Assignment 3 : Report on Evaluation of Content Analysis on Text Retrieval Conference (TREC) Polar Dynamic Domain Dataset

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

MIMETYPE DETECTION : GOOD OR BAD

We think the MimeType Detection performed really good. It was able to discern many files correctly. We were also able to detect new video files from octet-stream files because of the new magics added in Assignment one. We think the detection was "good" because (Define "good") :

1. It was quick as it was based on magics.
2. Only small fraction of the dataset was identified as octet-stream.
3. It was able to detect complex mimetypes like application/vnd.openxmlformats-officedocument.spreadsheetml.sheet.

TEXT EXTRACTION BY PARSERS:

We think the parsers are extracting "right" text. By "right" text we mean (Define "right"):

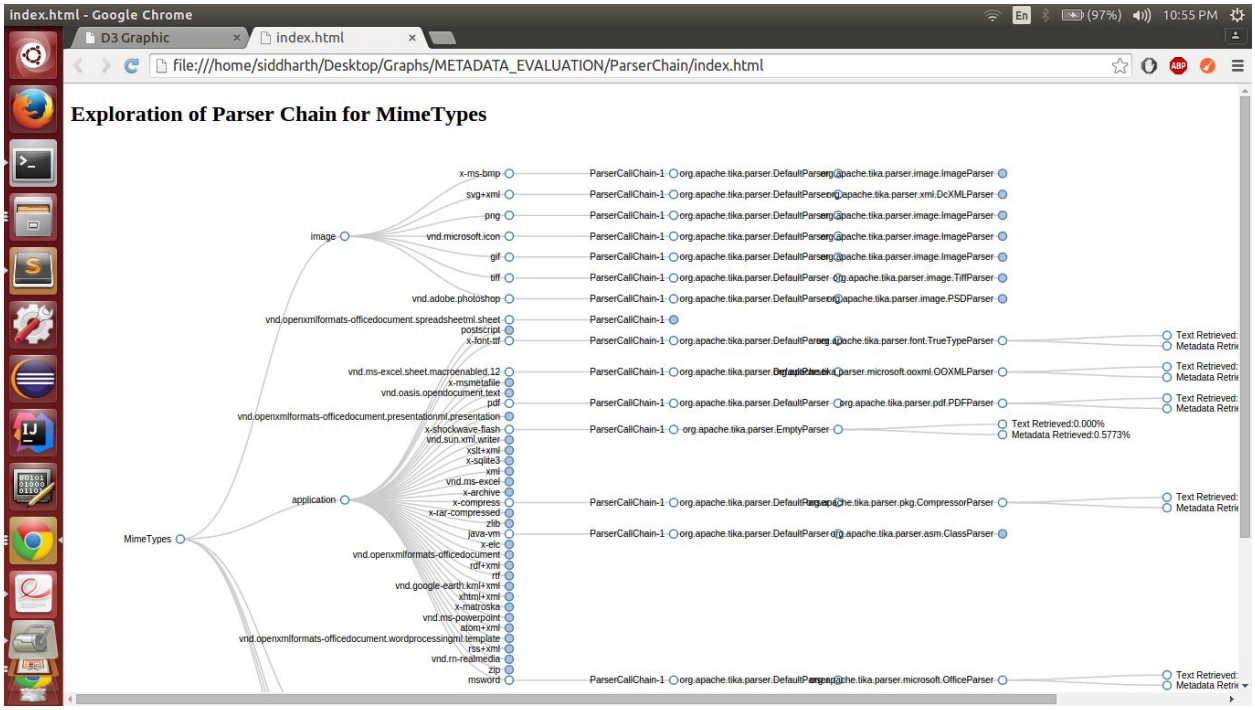
1. We are extracting text using TagRatio Parser. The input to TagRatio parser is the Xhtml Content Identified by Tika.
2. Thus, we are not only extracting Text , but we are also extracting important MetaData information.
3. Also we were able to extract many measurement and geolocations from the extracted text from parsers. If our parsers were not extracting "right" text, we would not have extracted location and measurement information from them.

