Assignment 1 : Report on MIME Diversity in the Text Retrieval Conference (TREC)

Team : 18 (Siddharth Bhayani [sbhayani@usc.edu], Nimesh Jain [nimeshja@usc.edu] , Suhas Suresh[suhassur@usc.edu])

OBSERVATIONS REGARDING DATASET :

Following are the observations regarding the TREC

1. The the dataset had several files which were empty which were detected as application/octet-stream.
2. “full-polardump” bucket had files with no extensions. If the extension was provided, tika’s output would have been more accurate because it detects through file extensions too. Since we didn’t had extensions, detection was mainly done using magic bytes in tika’s mime repository.
3. The data used for analysis was approximately 60 GB of S3 bucket(approx. 1.36M files)
4. New mimetypes were detected as compared to the mimetype distribution given on the dataset link (https://github.com/chrismattmann/trec-dd-polar/) . These types are:
    a. application/zlib
    b. application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
    c. application/x-tika-ooxml-protected
    d. application/x-grib
    e. application/vnd.apple.keynote
    f. application/vnd.openxmlformats-officedocument.presentationml.presentation
    g. application/vnd.openxmlformats-officedocument
    h. application/vnd.ms-excel.sheet.macroenabled.12
    i. application/x-mspublisher
    j. application/vnd.openxmlformats-officedocument

    This means that Tika’s mime type repository was enriched and updated later.

MIME TYPE SELECTION :

First we tried and selected some generic list of mime types. However, for some mime types we were not able to get any magic bytes and so we selected other mime types. We also took into consideration the amount of data we had for each mime type while selecting the mime types for analysis.

MimeTypes that we selected for our analysis are :

1. application/postscript
2. application/x-sh
3. application/vnd.openxmlformats-officedocument.wordprocessingml.document
4. application/xhtml+xml
5. application/vnd.ms-word
6. application/x-elc
7. application/x-msdownload
8. application/gzip
9. application/vnd.ms-excel

10. application/x-tika-msoffice
11. application/octet-stream
12. audio/mpeg
13. image/png
14. image/jpeg
15. video/quicktime

Identifications of MimeTypes with application/octet-stream:

- Of the 60GB, approx. 89675 files were detected as application/octet-stream. We were able to identify 2 different mimetypes from the octet-stream data.

- 89406 files in the octet-stream were empty files and we were able to detect 18 files as quicktime based on the new magic byte updated in mimetype.xml.

- It also contained some javascript and css file but we weren't able to detect.

- We added a new mimeType application/x-zerosize which detected the empty files.

BFA, BFC, BFCC and FHT Analysis and addition of magic byte:

We were able to analyze and infer knowledge from BFA, BFC, BFCC & FHT by looking at their corresponding d3 graphs. We got flat distribution for some mime types like png mimetype where almost all byte frequency had same frequency value and prominent spikes in some other file types for BFA like byte 10, 48 and 101 for postscript mime type. However, we also got false negatives eg. the frequency of byte 32 (space) was high in almost all the mimetypes.

BFC graphs was useful in validating our BFA fingerprints. It showed high correlations for bytes with high frequency too. The disadvantage which we saw for BFC from d3 graphs is that some bytes whose frequency was 0 had high correlation and as such there were many byte values having very high frequency.

BFCC was useful in identifying the range of bytes which are highly correlated and less correlated. However, we were not able to identify specific bytes, as it was difficult to look through every pair of high correlation.

FHT Analysis helped us to identify exact byte signature that appears in Head and Tail of the files. All the magics that we have identified are from the FHT Analysis of the corresponding mime type.

We added a new magic for quicktime mime type through which we were able to identify 18 files as quicktime video files from octet-stream files.

Below are the magics for all the mime types that we had selected. We had earlier selected different mime types, but changed them because we were not able to identify new magics that would enhance tika's mime type repository. For eg: video/mpeg, image/vnd.microsoft.icon,text/pdf, text/html.

Also for some mime types, we were able to detect only one new mime type because tika already had the knowledge of other significant header bytes. eg: application/x-tika-msoffice

For some mimetype , we could only add one new magic because as per the file format specification and description rest all bytes after that offset were variable. eg: application/gzip (http://www.zlib.org/rfc-gzip.html), image/jpeg (https://en.wikipedia.org/wiki/JPEG)

For video/quicktime, we added magics from FHT and also after reading the specification of the file structure. We were able to identify quicktime video files from octet-stream data through these new magic bytes.

