CS5354
UNIX TOOL PROGRAMMING

grep

- A. Sawarkar

Regular Expressions
grep and egrep

Previously

- Basic UNIX Commands
  - Files: rm, cp, mv, ls, ln
  - Processes: ps, kill
- Unix Filters
  - **cat**, **head**, **tail**, tee, **wc**
  - **cut**, paste
  - **find**
  - **sort**, **uniq**
  - tr

Today

- Regular Expressions
  - – Allow you to search for text in files
  - – **grep** command
- Stream *manipulation*:
  - – **sed**

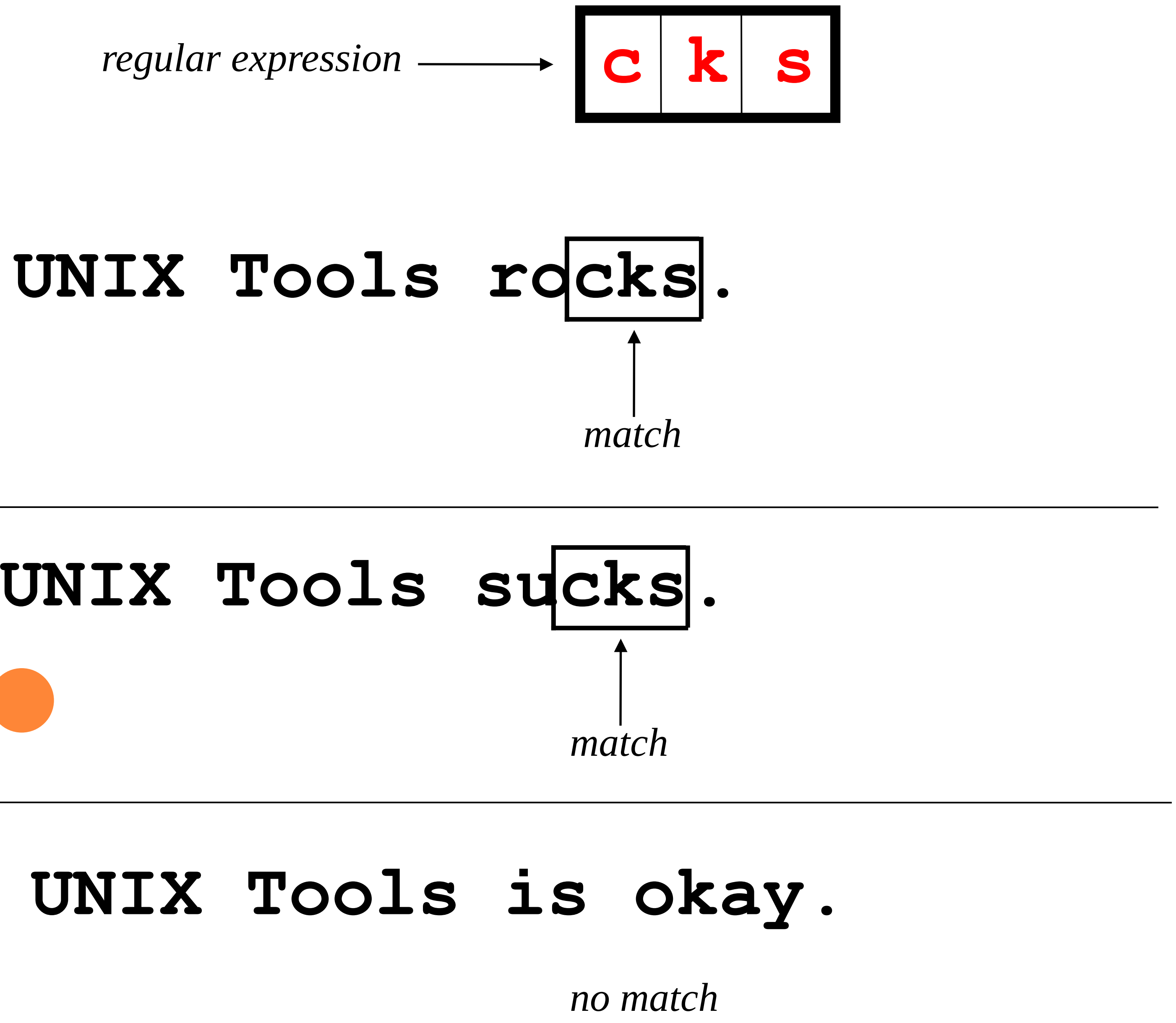
Regular Expressions

What Is a Regular Expression?

- A regular expression (*regex*) describes a set of possible input strings.
- *Regular expressions* is a fundamental concept in Computer Science called *finite automata* theory
- *Regular expressions* of Unix
  - **vi**, **ed**, **sed**, and **emacs**
  - **awk**, **tcl**, **perl** and **Python**
  - **grep**, **egrep**, **fgrep**
  - **compilers**

Regular Expressions

- The simplest regular expressions are a string of literal characters to match.
- The string ***matches*** the regular expression if it contains the substring.



regular expression → c k s

UNIX Tools rocks.

match

UNIX Tools sucks.

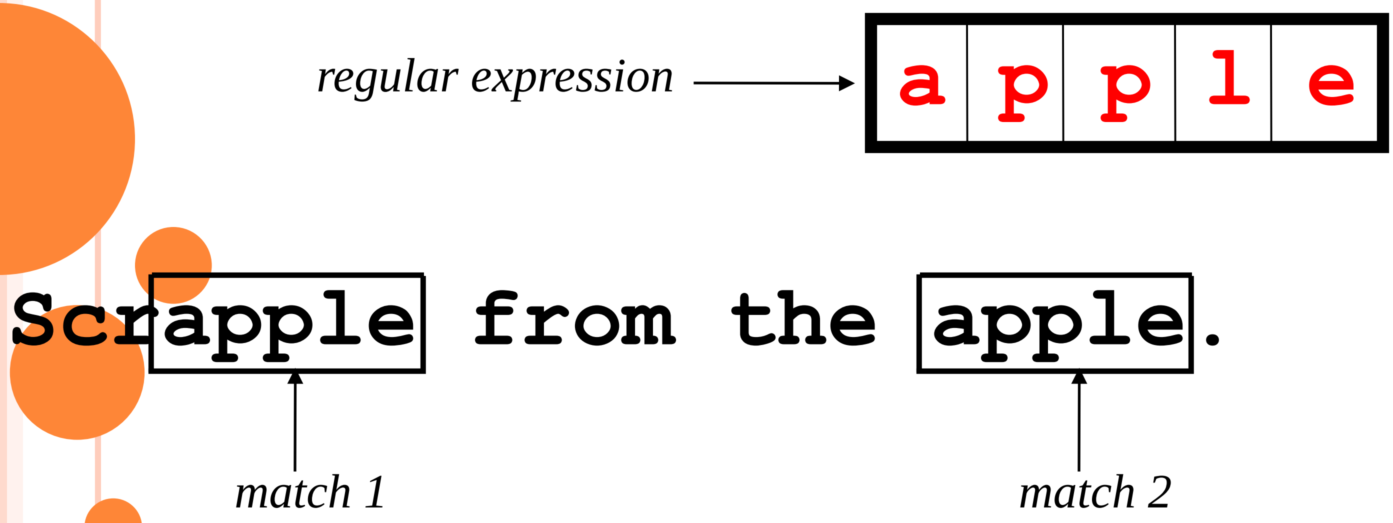
match

UNIX Tools is okay.

no match

Regular Expressions

- A regular expression can match a string in more than one place.



regular expression

a p p l e

Scrapple from the apple.

match 1

match 2

grep

- **`grep`** comes from the **`ed`** (Unix text editor) search command "**g**lobal **r**egular **e**xpression **p**rint" or **`g`**/*re*/**`p`**
- This was such a useful command that it was written as a standalone utility
- There are two other variants, ***egrep*** and ***fgrep*** that comprise the *grep* family
- *grep* is the answer to the moments where you know you want the file that contains a specific phrase but you can't remember its name

Family Differences

- **grep -** uses regular expressions for pattern matching
- **fgrep -** file grep, does not use regular expressions, only matches fixed strings but can get search strings from a file
- **egrep -** extended grep, uses a more powerful set of regular expressions but does not support backreferencing, generally the fastest member of the grep family
- **agrep** – approximate grep; not standard

Grep Syntax

grep : Searching for a pattern

-scan its input for pattern and display lines containing the pattern.

Syntax

```
grep [options] pattern filename(s)
```

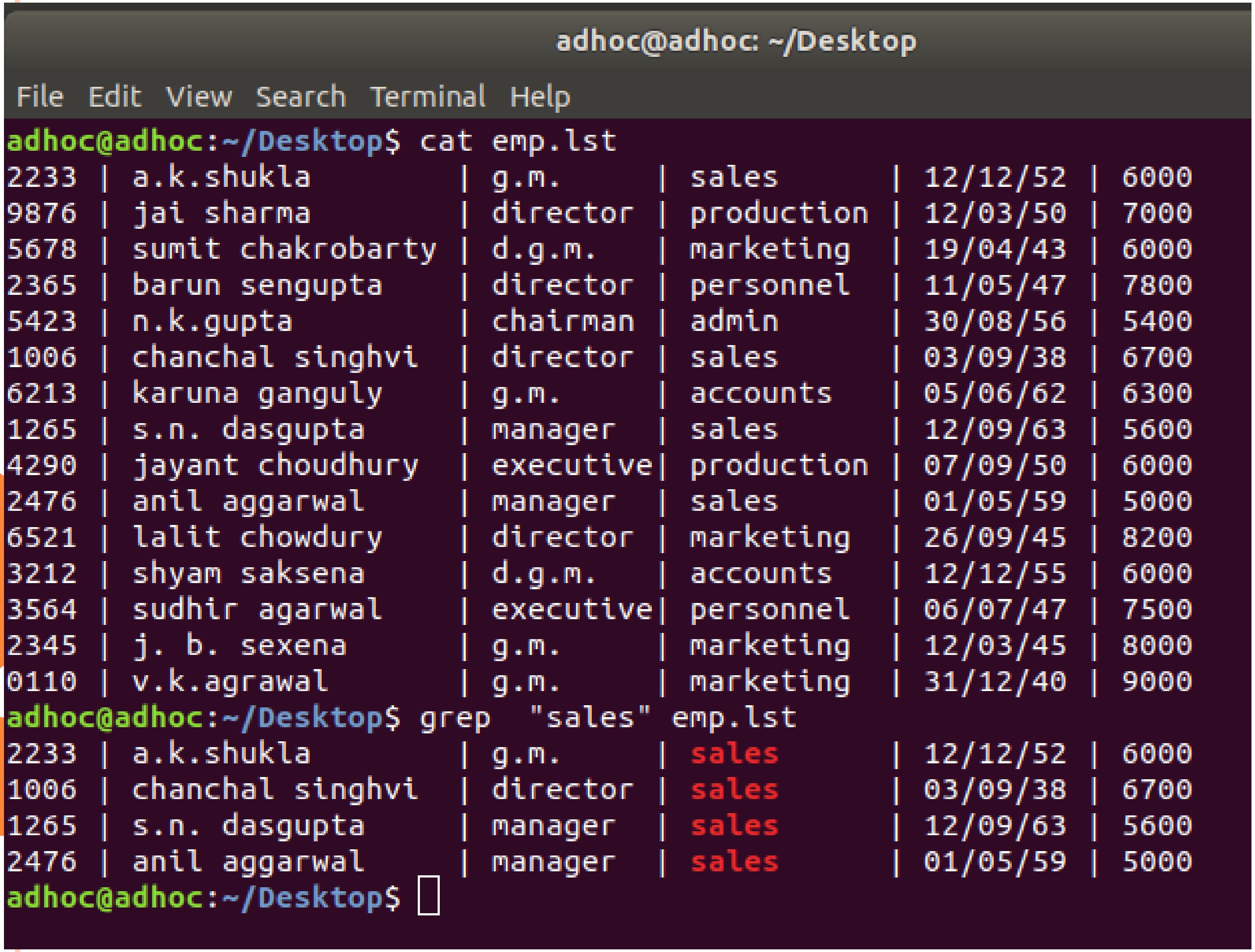
grep [-hilnv] [-e expression] [filename]

e.g

```
$ grep “sales” emp.lst
```

[display lines contain “sales” from file emp.lst.