PARSER SELECTION:

We think the Parser Selection Strategy work fine and we are selecting the "right" Parsers. By "right" parsers we mean (Define "right"):

1. Selection of specific Parsers for Proprietary files. (eg: org.apache.tika.parser.microsoft.OfficeParser is used for msword, etc)
2. Selection of a general parsers for Open format files (eg: org.apache.tika.parser.image.ImageParser for png & gif files)
3. Also, TIka Selects Parsers on the basis of MimeType. Since it is mimetype detection works good and has a good mimetypes to parser mapping, it's parser selection is good.

Below is the screenshot of the Parser Chain Selection for mimetypes :

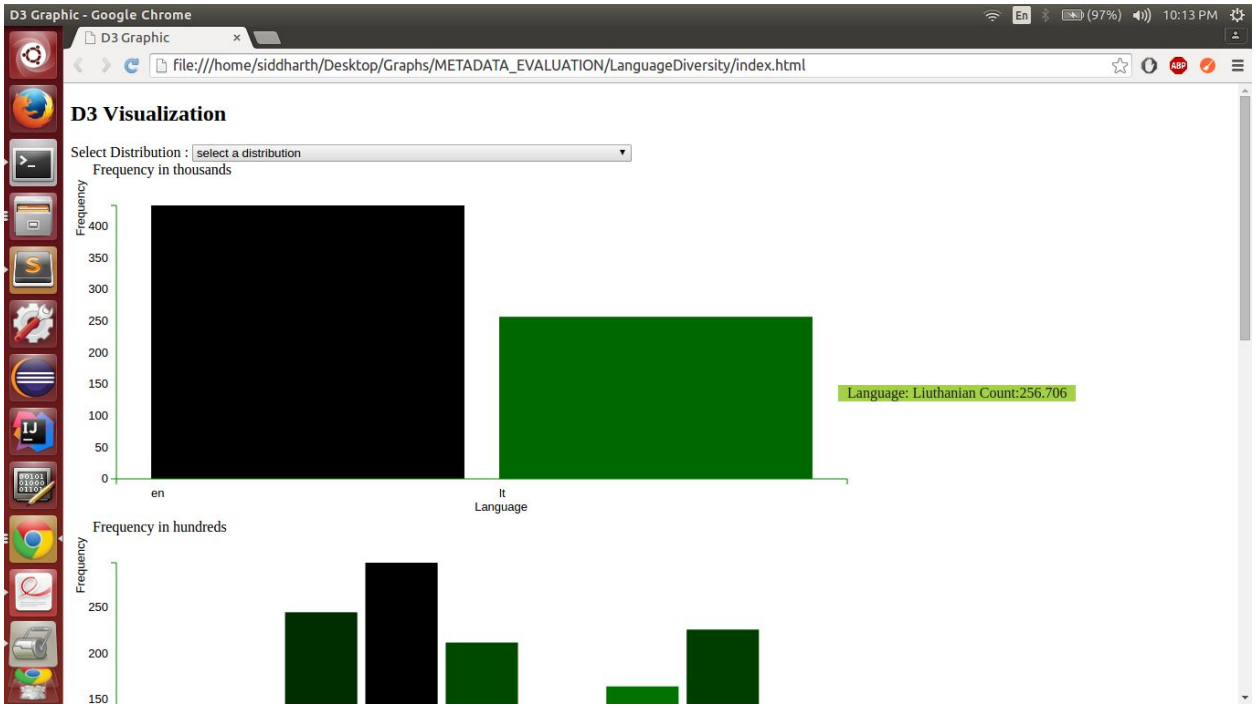


LANGUAGE IDENTIFICATION:

Language Detection is working satisfactorily. It's not very accurate though. For many mimetypes it identified the language as Lithuanian, however it was in English. Also for empty input, it detects Language as Lithuanian.

There is good diversity of Language in the datasets. We were able to detect files in 28 languages. However, most of the files were in English. The 2nd highest was Lithuanian. However, we think it was because of the above stated reasons.
We didn't found any mixed language documents. However, we think the accuracy of the identification.

Below is the screenshot of the language diversity over the entire dataset.

