Indexing OCR PDFs

By default, MindTouch does not filter OCR PDFs for indexing. Thankfully, it's fairly simple to implement. The first step is to download and install PDFMiner. This is a Python script and requires Python 2.4 or newer.

```
[root@dev ~]# wget http://pypi.python.org/packages/source/p/pdfminer/
pdfminer-20100213.tar.gz
[root@dev ~]# tar zvxf pdfminer-20100213.tar.gz
[root@dev ~]# cd pdfminer-20100213
[root@dev pdfminer-20100213]# python setup.py install
```

This will install PDFMiner as /usr/bin/pdf2txt.py

Now, we have to replace the existing filter with the new one. We will first backup the existing filter by renaming it to pdf2text.bak and then we'll download the new filter using wget:

```
[root@dev pdfminer-20100213]# cd /var/www/dekiwiki/bin/filters
[root@dev filters]# mv pdf2text pdf2text.bak
[root@dev filters]# wget http://developer.mindtouch.com/@api/deki/files/5890/=pdf2text
[root@dev filters]# chown dekiwiki.apache pdf2text
[root@dev filters]# chmod 755 pdf2text
```

You are now done! If you have existing OCR PDFs that need to be indexed, you must rebuild your search index by going to the Control Panel->Cache Management and select Rebuild Search Index. Otherwise, as you add your OCR PDFs, they will automatically be converted and indexed as usual.

NOTE: This will convert normal PDFs as well as OCR PDFs.

Credits to Yusuke Shinyama for writing pdf2txt.py. More information can be found at http://www.unixuser.org/~euske/pytho...ner/index.html