Homework: Crawling the World Wide Web with Apache Nutch
Due: April 16, 2014

---

1. Objective

In Homework: Tika you used Tika parsers to extract text from the provided PDF files and helped search for UFOs in the FBI's vault dataset.

But it's time to take a step back. Let's consider how did we obtain all those PDF files in the first place? The short answer is that we crawled the FBI's *Vault* site http://vault.fbi.gov/ and downloaded all of the relevant PDFs from that site, dropped them in a directory, and then packaged the whole directory into a tarball.

We used Apache Nutch (http://nutch.apache.org/) to collect the `vault.tar.gz` dataset. Now, you get to do it too! You'd think getting all the PDF files would be a snap. That is, of course, if they were named with URLs that ended with `.pdf`. And of course, provided that the PDFs were all available from a similar directory structure, like `http://vault.fbi.gov/pdfs/xxx.pdf`. Unfortunately, they're not.

Your goal in this assignment is to download, install and leverage the Apache Nutch web crawling framework to obtain all of the PDFs from a subset of the FBI's *Vault* website that we have mirrored at http://baron.pagemewhen.com/~chris/vault/vault.fbi.gov/ and then to package those PDFs and in order to create your own `vault.tar.gz` file.

1.1 Team Work

This assignment can be done as a team of up to four students.

See Section 6.1 for details on registering your team with Professor Mattmann. You must register your teams by 4/1/2014. Please only send one email per team.

2. Preliminaries

You will need about **11GB** of hard disk space to accommodate the data that you will crawl from our subset of the FBI website, and an additional **11GB** for the PDF files that you will subsequently extract from there, as well as space for Ubuntu OS and other overhead.

For this assignment, you will need a Linux OS (such as Ubuntu, separate instructions provided) or Cygwin if installing on Windows. We expect you to have basic understanding of Linux and/or Cygwin and assume you have a recent version of the OS installed.

3. Crawling using Apache Nutch

3.1 Downloading and Installing Apache Nutch

To get started with Apache Nutch, grab the latest release, available from:
http://www.apache.org/dyn/closer.cgi/nutch/

The latest release of Nutch that we will use in this assignment is Nutch 1.6.
Grab the binary distribution `bin.tar.gz`, and unpack it in a location where you have read, write, and execute permissions – for example, your home directory is ideal for this.

Make sure to `chmod +x bin/nutch`, and then to test it out by running `bin/nutch` and making sure you see command line output.

Once you have unpacked and set up Nutch, you're ready to get going!

3.2 Configuring Nutch

The best way to get started on configuring Nutch for your Web crawl is to read the Section 3 and 3.1 of the Quick Start on the wiki:
http://wiki.apache.org/nutch/NutchTutorial

Pay particular attention to the configuration files that you will need to change/update.
Note: You **must** set the agent name ***and*** its related properties appropriately (this will be graded!).

As in previous assignments, you are advised to do any and all testing using a small crawl. You do not want to spend several hours running a crawl only to find that you botched it and wasted your time.

Caution: Be very careful about the scope of your crawl! If your crawl is not properly confined to our mirror site and extends beyond to FBI websites, or worse, the entire Internet, you may end up filling up all available disk space and/or encounter other technical issues.

See Appendix C for details about storing your crawl data on a different drive or partition – this is necessary if you are booting Ubuntu from USB; it is optional but possibly useful in other cases.

3.2.1 Politeness

Crawlers can retrieve data much quicker and in greater depth than human users, so they can have a crippling impact on the performance of a site. A server would naturally have a hard time keeping up with requests from multiple crawlers – as some of you may have already discovered while doing the earlier crawler assignment!

Except this time *all* of you are going to be torturing the same server. Please be kind to it – if it dies under your bombardment, none of you will be able to complete the assignment. =(

To avoid overloading the server, you should set an appropriate interval between successive requests to the same server.

4. Extracting the PDFs from the Sequence File compressed format

Nutch uses a compressed "`SequenceFile`" binary format to represent the data that it downloads. Data is stored in `SequenceFiles` which are part of "segments" on disk, some splittable portion of the crawl. Nutch performs this operation both for efficiency, but also to allow for distribution to allow multiple fetches to be distributed out on the cluster using the Apache Hadoop Map-Reduce Framework.

After successful crawling of FBI vault site you should have a crawl folder with a bunch of segments in it, and inside of those segments, `SequenceFiles` with the content stored inside.

