**Report Assignment 3 – Team 11**

**Subject: CSCI 599 - Content Detection and Big Data Analysis**

**Data Set used –** CCA and Polar full dump

1. **Did the crawler find the most relevant pages? Why, why not?**

For most cases the crawler seems to find the relevant pages. The content type of the response from the server matches the content type of the request sent. Also some of the pages seem to return relevant content which was found by doing Named Entity Recognition.

1. **Is your MIME detection good? Define “good”.**

The MIME detection performs very well overall. However, we have encountered few cases where image/gif file was detected as text/plain. We came to know about this while doing text extraction.

1. **Are your parsers extracting the right text? Define “right”.**

The parsers seem to extract the right text except for one of the MIME type (application/tikamsoffice). The general parser for Microsoft office was not able to extract any text from these files.

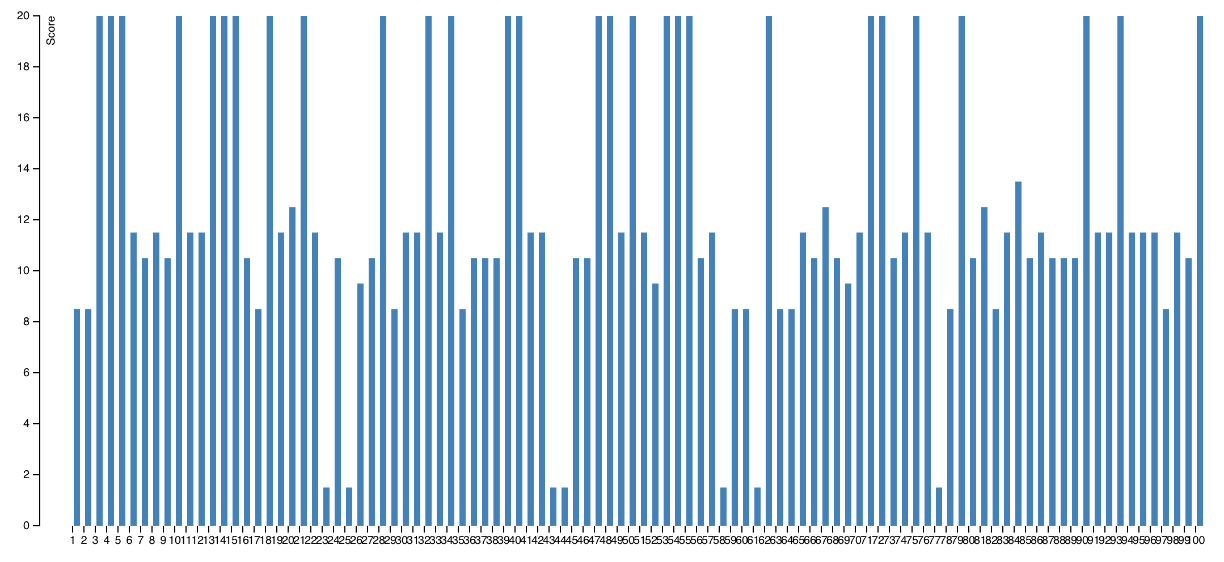
On the other hand, we checked extracted text from some xml files, word files, excel files, pdf files against actual files. It was perfect extraction.

1. **Are we selecting the right parser? Define “right”.**

Out of six mime types that we experimented with, five of them seem to call the right parser. By right, we mean for application/pdf files PDFparser was called and so on.

However, for mime type application/tikamsoffice, the parser being called was not able to extract any text from files. So it seems a bad parser for that mime type.

1. **Is your Metadata appropriate? What’s missing?**



From the above image, we can say that more than half of files seem to have more than average metadata and some files have very good metadata content (that is, high score). There are some files which have bad metadata and as a result have a low metadta.

Also, we checked metadata files for images and publications to see if the properties of metadata are relevant to file type as in resolution for image and author for publication etc. and we did not see any mismatch properties for any mime type.However, in some MS Excel files there were some properties with blank values.

1. **How well is my language detection performing? Comment based on the diversity of the languages derived in this assignment. Are there mixed languages? Did it affect your accuracy?**

The language detection performs fairly accurately. The language detector was able to find 25 different languages - *ca,da,de,el,en,eo,es,et,fa,fi,fr,gl,hu,it,lt,nl,no,pl,pt,ru,sk,sv,th,uk*. Yes, there are files with mixed languages. The language detector just gives out the language of the file which had more information and higher priority.

Also the following anomalies were observed.

1. It identifies a language but then there is no content in the page. In other words, the file size is 0 kb
2. It identifies the wrong language. For example, the page is in English but it detected Hungarian. The page is in English but detects Estonian.
3. **Do your Named Entities make sense?**

With raw results, named entity values contains lots of messy tokens such as html tags. Moreover, it gets lots of wrong result. We have run a filter which only extract meaningful results and remove messy html tags.

Also, many NER toolkits recognize “$” symbols as good indications for NER\_MONEY types which leads to a large amount of false positive.  Many extractions is also one-character tokens. Therefore, we also remove these tokens.

After the filtering, we get much more meaningful named entities and results can be viewed in d3.

Overall, we suspect a large amount of false positive extractions due to the nature of noisy data on the web. However, these issues can be addressed using some heuristic as mentioned above.

1. **Analyze whether your new joint agreement produces any update NER for your metadata records.**

Combining multiple NER toolkits helps find more named entities through the whole data corpus. As you can see from d3 file, there are differences in number of different toolkits. Therefore, using the max joint of these toolkits gave us a better results.

**Extra credit:**

We made a pull request to integrate Grobid Quantities into Tika. The link to the <https://github.com/apache/tika/pull/119>.

**Video:**

Also we uploaded the video walking through the answers to the above questions using the D3 visualizations and knowledge gained from this assignment. The link to the video: [https://youtu.be/qci68dyO16o](https://l.facebook.com/l.php?u=https%3A%2F%2Fyoutu.be%2Fqci68dyO16o&h=hAQGB03kG)

**Details about running the programs:**

We have added more details on how the run the files and view the visualizations in the Readme file.