$ grep “sales” emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep  "sales" emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
adhoc@adhoc:~/Desktop$ 
```

Grep with multiple file

```
$grep “director” emp1.lst  emp2.lst
```

[Quoting (‘ ‘) is essential when the pattern contains multiple word .

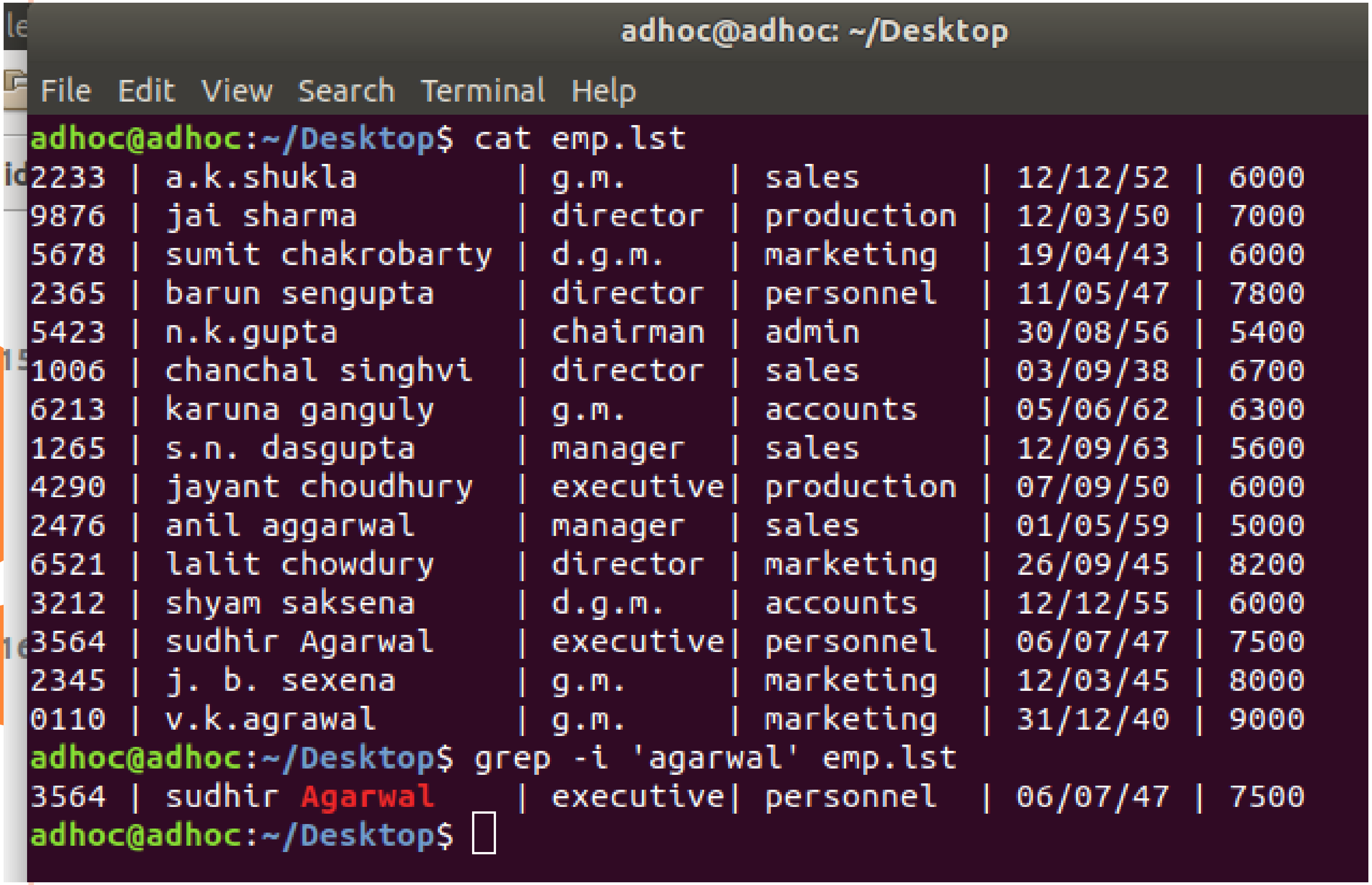
e.g $ grep ‘jai sharma’ emp.lst]

grep options:

-i	Ignore case for matching
-v	Doesn’t display lines matching expression
-n	Display line numbers along with lines
-c	Display count of number of occurrence
-l	Display list of filename only
-e	expression
-x	Match pattern with entire line
-f	File take pattern from file one per line
-E	Treats patterns as an extended regular expr
-F	Match multiple fixed string

-i : Ignoring Case

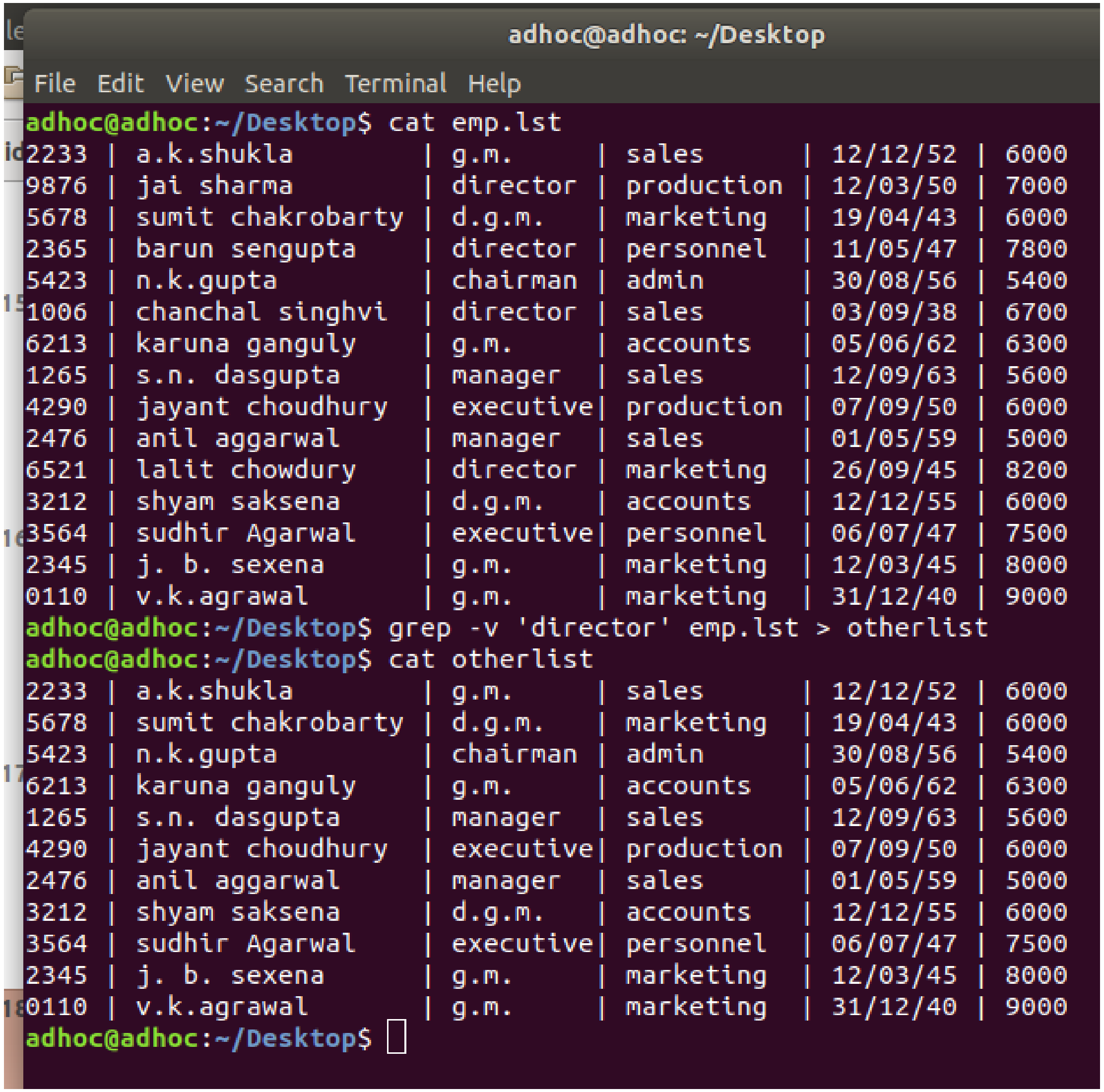
$ grep -i ‘agarwal’ emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep -i 'agarwal' emp.lst
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
adhoc@adhoc:~/Desktop$ 
```

-v: Deleting line (Inverse)

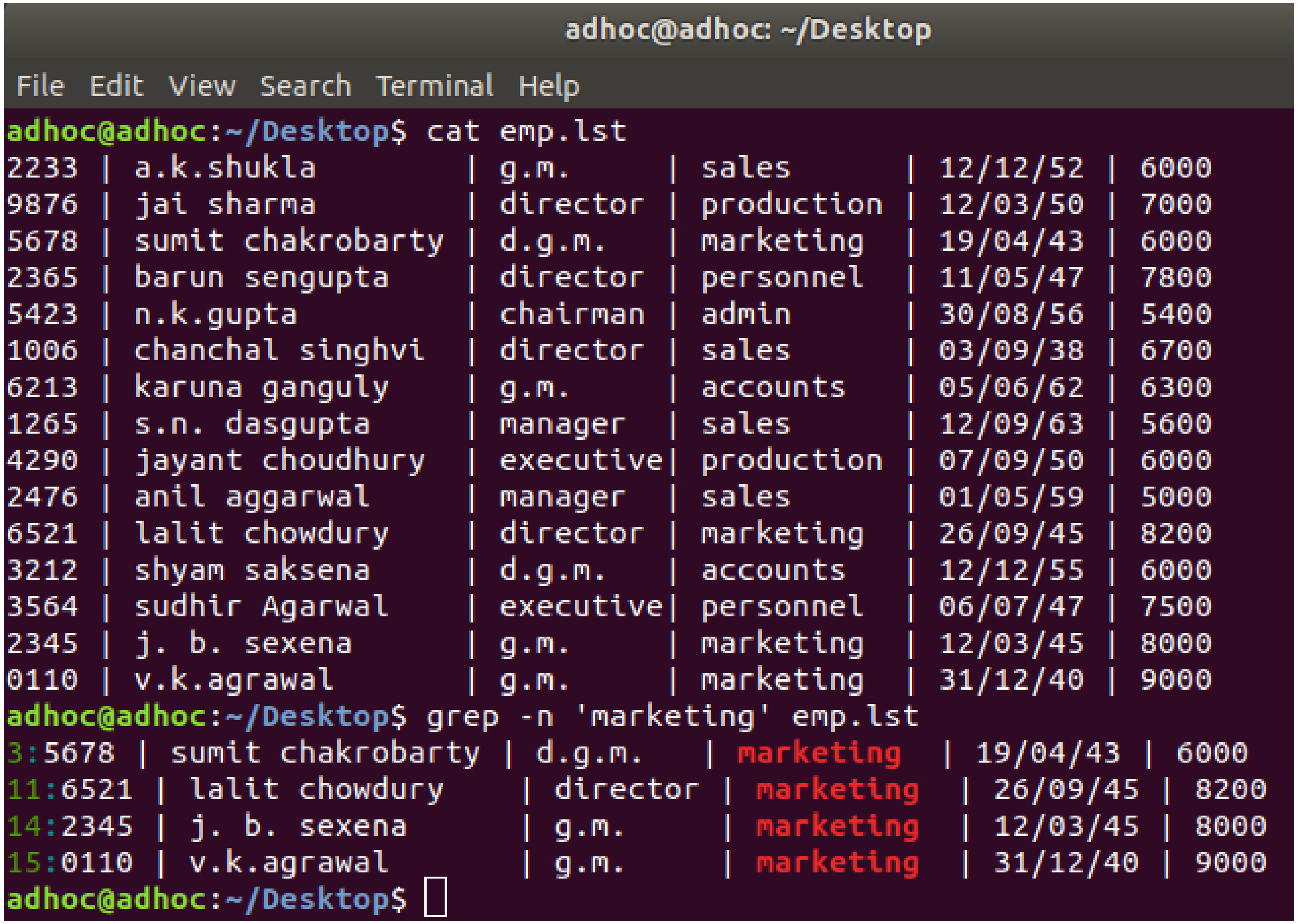
$ grep -v ‘director’ emp.lst > otherlist   [create]



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep -v 'director' emp.lst > otherlist
adhoc@adhoc:~/Desktop$ cat otherlist
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ 
```

-n : Displaying Line numbers

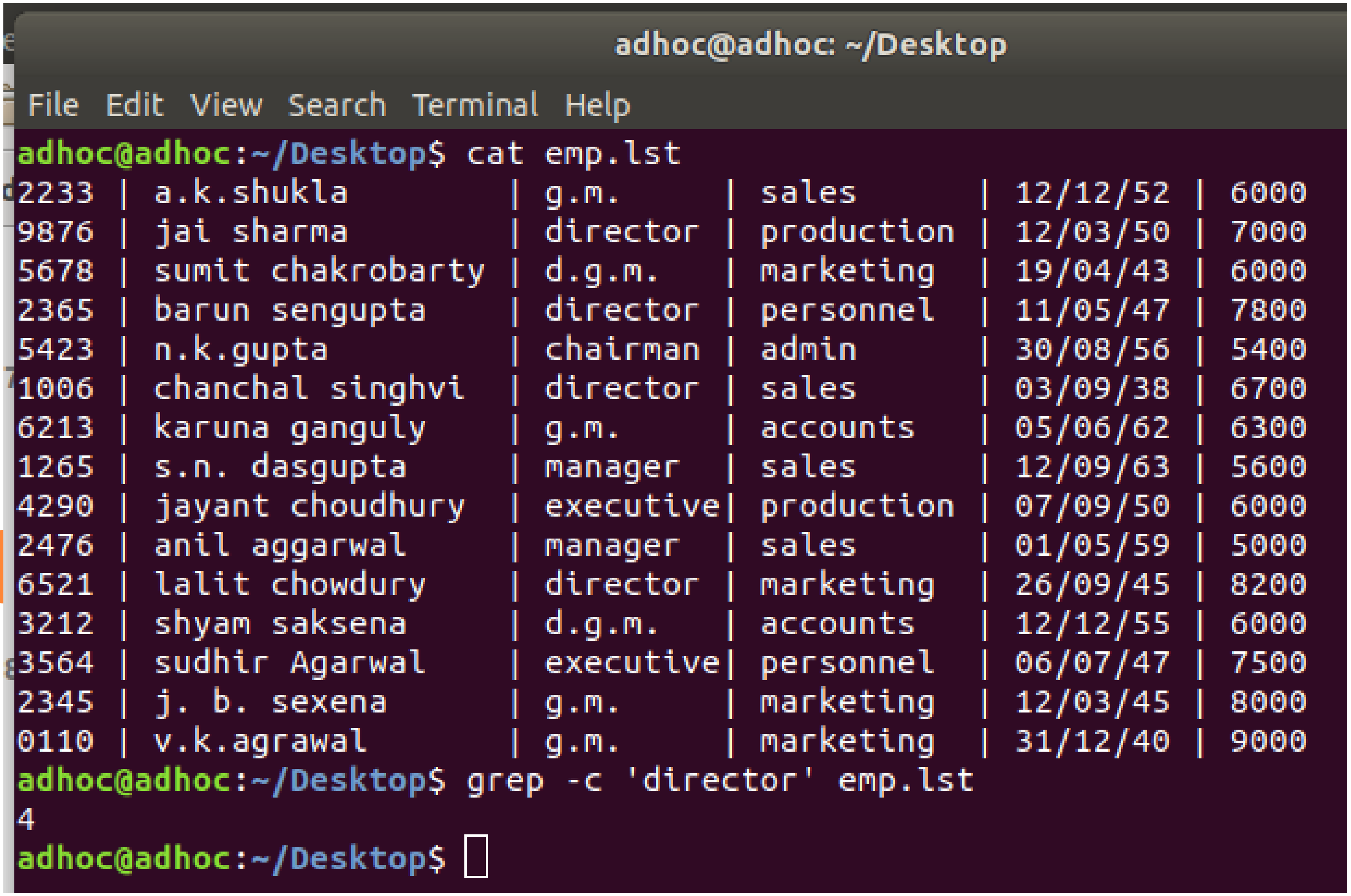
$grep -n marketing’ emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep -n 'marketing' emp.lst
3:5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
11:6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
14:2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
15:0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ 
```