To get some idea of how to extract data from Nutch `SequenceFiles`, have a look at this blog post:
http://www-scf.usc.edu/~csci572/2012Spring/homework/2/allenday20080829.html

Your goal in this portion of the assignment is to write a Java program `PDFExtractor.java` that will extract the PDF files out of the SequenceFile format, and to write the PDFs to disk. Your program should be executed with arguments indicating the source (crawl data) directory and the output (PDF files) directory, like this:

```
java PDFExtractor <crawl dir> <output dir>
```

The result of this command will be to extract all PDF files from <crawl dir>, and to output them as named individual PDF files to the directory path identified by <output dir>.

You may find it helpful to refer to:
Nutch API docs: http://nutch.apache.org/apidocs-1.6/index.html
Hadoop API docs: http://hadoop.apache.org/common/docs/r1.0.0/api/index.html

5. Hints

- Configuring Nutch to perform this task will likely be one of the trickiest parts of the assignment. Pay particular attention to the RegexURLFilter required to crawl the site, and to the crawl command parameters.
- Please use a wired connection if possible.
- Crawling takes time! You will very likely find yourself leaving the crawler to run overnight (or more), so make sure to plan accordingly.
- After successful crawl, total size of crawl folder would be 10~11GB
- Once you have completed your crawl, the PDF extraction step may be done in any environment you prefer – Java program `PDFExtractor.java` ought to be cross-platform, so there is no strict requirement for this step to be done in Linux.
- Successful PDF extract should have ~2000 files (10~11GB total size)
- This should be obvious: you can test your extracted PDF files by opening them in Adobe Reader.

6. Submission Guidelines

6.1 Team Registration

To register a team to work together on this assignment, send an email to the Dr. Mattmann as follows:

```
To: chris.mattmann@gmail.com
Subject: Nutch HW team

last_name_1, first_name_1 <username1@usc.edu>
last_name_2, first_name_2 <username2@usc.edu>
```

The email subject and body content format must be exactly as prescribed; stating your names and USC email addresses – no more and no less. It suffices for one team member to send this registration email; both students can expect to receive a confirmation in response from the TA.

You ***must*** register your by 4/1/2014.

6.2 Submission

This assignment is to be submitted ***electronically, by 12pm PT*** on the specified due date, via Gmail to chris.mattmann@gmail.com. Use the subject line: CSCI 572: <Horowitz|Mattmann>: Spring 2014: Nutch Homework: <Your Lastname>: <Your Firstname>. So if your name was Lord Voldemort, and you were taking Dr. Horowitz's course, you would submit an email to chris.mattmann@gmail.com with the subject "CSCI 572: Horowitz: Spring 2014: Nutch Homework: Voldemort: Lord" (no quotes).

- Only the first member of a team needs to make the assignment submission. (We highly suggest that the other member insist on re-downloading the submission and testing it to make sure there is no corrupted file or other technical problems.)
- Include all Nutch configuration files and/or code that you needed to change or implement in order to crawl the FBI Vault website. Avoid including unnecessary files. Place in a directory `nutchconfig`.
- Include all source code and external libraries needed to extract the PDFs from Sequence File compressed format. Put these in a directory `pdfextract`.
    - All source code is expected to be commented, to compile, and to run. You should have (at least) one Java source file: `PDFExtractor.java`, containing your `main()` function. You should also include other java source files that you added, if any. Do **not** submit `*.class` files. We will compile your program from submitted source.
    - There is no need to submit jar files for Tika, Nutch and/or Hadoop. If you have used any other external libraries, you should include those jar files in your submission.
    - Prepare a `readme.txt` containing a detailed explanation of how to compile and execute your program. Be especially specific on how to include other external libraries if you have used them.
- Also include your name, USC ID number and email address in the `readme.txt`.
- Do **not** include your crawled data or extracted PDF files.
- Compress all of the above (`nutchconfig` folder, `pdfextract` folder and `readme.txt`) into a single zip archive and name it according to the following filename convention:

    <lastname>_<firstname>_CSCI572_HW_NUTCH.zip

    Use only standard zip format. Do **not** use other formats such as zipx, rar, ace, etc.

Important Notes:

- Make sure that you have attached the file the when submitting. Failure to do so will be treated as non-submission.
- Successful submission will be indicated in the assignment's submission history. We advise that you check to verify the timestamp, download and double check your zip file for good measure.

6.3 Late Assignment Policy

- -10% if submitted within the first 24 hours
- -15% for each additional 24 hours or part thereof