smallwp user docs

This is a little program that gives you an offline version of the wikipedia in as little space as possible (well, probably you can do better, but for now, it should do). It does not do images. It can work off both HTML dumps and XML dumps with wiki markup.

Obtaining the data

Up front: There's some language-dependent aspects of the software. Right now, Asturian, German, English, and languages that use namespaces like English or German are supported. see README.HACKING for how to add support to others.

At this point, you needed to decide whether you want to use an XML-based or a ZIM-based one.

ZIM dumps (https://dumps.wikimedia.org/other/kiwix/zim/wikipedia/) have the advantage that rendering is basically like wikipedia does it. Math images are not supported, and it takes quite a bit more space than XML dumps. There are dumps including images – we'll think about them.

XML dumps would be smaller, can do math images and generally were "the right thing", but even with Wikipedia as of 10 years ago rendering frequently is shaky. These days, with wikidata and friends, it's probably not practical anymore, and I'm not sure there are XML dumps of the type we'd like any more. So, honestly, we mention this option for historical interest only.

Create a directory /var/share/wikipedia (if you want it somewhere else, set basedir in the config), move your ZIM dump there and then say:

```
sudo splitWp <name of the dump>
```

(of course, you only need sudo if you didn't give your user write permissions in the basedir). This will create some index-type files, which may take a while, in particular on slowar machines. The directory structure created is portable, though, so you can build it on a fast machine and move it to a slow one.

If you cannot sudo or do not want to, change the basedir in ~/.smallwp (see Configuration below) to something you can write to and leave out the sudo.

Usage

Run wpGui (or wpTui, if you don't run X) and point your browser to http://localhost:8780[1] .

You can browse around in the index or use the right search field with regular expressions (which is currently quite slow, though). The most convenient access method proabably is the Title Search, though: Just enter a few characters (case sensitively!), and you'll get a list of up to 40 titles that start with the letters you entered.

For now, you should run wpGui from a terminal -- you'll see tracebacks and the like, which right now may explain quite a bit. . .

Configuration

Smallwp is configured through an INI-style configuration file ~/.smallwp (if you don't like this location, you can change it using the SMALLWPCONF environment variable). To see what you can set in there, say:

```
splitWp -H
```

This will output something like this:

Section [general]

Configuration of smallwp

- address: string; defaults to '127.0.0.1' -- Address the server should listen on
- basedir: string; defaults to '/var/share/wikipedia' -- Path to root of wikipedia data
- customCSS: string; defaults to '' -- $HOME-relative path to a custom css file
- doTemplates: boolean; defaults to 'False' -- Attempt to render templates? (slow, buggy, but probably worth it)

- mathOutputResolution: integer; defaults to '110' -- Resolution (in dpi) of math images
- maxTitleMatches: integer; defaults to '1000' -- Maximal number of matches in title search
- mediaBase: string; defaults to 'http://upload.wikimedia.org/wikipedia/commons/' -- URL fragment to prepend to image links and the like in HTML dumps
- port: integer; defaults to '8780' -- Port the server should listen on
- sitename: string; defaults to 'Small Wikipedia Server' -- Name of wikipedia instance for display purposes

I do not recommend changing the address unless you know what you're doing and trust my code. I believe there are no obvious security holes in here, but then again I haven't really checked. The default only opens up access to your own machine (which may still be a security issue if I messed up and you have evil users on your machine).

An important setting is doTemplates. By default, it's false since the code that supports templates is not finished yet. On the other hand, templates work for many pages. They are slow though, because for every template, the system has to decrypt 1.5 megabyte on average, which slows things down quite a bit. Still, try setting it to True and see if you like what you see.

You also may want to change is mathOutputResolution -- the pngs for math get larger as you increase this. Also, this would be an obvious candidate for having a GUI control. . .

If you do not like the CSS built into smallwp, you can set your own local css using customCSS. I recommend starting from the built-in css, so start your server and say:

```
cd
wget -O .wpcss.css http://localhost:8780/css/smallwp.css
```

Then edit .wpcss.css to your liking and edit ~/.smallwp, adding a line like:

```
customCSS=.wpcss.css
```

If you must use any non-ASCII within your custom CSS, use latin1 encoding.

[1] it wouldn't be hard to provide a headless server if you want to run the thing all the time. It just happens that I don't. If you want this, ask (or do it yourself. . .).