-c :counting line containg pattern

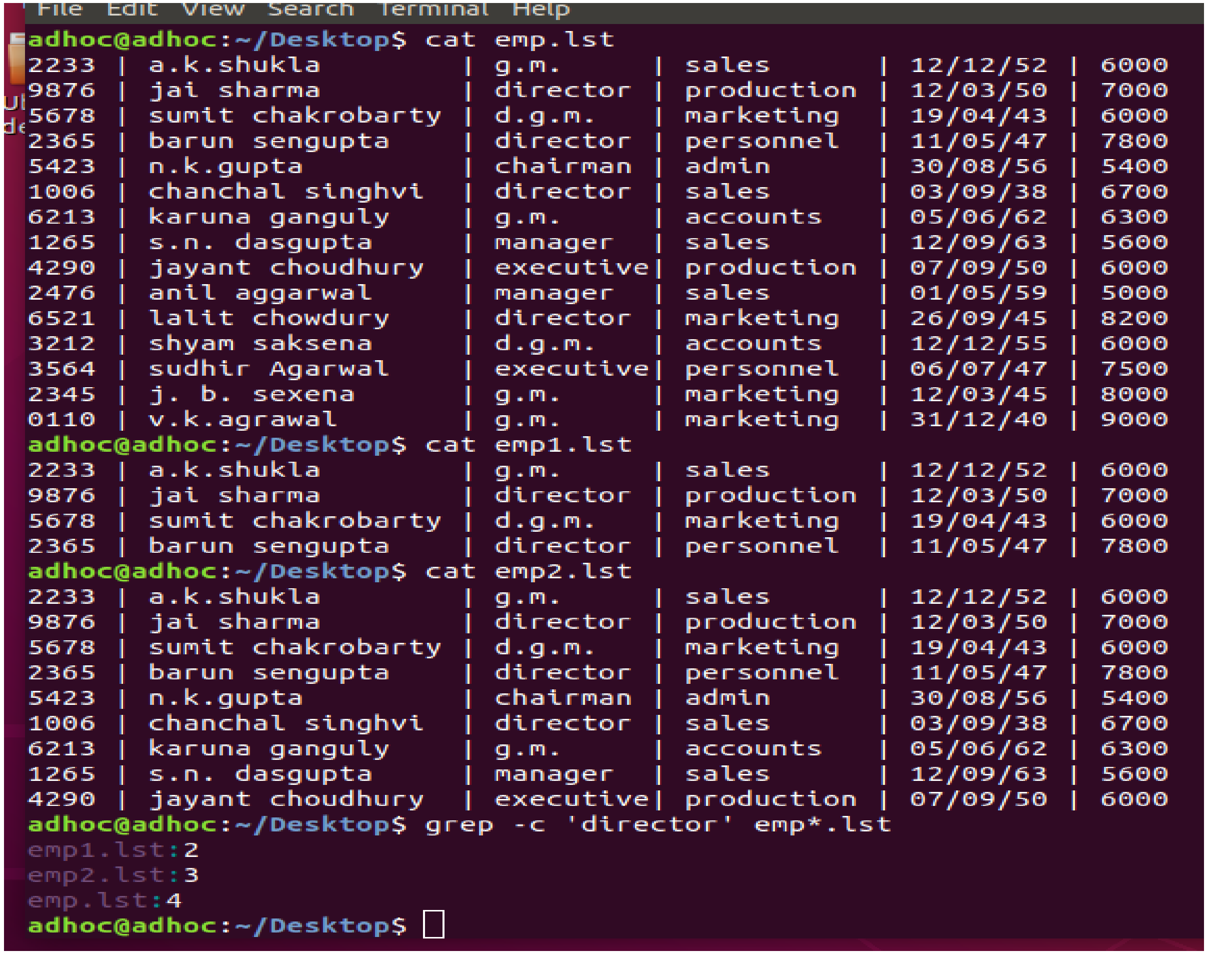
$grep -c ‘director’ emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep -c 'director' emp.lst
4
adhoc@adhoc:~/Desktop$ 
```

-c :counting line containg pattern
More than one files

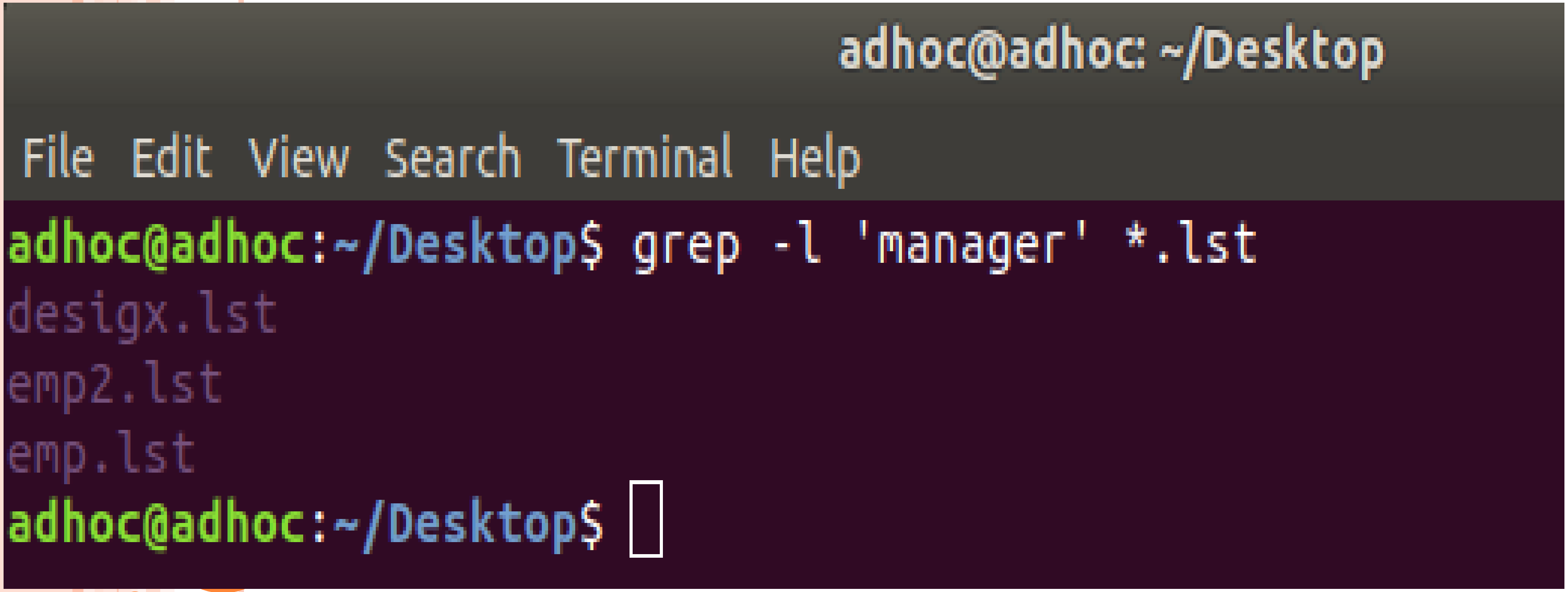
$grep -c ‘director’ emp*.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ cat emp1.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
adhoc@adhoc:~/Desktop$ cat emp2.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
adhoc@adhoc:~/Desktop$ grep -c 'director' emp*.lst
emp1.lst:2
emp2.lst:3
emp.lst:4
adhoc@adhoc:~/Desktop$
```

-l:Displaying filenames

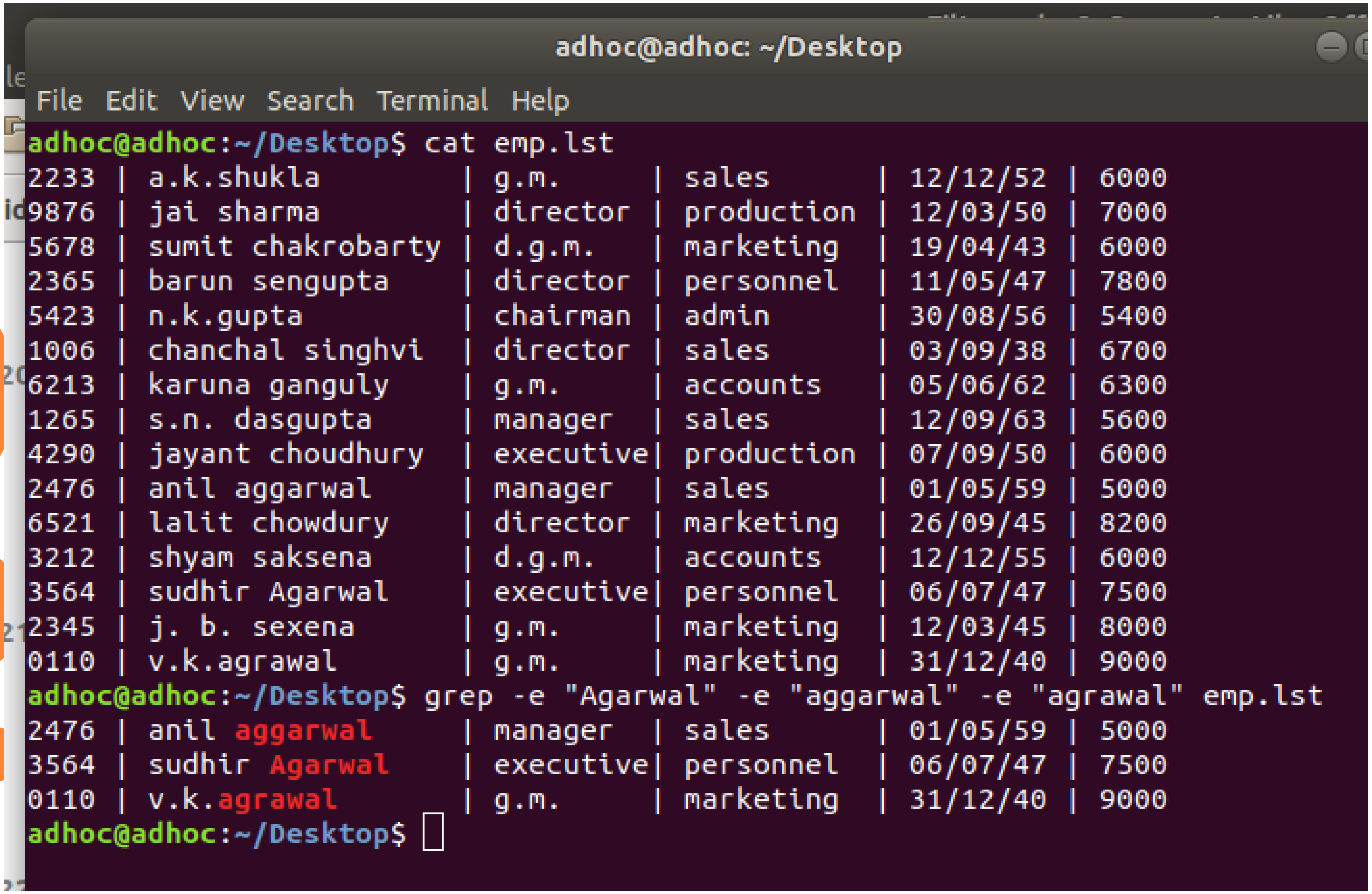
$grep -l ‘manager’ *.lst



```
adhoc@adhoc:~/Desktop$ grep -l 'manager' *.lst
desigx.lst
emp2.lst
emp.lst
adhoc@adhoc:~/Desktop$
```