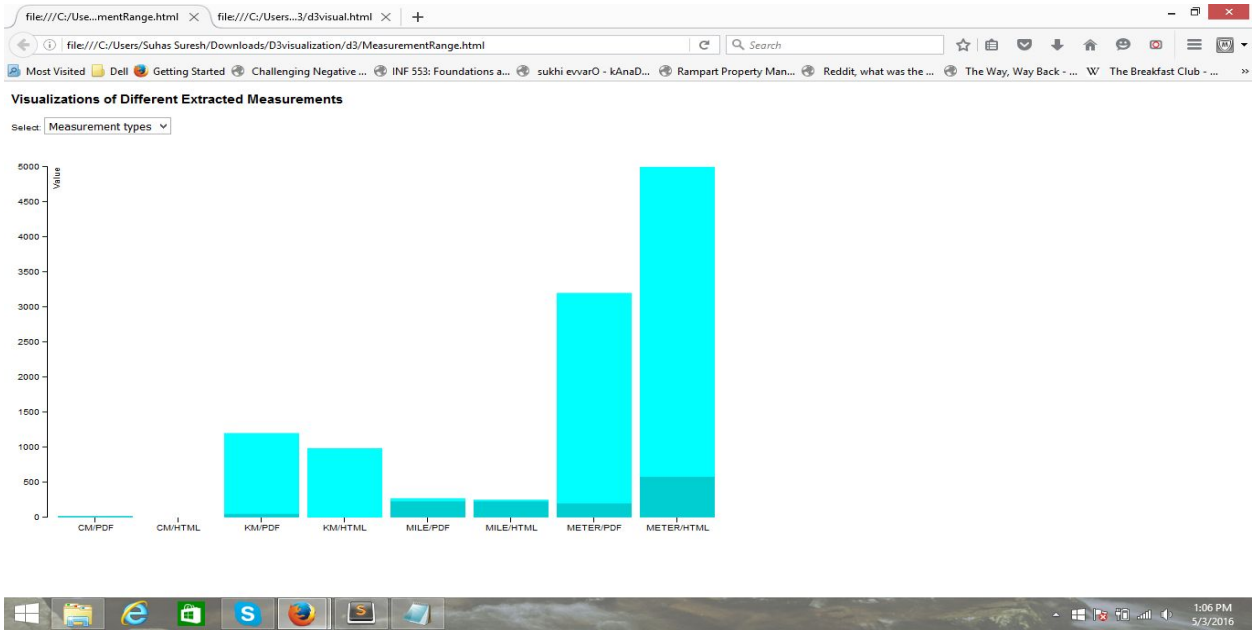


METADATA EVALUATION:

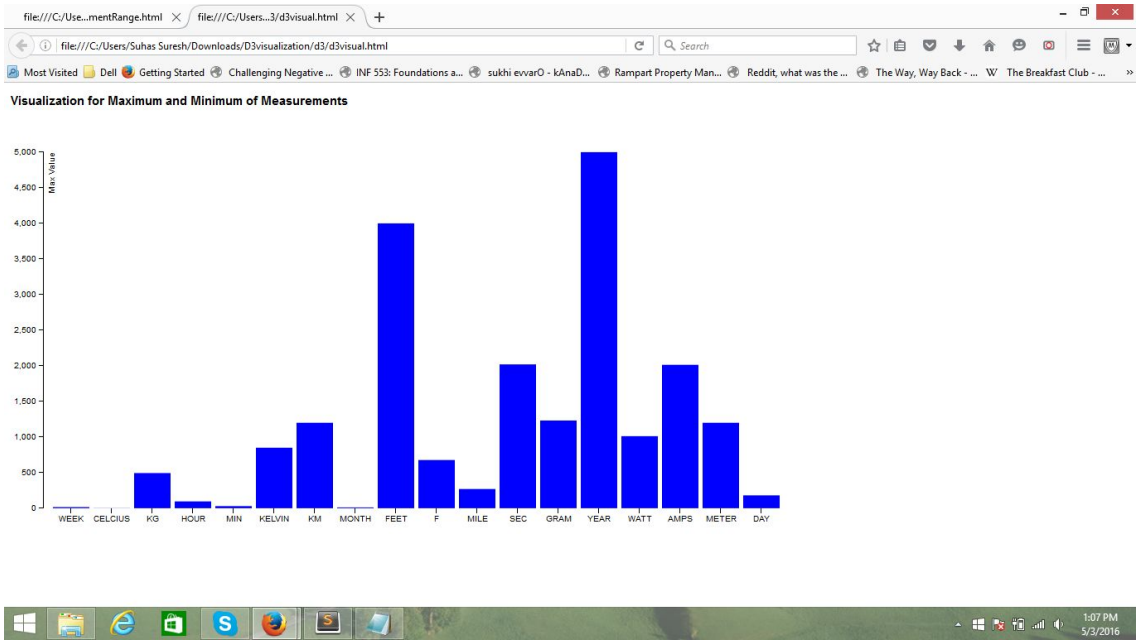
By comparing different parser we found that the metadata extracted by some parsers were far less than the metadata extracted by other NER parsers. For some of the files for CCA data when we parsed through TIKA parser the the metadata we got quite few results. For all the metadata the average size of metadata to the content size was approximately 35%.

