

Forum:[Hack Forums](#)

Thread: [Language detection with gzip](#)

glob - January 2, 2014, 12:44 am

I saw this a while back when I saw a post on a dailyprogrammer thread (which was based around language detection) that linked to this lecture by Peter Norvig:

Lecture ([Click to View](#))

[Video: <http://www.youtube.com/watch?v=dIEStL7riMo>]

The basic idea of this is that you have your text you want to determine the language of, and a variety of text files containing words or sentences of different languages.

If you compare the gzipped original language files with the gzipped language files concatenated with the text you want to verify the language of and record the difference in bytes for each language file, the right language will have the least difference in bytes.

An explanation of why it works is in the lecture above.

For this tutorial, the language files I will be using (and I am only using a few for simplicity) are:[EN - Huckleberry Finn](#)

[FR - Le tour de l'Espagne en Automobile](#)

[DE - Das Leben Tolstois](#)

I renamed the files EN.txt, FR.txt and DE.txt respectively, and placed them in a subdirectory named /LanguageFiles

Instead of this being a tutorial outlining what I write line by line and then posting the full python code at the end I have decided to comment the (probably overthought and overcomplicated) source and just post that. I will also include a download to a zip file containing the full project directory including the three language files.

This is how I chose to write my implementation: [Pastebin Link](#)

[Link to full project directory download and individual file download from my website](#)

References:<http://www.youtube.com/watch?v=dIEStL7riMo>

<https://ileriseviye.wordpress.com/2011/12/19/language-detection-in-the-bash-command-line-using-gzip-aka-comprehension-is-compression/>

Notes: This does not have a 100% success rate (Especially if you start adding more and more

languages that are similar to each other)

Longer text = higher success rate

Yugen - *January 2, 2014, 8:58 am*

Very nice thread, thanks for taking the time to make it.

Skyla Blue - *January 2, 2014, 7:15 pm*

Interesting tutorial. Better than the majority of "tutorials" in this section.