-e : Matching multiple patterns

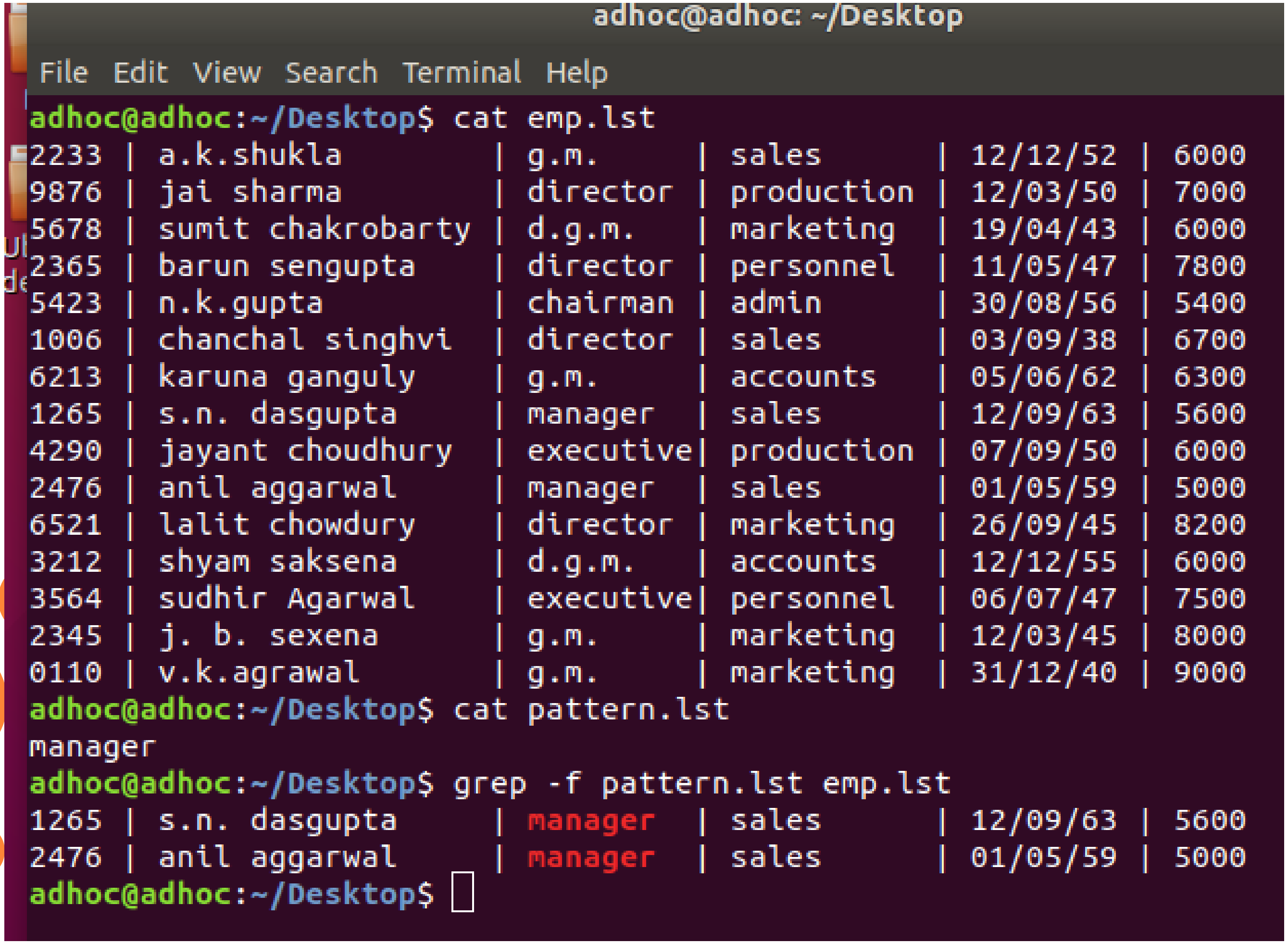
$grep -e “Agarwal” -e “aggarwal” -e “agrawal” emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla         | g.m.      | sales       | 12/12/52 | 6000
9876 | jai sharma         | director  | production  | 12/03/50 | 7000
5678 | sumit chakrobarty  | d.g.m.    | marketing   | 19/04/43 | 6000
2365 | barun sengupta     | director  | personnel   | 11/05/47 | 7800
5423 | n.k.gupta          | chairman  | admin       | 30/08/56 | 5400
1006 | chanchal singhvi   | director  | sales       | 03/09/38 | 6700
6213 | karuna ganguly     | g.m.      | accounts    | 05/06/62 | 6300
1265 | s.n. dasgupta      | manager   | sales       | 12/09/63 | 5600
4290 | jayant choudhury   | executive| production  | 07/09/50 | 6000
2476 | anil aggarwal      | manager   | sales       | 01/05/59 | 5000
6521 | lalit chowdury     | director  | marketing   | 26/09/45 | 8200
3212 | shyam saksena      | d.g.m.    | accounts    | 12/12/55 | 6000
3564 | sudhir Agarwal     | executive| personnel   | 06/07/47 | 7500
2345 | j. b. sexena       | g.m.      | marketing   | 12/03/45 | 8000
0110 | v.k.agrawal        | g.m.      | marketing   | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep -e "Agarwal" -e "aggarwal" -e "agrawal" emp.lst
2476 | anil aggarwal      | manager   | sales       | 01/05/59 | 5000
3564 | sudhir Agarwal     | executive| personnel   | 06/07/47 | 7500
0110 | v.k.agrawal        | g.m.      | marketing   | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ 
```

-f : Taking pattern from a file

$grep -f pattern.lst emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ cat pattern.lst
manager
adhoc@adhoc:~/Desktop$ grep -f pattern.lst emp.lst
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
adhoc@adhoc:~/Desktop$ 
```

Regular Expressions

When the **pattern** have to match a specific with in the form of expression like

***, g*, .*, [pqr] ,[1-3]**..etc

Then this pattern matching called **Regular Expressions**

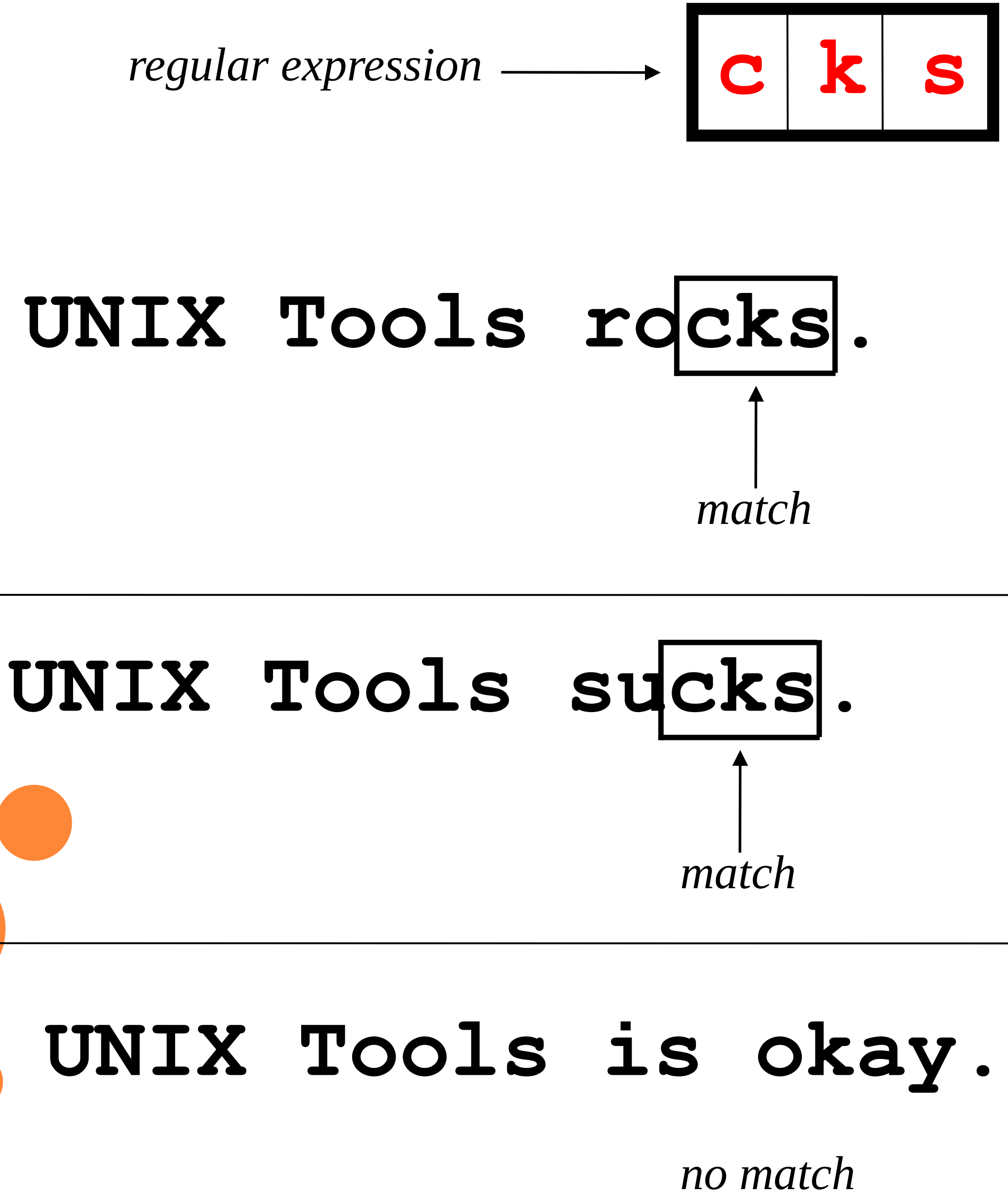
Protecting Regex Metacharacters

Since many of the special characters used in regexs also have special meaning to the shell, it's a good idea to get in the habit of single quoting your regexs

- This will protect any special characters from being operated on by the shell
- If you habitually do it, you won't have to worry about when it is necessary

Symbol or Expression	Matches
*	Zero or more occurrence of previous character
g*	Nothing or g,gg,ggg,...
.	A single character
.*	Nothing or any number of character
[pqr]	Single character from p or q or r
[c1-c2]	Single character with range c1 to c2
[1-3]	Single digit between 1 to 3
[^pqr]	Single character which not p ,q or r
[^a-zA-Z]	Single character non alphabet

Symbol or Expression	Matches
^pat	Beginning of line
Pat$	At end
Bash$	Bash at end
^bash$	Bash is only word line
^$	Line contains nothing



regular expression → `cks`

UNIX Tools rocks.

match

UNIX Tools sucks.

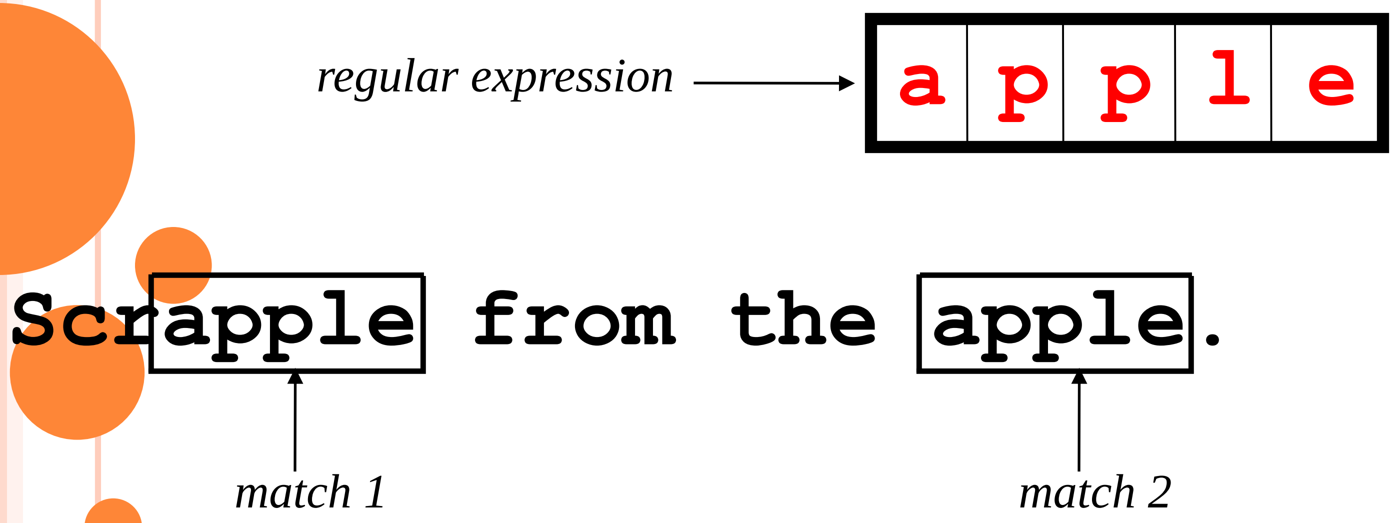
match

UNIX Tools is okay.

no match

Regular Expressions

- A regular expression can match a string in more than one place.



regular expression

a p p l e

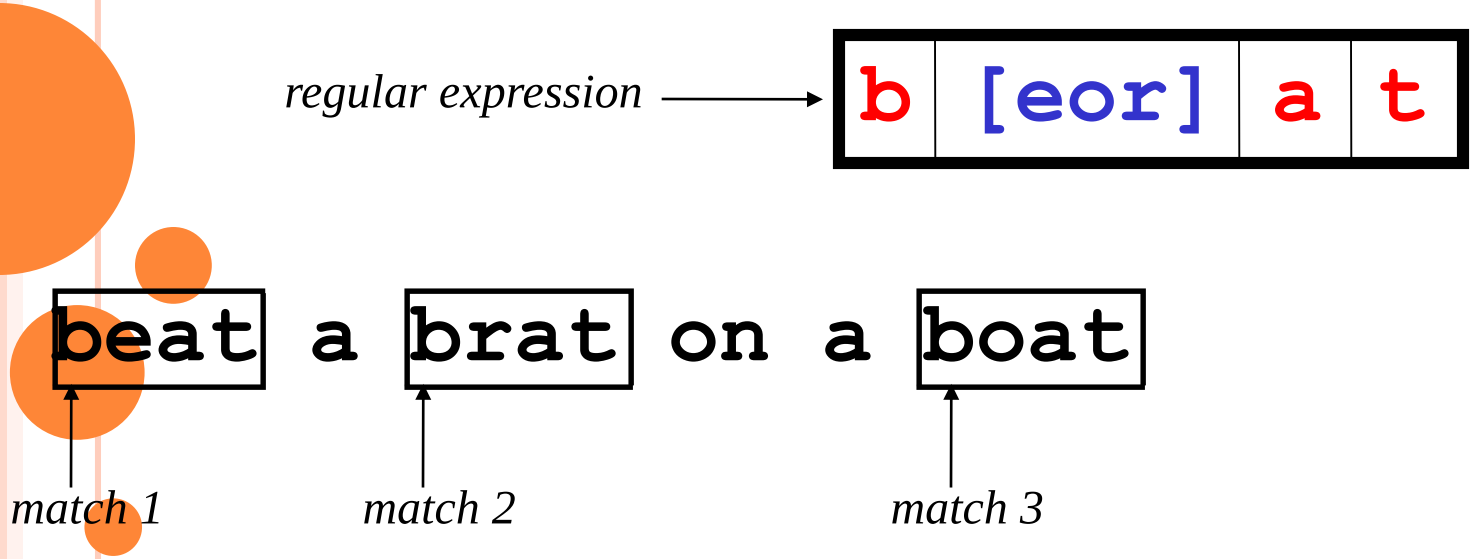
Scrapple from the apple.

match 1

match 2

Character Classes

- Character classes `[]` can be used to match any specific set of characters.



