There were thousands of pages that we were unable to fetch across our files, and there were dozens of possible reasons.

We have compiled a list of 100 difficult-to-fetch URLs in the file unfetched_urls.pdf.

In this, there are 10 common reasons why we are unable to fetch URLs.

1. _pst_:notfound(14)

Files with this error message are HTTP 404 Not Found. The most likely cause for these is that the files have been removed or permanently moved elsewhere.

2. _pst_:robots_denied(18)

This is probably the most self-explanatory. This occurs when the robots.txt for a given website has prevented access to that file.

3. _pst_:failed(2)

This is an error code that we encountered primarily with ftp files. It happens when a dialog box appears and you need to either enter username and password or else request guest access. Because we didn't have Selenium enabled on Crawl 1 we were unable to get to this data.

4. _pst_:exception(16)

This error message applies to a variety of potential issues, many of which are Java exceptions.

a) java.net.SocketTimeoutException: Read timed out

b) java.net.SocketTimeoutException: connect timed out

c) java.net.ConnectionException: Connection refused

These three are similar. Read timed out is what seems to be a genuine timeout (i.e. if we were to try again later we might succeed without atimeout. Alternatively it might just be a slow server). The "connect timed out" and "Connection refused" seem to indicate a more permanent problem.

d) java.net.UnknownHostException

This is associated with DNS lookup errors

e) Http code= 500 (Internal server error)

This is a problem on the server side. It is likely to be a temporary thing such as maintenance.

5. _pst_:access_denied(17)

These are pages in which authentication is required. (Similar to the _pst_:failed(2) situation mentioned above). Again Selenium would help us access some of this data.

6. _pst_:moved(12) and _pst_:temp_moved(13)

These are associated with pages which redirect elsewhere. It is not that we are unable to fetch these files per se, it is that the file content has moved to a different address. We'll try to fetch it in the next round.

7. Parse Tika Error

There are some files which we were able to fetch but were unable to parse because we did not have to corresponding tika parser. Crawling with Tesseract and GDAL may help us parse some of these files.