Below are the magics that we have identified for the mime types:

1. application/postscript:

```
<magic priority="50">
<match value="\045\041\120\123\055\101\144\157" type="string" offset="0" />
 <match value="\056\060\012\045\045" type="string" offset="12" />
</magc>
```

2. application/x-sh:

```
<magic priority="50">
 <match value="bin" offset="3" type="string"/>
 <match value="bin" type="string" offset="4"/>
 <match value="bash" type="string" offset="7"/>
 <match value="bash" type="string" offset="8"/>
</magic>
```

3. application/vnd.openxmlformats-officedocument.wordprocessingml.document

```
<magic>
 <match type="string" offset="0" value="\120\113\003\004\024\000"/>
 <match type="string" offset="9" value="\010\000\000\000\041\000"/>
 <match type="string" offset="24" value="\000\000\032\000\010\002\133\103"/>
</magic>
```

4. application/xhtml+xml

```
<magic priority="20">
<match value="\074\041\104\117\103\124\131\120\105\040\150\164\155\154" offset="0" type="string"/>
</magic>
<magic priority="40">
<match value="\055\047\047\127\063\103\047\047\054\124\104\024\072\060\124\115\114" offset="23" type="string"/>
</magic>
```

5. application/vnd.ms-word

```
<match value="\076\000\003\000\376\377\011\000" type="string" offset="24"/>
<match value="\073\000\003\000\376\377\011\000" type="string" offset="24"/>
```

6. application/x-elc

```
<magic priority="50">
 <match value="\012\050\146\165\156\143\164\151\157\156\050\044\051" type="string" offset="0"/>
 <match value="\012\050\146\165\156\143\164\151\157\156\040\050\044\051" type="string" offset="0"/>
</magic>
```

7.application/x-msdownload

```
<match offset="24" type="string" value="\100"/>
<match offset="12" type="string" value="\377\377"/>
```

8. application/gzip

```
<match value="\010" type="string" offset="2"/>
<match value="\x08" type="string" offset="2"/>
```

9.application/vnd.ms-excel

```
<magic priority="50">
 <match type="string" offset="24" value="\076\000\003\000\376\377\011\000"/>
</magic>
```

10. application/x-tika-msoffice

```
<match value="\076\000\003\000\376\377\011\000" type="string" offset="24"/>
```

11. audio/mpeg

```
<match value="\063" type="string" offset="2"/>
<match value="\220" type="string" offset="2"/>
```

12. image/png

```
<magic priority="50">
 <match value="\000\000\000\015" type="string" offset="8"/>
 <match value="\111\110\104\122" type="string" offset="8"/>
</magic>
```

13. image/jpeg

```
<match value="\340" type="string" offset="3"/>
```

14. video/quicktime

```
<!--As per specification. Saw little values in the FHT.-->
<match value="ftyp" type="string" offset="0"/>
<match value="pnot" type="string" offset="4"/>
<match value="\167\151\144\145" type="string" offset="4"/>
```

OBSERVATIONS AFTER UPDATION OF MIMETYPE XML:

The statistics after running the tika with updated mimetypes.xml are :

- Information regarding new mime types:
    - Mime : x-aiff Count : 9
    - Mime : x-zerosize Count : 89406
    - Mime : x-tex Count : 11
    - Mime : x-tika-msworks-spreadsheet Count : 1634
    - Mime : x-grib Count : 3
    - Mime : vnd.ms-htmlhelp Count : 1
    - Mime : x-java-source Count : 4

- Information regarding mimetypes with increased count:
    - Mime : msword Count : 21
    - Mime : vnd.openxmlformats-officedocument.spreadsheetml.sheet Count : 1
    - Mime : atom+xml Count : 2
    - Mime : mpeg Count : 483
    - Mime : x-sh Count : 3
    - Mime : quicktime Count : 17
    - Mime : vnd.openxmlformats-officedocument.wordprocessingml.document Count : 1354
    - Mime : x-tika-ooxml Count : 1
    - Mime : java-vm Count : 4

- Information regarding mimetypes with reduced count:
    - Mime : octet-stream Count : 89427
    - Mime : png Count : 13
    - Mime : x-msdownload Count : 8
    - Mime : jpeg Count : 69
    - Mime : vnd.ms-excel Count : 106
    - Mime : xhtml+xml Count : 45
    - Mime : x-tika-msoffice Count : 1549
    - Mime : vnd.microsoft.icon Count : 1
    - Mime : plain Count : 503
    - Mime : zip Count : 1354

From here we can see that , the frequency count of x-tika-msworks-spreadsheet in the given dataset increased and the frequency count of x-msdownload & x-tika-msoffice decreased after which are the subtypes of our selected mime-type x-tika-msoffice. However, we couldn't identify the reason behind the decrease in the zip files and the exact amount of increase in

vnd.openxmlformats-officedocument.wordprocessingml.document file count. When we ran the updated tika only on the application/octet-stream data, we got 89406 x-zerosize file and 20 quicktime video files. This metric was consistent even when we ran updated tika over entire dataset.

