|  |  |
| --- | --- |
| **Visualization File Name** | **Description** |
| UrlKeywordToNer-MostFrequentKeyword.html | This visualization shows the mapping between most frequent URL Keywords and Extracted Named Entities.  We use the extracted information to study the relationship between the Request URL keywords and extracted named entities.  For readability, we show 10 most frequent named entities of each class from 30 most frequent normalized keywords. |
| UrlKeywordToNer-MostNerExtracted.html | The request URL, the contents of the retrieved pages, and the named entities extracted have stark resemblance amongst them. |
| SizeRatioSummary.html | **This visualization shows Size Ratio of Solr Index to Original File Size (Grouped By MIME Type)**  It depicts the various MIME types vs. sizes of original files and Solr index. Generally, the Solr index is much larger compared to the original file, except for certain MIME types like application/gzip and application/msword. |
| SizeSummary.html | **Here, we see the File Size Summary Analysis**  The visualization depicts the average file size of the Common Crawl Dataset grouped by MIME types. |
| ParserVsCount.html | **Parser Hierarchy vs Document Count by Type**  This visualization depicts the various parsers used by Tika and the count of the files identified by them. |
| ParserVsMetadataRetrieved.html | **Parser Hierarchy vs Size of Metadata Retrieved Ratio by Type**   * The ratio is often greater than one. For instance, * text files parsed by TXTParser and gif and png * files parsed by ImageParser. * However, tiff files from TiffParser has very small * ratio because tiff files are usually much larger than * gif or png. * Even if the files are empty, we can extract some * information as metadata (for instance, files parsed * by EmptyParser). |
| ParserVsTextRetrieved.html | **Parser Hierarchy vs Size of Text Retrieved Ratio by Type**   * Text files parsed by TXTParser usually have * the ratio of 1. * Gzip files parsed by CompressorParser and, * Zip files parsed by PackageParser usually have the * ratio more than 1. |
| LanguageCompareFilter.html | **This visualization depicts the Number of Documents vs. Detected Language**  We compared Tika language detection feature with Optimaize language detector library. The visualization shows that around 300k files are detected to be *lt-Lithuanian* by Tika but Optimaize only detect 42 files. Also, around 300k files are detected to be *unknown* by Optimaize but Tika only detect 8 files. We think this might be related to detecting empty file.  We try to run the language detection again by filtering out the files that have too few text retrieved (<1000 characters). The result shows that number of documents detected to be *lt-Lithuanian* by Tika is significantly lower to 20k and number of documents detected to be *unknown* by Optimaize is lower to 7k.  *That means amount of text plays an important role for language detection accuracy.* |
| MixedLanguage.html | **Tika Mixed Language Detection**  We tried to split extracted text into chunks and then use Tika language identifier to detect language of each chunk. We then compare detected language from the full text to the majority of detected language of each chunks. If they are not the same, it possibly indicates that the file contains multiple languages. |
| WordCloudD3.html | **Word Cloud Representing Most Frequent terms in the Dataset**  This visualization shows the most frequent terms identified by measurement parser, SWEET ontology parser, GeoTopic parser and other metadata extractors. |
| GrobidQuantityTTR.html | **Horizontal Bar Chart showing Frequency of Occurrence of different measurement units identified by the Grobid Quantity Parser**  This visualization was created by a Tika NER implementation that invokes Grobid Quantities via its REST service. |
| NERAgreementFullText.html | NLTK, CoreNLP, and openNLP are Named Entity Recognizers used to extract named entities from the data. After setting them up, the NER agreement list was obtained by accumulating most common named entities extracted from all 3 recognizers. We configured the parser to apply tag Ratio analysis to the extracted text before doing named entity recognition. The visualization is made to display 200 most common named entities extracted by the recognizers. |
| Measurement\_ByDomain.html | **Zoomable Circle Packing D3 Showing vaious Measurement Quantities(Grouped By Domain)** |
| Measurement\_ByMIME.html | **Zoomable Circle Packing D3 Showing vaious Measurement Quantities(Grouped By MIME Type)** |
| measurementCountBarChart.html | **Bar Chart showing Frequency of Occurrence of different measurement units** |
| Measurement\_Range.html | **Dendogram showing Minimum, Maximum, and Mean values of Measurement Units** |