NAMED ENTITY EXTRACTION:

We parsed the entire dataset to find named entities. The parsers that we employed could get us Quantities, using which we obtained measurements. The dataset contained various measurements including CM, KM, Meter, Inches, Gram, Kilograms, Celsius, Fahrenheit, Amps, KiloWatts, etc. Certain measurements like CM (centimeter) wasn't found in HTML files at all. On PDF's files, the count of CM measurement was in 20's. Given below is a screenshot of various measurements in different MIME types.

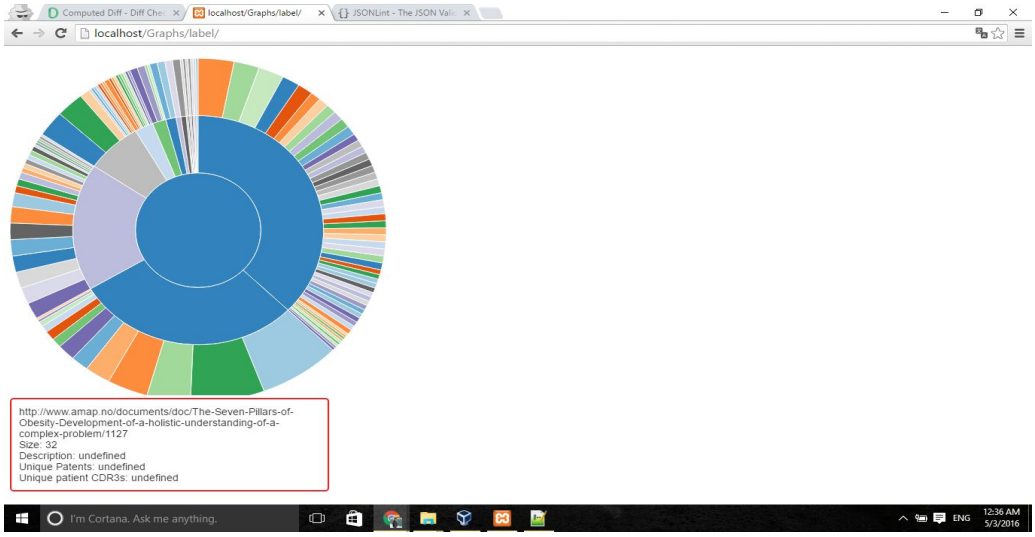


The named entities worked satisfactorily fine. The occurrence of different measurements are plotted on a bar graph and they are accurate. We find that the measurement 'Year' occurs maximum times and 'Celsius' minimum times in the dataset. It can also be seen that 'Miles' is equally present in both PDF's and HTML's. Given below is the screenshot of Max-Min occurrence of measurement from the whole dataset.



CLASSIFICATION PATH FROM REQUEST TO RESPONSE:DID THE CRAWLER FIND MOST RELEVANT PAGES?