regular expression →

`b [eor] a t`

`beat a brat on a boat`

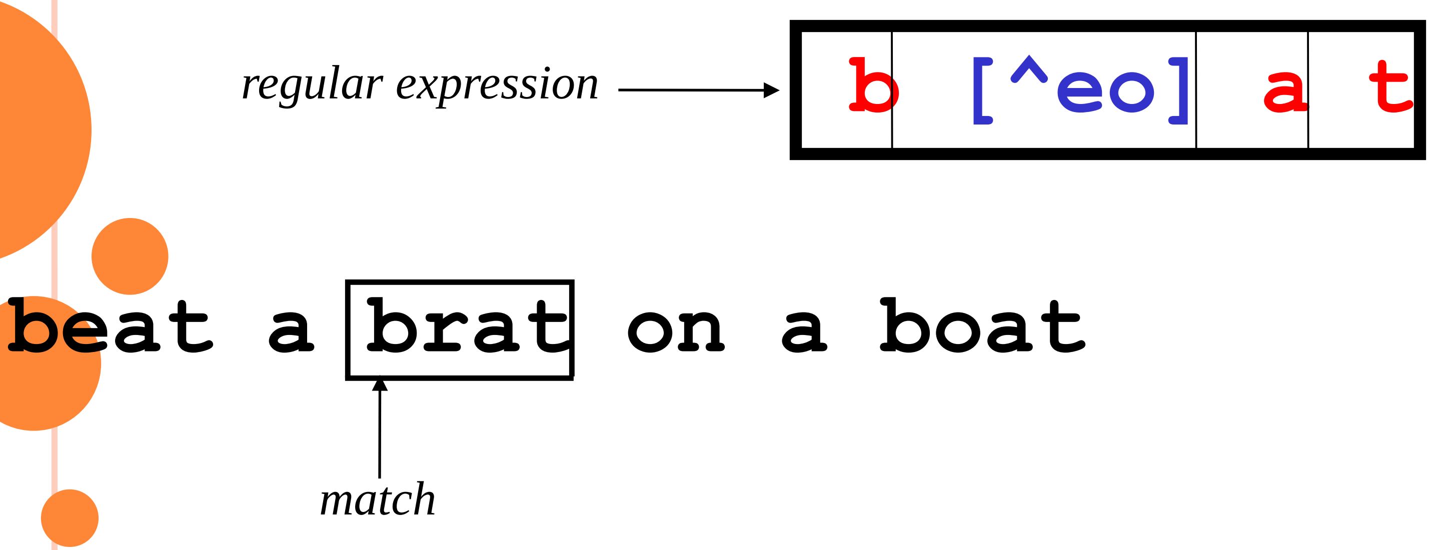
match 1

match 2

match 3

Negated Character Classes

- Character classes can be negated with the `[^]` syntax.



regular expression

`b [^eo] a t`

`beat a brat on a boat`

match

More About Character Classes

- `[aeiou]` will match any of the characters `a`, `e`, `i`, `o`, or `u`
- `[kK]orn` will match `korn` or `Korn`

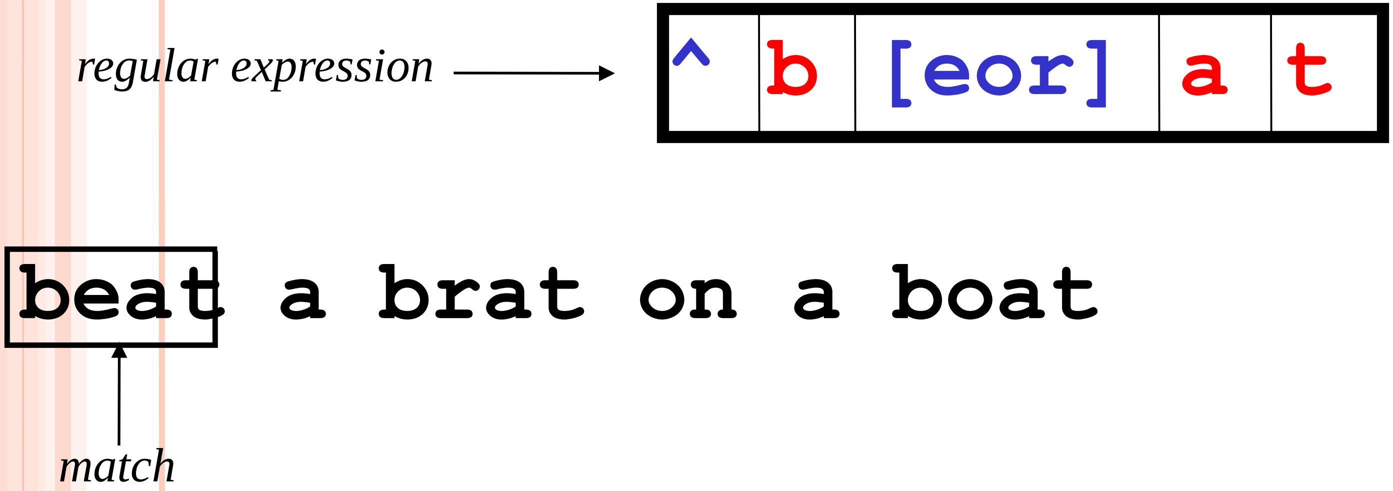
- Ranges can also be specified in character classes
  - `[1-9]` is the same as `[123456789]`
  - `[abcde]` is equivalent to `[a-e]`
  - You can also combine multiple ranges
    - `[abcde123456789]` is equivalent to `[a-e1-9]`
  - Note that the `-` character has a special meaning in a character class ***but only*** if it is used within a range,
    `[-123]` would match the characters `-`, `1`, `2`, or `3`

Named Character Classes

- Commonly used character classes can be referred to by name (*alpha*, *lower*, *upper*, *alnum*, *digit*, *punct*, *cntrl*)
  - Syntax `[:`***name***`:]`
    - `[a-zA-Z]` `[[:alpha:]]`
    - `[a-zA-Z0-9]` `[[:alnum:]]`
    - `[45a-z]` `[45[:lower:]]`
- Important for portability across languages

Anchors

- Anchors are used to match at the beginning or end of a line (or both).
- **^** means beginning of the line
- **$** means end of the line

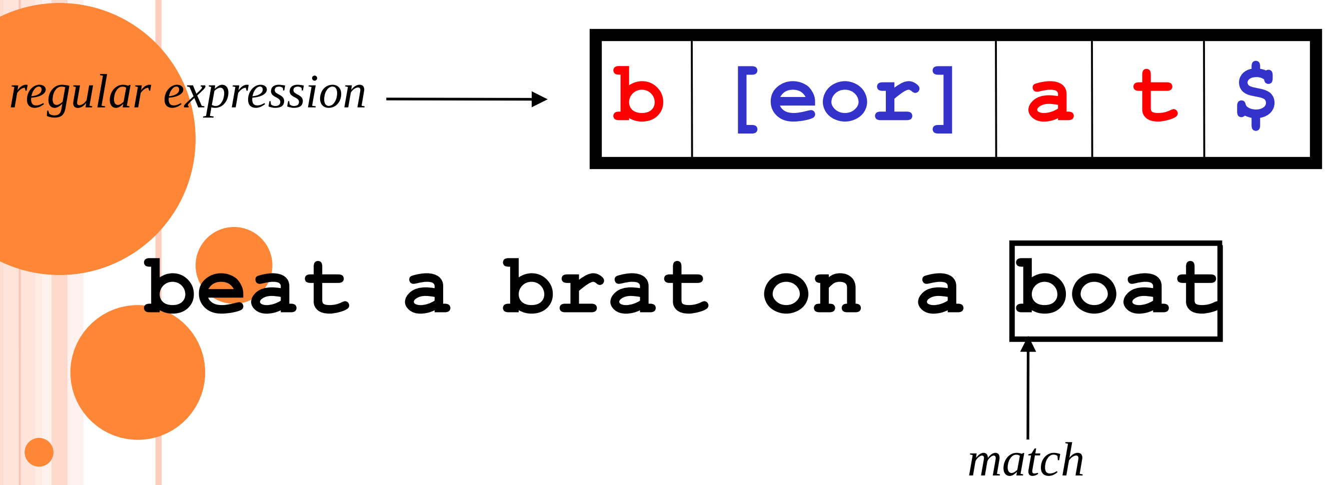


regular expression →

^	b	[eor]	a	t

beat a brat on a boat

match



regular expression →

b	[eor]	a	t	$

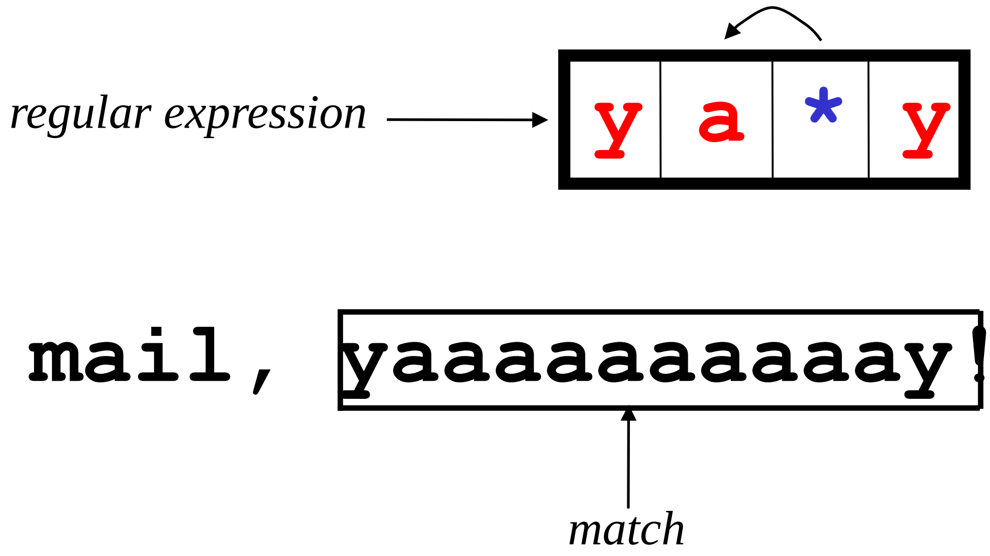
beat a brat on a **boat**

match

^word$ ^$

Repetition

- The * is used to define **zero or more** occurrences of the *single* regular expression preceding it.

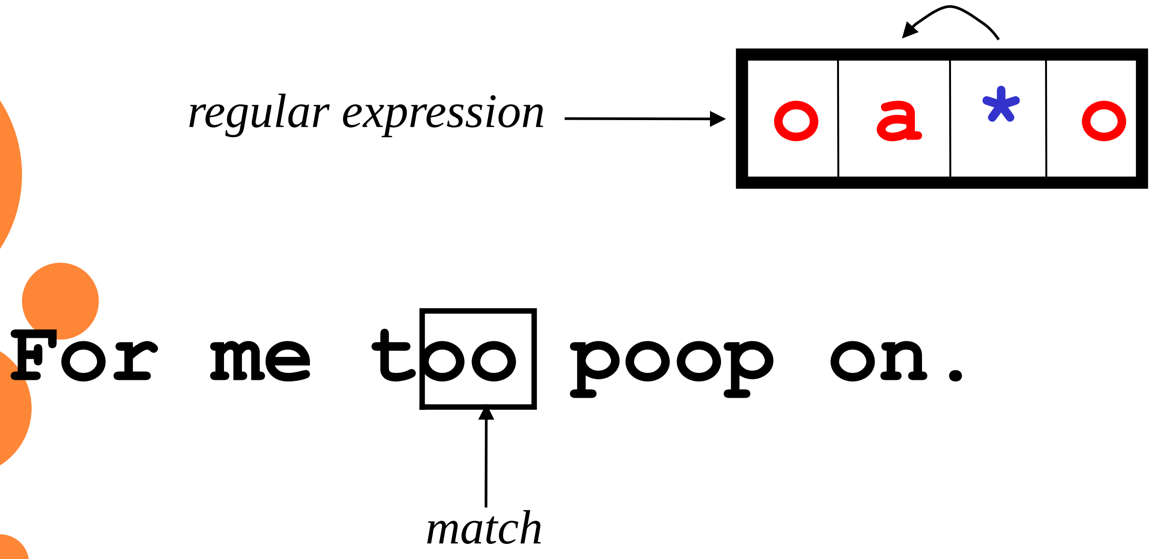


regular expression

y a * y

I got mail, yaaaaaaaaay!

match



regular expression

o a * o

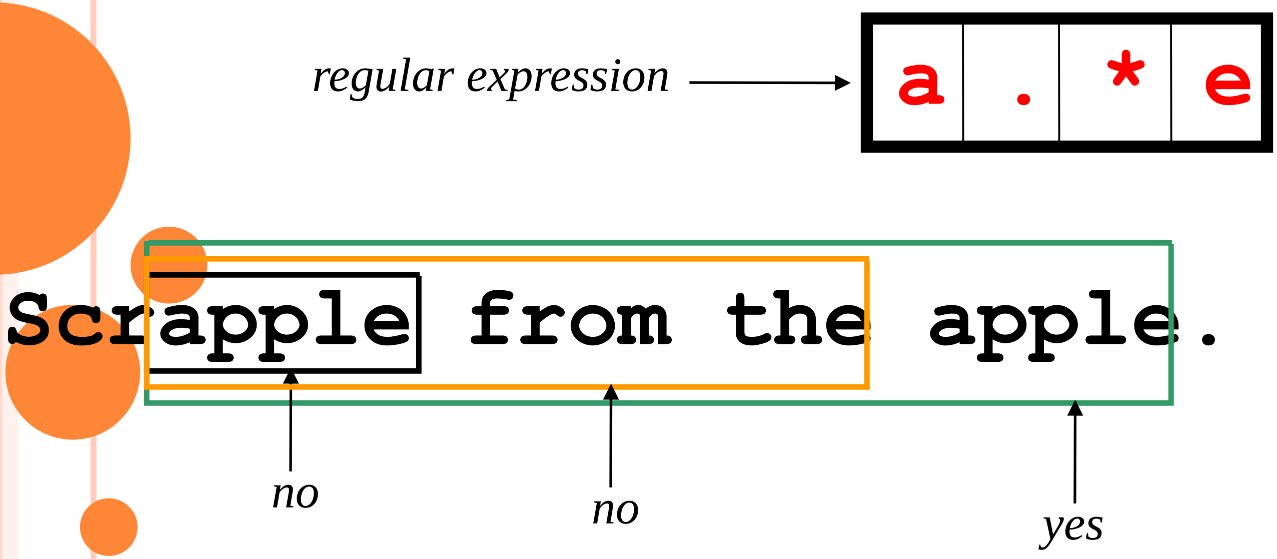
For me too poop on.

