

## Find the colleges in the ranklist (grep, pipe and wc)

WE'LL COVER THE FOLLOWING ^

- Do you want to know more?

Let's now proceed to our first analysis: To list all the lines in the data file that contain the phrase "college", we need to introduce you with the command `grep` (global regular expression print). Let's first watch the following video lecture:



In a nutshell, `grep` allows you to look through all the lines in a file but only output those that match a pattern. In our case, we want to find all the lines in the dataset that contain "college". Here's how we do it:

```
grep -i "college" unirank.csv | csvlook
```



Here, the `grep` command takes two command-line arguments: the first is the pattern, and the second is the file in which we want to search for this pattern. If you run this command you should see some lines that contain the string “college”:

```
unrankingdata: hash
hellobigdata@hash:~$ grep -i "college" unirank.csv | csvlook
Dartmouth College      Hanover      NH      51438      4307      11
-----
Boston College          Chestnut Hill MA      51,296      9,192      31
College of William & Mary Williamsburg VA      41,718      6,301      32
University of Maryland--College Park College Park MD      32,045      27,443      60
Texas A&M University--College Station College Station TX      28,768      48,960      74
SUNY College of Environmental Science and Forestry Syracuse NY      17,620      1,839      99
St. John Fisher College Rochester NY      31,600      2,605      146
Edgewood College        Madison      MI      27,530      1,813      171
hellobigdata@hash:~$
```

Institutes containing "colleges" in the unirank.csv data set

Note that we have put `-i` option to make the matching case insensitive. Also, find that the logic by mistake identified two universities as college! due to the fact that their names contained the string (“college”). So, you need to be careful, while using `grep` in data analytics and particularly before reaching a decision!

## Do you want to know more? #



'grep' man page

