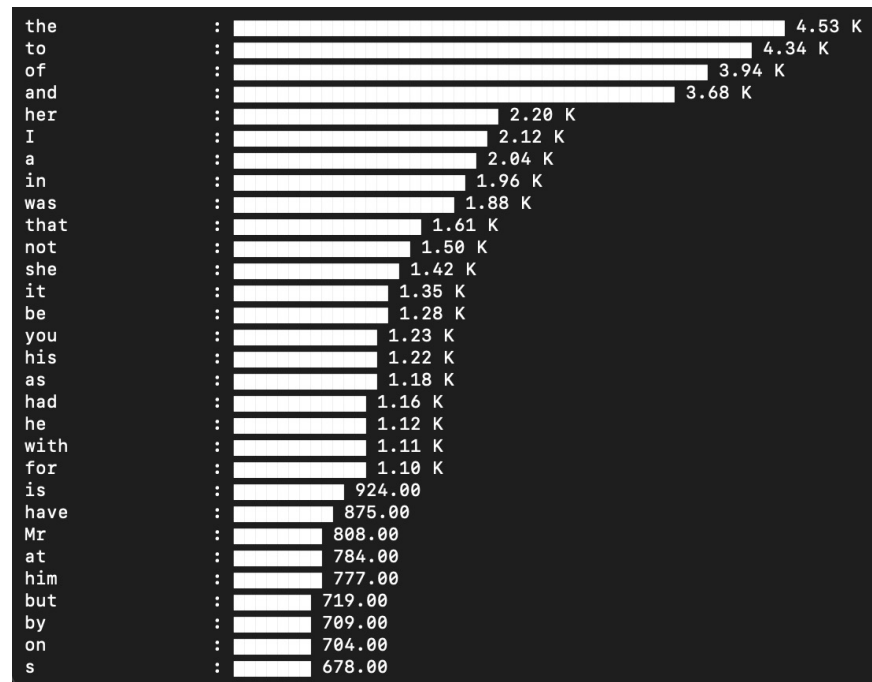
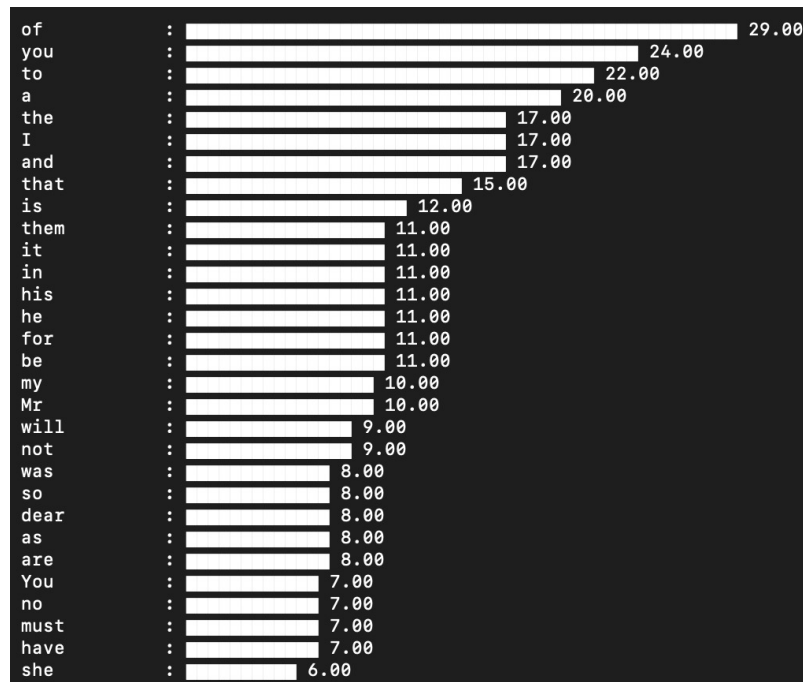
+2

the 56
and 26
of 22
a 20
it 19
was 18
she 15
in 15
her 14
to 12
one 11

awk '{print \$2, \$1}'

Pride and Prejudice

truncated vs. full corpus



The spirit of Unix (Linux)

The Power of the UNIX Command-Line

By - August 10, 2010

👁 70



One of the most novel and differentiating features of a UNIX system is its command line. With just a few keystrokes, including a bit of “glue”, you can use the command line to combine the finite set of UNIX utilities into innumerable, impromptu data transforms. These articles will teach you the basics of the UNIX shell and discover how you can use the command line:

- (1) [Command the power of the command line](#)
- (2) [Do everything right from the command line](#)

- wc
- sort
- uniq
- tr
- tail
- cat
- echo
- pr

Running shell scripts

Chmod 644 ← number (binary)					
Chmod 644 (<i>chmod a+rwX,u-X,g-wX,o-wX</i>) sets permissions so that, (U)ser / owner can read, can write and can't execute. (G)roup can read, can't write and can't execute. (O)thers can read, can't write and can't execute.					
	Owner Rights (u)		Group Rights (g)		Others Rights (o)
Read (4)	☒	1	☒	1	☒ 1
Write (2)	☒	1	☐	0	☐ 0
Execute (1)	☐	0	☐	0	☐ 0

Command:

- `chmod permissions filename`
- *permissions*: e.g. u+x (*user add execute*) or a **number**

Recall everything is binary:

- 110 = 6, 100 = 4
- 644 = 110100100 (*3 groups of binary*)