match

.*

Match length

- A match will be the longest string that satisfies the regular expression.



regular expression

`a` `.` `*` `e`

`Scrapple from the apple.`

no

no

yes

Repetition Ranges

- Ranges can also be specified
  - `{ }` notation can specify a range of repetitions for the immediately preceding regex
  - `{`***n***`}` means exactly *n* occurrences
  - `{`***n***`,}` means at least *n* occurrences
  - `{`***n***`,`***m***`}` means at least *n* occurrences but no more than *m* occurrences
- Example:
  - `.{0,}` same as `.*`
  - `a{2,}` same as `aaa*`

Sub-expressions

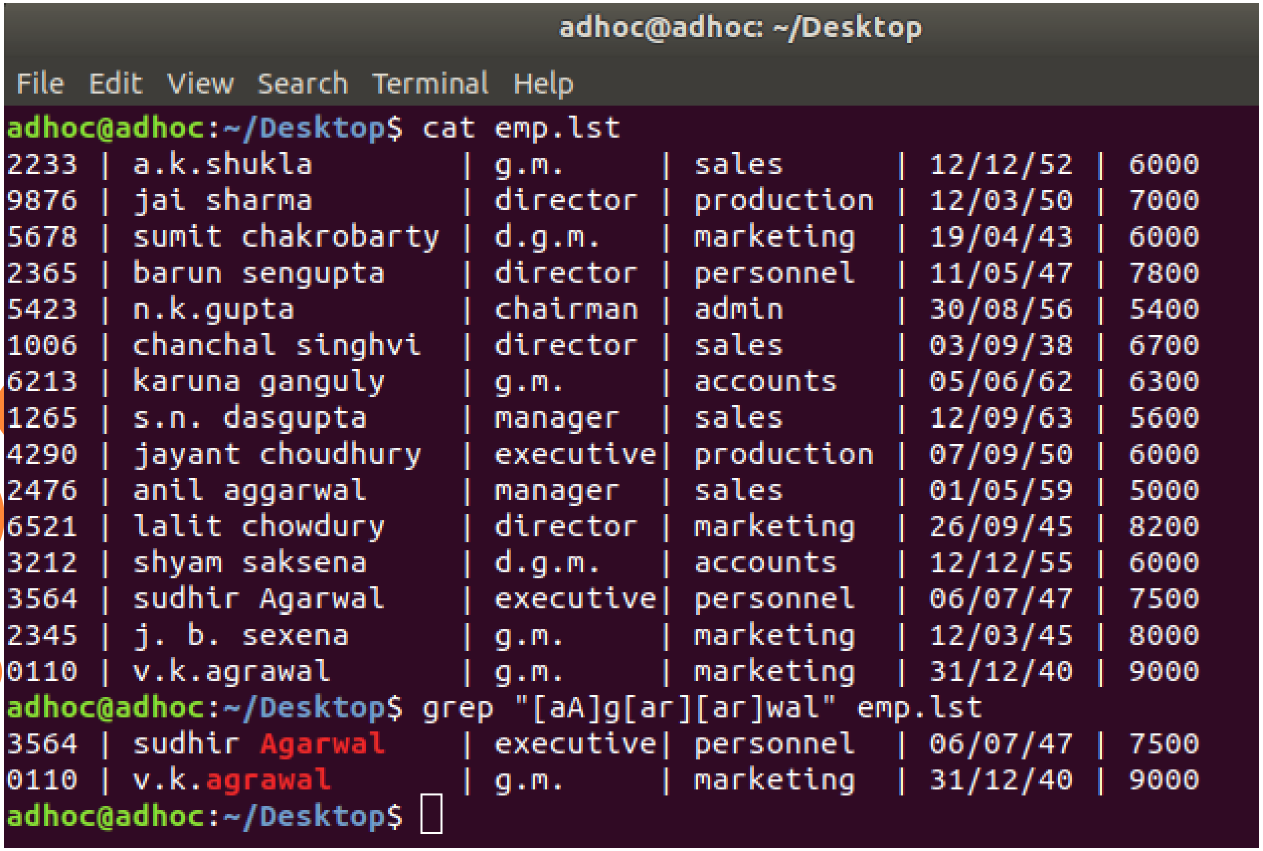
- If you want to group part of an expression so that `*` or `{ }` applies to more than just the previous character, use `( )` notation
- Subexpresssions are treated like a single character
  - `a*` matches 0 or more occurrences of `a`
  - `abc*` matches `ab`, `abc`, `abcc`, `abccc`, …
  - `(abc)*` matches `abc`, `abcabc`, `abcabcabc`, …
  - `(abc){2,3}` matches `abcabc` or `abcabcabc`

The character set

1. [ra]: either r or a

2.[aA] g [ar] [ar] wal : **agaa**wal, **agar**wal, **agra**wal, **agrr**wal, **Agaa**wal, **Agar**wal, **Agra**wal, **Agrr**wal

$ grep “[aA] g [ar] [ar] wal” emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep "[aA]g[ar][ar]wal" emp.lst
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ 
```

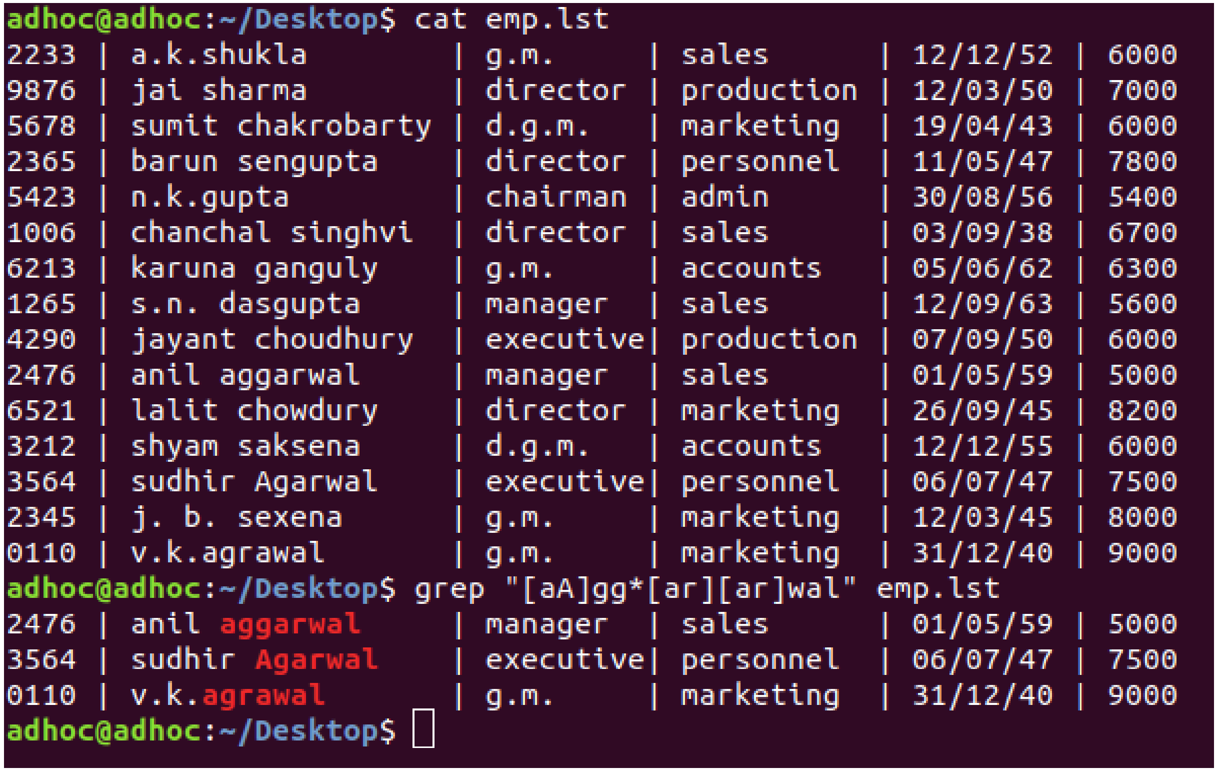
The *(Immediately prceding character)

g* : Zero or many number of g.

e.g – [aA]   gg*   [ar][ar]wal

{g,gg,gg…}

$grep “[aA]gg*[ar][ar]wal” emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla       | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma       | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta   | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta        | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly   | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta    | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury | executive| production | 07/09/50 | 6000
2476 | anil aggarwal    | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury   | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena    | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal   | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena     | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal      | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep "[aA]gg*[ar][ar]wal" emp.lst
2476 | anil aggarwal    | manager  | sales      | 01/05/59 | 5000
3564 | sudhir Agarwal   | executive| personnel  | 06/07/47 | 7500
0110 | v.k.agrawal      | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ 
```

The Dot(.)

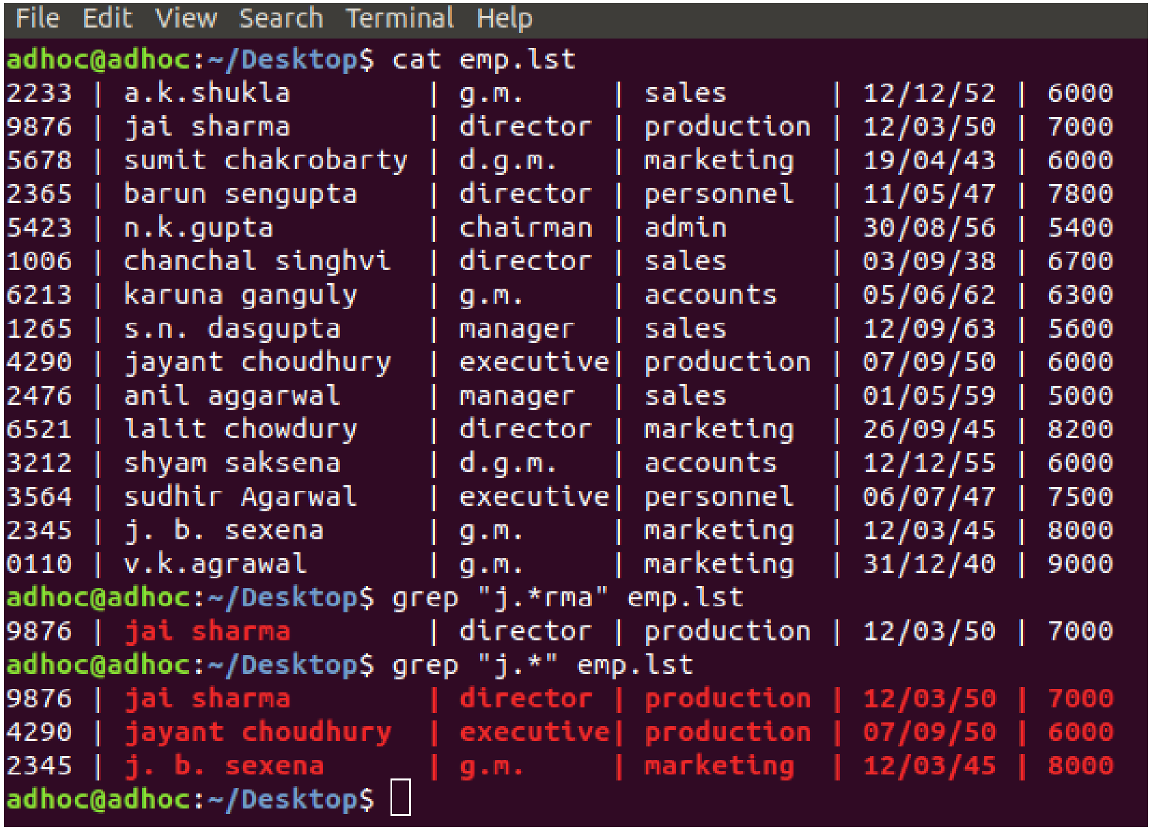
Matches a single character (like shell use ? Same as here is dot)

