

EZ-Bash: Designing a Metasearch Engine for Bash Command Queries

Arpit Dev Mathur, Arshdeep S Sabharwal, Ken Wang Mentor: Amirhossein Aleyasen

Motivation

Often, a new programmer in CS finds himself faced with a tar file (as an example) and he is asked to unzip it. The command to unzip it is incredibly complicated and not intuitive at all. Linux/Unix based systems have plenty of these commands that a novice would not be aware of. Our motivation is to bridge that information gap with our interface. We plan on adding this functionality into terminal so the user does not have to google for the correct command and has a set of scraped and ranked commands ready to choose from.

Linux console Interface

Google search for "linux find total disk space".

Terminal output showing search results for "linuxHelper "find total disk space used personally by the current user"":

```
stackoverflow.com
1) df(1) | awk '{ $3 ~ /[0-9]+/ { print $4 }' | tail(1) (vote:5)
2) df -Ph $PWD | tail -1 | awk '{ print $3}' (vote:0)
superuser.com
1) df -a | df -h (vote:12)
2) du -df -h --max-depth=1 (vote:2)
3) du -sh folder | df -h (vote:1)
4) du -sh folder | df -h (vote:1)
5) df -h | cd du -sh * --max-depth du df -i (vote:51)
```

Search results for "df -h" (Disk Free) will show the free space on each of the mounted file systems.

51 So cd to the filesystem that's full, and du -sh * (Disk Usage) will show the total space used by each of the files/directories in the current working directory. The --max-depth option for du may also be useful here.

Finding exactly what is responsible for using all the space can be somewhat of an art - This answer lists

Use CSS selector to extract the code

Code snippet extracted from the search result:

```
<code>df -h</code>
```

Project Description

The program will google the phrase on the professional coding websites (e.g. stackoverflow), where somebody probably has already asked the same question, extract the proper answer codes and give the suggested answers to users. We use several methods for validating and ranking the results (e.g. number of votes, programming language detection methods) to make sure they're appropriate results. The project can be also expanded to other programming languages (e.g. Java and Python.)

What We Have Done

- Implemented a system to retrieve websites with possible solutions.
- Scraped the web pages to retrieve relevant code snippets.
- Filtered relevant code snippets to remove duplicates as well as redundant syntax.
- Developed console-based and web-based user Interfaces

Ranking Results

We used a classifier to rank results. The classifier uses the following features and some manual training to rank results:

- length of command
- votes given to command
- position on page
- number of commands on page
- language of command

Technology Stack

To give our search tool persistency of data beyond system shutdown and restart, we implemented a MongoDB database to store relevant data. The data we stored included:

- Sources : The source websites that include Linux commands like cyberciti.biz, stackoverflow.com
- Code Selectors : The selectors we used to parse webpages.

Web Interface

Linux Assist found a total of 10 results for remove.

Commands displayed:

- rm -rf .
- rm -rf /Users/arshsab/Programming/~
- rm -rf .
- mkdir /data/db
- sudo service apache2 restart
- mv /path/to/source /path/to/dest
- mkdir /data/db
- sudo service nginx start
- git merge -s ours email
- git checkout -b old-state 0d1d7fc32

- Web interface allows the user to search for command without installing the software
- Uses the same backend logic as the command line interface to generate and rank results.
- Built using the play framework, jQuery, isotope.js, and bootstrap in order to create and format webpages.