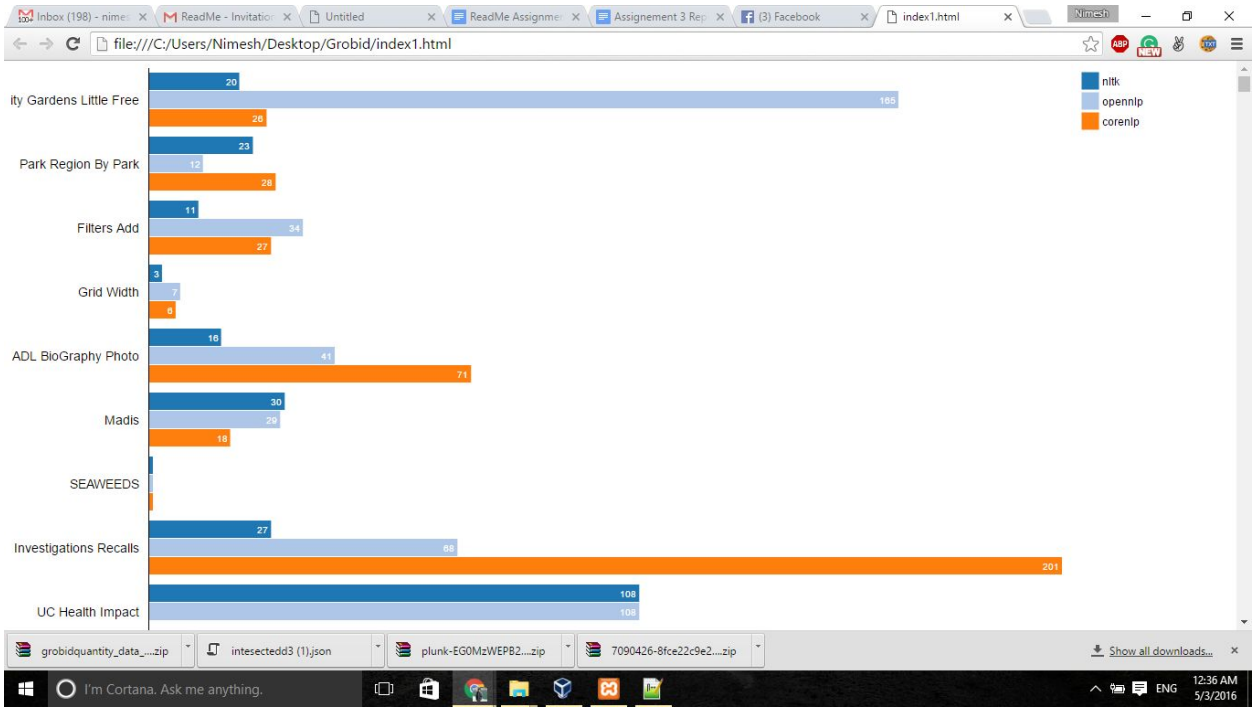
We classified the request based on host and url and the response was classified as the frequency of occurrence of the entities present in the request url. We extracted the content of the response body from the Tag Ratio algorithm and found the occurrence of words in request present in the output of the content of Tag Ratio Algorithm. We found that for some files the matches were absolutely 0. In other words the entities in request url did not match with any named entity in response body. There were several cases in which the crawler did return the most relevant pages where we found the count of entities in url was much higher in response body. Hence, the crawler did provide relevant pages but at times we found that the pages we found were irrelevant. The reason may be because the keywords are present in comment based on our analysis. The d3 provides more visual information.



GROBID QUANTITIES - Named Entity Recognition -CompositeNERParser

Created a python code to call the REST API for the grobid quantity running on the localhost. Executed this code on pdf, plain and xml files of the polar mime diversity.

We created a python code which finds the intersection of the entities parsed by OpenNlp, CoreNlp and NLTK. These are the common agreed entities parsed by all the 3 parsers. Then we find the count of these intersected entities in corpus retrieved from each of these 3 NER parsers. We created a d3 visualization of each intersected entity of how much of that entity been parsed by each of the parser.



Youtube URL : https://youtu.be/lbbjsnuaNoQ