e.g

2... -> Match four character begin with 2

2??? -> Match four character begin with 2 in **SHELL**

$grep “J.*rma” emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla          | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma          | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty   | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta      | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta           | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi    | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly      | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta       | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury    | executive| production | 07/09/50 | 6000
2476 | anil aggarwal       | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury      | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena       | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal      | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena        | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal         | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep "j.*rma" emp.lst
9876 | jai sharma          | director | production | 12/03/50 | 7000
adhoc@adhoc:~/Desktop$ grep "j.*" emp.lst
9876 | jai sharma          | director | production | 12/03/50 | 7000
4290 | jayant choudhury    | executive| production | 07/09/50 | 6000
2345 | j. b. sexena        | g.m.     | marketing  | 12/03/45 | 8000
adhoc@adhoc:~/Desktop$ 
```

Specifying Pattern Locations (^ and $)

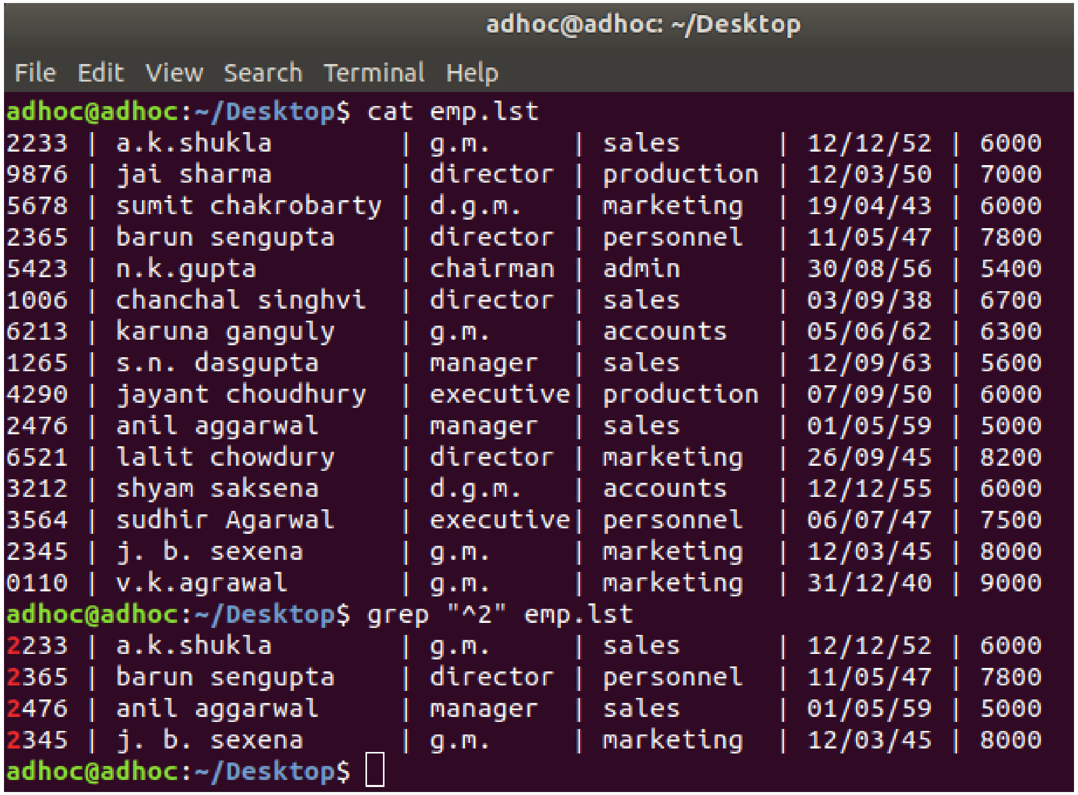
^: beginning of line

$: end of line

e.g Show employee details whose emp_id **beginning with 2**

2…

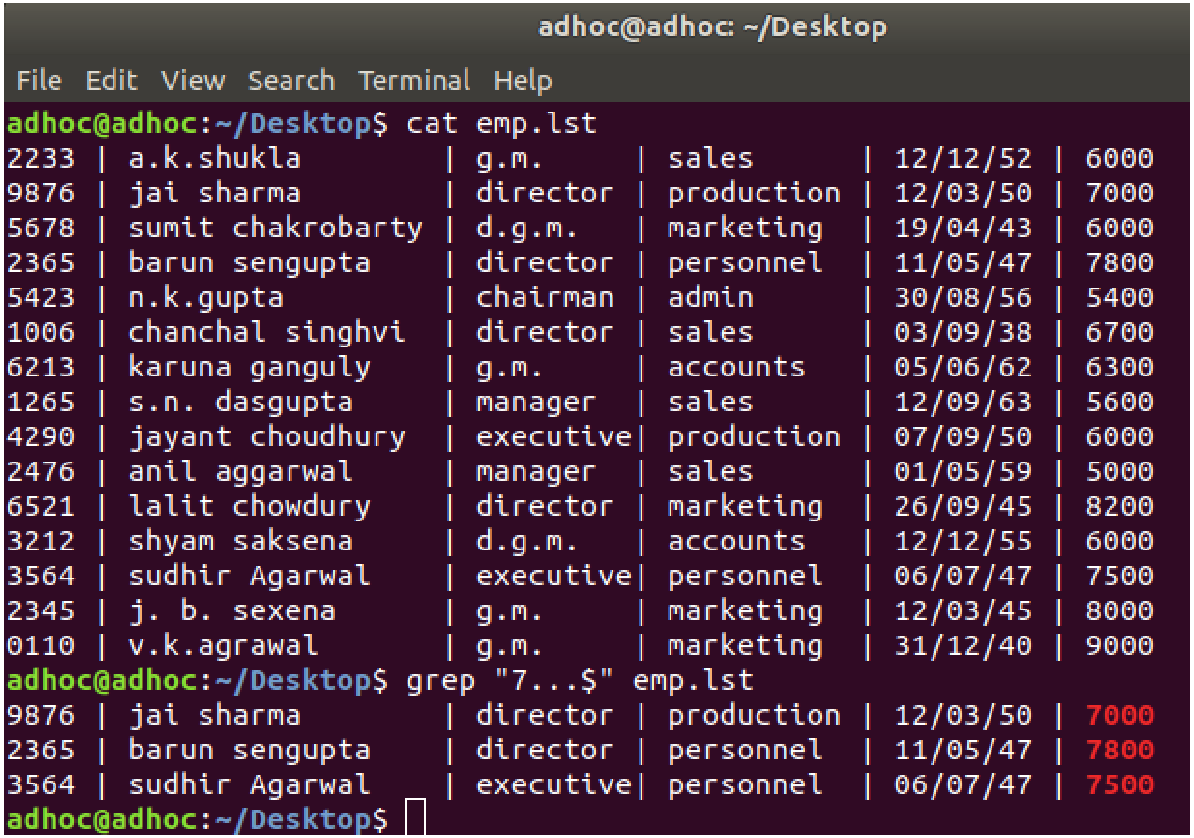
$grep "^2" emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla         | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma         | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty  | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta     | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta          | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi   | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly     | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta      | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury   | executive| production | 07/09/50 | 6000
2476 | anil aggarwal      | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury     | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena      | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal     | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena       | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal        | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep "^2" emp.lst
2233 | a.k.shukla         | g.m.     | sales      | 12/12/52 | 6000
2365 | barun sengupta     | director | personnel  | 11/05/47 | 7800
2476 | anil aggarwal      | manager  | sales      | 01/05/59 | 5000
2345 | j. b. sexena       | g.m.     | marketing  | 12/03/45 | 8000
adhoc@adhoc:~/Desktop$ 
```

e.g Show employee details whose salary **lies between 7000 and 7999**

$grep “7**...**$” emp.lst

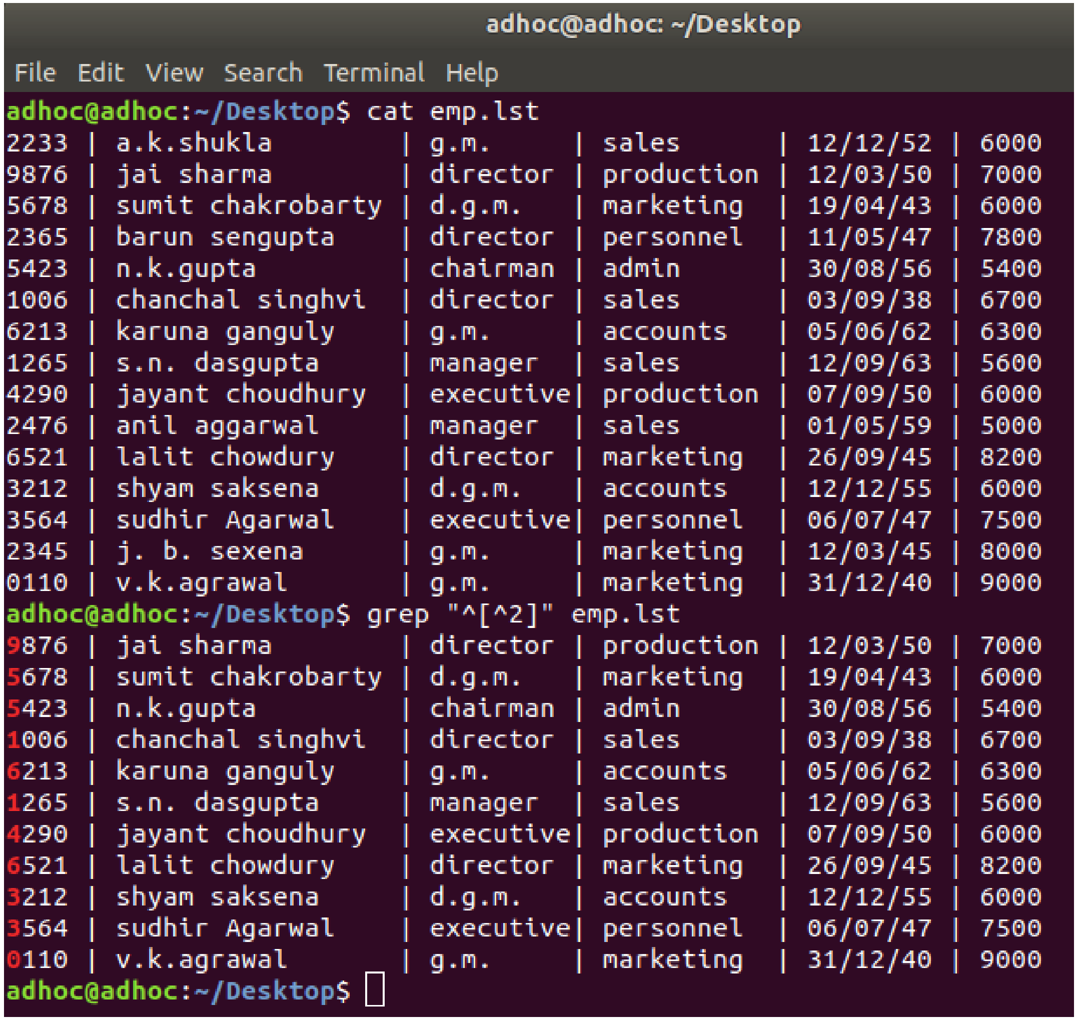


```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep "7...$" emp.lst
9876 | jai sharma        | director | production | 12/03/50 | 7000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
adhoc@adhoc:~/Desktop$
```

e.g

Show employee details whose emp_id **does not start with 2**

$grep^[^2]” emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep "^[^2]" emp.lst
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
6521 | lalit chowdury    | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$
```

Escaping Special Characters

- Even though we are single quoting our regexs so the shell won’t interpret the special characters, some characters are special to **grep** (eg * and **.**)
- To get literal characters, we *escape* the character with a \ (backslash)
- Suppose we want to search for the character sequence **`a*b*`**
  - Unless we do something special, this will match zero or more ‘a’s followed by zero or more ‘b’s, *not what we want*
  - **`a\*b\*`** will fix this - now the asterisks are treated as regular characters

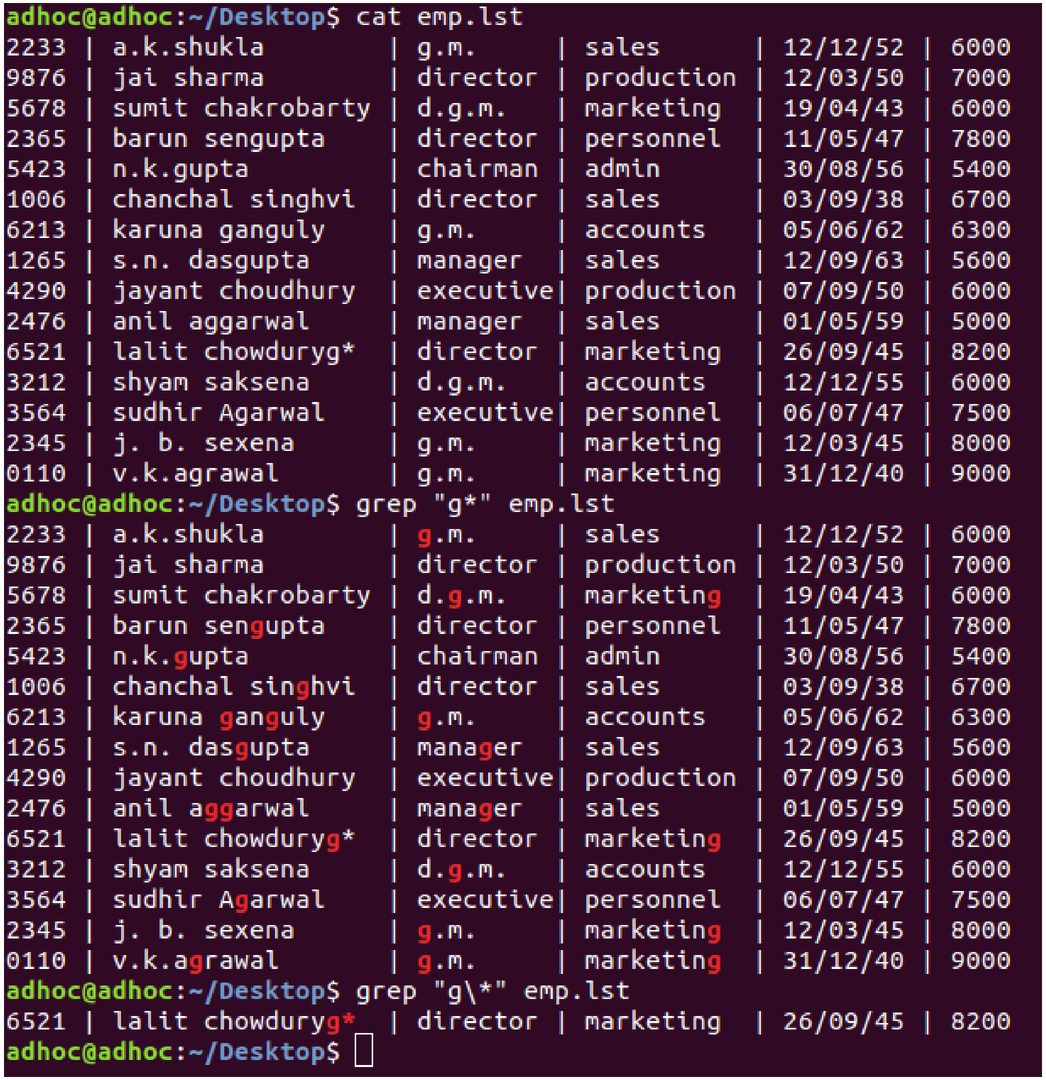
Escaping

e.g

Want to search g* in emp.lst

If $ grep “g*” emp.lst :> Find all **g**

$grep **“\g*”** emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowduryg*  | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep "g*" emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowduryg*  | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep "g\*" emp.lst
6521 | lalit chowduryg*  | director | marketing  | 26/09/45 | 8200
adhoc@adhoc:~/Desktop$
```

Extended R.E

Egrep: Alternation

- Regex also provides an alternation character `|` for matching one or another subexpression
  - `(T|Fl)an` will match ‘Tan’ or ‘Flan’
  - `^(From|Subject):` will match the From and Subject lines of a typical email message
    - It matches a beginning of line followed by either the characters ‘From’ or ‘Subject’ followed by a ‘:’
- Subexpressions are used to limit the scope of the alternation
  - `At(ten|nine)tion` then matches “Attention” or “Atninetion”, not “Atten” or “ninetion” as would happen without the parenthesis - `Atten|ninetion`