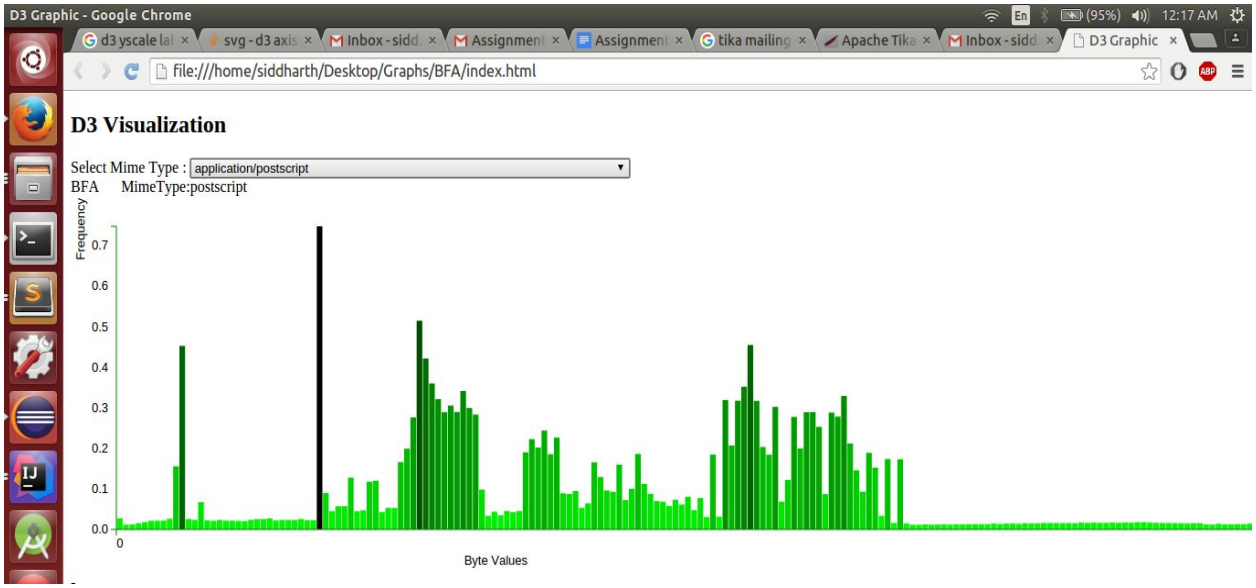
BROADER OBSERVATIONS :

Tika was not able to discern the mime types because it didn't had a mechanism to identify files of zero length and detect them as application/x-zerosize.
Some of the Tika's magic signature are only through analysis of FHT and not as per the corresponding file format specification. We were able to detect 18 quicktime video files, by adding a new magic after reading the file format specification.

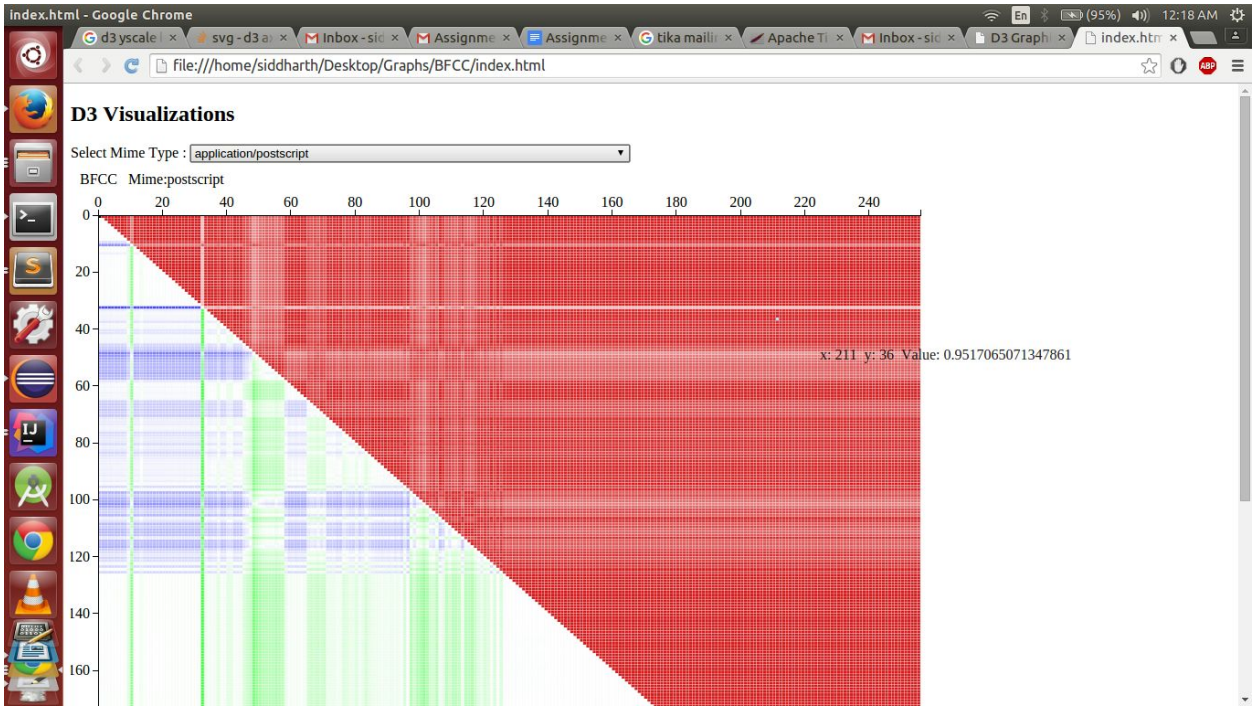
SCREENSHOTS OF D3 Graphs:

1. Frequency Distribution

2. BFA and BFC Distribution



3. BFCC Analysis



4. FHT Analysis