lab - Command Line Wikipedia

Wikipedia is a convenient source of information, a quick way to get a brief introduction or overview of a subject or a famous person. We normally think of Wikipedia as a browser-based resource, and so it is - but what can we do with it from the command line? Wouldn't it be convenient to be able to get some of that information with a simple command-line request? Let's build such a tool for this lab!

Example:

```
$ pedia Jane Austen
Jane Austen (UK: /ˈdʒɪn/; 16 December 1775 – 18 July 1817) was an English
novelist known primarily for her six major novels, which interpret, critique and
comment upon the British landed gentry at the end of the 18th century. Austen's
plots often explore the dependence of women on marriage in the pursuit of
favourable social standing and economic security. Her works critique the novels of
sensibility of the second half of the 18th century and are part of the transition
to 19th-century literary realism.[2][b] Her use of biting irony, along with her
realism and social commentary, have earned her acclaim among critics and scholars.
$
$ pedia skolem normal form
Every first-order formula may be converted into Skolem normal form while not
changing its satisfiability via a process called Skolemization (sometimes spelled
Skolemnization). The resulting formula is not necessarily equivalent to the
original one, but is equisatisfiable with it: it is satisfiable if and only if the
original one is satisfiable.[1]
```

As you can see from the examples, we're not showing any html tags - we've stripped those from the text that comes back, but otherwise we're just dumping the non-html text, the content, so we may get some Unicode oddities and we don't do formatting, so footnotes look like normal text.. But our purpose is to get at the information. This is "good enough".

So how do we do it? Look up a topic in Wikipedia - check out the URL for the page that has the information you're after. The works you used for your search would be the ones supplied on a command line invocation of the script. Can you see how to map those to a URL?

When to stop: Notice that the examples don't dump the entire wikipedia page. We're just after the introductory information. How much is that? If you're using Firefox, point it to a Wikipedia page and type control-U to see the page source. Try it on a few pages and see if you see a common pattern - something that might indicate the start and end of that introductory content. For example, begin with the first paragraph tag ("<p>").

Now you're set. Write a script that will: take words from the command line and convert them to a URL; download the page pointed to by that URL, pulling out the html and leaving just plain text. Stop when you find a good stopping point, as demonstrated in the examples, above.