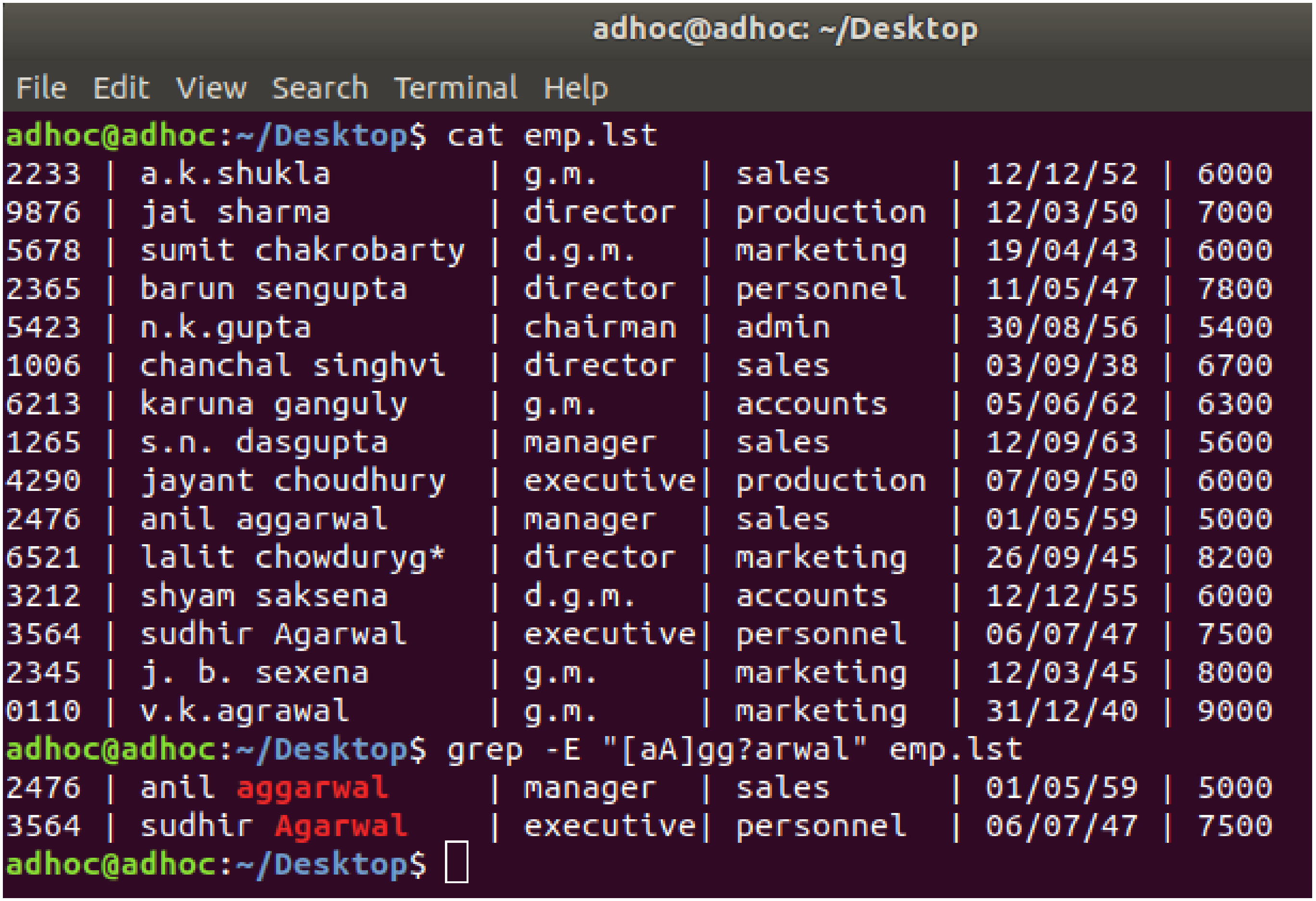
Egrep: Repetition Shorthands

- The `*` (star) has already been seen to specify zero or more occurrences of the immediately preceding character
- `+` (plus) means “one or more”
    - `abc+d` will match ‘abcd’, ‘abccd’, or ‘abccccccd’ but will not match ‘abd’
    - Equivalent to `{1,}`

Egrep: Repetition Shorthands

- The '`?`' (question mark) specifies an **optional** character, the single character that immediately precedes it
  - `July?` will match 'Jul' or 'July'
  - Equivalent to `{0,1}`
  - Also equivalent to `(Jul|July)`
- The `*`, `?`, and `+` are known as *quantifiers* because they specify the quantity of a match
  - Quantifiers can also be used with subexpressions
    - `(a*c)+` will match 'c', 'ac', 'aac' or 'aacaacac' but will not match 'a' or a blank line

$grep -E “[aA]gg?arwal” emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla         | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma         | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty  | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta     | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta          | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi   | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly     | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta      | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury   | executive| production | 07/09/50 | 6000
2476 | anil aggarwal      | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowduryg*   | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena      | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal     | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena       | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal        | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep -E "[aA]gg?arwal" emp.lst
2476 | anil aggarwal      | manager  | sales      | 01/05/59 | 5000
3564 | sudhir Agarwal     | executive| personnel  | 06/07/47 | 7500
adhoc@adhoc:~/Desktop$ 
```

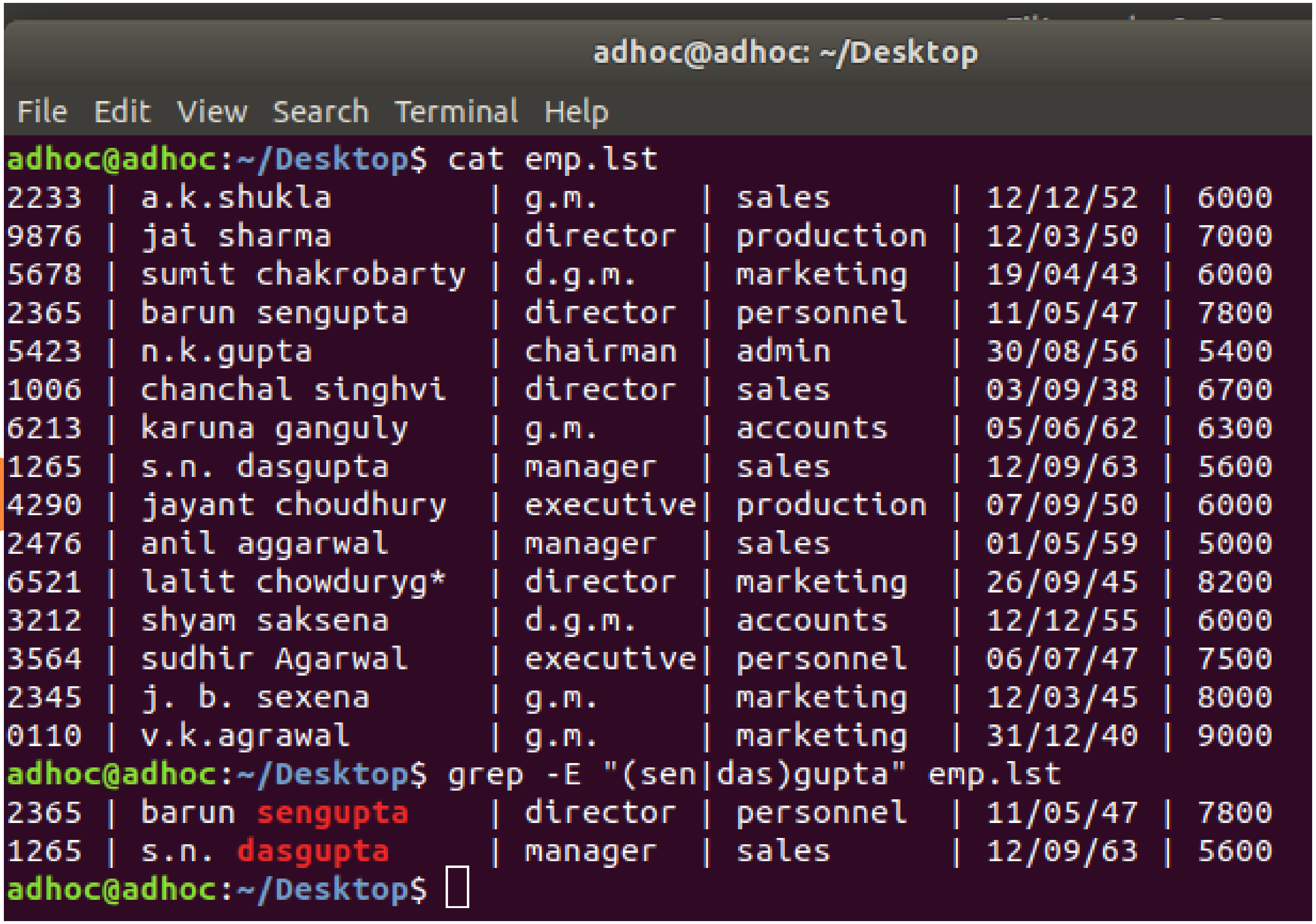
#inclue_<stdio.h>, #inclue_ _<stdio.h>, #inclue_ _ _<stdio.h>,
#inclue _ _ _ _ _ <stdio.h>

:> #incluide_+<stdio.h>

Matching multiple patterns(| ,( ))

$grep -E ‘Kishan|Ravikishan” emp.lst

$grep -E “(| Ravi)kishan” emp.lst



```
adhoc@adhoc:~/Desktop$ cat emp.lst
2233 | a.k.shukla        | g.m.     | sales      | 12/12/52 | 6000
9876 | jai sharma        | director | production | 12/03/50 | 7000
5678 | sumit chakrobarty | d.g.m.   | marketing  | 19/04/43 | 6000
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
5423 | n.k.gupta         | chairman | admin      | 30/08/56 | 5400
1006 | chanchal singhvi  | director | sales      | 03/09/38 | 6700
6213 | karuna ganguly    | g.m.     | accounts   | 05/06/62 | 6300
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
4290 | jayant choudhury  | executive| production | 07/09/50 | 6000
2476 | anil aggarwal     | manager  | sales      | 01/05/59 | 5000
6521 | lalit chowduryg*  | director | marketing  | 26/09/45 | 8200
3212 | shyam saksena     | d.g.m.   | accounts   | 12/12/55 | 6000
3564 | sudhir Agarwal    | executive| personnel  | 06/07/47 | 7500
2345 | j. b. sexena      | g.m.     | marketing  | 12/03/45 | 8000
0110 | v.k.agrawal       | g.m.     | marketing  | 31/12/40 | 9000
adhoc@adhoc:~/Desktop$ grep -E "(sen|das)gupta" emp.lst
2365 | barun sengupta    | director | personnel  | 11/05/47 | 7800
1265 | s.n. dasgupta     | manager  | sales      | 12/09/63 | 5600
adhoc@adhoc:~/Desktop$ 
```

Practical Regex Examples

- Variable names in C
  - `[a-zA-Z_][a-zA-Z_0-9]*`
- Dollar amount with optional cents
  - `\$[0-9]+(\.[0-9][0-9])?`
- Time of day
  - `(1[012]|[1-9]):[0-5][0-9] (am|pm)`
- HTML headers <h1> <H1> <h2> …
  - `<[hH][1-4]>`

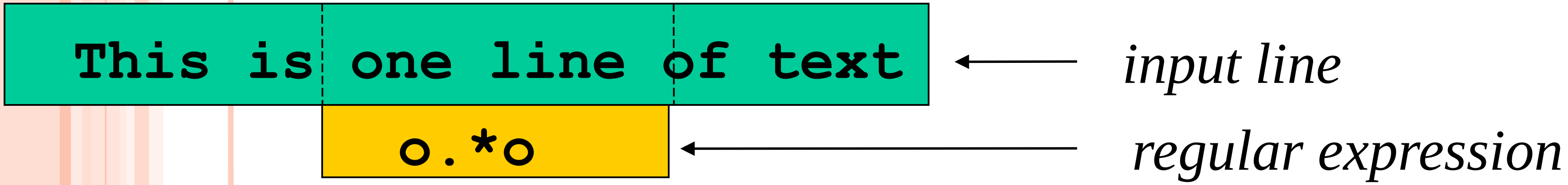
grep Examples

- `grep 'men' File_name`
- `grep 'fo*' File_name`
- `egrep 'fo+' File_name`
- `egrep -n '[Tt]he' File_name`
- `fgrep 'The' File_name`
- `egrep 'NC+[0-9]*A?' File_name`
- `fgrep -f expfile File_name`

- Find all lines with signed numbers

```
$ egrep ’[-+][0-9]+\.?[0-9]*’ *.c
bsearch. c: return -1;
strcmp. c: return -1;
strcmp. c: return +1;
```

- **egrep** has its limits: For example, it cannot match all lines that contain a number divisible by 7.



This is one line of text

o.*o

input line

regular expression

x	Ordinary characters match themselves	fgrep, grep, egrep
	(NEWLINES and metacharacters excluded)	
xyz	Ordinary strings match themselves	
\m	Matches literal character *m*	grep, egrep
^	Start of line	
$	End of line	
.	Any single character	
[xy^$x]	Any of x, y, ^, $, or z	
[^xy^$z]	Any one character other than x, y, ^, $, or z	
[a-z]	Any single character in given range	
r*	zero or more occurrences of regex r	
r1r2	Matches r1 followed by r2	
\(r\)	Tagged regular expression, matches r	grep
\n	Set to what matched the *n*th tagged expression (n = 1-9)	
\{n,m\}	Repetition	
r+	One or more occurrences of r	egrep
r?	Zero or one occurrences of r	
r1|r2	Either r1 or r2	
(r1|r2)r3	Either r1r3 or r2r3	
(r1|r2)*	Zero or more occurrences of r1|r2, e.g., r1, r1r1, r2r1, r1r1r2r1,...)	
{n,m}	Repetition	

Quick Reference